

Вариант 1

- Лара Крофт исследует древние храмы, где 80% ловушек срабатывают в помещениях с мозаикой на полу, а 20% срабатывают в помещениях без мозаики. Известно, что в этих храмах мозаика встречается в 30% помещений.

Какой вывод верен?

- Вероятность встретить ловушку в комнате с мозаикой выше, чем в комнатах без мозаики
 - Ловушки не размещают в комнатах без мозаики
 - Наличие мозаики не влияет на вероятность нахождения ловушки
 - Данных недостаточно
- В музее хранились только подлинные предметы. Воры украли некоторые экспонаты и заменили их фальшивыми. После кражи 90% от всех золотых предметов остались подлинными, а 20% от всех керамических предметов — фальшивыми.

Изначально в музее хранились только золотые и керамические предметы, причем золотых предметов было в 3 раза больше, чем керамических.

Какая доля музейных экспонатов оказалась фальшивой? Ответ запишите в процентах и округлите до десятых.

- Лара взламывает сейф с украденными артефактами. Она знает:
 - пароль к сейфу состоит из четырех цифр (от 0 до 9);
 - для него используются только четные цифры, ни одна из них не встречается дважды, а сумма первых двух цифр равна сумме последних двух.

Сколько возможных комбинаций у такого замка?

- В банке работают три независимые команды аналитиков по 50 человек. Среди сотрудников каждой команды 20% — стажеры, 30% — начинающие сотрудники и 50% — опытные аналитики. Стажер выполняет одну задачу за час, начинающий сотрудник — за 45 минут, опытный сотрудник — за полчаса. Коллеги могут делить задачи между собой, то есть каждому сотруднику может быть поручено выполнение не только целой задачи (нескольких целых задач), но и частей каких-то задач (например, половины задачи или десятой части задачи и т. д.).

Рабочий день каждой команды начинается, когда все сотрудники одновременно приступают к выполнению своих задач, при этом они работают одновременно, без перерывов, и рабочий день команды считается законченным, когда последний сотрудник выполнит свою последнюю задачу.

В первой команде все задачи распределяются равномерно между всеми сотрудниками — чтобы все сделали равное число задач.

Во второй команде каждый опытный сотрудник выполняет по 6 задач, а оставшиеся задачи делятся равномерно между оставшимися сотрудниками.

В третьей команде задачи распределяются так, чтобы все сотрудники закончили выполнять свои задачи одновременно.

Сегодня каждой команде поступило 200 задач. Сколько часов будет длиться рабочий день каждой команды? Ответы округлите до десятых.

5. Учительница собрала данные о времени, которое ученики 10-го класса тратят на выполнение домашнего задания по математике (в минутах). Гистограмма по этим данным показала, что распределение времени совпадает с нормальным. Среднее значение — 60 минут, стандартное отклонение — 15 минут.

Какова вероятность, что случайно выбранный ученик тратит на домашнее задание больше 75 минут?

- 1) 0.162
- 2) 0
- 3) 0.03
- 4) 0.32

Если учительница 50 раз подряд возьмет 25 случайных учеников и усреднит их время выполнения домашнего задания по математике, то как будет распределено среднее значение?

- 1) Нормально с $\mu = 60$, $\sigma = 3$
- 2) Нормально с $\mu = 60$, $\sigma = 15$
- 3) Равномерно~Экспоненциально

Учительница также собрала данные о времени выполнения домашнего задания по обществознанию. Оказалось, все ученики тратят на домашнее задание от 10 до 20 минут и любое время в этом интервале равновероятно. Какое это распределение?

- 1) Равномерное
- 2) Нормальное
- 3) Экспоненциальное
- 4) Биномиальное

6. В задаче идентификации рака важно не пропустить больных.

Модель может выдавать следующие результаты:

True Positive (TP) — пациент болен, модель правильно это определила.

True Negative (TN) — пациент здоров, и модель это подтвердила.

False Negative (FN) — пациент болен, но модель пропустила болезнь.

False Positive (FP) — пациент здоров, но модель ошибочно определила, что он болен.

С помощью этих результатов можно посчитать следующие метрики:

Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.

Recall — доля верно обнаруженных реальных больных $TP / (TP + FN)$.

Precision — доля истинно положительных срабатываний $TP / (TP + FP)$.

F1-score — гармоническое среднее между Recall и Precision $2 \times ((Precision \times Recall) / (Precision + Recall))$.

Какую метрику нужно оптимизировать?

1) Accuracy

2) Recall

3) Precision

4) F1-score

7. В задаче рекомендации фильмов важно предлагать пользователям только те фильмы, что им действительно понравятся.

Модель может выдавать следующие результаты:

True Positive (TP) — фильм пользователю понравится, и модель правильно это предсказала (рекомендовала его).

True Negative (TN) — фильм пользователю не понравится, и модель правильно это предсказала (не рекомендовала его).

False Negative (FN) — фильм пользователю понравится, но модель ошиблась и не рекомендовала его (упущенная возможность).

False Positive (FP) — фильм пользователю не понравится, но модель ошиблась и рекомендовала его (неудачная рекомендация).

С помощью этих результатов можно посчитать следующие метрики:

Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.

Recall — доля понравившихся пользователю фильмов, которые модель смогла обнаружить и рекомендовать $TP / (TP + FN)$.

Precision — доля удачных рекомендаций среди всех рекомендованных моделью фильмов $TP / (TP + FP)$.

F1-score — гармоническое среднее между Precision и Recall. $2 \times ((\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}))$.

Значения параметров модели: FP = 50, FN = 30, TP = 70, TN = 150.

Чему равен Accuracy? Ответ округлите до сотых.

8. Тимлид руководит коллективом из 50 человек: ежедневно он раздает всем сотрудникам одинаковое количество сопоставимых по сложности задач. Эти задачи необходимо выполнить до конца рабочего дня, если сотрудник заканчивает выполнение задач раньше — он может уйти домой, время окончания работы над задачами фиксируется. Тимлид решил проверить, что влияет на продуктивность его сотрудников. Он собрал данные за один день работы всех его сотрудников:

hours — время в часах, за которое были выполнены все задачи (чем быстрее были выполнены задачи, тем продуктивнее сотрудник), зависимая переменная.

age — возраст.

gender — пол сотрудника (gender = 1 для мужчин, gender = 0 для женщин).

experience — опыт работы в годах.

coffee — количество перерывов на кофе за день.

friend — присутствие друга за соседним рабочим столом (friend = 1, если рядом сидит друг; friend = 0, если рядом сидит просто коллега, а не друг).

По собранным данным тимлид оценил линейную регрессию с hours в качестве зависимой переменной и собрал результаты в таблицу. Все коэффициенты оказались статистически значимыми.

Переменная	Оценка коэффициента
age	−0.7
age2	0.01
gender	−1.5
experience	−0.1
coffee	0.25
friend	2.1
gender * friend	−0.7
constant	5.2

Заполните пропуски или выберите верные варианты в утверждениях о коллективе тимлида, которые следуют из результатов оцененной им регрессии (каждый вывод ниже делается при прочих равных, то есть в предположении неизменных остальных переменных):

- 1) 35 лет — это возраст, при котором сотрудники наиболее продуктивны
- 2) Более высокую производительностью показывают мужчины/женщины
- 3) Каждый дополнительный год опыта позволяет сотруднику выполнять задачи на 6 минут / минуты быстрее
- 4) Каждый дополнительный перерыв на кофе увеличивает время выполнения задач на 15 минут / минуты
- 5) Если рабочее место сотрудника находится рядом с другом, то его продуктивность уменьшается/увеличивается
- 6) Присутствие друга рядом сильнее/слабее влияет на продуктивность женщин, чем на продуктивность мужчин

9. Продажи обогревателей зимой в 3 раза выше, чем летом. Это пример:

- 1) Тренда
- 2) Сезонности
- 3) Шума
- 4) Аномалии

10. Зоопарк хочет оптимизировать закупку корма для животных. Вам нужно собрать модель, которая предсказывает необходимую суточную порцию корма (в кг) для животного на основе его характеристик.

В вашем распоряжении есть датасет со следующими признаками:

weight — вес животного (в кг).

age — возраст животного (в годах).

activity_level — уровень активности (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).

metabolic_rate — уровень метаболизма (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).

ambient_temperature — средняя температура в вольере (в °C).

Целевая переменная daily_food_intake — суточная норма корма (в кг).

Вы построили линейную регрессию на данных о млекопитающих и получили следующие коэффициенты:

intercept (свободный член) = 0.5.

Коэффициент для weight = 0.02.

Коэффициент для age = -0.1.

Коэффициент для activity_level = 0.3.

Коэффициент для metabolic_rate = 0.4.

Коэффициент для `ambient_temperature` = -0.05 .

Вопрос 1. Какую суточную норму корма (в кг) предскажет модель для молодого леопарда со следующими параметрами? Ответ округлите до сотых.

`weight` = 50 кг

`age` = 2 года

`activity_level` = 3 (высокий)

`metabolic_rate` = 3 (высокий)

`ambient_temperature` = 25 °C

Вопрос 2. Отдел закупок должен заказать корм на месяц (30 дней) для этого леопарда, основываясь на прогнозе модели. Известно, что модель на тестовых дневных данных давала прогнозы с $MSE = 0.25$.

Чтобы избежать ситуации нехватки корма, отдел закупок решил заказать количество корма, которое с вероятностью 95% покрывает потребности животного с учетом погрешности модели.

Какой минимальный запас корма (в кг) необходимо заказать, если за основу взять типичную величину ошибки прогноза? Учтите, что заказывать можно только целое число килограммов корма. Распределение расходов корма считайте нормальным, ошибки в разные дни независимы.

11. Интернет-магазин хочет предсказать время выполнения заказа (в минутах) от момента оформления до момента доставки. В распоряжении есть исторические данные.

Признаки:

`product_count` — количество товаров в заказе (шт.).

`distance` — расстояние до клиента (км).

`is_weekend` — выходной день (0 — нет, 1 — да).

Целевая переменная: `delivery_time` — время выполнения заказа (минуты).

После построения линейной регрессии получилось следующее уравнение:

$$\text{delivery_time} = 17 + 2,3 \times \text{product_count} + 7 \times \text{distance} + 14 \times \text{is_weekend}$$

Как интерпретировать коэффициент 7 для признака `distance`, если все коэффициенты значимы на 5% уровне значимости?

- 1) Расстояние определяет 7% времени доставки
- 2) При увеличении расстояния на 7 км время доставки увеличивается в среднем на 1 минуту
- 3) При увеличении расстояния на 1 км время доставки увеличивается в среднем на 7 минут

4) При увеличении расстояния в четыре раза время доставки увеличится в двадцать восемь раз

12. Магазин тестирует скидки:

- A. 5%: +10% продаж.
- Б. 10%: +25% продаж.
- В. 15%: +30% продаж.

Какая скидка дает максимальную прибыль, если наценка — 300%?

13. Скачать датасет — [файл](#)

Университет собрал данные об обучении за последний семестр:

Student_id — ID студента.

Gender — пол студента.

Type — тип обучения (local / exchange).

Age — возраст студента.

Department — направление обучения.

Attendance — посещаемость (в %).

Total_Score — общий балл.

Grade — оценка.

Study_Hours_per_Week — среднее количество часов обучения в неделю.

Scholarship — стипендия (да/нет).

Sleep_Hours_per_Night — среднее количество часов сна.

Вопрос 1. Различается ли общий балл у студентов между местными студентами и студентами, приехавшими учиться по обмену? Рассчитайте модуль разницы средних и напишите, является ли она статистически значимой.

Вопрос 2. Различается ли доля студентов со стипендией между местными студентами и студентами, приехавшими учиться по обмену? Рассчитайте модуль разницы долей и напишите, является ли она статистически значимой.

Для всех вопросов $\alpha = 0.05$. Ответ округлите до сотых.

14. Перед вами график мировых цен на сахар. Соотнесите возможные события с отмеченными точками на графике, которые могли стать причиной такого движения цены.



2021 — Опасения по поводу сокращения производства в Бразилии

2023 — Запрет экспорта на сахар из Индии — второго по величине экспортера сахара

2025 — Богатый урожай в основных странах-изготовителях

15. Скачать датасет — [файл](#)

Вам представлен набор данных о кассовых сборах фильмов, а также дополнительная информация о каждом фильме. Кинокомпании используют такие данные для прогнозирования успеха новых проектов. В заданиях ниже вам предстоит исследовать взаимосвязь между различными факторами.

Вопрос 1. Рассчитайте следующие показатели для фильмов, произведенных в Канаде:

- Медианный бюджет фильмов.
- Средний рейтинг фильмов и стандартное отклонение. Ответ округлите до сотых.
- Доля прибыльных фильмов ($\text{gross} > \text{budget}$) в группе с бюджетом ниже медианы. Ответ округлите до процентов. В расчет берите только фильмы, у которых есть оба показателя.
- Компания с наибольшей средней прибылью на фильм. Прибыль можно рассчитать, только если есть данные и по бюджету, и по сборам.

Вопрос 2. Выполните корреляционный анализ для фильмов, произведенных в Канаде.

Рассчитайте и округлите ответы до сотых:

- Коэффициент корреляции между бюджетом и сборами.
- Коэффициент корреляции между рейтингом и прибылью.

16. Вы разрабатываете модель для автоматической модерации комментариев на платформе. Модель должна уметь:

- Находить все оскорбительные сообщения (минимизировать пропуски).
- Не блокировать нормальные комментарии (минимизировать ложные срабатывания).

Для вас протестировали три модели с разными характеристиками:

Модель	TP (Заблокировано оскорбительных)	FP (Заблокировано нейтральных)	TN (Пропущено нейтральных)	FN (Пропущено оскорбительных)
A	241	95	858	43
Б	270	238	715	14
В	199	29	924	85

Какую модель следует выбрать для детской платформы, если более важно минимизировать количество оскорбительных комментариев?

- 1) А
- 2) Б
- 3) В

Какую модель следует выбрать для соцсети, если нужно минимизировать число жалоб и известно, что:

Каждая ложная блокировка (FP) приводит к 1 жалобе пользователя.
Каждый пропущенный оскорбительный комментарий (FN) — к 3 жалобам.

- 1) А
- 2) Б
- 3) В

17. Скачать датасет — [файл](#)

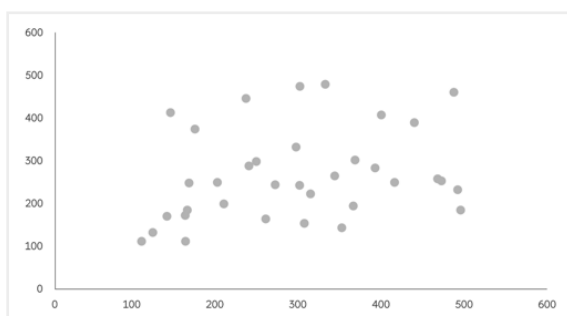
Вам предоставлен датасет с логами действий пользователей в мобильном приложении.

Выделите сессии пользователя по правилу: если более 30 минут не было никаких действий, то создается новая сессия. Для каждого пользователя пронумеруйте сессии, используя это правило.

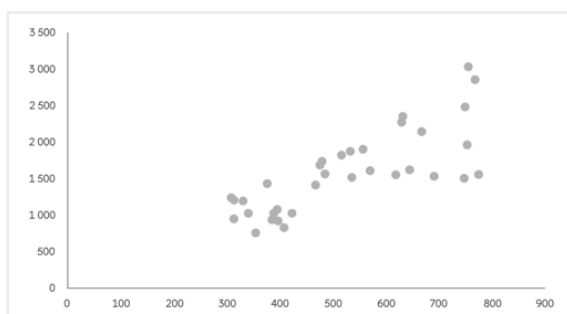
Найдите пользователя с максимальным числом сессий. В ответе запишите ID пользователя.

18. Соотнесите коэффициент корреляции и диаграмму рассеяния.

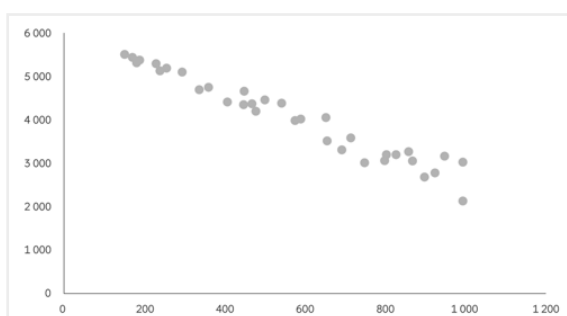
1)



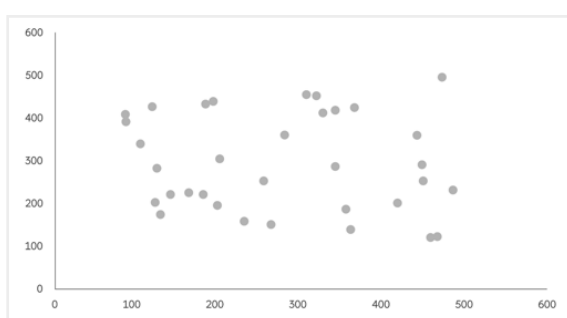
2)



3)



4)



A. 0.26

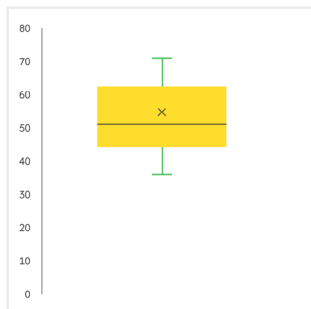
Б. 0.72

В. -0.97

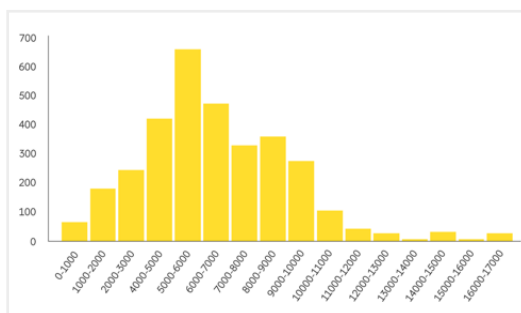
Г. -0.12

19. Вы редактировали журнал «Наглядная статистика» и обнаружили, что заголовки явно не соответствуют графикам. Соотнесите график и его правильный заголовок:

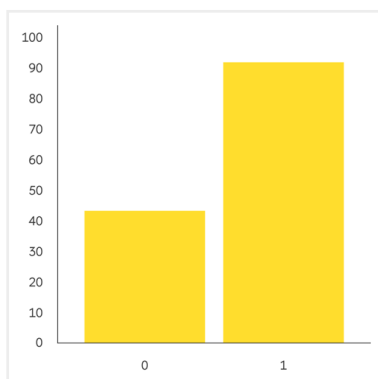
1)



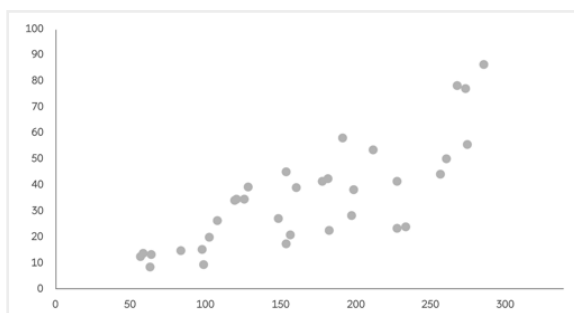
2)



3)



4)



А. Вес учащихся 9 класса

Б. Средний чек в театре

В. Распределение школьников по итогам тура олимпиады (прошел в финал / не прошел в финал)

Г. Стоимость загородного дома (в млн рублей) в зависимости от его площади

20. В парке развлечений два самых популярных аттракциона — «Космическая гонка» и «Супергигант». Они абсолютно одинаковы по своим характеристикам (спрос, время ожидания, длина очереди и т. д.). Цена билета на каждый — 250 ₽ в любое время. Среднее число посетителей в месяц — 12 000 на каждый аттракцион.

Если клиенты попадают на аттракцион во время пиковой нагрузки, им приходится ждать в очереди 80 минут. Обычное время ожидания — 20 минут.

Распределение посещаемости на каждом аттракционе выглядит так:

- Пиковая нагрузка (30% времени): 5 000 посетителей.
- Обычное время (70% времени): 7 000 посетителей.

Чтобы оптимизировать доход и управлять спросом, администрация хочет сравнить две стратегии ценообразования:

- «Космическая гонка» — фиксированный повышенный тариф (400 ₽ всегда).
- «Супергигант» — динамический тариф: обычное время — 300 ₽, час пик — 500 ₽.

Если время ожидания превышает 60 минут, парк выдает клиенту ваучер на напиток стоимостью 100 ₽ и на хот-дог стоимостью 200 ₽, при этом вероятность приобретения напитка за ваучер — 90%, а хот-дога — 60%.

Показатель	Было (250 ₽ всегда, «Космическая гонка» и «Супергигант» вместе)	«Космическая гонка» (400 ₽ всегда)	«Супергигант» (300/500 ₽)
Цена в час пик	250	400	500
Цена в обычное время	250	400	300
Прокатившихся в час пик	10 000	4 700	4 200
Прокатившихся в обычное время	14 000	6 200	6 600
Среднее время ожидания (час пик)	80 мин., равномерно распределено от 60 до 100 мин.	60 мин., равномерно распределено от 40 до 80 мин.	40 мин., равномерно распределено от 20 до 60 мин.
Время ожидания (обычное время)	20 мин.	15 мин.	20 мин.

Вопрос 1. Какой вариант тарификации оптимальный с точки зрения дохода?

- 1) Динамический
- 2) Фиксированный повышенный

Сколько денег дополнительно получит парк по сравнению со старым тарифом, если внедрить его на оба аттракциона?

Вопрос 2. Какой вариант новой тарификации следует выбрать для обоих аттракционов при уровне значимости 95%? Укажите вариант тарифа:

- 1) Фиксированный повышенный тариф
- 2) Динамический тариф
- 3) Разницы нет

Вопрос 3. А при уровне значимости 99%?

- 1) Динамический тариф
- 2) Разницы нет
- 3) Фиксированный повышенный тариф

Вопрос 4. Укажите Z статистику, используемую для принятия решения.

Ответ округлите до сотых. При расчетах не округляйте промежуточные результаты до получения финального ответа.

Вариант 2

- Археолог Джонс исследует пирамиды. Известно, что 60% случаев активации скрытых механизмов приходится на залы с иероглифами на стенах, а 40% — на залы без иероглифов.

Какой вывод верен?

- Механизмы не размещают в залах без иероглифов
 - Наличие иероглифов не влияет на вероятность активации механизма
 - Вероятность активации механизма в зале с иероглифами выше, чем в зале без них
 - Данных недостаточно
- На фармацевтическом складе хранились только оригинальные лекарства. Из-за ошибки логистов партию товара смешали с более дешевыми аналогами (дженериками). После проверки выяснилось, что среди всех жаропонижающих средств 95% — это оригинальные препараты, а среди всех обезболивающих средств 15% оказались дженериками.

Изначально на складе хранились только жаропонижающие и обезболивающие препараты, причем жаропонижающих было в 2 раза больше, чем обезболивающих.

Какая доля всего товара на складе оказалась дженериками? Ответ запишите в процентах и округлите до десятых.

- Хакер Нео взламывает терминал доступа к секретной базе данных. Из перехваченных данных он узнал:

- пароль состоит из четырех цифр (от 0 до 9);
- в пароле используются только нечетные цифры, и все они различны;
- сумма цифр на первой и третьей позициях равна сумме цифр на второй и четвертой позициях.

Сколько возможных комбинаций пароля соответствует этим критериям?

- В банке работают три независимые команды аналитиков по 70 человек. Среди сотрудников каждой команды 10% — стажеры, 40% — начинающие сотрудники и 50% — опытные аналитики. Стажер выполняет одну задачу за полтора часа, начинающий сотрудник — за 45 минут, опытный сотрудник — за полчаса.

Коллеги могут делить задачи между собой, то есть каждому сотруднику может быть поручено выполнение не только целой задачи (нескольких целых задач), но и частей каких-то задач (например, половины задачи или десятой части задачи и т. д.).

Рабочий день каждой команды начинается, когда все сотрудники одновременно приступают к выполнению своих задач, при этом они работают одновременно, без перерывов, и рабочий день команды считается законченным, когда последний сотрудник выполнит свою последнюю задачу.

В первой команде все задачи распределяются равномерно между всеми сотрудниками — чтобы все сделали равное число задач.

Во второй команде каждый опытный сотрудник выполняет по 5 задач, а оставшиеся задачи делятся равномерно между оставшимися сотрудниками.

В третьей команде задачи распределяются таким образом, чтобы все сотрудники закончили выполнять свои задачи одновременно.

Сегодня каждой команде поступило 280 задач. Сколько часов будет длиться рабочий день каждой команды? Ответы округлите до десятых.

5. Тренер собрал данные о времени, за которое спортсмены бегут стометровку (в секундах). Гистограмма по этим данным показала, что распределение времени совпадает с нормальным. Среднее значение — 12.5 секунд, стандартное отклонение — 0.5 секунд.

Какова вероятность, что случайно выбранный спортсмен пробежит стометровку медленнее 13.5 секунды?

- 1) 0.023
- 2) 0
- 3) 0.16
- 4) 0.32

Если тренер 100 раз подряд возьмет 16 случайных спортсменов и усреднит их результат, то как будет распределено среднее значение?

- 1) Нормально с $\mu = 12.5$, $\sigma = 0.125$
- 2) Нормально с $\mu = 12.5$, $\sigma = 0.5$
- 3) Равномерно~Экспоненциально

Тренер также собрал данные о времени пробега дистанции 400 метров. Оказалось, все спортсмены показывают результат от 50 до 60 секунд и любое время в этом интервале равновероятно. Какое это распределение?

- 1) Равномерное
- 2) Нормальное
- 3) Экспоненциальное
- 4) Биномиальное

6. При разработке спам-фильтра для почтового сервиса важно, чтобы пользователи не пропускали важные письма, ошибочно помеченные как спам.

Модель может выдавать следующие результаты:

True Positive (TP) — письмо является спамом, модель правильно это определила.

True Negative (TN) — письмо не спам, и модель это подтвердила.

False Negative (FN) — письмо является спамом, но модель пропустила письмо.

False Positive (FP) — письмо не спам, но модель ошибочно пометила его как спам.

С помощью этих результатов можно посчитать следующие метрики:

Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.

Recall — доля верно обнаруженного спама $TP / (TP + FN)$.

Precision — доля истинно положительных срабатываний $TP / (TP + FP)$.

F1-score — гармоническое среднее между Recall и Precision $2 \times ((Precision \times Recall) / (Precision + Recall))$.

Какую метрику нужно оптимизировать?

- 1) Accuracy
- 2) Recall
- 3) Precision
- 4) F1-score

7. В задаче идентификации рака важно не пропустить больных. Модель может выдавать следующие результаты:

True Positive (TP) — пациент болен, модель правильно это определила.

True Negative (TN) — пациент здоров, и модель это подтвердила.

False Negative (FN) — пациент болен, но модель пропустила болезнь.

False Positive (FP) — пациент здоров, но модель ошибочно определила, что он болен.

С помощью этих результатов можно посчитать следующие метрики:

Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.

Recall — доля верно обнаруженных реальных больных $TP / (TP + FN)$.

Precision — доля ложных тревог $TP / (TP + FP)$.

F1-score — гармоническое среднее между Recall и Precision $2 \times ((Precision \times Recall) / (Precision + Recall))$.

Значения параметров модели: $FP = 10$, $FN = 20$, $TP = 30$, $TN = 40$.

Чему равен precision? Ответ округлите до сотых.

8. Руководитель центра обратной связи командует коллективом из 60 человек. Ежедневно центру приходят письма с жалобами и предложениями от клиентов, на которые необходимо отвечать. Руководитель предварительно фильтрует и сортирует эти письма так, чтобы всем сотрудникам нужно было ответить на одинаковое количество писем с сопоставимыми по сложности запросами. Время, которое проходит от момента получения сотрудником письма до предоставления ответа с его стороны, фиксируется. Руководитель решил проверить, что влияет на продуктивность его сотрудников. Он собрал данные за один день работы всех его сотрудников:

minutes — время в минутах, за которое сотрудник отвечает на письмо клиента (чем быстрее был предоставлен ответ, тем продуктивнее сотрудник), зависимая переменная.

experience — опыт работы в годах.

gender — пол сотрудника (gender = 1 для мужчин, gender = 0 для женщин).

education — количество лет образования.

snack — количество перерывов для перекуса за день.

laptop — факт работы за ноутбуком (laptop = 1, если сотрудник работает за ноутбуком; laptop = 0, если сотрудник работает за стационарным компьютером).

По собранным данным руководитель оценил линейную регрессию с minutes в качестве зависимой переменной, все коэффициенты оказались статистически значимыми, результаты представлены в таблице.

Переменная	Оценка коэффициента
experience	−0.6
experience2	0.02
gender	−1.2
education	−0.25
snack	0.5
laptop	1.7
gender * laptop	0.4
constant	4.3

Заполните пропуски или выберите верные варианты в утверждениях о коллективе центра обратной связи, которые следуют из результатов оцененной регрессии (каждый вывод ниже делается при прочих равных, то есть в предположении неизменных остальных переменных):

- A. Сотрудники с опытом работы 15 лет наиболее продуктивны
- B. Более высокую производительностью показывают мужчины/женщины
- C. Каждый дополнительный год образования позволяет сотруднику отвечать на письмо в среднем на 15 секунд / секунды быстрее

- D. Каждый дополнительный перерыв на перекус увеличивает среднее время ответа на письмо на 30 секунд / секунды
- E. Если сотрудник работает за ноутбуком, то его продуктивность уменьшается/увеличивается
- F. Работа за ноутбуком вместо стационарного компьютера слабее/сильнее влияет на продуктивность женщин, чем на продуктивность мужчин

9. Датчик температуры на метеостанции передает показания каждую секунду. На графике видно, что помимо плавного суточного хода температуры присутствуют мелкие, случайные и незначительные колебания показаний в пределах ± 0.2 °C. Эти колебания не связаны с реальным изменением погоды. Это пример:

- 1) Тренда
- 2) Сезонности
- 3) Шума
- 4) Аномалии

10. Зоопарк хочет оптимизировать закупку корма для животных. Вам нужно собрать модель, которая предсказывает необходимую суточную порцию корма (в кг) для животного на основе его характеристик.

В вашем распоряжении есть датасет со следующими признаками:

weight — вес животного (в кг).

age — возраст животного (в годах).

activity_level — уровень активности (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).

metabolic_rate — уровень метаболизма (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).

ambient_temperature — средняя температура в вольере (в °C).

Целевая переменная daily_food_intake — суточная норма корма (в кг).

Вы построили линейную регрессию на данных о млекопитающих и получили следующие коэффициенты:

intercept (свободный член) = 0.5.

Коэффициент для weight = 0.03.

Коэффициент для age = 0.15.

Коэффициент для activity_level = 0.2.

Коэффициент для metabolic_rate = 0.25.

Коэффициент для ambient_temperature = -0.04.

Вопрос 1. Какую суточную норму корма (в кг) предскажет модель для бурого медведя со следующими параметрами? Ответ округлите до сотых.

weight = 200 кг

age = 5 лет

activity_level = 2 (средний)

metabolic_rate = 2 (средний)

ambient_temperature = 15 °C

Вопрос 2. Отдел закупок должен заказать корм на месяц (30 дней) для этого медведя, основываясь на прогнозе модели. Известно, что модель на тестовых дневных данных давала прогнозы с $MSE = 0.36$. Чтобы избежать ситуации нехватки корма, отдел закупок решил заказать количество корма, которое с вероятностью 95% покрывает потребности животного с учетом погрешности модели.

Какой минимальный запас корма (в кг) необходимо заказать, если за основу взять типичную величину ошибки прогноза? Учтите, что заказывать можно только целое число килограммов корма. Распределение расходов корма считайте нормальным, ошибки в разные дни независимы.

11. Служба доставки еды хочет предсказать время доставки заказа (в минутах). В распоряжении есть исторические данные.

Признаки:

order_size — размер заказа (усл. ед.).

distance — расстояние до клиента (км).

is_peak — час пик (0 — нет, 1 — да).

Целевая переменная: delivery_time — время доставки (минуты).

После построения линейной регрессии получилось следующее уравнение:

$$\text{delivery_time} = 10 + 5 \times \text{order_size} + 8 \times \text{distance} + 20 \times \text{is_peak}$$

Все коэффициенты значимы на 5% уровне значимости. Вопрос: как интерпретировать коэффициент 5 для признака order_size?

- 1) При увеличении размера заказа на 1 усл. ед. время доставки увеличивается в среднем на 5 минут
- 2) При увеличении размера заказа на 5 усл. ед. время доставки увеличивается в среднем на 1 минуту
- 3) Размер заказа определяет 5% времени доставки
- 4) При увеличении размера заказа в четыре раза время доставки увеличится в двадцать раз

12. Кофейня тестирует скидки на кофе с себестоимостью 50 Р:

- А. 5%: +15% продаж.
- Б. 10%: +35% продаж.
- В. 15%: +50% продаж.

Какая скидка дает максимальную прибыль, если наценка на чашку кофе до скидки составляет 100%?

13. Скачать датасет — [файл](#)

Университет собрал данные об обучении за последний семестр:

Student_id — ID студента.

Gender — пол студента.

Type — тип обучения (local / exchange).

Age — возраст студента.

Department — направление обучения.

Attendance — посещаемость (в %).

Total_Score — общий балл.

Grade — оценка.

Study_Hours_per_Week — среднее количество часов обучения в неделю.

Scholarship — стипендия (да/нет).

Sleep_Hours_per_Night — среднее количество часов сна.

Вопрос 1. Различается ли общий балл у студентов женского и мужского пола?

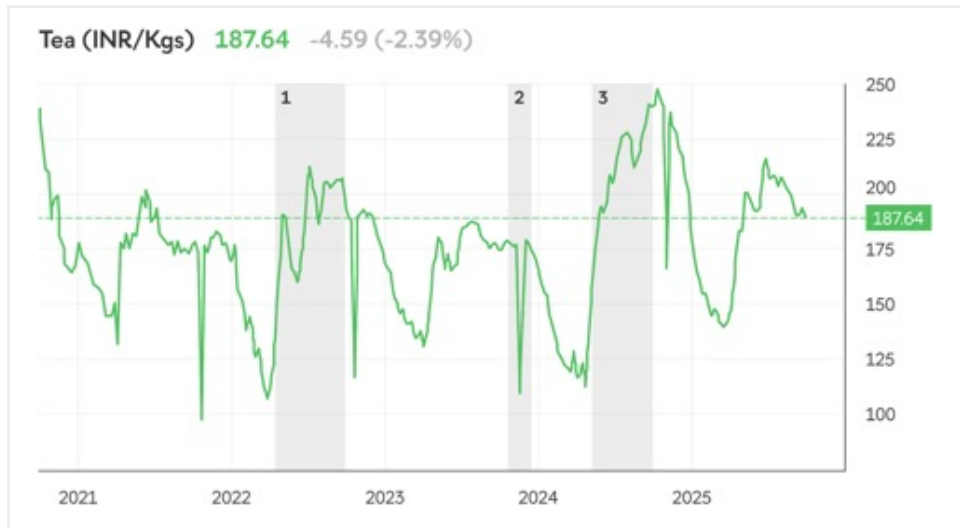
- а) Рассчитайте модуль разницы средних
- б) и напишите, является ли она статистически значимой.

Вопрос 2. Различается ли доля студентов со стипендией у студентов женского и мужского пола?

- а) Рассчитайте модуль разницы долей
- б) и напишите, является ли она статистически значимой.

Для всех вопросов $\alpha = 0.05$. Ответ округлите до сотых.

14. Перед вами график мировых цен на чай. Соотнесите возможные события с отмеченными точками на графике, которые могли стать причиной такого движения цены.



2022 — Истощение запасов старого урожая с последующей стабилизацией по мере возобновления поставок

2023 — Пик предложения в связи с одновременным урожаем в ключевых странах-производителях

2024 — Мода на чай и чайные напитки

15. Скачать датасет — [файл](#)

Вам представлен набор данных о кассовых сборах фильмов, а также дополнительная информация о каждом фильме. Кинокомпании используют такие данные для прогнозирования успеха новых проектов. В заданиях ниже вам предстоит исследовать взаимосвязь между различными факторами.

Вопрос 1. Рассчитайте следующие показатели для фильмов, произведенных во Франции:

- Медианный бюджет фильмов.
- Средний рейтинг фильмов и стандартное отклонение. Ответ округлите до сотых.
- Доля прибыльных фильмов ($\text{gross} > \text{budget}$) в группе с бюджетом ниже медианы. Ответ округлите до процентов. В расчет берите только фильмы, у которых есть оба показателя.
- Компания с наибольшей средней прибылью на фильм. Прибыль можно рассчитать, только если есть данные и по бюджету, и по сборам.

Вопрос 2. Выполните корреляционный анализ для фильмов, произведенных во Франции.

Рассчитайте и округлите ответы до сотых:

- Коэффициент корреляции между бюджетом и сборами.
- Коэффициент корреляции между рейтингом и прибылью.

16. Вы разрабатываете модель для автоматической модерации комментариев на платформе. Модель должна уметь:

- Находить все оскорбительные сообщения (минимизировать пропуски).
- Не блокировать нормальные комментарии (минимизировать ложные срабатывания).

Для вас протестировали три модели с разными характеристиками:

Модель	TP (Заблокировано оскорбительных)	FP (Заблокировано нейтральных)	TN (Пропущено нейтральных)	FN (Пропущено оскорбительных)
A	219	48	903	53
Б	303	197	764	28
В	273	61	870	75

Какую модель следует выбрать для детской платформы, если более важно минимизировать количество оскорбительных комментариев?

- 1) А
- 2) Б
- 3) В

Какую модель следует выбрать для соцсети, если нужно минимизировать число жалоб и известно, что:

Каждая ложная блокировка (FP) приводит к 1 жалобе пользователя.
Каждый пропущенный оскорбительный комментарий (FN) — к 4 жалобам.

- 1) А
- 2) Б
- 3) В

17. Скачать датасет — [файл](#)

Вам предоставлен датасет с логами действий пользователей в мобильном приложении.

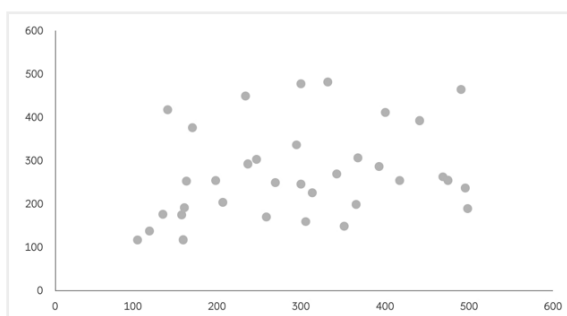
Выделите сессии пользователя по правилу: если более 15 минут не было никаких действий, создается новая сессия. Для каждого пользователя пронумеруйте сессии, используя это правило.

Найдите пользователя с максимальным числом сессий.

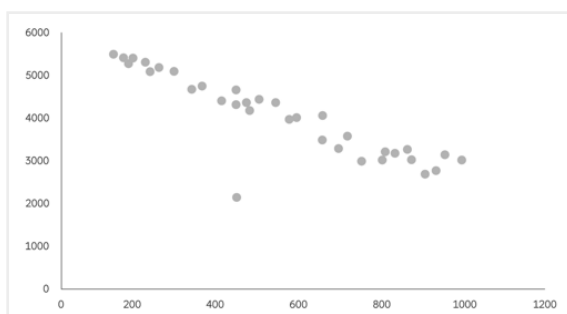
В ответе запишите последнее действие пользователя с максимальным количеством сессий.

18. Соотнесите коэффициент корреляции и диаграмму рассеяния.

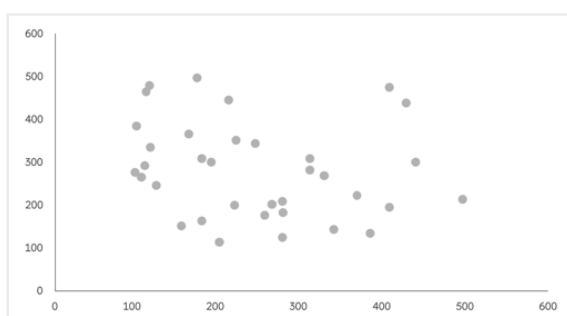
1)



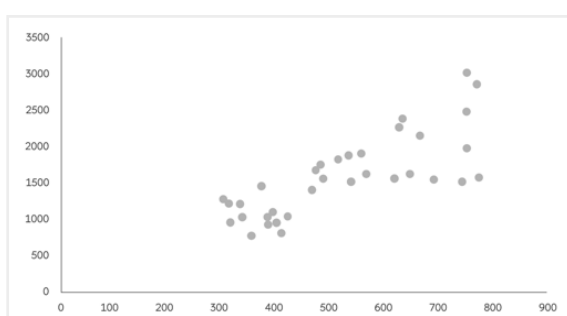
2)



3)



4)



A. 0.26

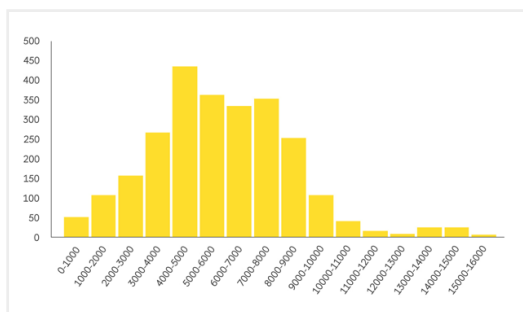
Б. -0.93

В. -0.21

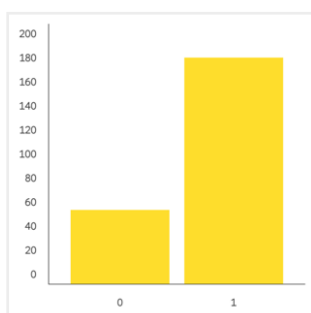
Г. 0.72

19. Вы редактировали журнал «Наглядная статистика» и обнаружили, что заголовки явно не соответствуют графикам. Соотнесите график и его правильный заголовок:

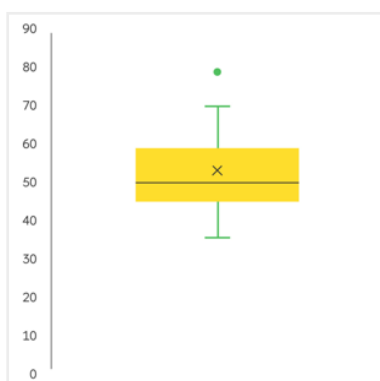
1)



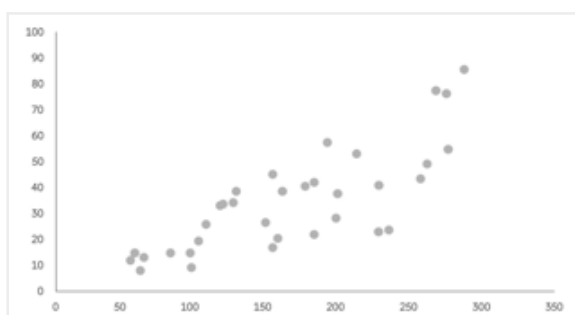
2)



3)



4)



А. Средний чек в театре

Б. Распределение школьников по итогам тура олимпиады (прошел в финал / не прошел в финал)

В. Вес учащихся 9 класса

Г. Стоимость загородного дома (в млн рублей) в зависимости от его площади

20. В парке развлечений два самых популярных аттракциона — «Космическая гонка» и «Супергигант». Они абсолютно одинаковы по своим характеристикам (спрос, время ожидания, длина очереди и т. д.). Цена билета на каждый — 350 ₽ в любое время. Среднее число посетителей в месяц — 15 000 на каждый аттракцион.

Если клиенты попадают на аттракцион во время пиковой нагрузки, им приходится ждать в очереди 90 минут. Обычное время ожидания — 30 минут.

Распределение посещаемости на каждом аттракционе выглядит так:

- Пиковая нагрузка (25% времени): 7 000 посетителей.
- Обычное время (75% времени): 8 000 посетителей.

Чтобы оптимизировать доход и управлять спросом, администрация хочет сравнить две стратегии ценообразования:

- «Космическая гонка» — фиксированный повышенный тариф (500 ₽ всегда).
- «Супергигант» — динамический тариф: обычное время — 400 ₽, час пик — 600 ₽.

Если время ожидания превышает 60 минут, парк выдает клиенту ваучер на напиток стоимостью 150 ₽ и на хот-дог стоимостью 250 ₽, при этом вероятность приобретения напитка за ваучер — 80%, а хот-дога — 70%.

Показатель	Было (350 ₽ всегда, «Космическая гонка» и «Супергигант» вместе)	«Космическая гонка» (500 ₽ всегда)	«Супергигант» (400/600 ₽)
Цена в час пик	350	500	600
Цена в обычное время	350	500	400
Прокатившихся в час пик	14 000	6 500	6 200
Прокатившихся в обычное время	16 000	7 400	7 600
Среднее время ожидания (час пик)	90 мин., равномерно распределено от 60 до 120 мин.	60 мин., равномерно распределено от 50 до 70 мин.	40 мин., равномерно распределено от 25 до 55 мин.
Время ожидания (обычное время)	30 мин.	15 мин.	20 мин.

Вопрос 1. Какой вариант тарификации оптимальный с точки зрения дохода?

- 1) Динамический
- 2) Фиксированный повышенный

Сколько денег дополнительно получит парк по сравнению со старым тарифом, если внедрить его на оба аттракциона?

Вопрос 2. Какой вариант новой тарификации следует выбрать для обоих аттракционов при уровне значимости 95%? Укажите вариант тарифа:

- 1) Фиксированный повышенный тариф
- 2) Динамический тариф
- 3) Разницы нет

Вопрос 3. А при уровне значимости 99%?

- 1) Динамический тариф
- 2) Разницы нет
- 3) Фиксированный повышенный тариф

Вопрос 4. Укажите Z статистику, используемую для принятия решения.

Ответ округлите до сотых. При расчетах не округляйте промежуточные результаты до получения финального ответа.

Вариант 3

- Команда хакеров пытается найти уязвимости в сети. 95% успешных атак происходят с серверов, находящихся в облаке Cloud-182, а 5% — с серверов в других облаках. Известно, что в этой сети 95% всех серверов размещены в Cloud-182.

Какой вывод верен?

- Атаковать серверы с Cloud-182 эффективнее
 - Атаковать серверы не с Cloud-182 эффективнее
 - В Cloud-182 лучше защищают серверы
 - Вероятность успешной атаки не зависит от расположения сервера
- Фермер собрал урожай с двух полей: на одном росла только пшеница, на другом — только рожь. При транспортировке часть зерна на склад его случайно смешали с прошлогодним урожаем более низкого качества. После обнаружения инцидента провели анализ, который показал, что 98% от всей пшеницы — это свежее зерно, а 5% от всей ржи — старое зерно.

Известно, что пшеницы фермер собрал в 4 раза больше, чем ржи.

Какая доля всего зерна на складе оказалась старого урожая? Ответ запишите в процентах и округлите до десятых.

- Для запуска межпланетного зонда ученым необходимо ввести код инициализации. Из технического мануала известно:
 - код состоит из четырех символов, каждый из которых является цифрой (от 0 до 9);
 - используются только простые числа и ноль (2, 3, 5, 7, 0), причем каждая цифра встречается ровно один раз;
 - сумма двух левых цифр кода должна быть равна сумме двух правых цифр.

Сколько существует различных кодов, подходящих для запуска?

- В банке работают три независимые команды аналитиков по 40 человек. Среди сотрудников каждой команды 20% — стажеры, 30% — начинающие сотрудники и 50% — опытные аналитики. Стажер выполняет одну задачу за час, начинающий сотрудник — за 45 минут, опытный сотрудник — за 15 минут.

Коллеги могут делить задачи между собой, то есть каждому сотруднику может быть поручено выполнение не только целой задачи (нескольких целых задач), но и частей каких-то задач (например, половины задачи или десятой части задачи и т. д.).

Рабочий день каждой команды начинается, когда все сотрудники одновременно приступают к выполнению своих задач, при этом они работают одновременно, без перерывов, и рабочий день команды считается законченным, когда последний сотрудник выполнит свою последнюю задачу.

В первой команде все задачи распределяются равномерно между всеми сотрудниками — чтобы все сделали равное число задач.

Во второй команде каждый опытный сотрудник выполняет по 7 задач, а оставшиеся задачи делятся равномерно между оставшимися сотрудниками.

В третьей команде задачи распределяются таким образом, чтобы все сотрудники закончили выполнять свои задачи одновременно.

Сегодня каждой команде поступило 260 задач. Сколько часов будет длиться рабочий день каждой команды? Ответы округлите до десятых.

5. Инженер по качеству собрал данные о массе готовых шоколадных батончиков (в граммах). Гистограмма по этим данным показала, что распределение массы совпадает с нормальным. Среднее значение — 50 грамм, стандартное отклонение — 2 грамма.

Какова вероятность, что случайно выбранный батончик весит меньше 46 грамм?

- 1) 0.023
- 2) 0.16
- 3) 0
- 4) 0.32

Если инженер 30 раз подряд возьмет 36 случайных батончиков и усреднит их массу, то как будет распределено среднее значение?

- 1) Нормально с $\mu = 50$, $\sigma = 0.333$
- 2) Нормально с $\mu = 50$, $\sigma = 2$
- 3) Равномерно~Экспоненциально

Инженер также проверил время изготовления одного батончика на конвейере. Оказалось, изготовление занимает от 5 до 10 секунд и любое время в этом интервале равновероятно. Какое это распределение?

- 1) Равномерное
- 2) Нормальное
- 3) Экспоненциальное
- 4) Биномиальное

6. Стриминговый сервис хочет доработать алгоритм, который определяет, понравится ли пользователю фильм (лайк/дизлайк). Для этого алгоритм анализирует его прошлые оценки. Важно, чтобы прогноз алгоритма максимально часто совпадал с фактической реакцией пользователя. Одинаково плохо, если система порекомендует неинтересный фильм (пользователь потратит время впустую) или пропустит интересный (пользователь не получит ценности и будет реже пользоваться сервисом).

Модель может выдавать следующие результаты:

True Positive (TP) — фильм пользователю понравится, и модель правильно это предсказала (рекомендовала его).

True Negative (TN) — фильм пользователю не понравится, и модель правильно это предсказала (не рекомендовала его).

False Negative (FN) — фильм пользователю понравится, но модель ошиблась и не рекомендовала его (упущенная возможность).

False Positive (FP) — фильм пользователю не понравится, но модель ошиблась и рекомендовала его (неудачная рекомендация).

С помощью этих результатов можно посчитать следующие метрики:

Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.

Recall — доля верно порекомендованных фильмов $TP / (TP + FN)$.

Precision — доля верных рекомендаций $TP / (TP + FP)$.

F1-score — гармоническое среднее между Recall и Precision $2 \times ((Precision \times Recall) / (Precision + Recall))$.

Какую метрику нужно оптимизировать?

- 1) Accuracy
- 2) Recall
- 3) Precision
- 4) F1-score

7. При разработке спам-фильтра для почтового сервиса важно, чтобы пользователи не пропускали важные письма, ошибочно помеченные как спам.

True Positive (TP) — письмо является спамом, модель правильно это определила.

True Negative (TN) — письмо не спам, и модель это подтвердила.

False Negative (FN) — письмо является спамом, но модель пропустила письмо.

False Positive (FP) — письмо не спам, но модель ошибочно пометила его как спам.

С помощью этих результатов можно посчитать следующие метрики:

Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.

Recall — доля верно обнаруженного спама $TP / (TP + FN)$.

Precision — доля истинно положительных срабатываний $TP / (TP + FP)$.

F1-score — гармоническое среднее между Recall и Precision $2 \times ((Precision \times Recall) / (Precision + Recall))$.

Значения параметров модели: $FP = 20$, $FN = 70$, $TP = 80$, $TN = 80$.

Чему равен Recall? Ответ округлите до сотых.

8. Тимлид руководит коллективом из 70 человек: каждый понедельник он выдает всем сотрудникам по одной сопоставимой по сложности задаче. Задачу необходимо выполнить до конца рабочей недели — то есть за 5 рабочих дней, в каждом из которых по 8 рабочих часов. Если сотрудник заканчивает выполнение своей задачи раньше — он может не работать оставшиеся дни и оставшееся время в день закрытия задачи, момент окончания работы над задачей фиксируется. Тимлид решил проверить, что влияет на продуктивность его сотрудников. Он собрал данные за одну неделю работы всех его сотрудников:

days — время в 8-часовых рабочих днях, за которое была выполнена его задача (чем быстрее были выполнены задачи, тем продуктивнее сотрудник), зависимая переменная.

age — возраст.

gender — пол сотрудника (gender = 1 для мужчин, gender = 0 для женщин).

study — количество лет обучения в аспирантуре (высшая ступень образования).

call — количество рабочих созвонов, на которых присутствует сотрудник.

online — удаленная работа (online = 1, если сотрудник работает удаленно; online = 0, если сотрудник работает из офиса).

По собранным данным тимлид оценил линейную регрессию с days в качестве зависимой переменной, все коэффициенты оказались статистически значимыми, результаты представлены в таблице.

Переменная	Оценка коэффициента
age	−0.6
age2	0.01
gender	−0.9
study	−0.5
call	0.25
online	0.9
gender * online	−0.3
constant	4.5

Заполните пропуски или выберите верные варианты в утверждениях о коллективе тимлида, которые следуют из результатов оцененной им регрессии (каждый вывод ниже делается при прочих равных, то есть в предположении неизменных остальных переменных):

- A. 30 лет — это возраст, при котором сотрудники наиболее продуктивны
- B. Более высокую производительностью показывают мужчины/женщины
- C. Каждый дополнительный год обучения в аспирантуре позволяет сотруднику выполнять задачу на 4 часов / часа быстрее
- D. Каждый дополнительный рабочий созвон увеличивает время выполнения задачи на 2 часов / часа
- E. Если сотрудник работает удаленно, то его продуктивность уменьшается/увеличивается
- F. Удаленная работа сильнее/слабее влияет на продуктивность женщин, чем на продуктивность мужчин

9. В день проведения крупнейшей киберспортивной турнирной битвы стриминговые платформы зафиксировали десятикратный всплеск зрителей по сравнению с обычными днями. Это пример:

- 1) Тренда
- 2) Сезонности
- 3) Шума
- 4) Аномалии

10. Зоопарк хочет оптимизировать закупку корма для животных. Вам нужно собрать модель, которая предсказывает необходимую суточную порцию корма (в кг) для животного на основе его характеристик.

В вашем распоряжении есть датасет со следующими признаками:

weight — вес животного (в кг).

age — возраст животного (в годах).

activity_level — уровень активности (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).

metabolic_rate — уровень метаболизма (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).

ambient_temperature — средняя температура в вольере (в °C).

Целевая переменная daily_food_intake — суточная норма корма (в кг).

Вы построили линейную регрессию на данных о млекопитающих и получили следующие коэффициенты:

intercept (свободный член) = 1.2.

Коэффициент для weight = 0.01.

Коэффициент для age = -0.05.

Коэффициент для activity_level = 0.1.

Коэффициент для metabolic_rate = 0.15.

Коэффициент для ambient_temperature = -0.02.

Вопрос 1. Какую суточную норму корма (в кг) предскажет модель для панды со следующими параметрами? Ответ округлите до сотых.

weight = 100 кг

age = 4 года

activity_level = 1 (низкий)

metabolic_rate = 1 (низкий)

ambient_temperature = 20 °C

Вопрос 2. Отдел закупок должен заказать корм на месяц (30 дней) для этой панды, основываясь на прогнозе модели. Известно, что модель на тестовых дневных данных давала прогнозы с MSE = 0.16.

Чтобы избежать ситуации нехватки корма, отдел закупок решил заказать количество корма, которое с вероятностью 95% покроет потребности животного с учетом погрешности модели.

Какой минимальный запас корма (в кг) необходимо заказать, если за основу взять типичную величину ошибки прогноза? Учтите, что заказывать можно только целое число килограммов корма. Распределение расходов корма считайте нормальным, ошибки в разные дни независимы.

11. Сервис такси хочет предсказать стоимость поездки (в рублях). В распоряжении есть исторические данные.

Признаки:

duration — время поездки (мин.).

distance — расстояние (км).

surge_multiplier — коэффициент повышенного спроса (например, 1.0x, 1.5x, 2.0x).

Целевая переменная: trip_cost — стоимость поездки (руб.).

После построения линейной регрессии получилось следующее уравнение:

$$\text{trip_cost} = 50 + 15 \times \text{duration} + 25 \times \text{distance} + 100 \times \text{surge_multiplier}$$

Все коэффициенты значимы на 5% уровне значимости. Вопрос: как интерпретировать коэффициент 100 для признака surge_multiplier?

- 1) При увеличении коэффициента спроса на 1.0 стоимость поездки увеличивается в среднем на 100 ₺
- 2) При увеличении коэффициента спроса на 100% стоимость поездки увеличивается в среднем на 1 ₺
- 3) Коэффициент спроса определяет 100% стоимости поездки
- 4) При увеличении коэффициента спроса в четыре раза стоимость поездки увеличится в четыреста раз

12. Магазин электроники тестирует скидки на товар с себестоимостью 5 000 ₺:

- A. 5%: +15% продаж.
- B. 10%: +22% продаж.
- B. 15%: +28% продаж.

Какая скидка дает максимальную прибыль, если наценка на товар до скидки составляет 100%?

13. Скачать датасет — [файл](#)

Университет собрал данные об обучении за последний семестр:

Student_id — ID студента.

Gender — пол студента.

Type — тип обучения (local / exchange).

Age — возраст студента.

Department — направление обучения.

Attendance — посещаемость (в %).

Total_Score — общий балл.

Grade — оценка.

Study_Hours_per_Week — среднее количество часов обучения в неделю.

Scholarship — стипендия (да/нет).

Sleep_Hours_per_Night — среднее количество часов сна.

Вопрос 1. Различается ли общий балл между группой студентов, получающих стипендию, и группой студентов, не получающих стипендию? Рассчитайте модуль разницы средних и напишите, является ли она статистически значимой.

Вопрос 2. Различается ли доля студентов, приехавших по обмену, между группой студентов, получающих стипендию, и группой студентов, не получающих стипендию? Рассчитайте модуль разницы долей и напишите, является ли она статистически значимой.

Для всех вопросов $\alpha = 0.05$. Ответ округлите до сотых.

14. Перед вами график мировых цен на какао. Соотнесите возможные события с отмеченными точками на графике, которые могли стать причиной такого движения цены.



2020 — Закупка полезных шоколадных батончиков крупными корпорациями в США

2024 — Структурные проблемы и неблагоприятная погода в Западной Африке, которая поставляет 70% мирового объема какао

2025 — Поставка новой партии урожая

15. Скачать датасет — [файл](#)

Вам представлен набор данных о кассовых сборах фильмов, а также дополнительная информация о каждом фильме. Кинокомпании используют такие данные для прогнозирования успеха новых проектов. В заданиях ниже вам предстоит исследовать взаимосвязь между различными факторами.

Вопрос 1. Рассчитайте следующие показатели для фильмов, произведенных во Франции:

- Медианный бюджет фильмов.
- Средний рейтинг фильмов и стандартное отклонение. Ответ округлите до сотых.
- Доля прибыльных фильмов ($\text{gross} > \text{budget}$) в группе с бюджетом ниже медианы. Ответ округлите до процентов. В расчет берите только фильмы, у которых есть оба показателя.
- Компания с наибольшей средней прибылью на фильм. Прибыль можно рассчитать, только если есть данные и по бюджету, и по сборам.

Вопрос 2. Выполните корреляционный анализ для фильмов, произведенных во Франции.

Рассчитайте и округлите ответы до сотых:

- Коэффициент корреляции между бюджетом и сборами.
- Коэффициент корреляции между рейтингом и прибылью.

16. Вы разрабатываете модель для автоматической модерации комментариев на платформе. Модель должна уметь:

- Находить все оскорбительные сообщения (минимизировать пропуски).
- Не блокировать нормальные комментарии (минимизировать ложные срабатывания).

Для вас протестировали три модели с разными характеристиками:

Модель	TP (Заблокировано оскорбительных)	FP (Заблокировано нейтральных)	TN (Пропущено нейтральных)	FN (Пропущено оскорбительных)
A	178	54	849	55
Б	246	102	792	93
В	191	77	905	44

Какую модель следует выбрать для детской платформы, если более важно минимизировать количество оскорбительных комментариев?

- 1) А
- 2) Б
- 3) В

Какую модель следует выбрать для соцсети, если нужно минимизировать число жалоб и известно, что:

Каждая ложная блокировка (FP) приводит к 1 жалобе пользователя.
Каждый пропущенный оскорбительный комментарий (FN) — к 4 жалобам.

- 1) А
- 2) Б
- 3) В

17. Скачать датасет — [файл](#)

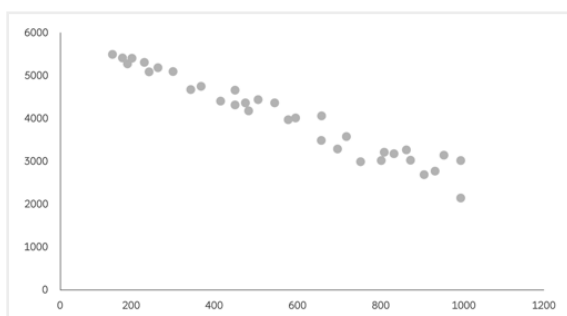
Вам предоставлен датасет с логами действий пользователей в мобильном приложении.

Выделите сессии пользователя по правилу: если более 40 минут не было никаких действий, то создается новая сессия. Для каждого пользователя пронумеруйте сессии, используя это правило.

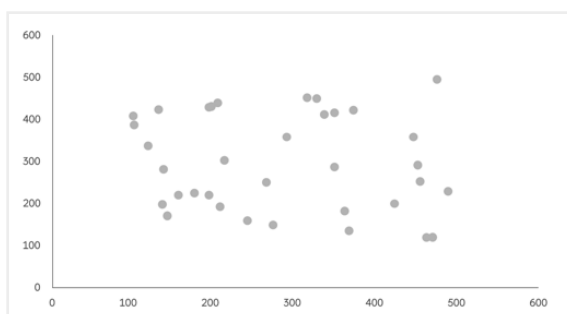
Найдите пользователя с максимальным числом сессий. В ответе запишите локацию, из которой был первый заход пользователя с максимальным количеством сессий.

18. Соотнесите коэффициент корреляции и диаграмму рассеяния.

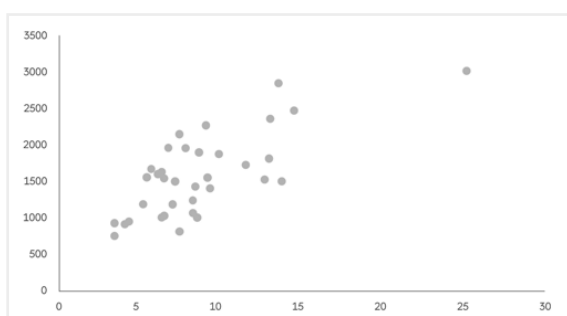
1)



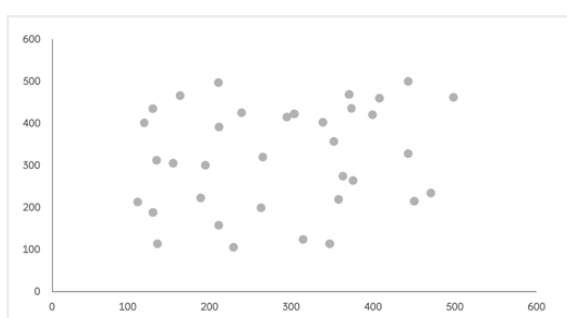
2)



3)



4)



A. -0.97

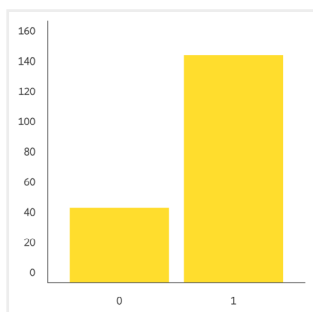
Б. -0.12

В. 0.69

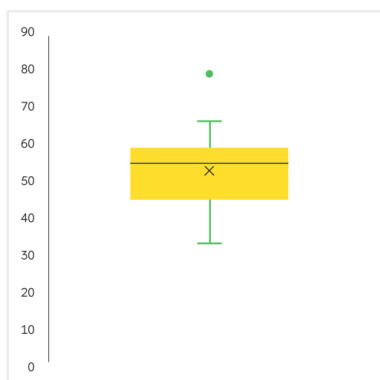
Г. 0.18

19. Вы редактировали журнал «Наглядная статистика» и обнаружили, что заголовки явно не соответствуют графикам. Соотнесите график и его правильный заголовок:

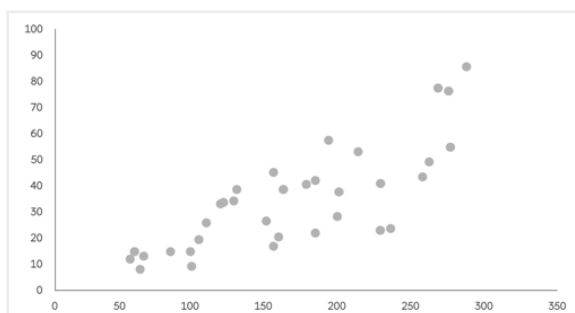
1)



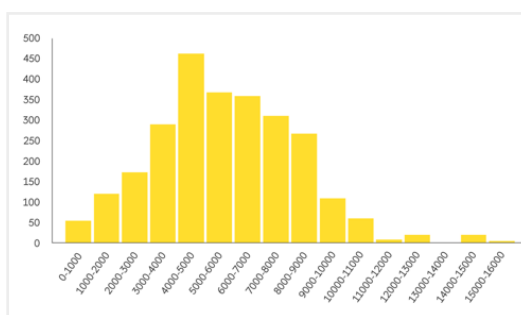
2)



3)



4)



- А. Распределение школьников по итогам тура олимпиады (прошел в финал / не прошел в финал)
- Б. Вес учащихся 9 класса
- В. Стоимость загородного дома (в млн рублей) в зависимости от его площади
- Г. Средний чек в театре

20. В парке развлечений два самых популярных аттракциона — «Космическая гонка» и «Супергигант». Они абсолютно одинаковы по своим характеристикам (спрос, время ожидания, длина очереди и т. д.). Цена билета на каждый — 300 ₽ в любое время. Среднее число посетителей в месяц — 11 000 на каждый аттракцион.

Если клиенты попадают на аттракцион во время пиковой нагрузки, им приходится ждать в очереди 80 минут. Обычное время ожидания — 20 минут.

Распределение посещаемости на каждом аттракционе выглядит так:

- Пиковая нагрузка (20% времени): 4 000 посетителей.
- Обычное время (80% времени): 7 000 посетителей.

Чтобы оптимизировать доход и управлять спросом, администрация хочет сравнить две стратегии ценообразования:

- «Космическая гонка» — фиксированный повышенный тариф (400 ₽ всегда).
- «Супергигант» — динамический тариф: обычное время — 350 ₽, час пик — 550 ₽.

Если время ожидания превышает 60 минут, парк выдает клиенту ваучер на напиток стоимостью 100 ₽ и на хот-дог стоимостью 200 ₽, при этом вероятность приобретения напитка за ваучер — 90%, а хот-дога — 60%.

Показатель	Было (300 ₽ всегда, «Космическая гонка» и «Супергигант» вместе)	«Космическая гонка» (400 ₽ всегда)	«Супергигант» (350/550 ₽)
Цена в час пик	300	400	550
Цена в обычное время	300	400	350
Прокатившихся в час пик	8 000	3 600	3 200
Прокатившихся в обычное время	14 000	6 200	6 500
Среднее время ожидания (час пик)	80 мин., равномерно распределено от 70 до 90 мин.	60 мин., равномерно распределено от 40 до 80 мин.	30 мин., равномерно распределено от 20 до 40 мин.
Время ожидания (обычное время)	25 мин.	15 мин.	20 мин.

Вопрос 1. Какой вариант тарификации оптимальный с точки зрения дохода?

- 1) Динамический
- 2) Фиксированный повышенный

Сколько денег дополнительно получит парк по сравнению со старым тарифом, если внедрить его на оба аттракциона?

Вопрос 2. Какой вариант новой тарификации следует выбрать для обоих аттракционов при уровне значимости 95%? Укажите вариант тарифа:

- 1) Фиксированный повышенный тариф
- 2) Динамический тариф
- 3) Разницы нет

Вопрос 3. А при уровне значимости 99%?

- 1) Динамический тариф
- 2) Разницы нет
- 3) Фиксированный повышенный тариф

Вопрос 4. Укажите Z статистику, используемую для принятия решения.

Ответ округлите до сотых. При расчетах не округляйте промежуточные результаты до получения финального ответа.

Разбор вопросов

№ 1

Лара Крофт исследует древние храмы, где 80% ловушек срабатывают в помещениях с мозаикой на полу, а 20% срабатывают в помещениях без мозаики. Известно, что в этих храмах мозаика встречается в 30% помещений. Какой вывод верен?

- а) Вероятность встретить ловушку в комнате с мозаикой выше, чем в комнатах без мозаики
- б) Ловушки не размещают в комнатах без мозаики
- в) Наличие мозаики не влияет на вероятность нахождения ловушки
- г) Данных недостаточно

Разбор задачи:

Согласно условию:

$$P(\text{срабатывание} \mid \text{мозаика}) = 0.8$$

$$P(\text{срабатывание} \mid \text{нет мозаики}) = 0.2$$

$$P(\text{мозаика}) = 0.3$$

Проверка утверждений:

- а) Вероятность встретить ловушку в комнате с мозаикой выше, чем в комнатах без мозаики

Сравниваем условные вероятности:

$$P(\text{срабатывание} \mid \text{мозаика}) = 0.8$$

$$P(\text{срабатывание} \mid \text{нет мозаики}) = 0.2$$

$$0.8 > 0.2 \rightarrow \text{утверждение верно.}$$

- б) Ловушки не размещают в комнатах без мозаики

Это означало бы $P(\text{срабатывание} \mid \text{нет мозаики}) = 0$, но по условию она равна 0.2 $\rightarrow$ ловушки есть $\rightarrow$ утверждение неверно.

- в) Наличие мозаики не влияет на вероятность нахождения ловушки

Если бы не влияло, $P(\text{срабатывание} \mid \text{мозаика}) \neq P(\text{срабатывание} \mid \text{нет мозаики})$.

Но эти вероятности по условию не равны, значит, утверждение неверно.

г) Данных недостаточно

Для ответа на вопрос данных достаточно

Ответ: а № 2

В музее хранились только подлинные предметы. Воры украли некоторые экспонаты и заменили их фальшивыми. После кражи 90% от всех золотых предметов остались подлинными, а 20% от всех керамических предметов — фальшивыми.

Изначально в музее хранились только золотые и керамические предметы, причем золотых предметов было в 3 раза больше, чем керамических.

Какая доля музейных экспонатов оказалась фальшивой? Ответ запишите в процентах и округлите до десятых.

Разбор задачи:

Введем переменные. Пусть x — количество керамических предметов. Тогда золотых предметов — $3x$.

Общее количество предметов: $3x + x = 4x$.

Найдем количество фальшивых предметов:

Фальшивых золотых: 10% от $3x = 0.1 \cdot 3x = 0.3x$.

Фальшивых керамических: 20% от $x = 0.2 \cdot x = 0.2x$.

Общее количество фальшивых: $0.3x + 0.2x = 0.5x$.

Найдем долю фальшивых от общего числа:

Доля = (Количество фальшивых) / (Общее количество) = $(0.5x) / (4x) = 0.5 / 4 = 0.125 = 12.5\%$.

Ответ: 12.5%

№ 3

Лара взламывает сейф с украденными артефактами. Она знает:

- пароль к сейфу состоит из четырех цифр (от 0 до 9);
- для него используются только четные цифры, ни одна из них не встречается дважды, а сумма первых двух цифр равна сумме последних двух.

Сколько возможных комбинаций у такого замка?

Разбор задачи:

Всего чётных цифр 5. Нужно отобрать такие числа, которые давали бы одинаковые

суммы и при этом не повторялись. Методом перебора получим, что одинаковые суммы из разных чисел дают следующие пары:

- 0 и 6, 2 и 4
- 0 и 8, 2 и 6
- 2 и 8, 4 и 6

Каждая пара дает 8 комбинаций. На примере первой пары:

- 0624
- 6024
- 0642
- 6042
- 2406
- 2460
- 4206
- 4260

Аналогично для других наборов чисел. Так как возможных вариантов наборов чисел всего три, то общее количество комбинаций будет $8 \times 3 = 24$

Ответ: 24

№ 4

В банке работают три независимые команды аналитиков по 50 человек.

Среди сотрудников каждой команды 20% — стажеры, 30% — начинающие сотрудники и 50% — опытные аналитики. Стажер выполняет одну задачу за час, начинающий сотрудник — за 45 минут, опытный сотрудник — за полчаса. Коллеги могут делить задачи между собой, то есть каждому сотруднику может быть поручено выполнение не только целой задачи (нескольких целых задач), но и частей каких-то задач (например, половины задачи или десятой части задачи и т. д.).

Рабочий день каждой команды начинается, когда все сотрудники одновременно приступают к выполнению своих задач, при этом они работают одновременно, без перерывов, и рабочий день команды считается законченным, когда последний сотрудник выполнит свою последнюю задачу.

В первой команде все задачи распределяются равномерно между всеми сотрудниками — чтобы все сделали равное число задач.

Во второй команде каждый опытный сотрудник выполняет по 6 задач, а оставшиеся задачи делятся равномерно между оставшимися сотрудниками.

В третьей команде задачи распределяются так, чтобы все сотрудники закончили выполнять свои задачи одновременно.

Сегодня каждой команде поступило 200 задач. Сколько часов будет длиться рабочий день каждой команды? Ответ запишите через точку с запятой в порядке возрастания

номеров команд.

Разбор задачи:

Команда 1: Равномерное распределение

Время работы будет определяться самым медленным сотрудником, то есть стажером.

При этом каждый сотрудник, в том числе стажер, получит $200/50 = 4$ задачи.

При скорости 1 задача в час стажер выполнит 4 задачи за 4 часа.

Команда 2: Опытные — по 6 задач, остальное — поровну другим

Количество опытных работников в команде составляет 50% от 50 человек, следовательно, в команде $50 \cdot 50\% = 25$ опытных сотрудников. Им достанется $25 \cdot 6 = 150$ задач, остальные 50 задач распределятся равномерно между оставшимися 25 сотрудниками. Следовательно, все начинающие сотрудники и стажеры получают по 2 задачи. Время работы в этой группе сотрудников также будет определяться самым медленным сотрудником. Стажер в этой команде сделает две задачи за два часа.

При этом опытным сотрудникам на 6 задач понадобится $6 \cdot 0.5ч = 3$ часа.

Значит, рабочий день команды будет длиться 3 часа.

Команда 3: Все заканчивают одновременно

За час опытный сотрудник выполнит 2 задачи, начинающий - $4/3$ задачи, стажёр - 1 задачу. Значит, при полной загрузке (чтобы никто не простаивал) 50 человек, из которых 10 стажёров, 15 начинающих и 25 опытных, за час способны выполнить: $2 \cdot 25 + 4/3 \cdot 15 + 1 \cdot 10 = 80$ задач.

На 200 задач у них уйдёт $200/80 = 2.5$ часа.

Ответ: 4; 3; 2.5

№ 5

Учительница собрала данные о времени, которое ученики 10-го класса тратят на выполнение домашнего задания по математике (в минутах). Гистограмма по этим данным показала, что распределение времени совпадает с нормальным. Среднее значение — 60 минут, стандартное отклонение — 15 минут.

Какова вероятность, что случайно выбранный ученик тратит на домашнее задание больше 75 минут?

- а) 0
- б) ~ 0.03
- в) ~ 0.16
- г) ~ 0.32

Если учительница 50 раз подряд возьмет 25 случайных учеников и усреднит их время выполнения домашнего задания по математике, то как будет распределено среднее значение?

- а) Нормально с $\mu = 60$, $\sigma = 15$

б) Нормально с $\mu = 60$, $\sigma = 3$

в) Равномерно

г) Экспоненциально

Учительница также собрала данные о времени выполнения домашнего задания по обществознанию. Оказалось, все ученики тратят на домашнее задание от 10 до 20 минут и любое время в этом интервале равновероятно. Какое это распределение?

а) Нормальное

б) Равномерное

в) Экспоненциальное

г) Биномиальное

Разбор задачи:

Стандартное отклонение (σ) — это мера «разброса» данных вокруг среднего. В нашей задаче это значение уже дано — 15 минут.

68% всех данных лежат в пределах одного стандартного отклонения от среднего, значит 68% учеников тратят на домашнюю работу $[60 - 15; 60 + 15] = [45; 75]$ —> от 45 до 75 минут.

Так как распределение симметрично, оставшиеся $100\% - 68\% = 32\%$ учеников делятся поровну между двумя «хвостам» колокола:

Левый хвост (те, кто делает быстрее 45 минут) — $\approx 16.6\%$.

Правый хвост (те, кто делает дольше 75 минут) — $\approx 16.6\%$.

Среднее будет равно 60, σ считается по формуле ($\sigma/\sqrt{n}$)

$$\sigma = 15 / \sqrt{25} = 15/5 = 3$$

Равную вероятность на любом участке интервала имеет только равномерное распределение

Ответ: в; б; б

№ 6

В задаче идентификации рака важно не пропустить больных.

Модель может выдавать следующие результаты:

- True Positive (TP) — пациент болен, модель правильно это определила.
- True Negative (TN) — пациент здоров, и модель это подтвердила.
- False Negative (FN) — пациент болен, но модель пропустила болезнь.
- False Positive (FP) — пациент здоров, но модель ошибочно определила, что он болен.

С помощью этих результатов можно посчитать следующие метрики:

- Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.
- Recall — доля верно обнаруженных реальных больных $TP / (TP + FN)$.
- Precision — доля истинно положительных срабатываний $TP / (TP + FP)$.
- F1-score — гармоническое среднее между Recall и Precision $2 \times ((Precision \times Recall) / (Precision + Recall))$.

Какую метрику нужно оптимизировать?

- а) Accuracy
- б) Recall
- в) Precision
- г) F1-score

Разбор задачи:

По условию задачи нам необходимо минимизировать False Negative. Именно Recall напрямую показывает, насколько хорошо модель находит всех больных, минимизируя пропуски (FN).

Ответ: б

№ 7

В задаче рекомендации фильмов важно предлагать пользователям только те фильмы, что им действительно понравятся.

Модель может выдавать следующие результаты:

- True Positive (TP) — фильм пользователю понравится, и модель правильно это предсказала (рекомендовала его).
- True Negative (TN) — фильм пользователю не понравится, и модель правильно это предсказала (не рекомендовала его).
- False Negative (FN) — фильм пользователю понравится, но модель ошиблась и не рекомендовала его (упущенная возможность).
- False Positive (FP) — фильм пользователю не понравится, но модель ошиблась и рекомендовала его (неудачная рекомендация).

С помощью этих результатов можно посчитать следующие метрики:

- Accuracy — общая доля верных прогнозов $(TP + TN) / (TP + TN + FP + FN)$.
- Recall — доля понравившихся пользователю фильмов, которые модель смогла обнаружить и рекомендовать $TP / (TP + FN)$.
- Precision — доля удачных рекомендаций среди всех рекомендованных моделью фильмов $TP / (TP + FP)$.
- F1-score — гармоническое среднее между Precision и Recall. $2 \times ((Precision \times Recall) / (Precision + Recall))$

(Precision + Recall)).

Значения параметров модели: $FP = 50$, $FN = 30$, $TP = 70$, $TN = 150$. Чему равен Accuracy?
Ответ округлите до сотых.

Разбор задачи:

Подставим значения в формулу Accuracy $(TP + TN) / (TP + TN + FP + FN)$:

$$(70 + 150) / (50 + 30 + 70 + 150) = 220 / 300 = 0.7333$$

Округлив это значение до сотых, получим 0.73

Ответ: 0.73

№ 8

Тимлид руководит коллективом из 50 человек: ежедневно он раздает всем сотрудникам одинаковое количество сопоставимых по сложности задач. Эти задачи необходимо выполнить до конца рабочего дня, если сотрудник заканчивает выполнение задач раньше — он может уйти домой, время окончания работы над задачами фиксируется. Тимлид решил проверить, что влияет на продуктивность его сотрудников. Он собрал данные за один день работы всех его сотрудников:

- hours — время в часах, за которое были выполнены все задачи (чем быстрее были выполнены задачи, тем продуктивнее сотрудник), зависимая переменная.
- age — возраст.
- gender — пол сотрудника (gender = 1 для мужчин, gender = 0 для женщин).
- experience — опыт работы в годах.
- coffee — количество перерывов на кофе за день.
- friend — присутствие друга за соседним рабочим столом (friend = 1, если рядом сидит друг; friend = 0, если рядом сидит просто коллега, а не друг).

По собранным данным тимлид оценил линейную регрессию с hours в качестве зависимой переменной и собрал результаты в таблицу. Все коэффициенты оказались статистически значимыми.

Переменная	Оценка коэффициента
age	-0.7
age ²	0.01
gender	-1.5
experience	-0.1
coffee	0.25
friend	2.1

gender * friend	-0.7
constant	5.2

Заполните пропуски или выберите верные варианты в утверждениях о коллективе тимлида, которые следуют из результатов оцененной им регрессии (каждый вывод ниже делается при прочих равных, то есть в предположении неизменных остальных переменных):

- A. _____ лет — это возраст, при котором сотрудники наиболее продуктивны
- B. Более высокую производительностью показывают **мужчины / женщины**
- C. Каждый дополнительный год опыта позволяет сотруднику выполнять задачи на _____ минут / минуты быстрее
- D. Каждый дополнительный перерыв на кофе увеличивает время выполнения задач на _____ минут / минуты
- E. Если рабочее место сотрудника находится рядом с другом, то его продуктивность **увеличивается / уменьшается**
- F. Присутствие друга рядом **сильнее / слабее** влияет на продуктивность женщин, чем на продуктивность мужчин

Разбор задачи:

Зависимость от возраста — квадратичная (в регрессии есть и возраст, и его квадрат), она выглядит как парабола ветвями вверх. Минимальные затраты времени (= максимальная продуктивность) достигаются в вершине параболы:
в $0.7 / 0.02 = 35$ лет.

При прочих равных, мужчина тратит на выполнение задач на 1.5 часа меньше, чем женщина. Значит, мужчины более производительны.

Каждый дополнительный год опыта уменьшает время выполнения на 0.1 часа = на 6 минут.

Каждый дополнительный перерыв на кофе увеличивает время выполнения на 0.25 часа = на 15 минут.

Присутствие друга увеличивает время выполнения задач — значит, продуктивность из-за присутствия друга уменьшается.

Присутствие друга увеличивает время выполнения задач на $2.1 - 0.7 * \text{gender}$, то есть, на 2.1 для женщин и на 1.4 - для мужчин. Значит, на продуктивность женщин присутствие рядом друга влияет сильнее.

Ответ: A. 35; B. Мужчины; C. 6; D. 15; E. уменьшается ; F. сильнее

№ 9

Продажи обогревателей зимой в 3 раза выше, чем летом. Это пример:

- а) Тренда

б) Сезонности

в) Шума

г) Аномалии

Разбор задачи:

Разберем каждый из терминов:

Тренд — это долгосрочное устойчивое направление изменения (рост или падение) данных. Не подходит, так как изменение является неустойчивым, зависит от сезона

Сезонность — регулярные и предсказуемые колебания, которые повторяются с фиксированным периодом (например, ежегодно). Верный ответ

Шум — случайные, несистематические колебания данных. Не подходит, так как такие колебания неслучайны

Аномалия — редкое, резкое отклонение от обычной модели данных. Для этого примера не подойдет, так как это стабильное и предсказуемое изменение

Ответ: б

№ 10

Зоопарк хочет оптимизировать закупку корма для животных. Вам нужно собрать модель, которая предсказывает необходимую суточную порцию корма (в кг) для животного на основе его характеристик.

В вашем распоряжении есть датасет со следующими признаками:

- weight — вес животного (в кг).
- age — возраст животного (в годах).
- activity_level — уровень активности (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).
- metabolic_rate — уровень метаболизма (категориальный признак: 1 — низкий, 2 — средний, 3 — высокий).
- ambient_temperature — средняя температура в вольере (в °C).
- Целевая переменная daily_food_intake — суточная норма корма (в кг).

Вы построили линейную регрессию на данных о млекопитающих и получили следующие коэффициенты:

- intercept (свободный член) = 0.5.
- Коэффициент для weight = 0.02.
- Коэффициент для age = -0.1.
- Коэффициент для activity_level = 0.3.

- Коэффициент для metabolic_rate = 0.4.
- Коэффициент для ambient_temperature = -0.05.

Разбор задачи:

Вопрос 1. Какую суточную норму корма (в кг) предскажет модель для молодого леопарда со следующими параметрами? Ответ округлите до сотых.

weight = 50 кг

age = 2 года

activity_level = 3 (высокий)

metabolic_rate = 3 (высокий)

ambient_temperature = 25 °C

Отдел закупок должен заказать корм на месяц (30 дней) для этого леопарда, основываясь на прогнозе модели. Известно, что модель на тестовых дневных данных давала прогнозы с $MSE = 0.25$.

Чтобы избежать ситуации нехватки корма, отдел закупок решил заказать количество корма, которое с вероятностью 95% покроет потребности животного с учетом погрешности модели.

Вопрос 2. Какой минимальный запас корма (в кг) необходимо заказать, если за основу взять типичную величину ошибки прогноза? Учтите, что заказывать можно только целое число килограммов корма. Распределение расходов корма считайте нормальным, ошибки в разные дни независимы.

Для ответа на первый вопрос подставим данные в условия значения в формулу:

$$\text{daily_food_intake} = 0.5 + 0.02 \cdot 50 + (-0.1) \cdot 2 + 0.3 \cdot 3 + 0.4 \cdot 3 + (-0.05) \cdot 25 = 0.5 + 1.0 - 0.2 + 0.9 + 1.2 - 1.25 = 2.15$$

Чтобы ответить на второй вопрос, нужно выполнить следующие шаги:

1. Посчитать среднее суммарное потребление за 30 дней:

$$\text{Mean_total} = 30 \cdot 2.15 = 64.5 \text{ кг}$$

2. Найти стандартную ошибку прогноза на один день:

$$\text{RMSE} = \sqrt{MSE} = \sqrt{0.25} = 0.5 \text{ кг}$$

3. Найти стандартную ошибку суммарного потребления за 30 дней. Поскольку ошибки независимы, дисперсии суммируются:

$$\text{Variance_total} = 30 \cdot (0.5)^2 = 30 \cdot 0.25 = 7.5$$

$$\text{SD_total} = \sqrt{\text{Variance}} = \sqrt{7.5} \approx 2.7386 \text{ кг}$$

4. Найти запас с учетом погрешности:

Z-значение для 95% вероятности ≈ 1.645 (односторонний квантиль нормального распределения)

$$\text{Total_food} = \text{Mean_total} + z * \text{SD_total}$$

$$\text{Total_food} = 64.5 + 1.645 * 2.7386 \approx 64.5 + 4.505 \approx 69.005 \text{ кг}$$

Так как мы можем заказывать только целое число килограммов, то это значение нужно округлить вверх. При округлении вниз вероятность покрыть потребности леопарда будет меньше 95%, и условие не выполнится.

Ответ: 2.15; 70

№ 11

Интернет-магазин хочет предсказать время выполнения заказа (в минутах) от момента оформления до момента доставки. В распоряжении есть исторические данные.

Признаки:

- product_count — количество товаров в заказе (шт.).
- distance — расстояние до клиента (км).
- is_weekend — выходной день (0 — нет, 1 — да).

Целевая переменная: delivery_time — время выполнения заказа (минуты).

После построения линейной регрессии получилось следующее уравнение:

$$\text{delivery_time} = 17 + 2,3 \times \text{product_count} + 7 \times \text{distance} + 14 \times \text{is_weekend}$$

Как интерпретировать коэффициент 7 для признака distance, если все коэффициенты значимы на 5% уровне значимости?

- а) Расстояние определяет 7% времени доставки
- б) При увеличении расстояния на 7 км время доставки увеличивается в среднем на 1 минуту
- в) При увеличении расстояния на 1 км время доставки увеличивается в среднем на 7 минут
- г) При увеличении расстояния в четыре раза время доставки увеличится в двадцать восемь раз

Разбор задачи:

- а) Неверно. Коэффициент 7 не является процентом.
- б) Неверно. Это обратная интерпретация. Коэффициент 7 означает, что при увеличении расстояния на 1 км время увеличивается на 7 минут, а не наоборот.
- в) Верно. При увеличении расстояния на 1 км время доставки увеличивается в среднем на 7 минут (при прочих равных условиях).
- г) Неверно. Модель линейная, поэтому при увеличении расстояния в 4 раза, время доставки увеличится на $7 * 4 = 28$ минут, а не в 28 раз.

Ответ: В № 12

Магазин тестирует скидки:

- А. 5%: +10% продаж.
- Б. 10%: +25% продаж.
- В. 15%: +30% продаж.

Какая скидка дает максимальную прибыль, если наценка — 300%?

Разбор задачи:

Нужно для каждого варианта скидки найти итоговую прибыль. Если наценка равна 300%, значит, себестоимость = 1, а полная цена товара: $1+3 = 4$.

Прибыль без скидки с одной единицы: $4 - 1 = 3$.

Общая формула прибыли на единицу:

Прибыль = (Цена со скидкой - Себестоимость) × Рост продаж

1. Скидка 5%:

Цена со скидкой: $4 \times 0.95 = 3.8$

Прибыль с 1 шт.: $3.8 - 1 = 2.8$

Рост продаж: 1.1

Относительная прибыль: $2.8 \times 1.1 = 3.08$

2. Скидка 10%:

Цена со скидкой: $4 \times 0.9 = 3.6$

Прибыль с 1 шт.: $3.6 - 1 = 2.6$

Рост продаж: 1.25

Относительная прибыль: $2.6 \times 1.25 = 3.25$

3. Скидка 15%:

Цена со скидкой: $4 \times 0.85 = 3.4$

Прибыль с 1 шт.: $3.4 - 1 = 2.4$

Рост продаж: 1.3

Относительная прибыль: $2.4 \times 1.3 = 3.12$

Скидка 10% дает наибольшую относительную прибыль, значит следует выбрать ее.

Ответ: Б

№ 13

Университет собрал данные об обучении за последний семестр:

- Student_id — ID студента.
- Gender — пол студента.
- Type — тип обучения (local / exchange).
- Age — возраст студента.
- Department — направление обучения.
- Attendance — посещаемость (в %).
- Total_Score — общий балл.
- Grade — оценка.
- Study_Hours_per_Week — среднее количество часов обучения в неделю.
- Scholarship — стипендия (да/нет).
- Sleep_Hours_per_Night — среднее количество часов сна.

Различается ли общий балл у студентов между местными студентами и студентами, приехавшими учиться по обмену? Рассчитайте модуль разницы средних и напишите, является ли она статистически значимой.

Различается ли доля студентов со стипендией между местными студентами и студентами, приехавшими учиться по обмену? Рассчитайте модуль разницы долей и напишите, является ли она статистически значимой.

Для всех вопросов $\alpha = 0.05$. Ответ округлите до сотых.

Разбор задачи:

Для ответа на первый вопрос нужно для начала разделить группы студентов на отдельные таблицы: в одной таблице будут местные студенты, в другой — студенты, приехавшие учиться по обмену. Затем найти средние значения с помощью формулы СРЗНАЧ. Либо изначально использовать СРЗНАЧЕСЛИ, и в условии к каждой группе использовать значения колонки Type.

Модуль разницы между этими значениями составляет 0.04.

Чтобы найти разницу доли студентов со стипендией и без между указанными группами, нужно определить долю студентов со стипендией в каждой группе. Это можно сделать функцией СЧЁТЕСЛИ/СЧЁТ, либо преобразовать значения в столбце Scholarship в 1/0 и найти среднее значение.

Модуль разницы между этими значениями также составит 0.04

Определить статзначимость можно с помощью функции ТТЕСТ на подготовленном датасете. При значении менее 0.05 (данный в условии уровень значимости) результат будет являться статистически значимым. При значении p value более 0.05 результат будет статистически незначимым.

Результат будет статзначимым для первого вопроса и нестатзначимым для второго вопроса.

Ответ: 0.04; нет; 0.04; да

№ 14

Перед вами график мировых цен на сахар. Соотнесите возможные события с отмеченными точками на графике, которые могли стать причиной такого движения цены.



1. 2021	а) Опасения по поводу сокращения производства в Бразилии
2. 2023	б) Богатый урожай в основных странах-изготовителях
3. 2025	в) Запрет экспорта на сахар из Индии — второго по величине экспортера сахара

Разбор задачи:

У нас на графике есть два участка с ростом цены (1 и 2) и один с падением (3). Разберем возможное влияние каждого события. Событие а) влияет на рост спроса, а следовательно, повышение цены. Событие б) влечет за собой рост предложения, а следовательно снижение цены. Событие в) сокращает предложение, и, следовательно, повышает цену на рынке.

Таким образом, у нас однозначно складывается пара 3б. Осталось разобраться, какое из событий влияет на цену сильнее. Здесь можно рассуждать следующими способами:

- В одном случае у нас описываются опасения, а в другом фактический запрет. Реальные мировые изменения, предположительно, больше повлияют на мировые цены, чем слухи и опасения
- При поиске новостей про экспорт сахара из Индии можно найти дату этого запрета, которая и укажет на нужный нам период — 2023

Ответ: 1а, 2в, 3б

№ 15

Вам представлен набор данных о кассовых сборах фильмов, а также дополнительная информация о каждом фильме. Кинокомпании используют такие данные для прогнозирования успеха новых проектов. В заданиях ниже вам предстоит исследовать взаимосвязь между различными факторами. (*dataset в папке*)

1. Рассчитайте следующие показатели для фильмов, произведенных в Канаде:
 - Медианный бюджет фильмов.
 - Средний рейтинг фильмов и стандартное отклонение. Ответ округлите до сотых.
 - Доля прибыльных фильмов ($\text{gross} > \text{budget}$) в группе с бюджетом ниже медианы. Ответ округлите до процентов ($0.174096 \rightarrow 17$). В расчет берите только фильмы, у которых есть оба показателя.
 - Компания с наибольшей средней прибылью на фильм. Прибыль можно рассчитать, только если есть данные и по бюджету, и по сборам.
2. Выполните корреляционный анализ для фильмов, произведенных в Канаде. Рассчитайте:
 - Коэффициент корреляции между бюджетом и сборами.
 - Коэффициент корреляции между рейтингом и прибылью.

Разбор задачи:

Ключевым шагом здесь будет собрать в новую таблицу только фильмы, которые относятся к указанной стране, иначе формулы, даже при активном фильтре, будут учитывать все значения таблицы.

- Медиану ищем с помощью формулы МЕДИАНА
- Средний рейтинг и стандартное отклонение — с помощью формул СРЗНАЧ и СТАНДОТКЛОН
- Чтобы найти долю прибыльных фильмов, нужно:
 - a. Посчитать прибыль каждого фильма, причем учесть, что обе переменные непустые
 - b. Найти количество прибыльных фильмов с помощью функции СЧЁТЕСЛИМН. Условия внутри формулы должны проверять, что бюджет фильма ниже найденной в первом пункте медианы, поле budget непустое, поле gross непустое, поле profit ($\text{gross} - \text{budget}$) больше нуля
 - c. Найти количество всех фильмов с помощью функции СЧЁТЕСЛИМН. Условия внутри формулы должны проверять, что бюджет фильма ниже найденной в первом пункте медианы, поле budget непустое, поле gross непустое
 - d. Разделить пункт b на пункт c
- Компанию с наибольшей прибылью ищем с помощью СРЗНАЧЕСЛИМН, где проверяем, что поле budget непустое, поле gross непустое, и что фильмы выпущены именно этой компанией. Среди полученных значений найти максимум и найти компанию, которой принадлежит максимальное значение

- Корреляцию можно найти с помощью формулы КОРРЕЛ

Ответ: 14000000; 6.13; 0.98; 55%; Twentieth Century Fox; 0.39; 0.17

№ 16

Вы разрабатываете модель для автоматической модерации комментариев на платформе. Модель должна уметь:

- Находить все оскорбительные сообщения (минимизировать пропуски).
- Не блокировать нормальные комментарии (минимизировать ложные срабатывания).

Для вас протестировали три модели с разными характеристиками:

Модель	ТР (Заблокировано оскорбительных)	FP (Заблокировано нейтральных)	TN (Пропущено нейтральных)	FN (Пропущено оскорбительных)
А	241	95	858	43
Б	270	238	715	14
В	199	29	924	85

1. Какую модель следует выбрать для детской платформы, если более важно минимизировать количество оскорбительных комментариев?
2. Какую модель следует выбрать для соцсети, если нужно минимизировать число жалоб и известно, что:
 - Каждая ложная блокировка (FP) приводит к 1 жалобе пользователя.
 - Каждый пропущенный оскорбительный комментарий (FN) — к 3 жалобам.

Разбор задачи:

В условии сказано, что нам нужно минимизировать количество оскорбительных комментариев, значит, нужно минимизировать пропуски оскорблений, то есть FN.

Считаем долю FN для каждой модели:

Модель А: $FN = 43 / 1237 = 0,0348$

Модель Б: $FN = 14 / 1237 = 0,0113$

Модель В: $FN = 85 / 1237 = 0,0687$

Наименьшая доля пропусков оскорбительных комментариев у модели Б

При выборе модели для соцсети нам нужно минимизировать число жалоб на блокировки, причем каждый FP дает 1 жалобу, каждый FN дает 3 жалобы.

Общие жалобы = $1 * FP + 3 * FN$

Считаем для каждой модели:

Модель А: $(1*95 + 3*43) / 1237 = (95 + 129) / 1237 = 0,181$ жалоб / пользователя

Модель Б: $(1 \cdot 238 + 3 \cdot 14) / 1237 = (238 + 42) / 1237 = 0,226$ жалоб / пользователя

Модель В: $(1 \cdot 29 + 3 \cdot 85) / 1237 = (29 + 255) / 1237 = 0,2296$ жалоб / пользователя

Наименьшее число жалоб у модели А

Ответ: Б; А

№ 17

Вам предоставлен датасет с логами действий пользователей в мобильном приложении.

Выделите сессии пользователя по правилу: если более 30 минут не было никаких действий, то создается новая сессия. Для каждого пользователя пронумеруйте сессии, используя это правило.

Найдите пользователя с максимальным числом сессий. В ответе запишите ID пользователя.

Разбор задачи:

Сначала нужно преобразовать поле с датой и временем. Для этого можно воспользоваться функциями ЛЕВСИМВ, ПРАВСИМВ и их комбинациями, чтобы достать год, месяц, день, часы, минуты и секунды события.

Затем отсортировать таблицу по юзеру и дате-времени логов юзера. Для каждого юзера пронумеровать сессии, проверяя разницу времени между новым событием и предыдущим событием. Если разница больше 30 минут, то прибавляем единицу к текущему количеству сессий.

После этого находим максимальное количество сессий на клиента (в этой задаче самое большое количество сессий - 8), и найти юзера, которому принадлежит это количество сессий. Это будет пользователь с id 44686.

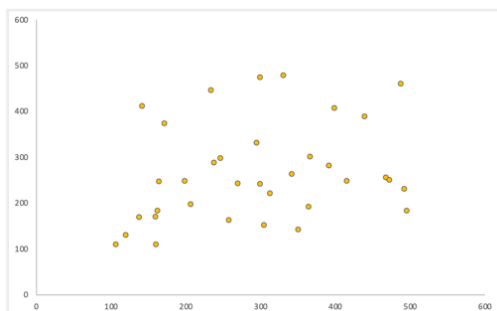
Ответ: 44686

№ 18

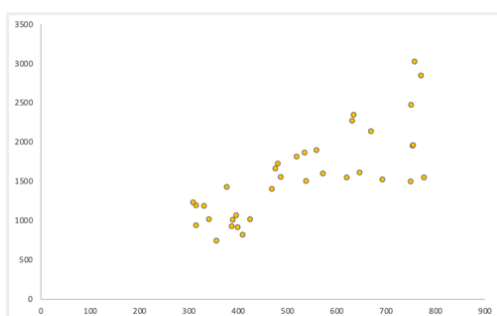
Соотнесите коэффициент корреляции и диаграмму рассеяния.

1. -0.97
2. 0.26
3. -0.12
4. 0.72

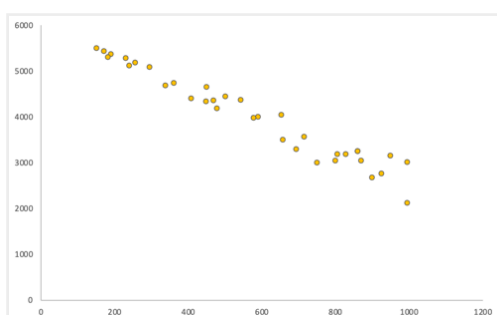
A)



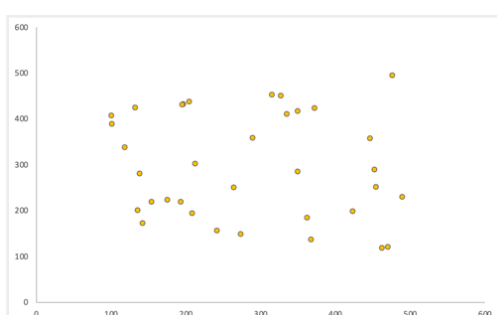
Б)



В)



Г)



Разбор задачи:

Чтобы соотнести график и коэффициент корреляции, нужно учитывать следующие особенности:

- Положительная корреляция имеет положительный наклон
- Чем коэффициент корреляции ближе к единице, тем более кучно расположены точки

Таким образом, коэффициенты 2 и 3 нам нужно распределить между А и Г, при этом на графике А общий наклон положительный, а на графике Г — отрицательный.

Коэффициенты 1 и 4 нужно распределить между графиками Б и В, при этом прямая на графике В имеет отрицательный наклон, а на графике Б — положительный

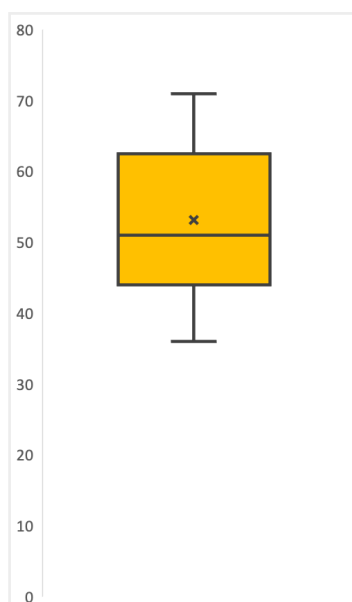
Ответ: 1В 2А 3Г 4Б

№ 19

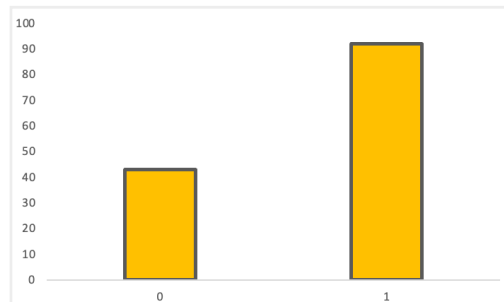
Вы редактировали журнал «Наглядная статистика» и обнаружили, что заголовки явно не соответствуют графикам. Соотнесите график и его правильный заголовок:

1. Средний чек в театре
2. Распределение школьников по итогам тура олимпиады (прошел в финал / не прошел в финал)
3. Стоимость загородного дома (в млн рублей) в зависимости от его площади
4. Вес учащихся 9 класса

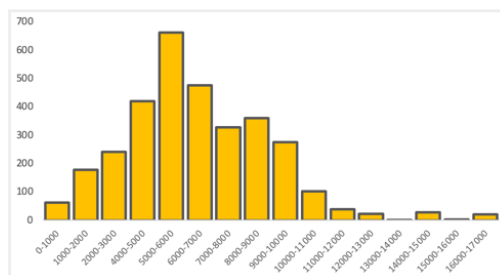
А)



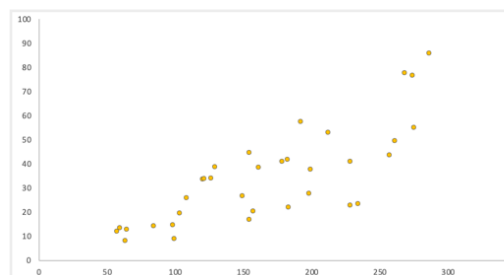
В)



Б)



Г)



Разбор задачи:

Распределим названия графиков между графиками.

1. Средний чек в театре — график А не подходит из-за низкий значений для стоимости билета в театр. Под такое описание подходят графики Б и Г. Посмотрим остальные пункты.

2. Распределение школьников по итогам тура олимпиады (прошел в финал / не прошел в финал) — биномиальное распределение, значит, нам подходит только график В
3. Стоимость загородного дома (в млн рублей) в зависимости от его площади. В теории мог бы подойти и график А, и график Г. Посмотрим на оставшийся пункт, чтобы методом исключения распределить оставшиеся пункты
4. Вес учащихся 9 класса — учитывая подписи осей, нам подходит только график А. Значения на графике Б слишком большие для килограммов и слишком маленькие для граммов, аналогично не подходят значения на графике Г.

Значит, для стоимости загородного дома остается Г, а для среднего чека — Б. Значения на этих график совпадают с реалистичными представлениями о распределении данных.

Ответ: 1Б 2В 3Г 4А

№ 20

В парке развлечений два самых популярных аттракциона — «Космическая гонка» и «Супергигант». Они абсолютно одинаковы по своим характеристикам (спрос, время ожидания, длина очереди и т. д.). Цена билета на каждый — 250 ₽ в любое время. Среднее число посетителей в месяц — 12 000 на каждый аттракцион.

Если клиенты попадают на аттракцион во время пиковой нагрузки, им приходится ждать в очереди 80 минут. Обычное время ожидания — 20 минут.

Распределение посещаемости на каждом аттракционе выглядит так:

- Пиковая нагрузка (30% времени): 5 000 посетителей.
- Обычное время (70% времени): 7 000 посетителей.

Чтобы оптимизировать доход и управлять спросом, администрация хочет сравнить две стратегии ценообразования:

- «Космическая гонка» — фиксированный повышенный тариф (400 ₽ всегда).
- «Супергигант» — динамический тариф: обычное время — 300 ₽, час пик — 500 ₽.

Если время ожидания превышает 60 минут, парк выдает клиенту ваучер на напиток стоимостью 100 ₽ и на хот-дог стоимостью 200 ₽, при этом вероятность приобретения напитка за ваучер — 90%, а хот-дога — 60%.

Показатель	Было (250 ₽ всегда, «Космическая гонка» и «Супергигант» вместе)	«Космическая гонка» (400 ₽ всегда)	«Супергигант» (300/500 ₽)
Цена в час пик	250	400	500
Цена в обычное время	250	400	300
Прокатившихся в час пик	10 000	4 700	4 200

Прокатившихся в обычное время	14 000	6 200	6 600
Среднее время ожидания (час пик)	80 мин., равномерно распределено от 60 до 100 мин.	60 мин., равномерно распределено от 40 до 80 мин.	40 мин., равномерно распределено от 20 до 60 мин.
Время ожидания (обычное время)	20 мин.	15 мин.	20 мин.

Если время ожидания превышает 60 минут, парк выдает клиенту ваучер на напиток стоимостью 100 ₽ и на хот-дог стоимостью 200 ₽, при этом вероятность приобретения напитка за ваучер — 90%, а хот-дога — 60%.

Вопрос 1. Какой вариант тарификации оптимальный с точки зрения дохода?

- а) Динамический
- б) Фиксированный повышенный

Сколько денег дополнительно получит парк по сравнению со старым тарифом, если внедрить его на оба аттракциона?

Анализ финансовой отчетности показал, что выручка от старого тарифа не покрывала ежемесячные фиксированные издержки на содержание аттракционов и парк терпел убытки. Оба новых тарифа позволяют выйти на положительный доход, но для администрации критически важно оставить конверсию в поездку максимальной.

За конверсию принимаем долю прокатившихся на аттракционах людей относительно спроса при старом тарифе.

Вопрос 2. Какой вариант новой тарификации следует выбрать для обоих аттракционов при уровне значимости 95%? Укажите вариант тарифа:

- а) Фиксированный повышенный тариф
- б) Динамический тариф
- в) Разницы нет

Вопрос 3. А при уровне значимости 99%?

- а) Фиксированный повышенный тариф
- б) Динамический тариф
- в) Разницы нет

Вопрос 4. Укажите Z статистику, используемую для принятия решения. Ответ округлите до сотых. При расчетах не округляйте промежуточные результаты до получения финального ответа.

Разбор задачи:

Вопрос 1. Посчитаем прибыль по каждому из тарифов:

Старый тариф (250₽ всегда)

Выручка с билетов:

$$(10\,000 \text{ чел.} + 14\,000 \text{ чел.}) \times 250 = 6\,000\,000 \text{ ₽}$$

Затраты на ваучеры:

$$\text{Напиток: } 100 \text{ ₽} \times 90\% = 90 \text{ ₽}$$

$$\text{Хот-дог: } 200 \text{ ₽} \times 60\% = 120 \text{ ₽}$$

Итого: 210 ₽ на человека, которому выдали ваучер.

Время ожидания в пик: равномерно от 60 до 100 мин, значит, все клиенты в час пик получают ваучеры

$$10\,000 \text{ чел.} \times 210 = 2\,100\,000 \text{ ₽}$$

Общая прибыль: 3 900 000 ₽

Фиксированный повышенный (400 ₽ всегда)

Выручка с билетов:

$$(4\,700 \text{ чел} + 6\,200 \text{ чел.}) \times 400 = 4\,360\,000 \text{ ₽}$$

Затраты на ваучеры:

Время ожидания в пик: равномерно от 40 до 80 мин, значит, половина клиентов в час пик получит ваучеры

$$(4\,700 \text{ чел} / 2) \times 210 = 2\,350 \times 210 = 493\,500 \text{ ₽}$$

Прибыль с аттракциона: 3 866 500 ₽

Динамический (пик 500 ₽, обычное 300 ₽)

Выручка с билетов:

$$4\,200 \text{ чел.} \times 500 + 6\,600 \text{ чел.} \times 300 = 2\,100\,000 \text{ ₽} + 1\,980\,000 \text{ ₽}$$

Затрат на ваучеры в этом случае нет.

Прибыль с аттракциона будет составлять 4 080 000 ₽.

Динамический тариф дает максимальную прибыль из всех предложенных вариантов тарификации. При масштабировании этого тарифа на оба аттракциона парк будет получать $4\,080\,000 \text{ ₽} \times 2 = 9\,160\,000 \text{ ₽}$

Разница между внедрением нового тарифа на оба аттракциона и старым тарифом составляет $9\,160\,000 \text{ ₽} - 3\,900\,000 \text{ ₽} = 4\,260\,000 \text{ ₽}$

Вопросы 2–4. Для ответа на эти вопросы нам нужно сравнить конверсии в двух вариантах новых тарифов.

При фиксированном повышенном тарифе на аттракционах прокатятся 10 900 человек из 12 000 максимально возможных. При динамическом тарифе аттракцион посетят 10 800 человек из максимально возможных 12 000. Рассчитаем Z значение для этих показателей по формуле:

$$(p_1 - p_2) / \sqrt{p_1(1-p_1) / N + p_2(1-p_2) / N}$$

Получим Z значение равное 2.18.

Чтобы понять статистическую значимость, нужно сравнить полученное значение с числом из таблицы напротив соответствующего уровня значимости.

Критические значения Z:

95% уровень значимости (двусторонний 1,96) → $Z > 1,96$ → разница значима на уровне 95%

99% уровень значимости (двусторонний 2,576) → $Z < 2,576$ → разница незначима на уровне 99%

Получается, что в первом случае нужно оставить фиксированный тариф, так как он дает большую конверсию. Во втором случае разница между конверсиями не является статистически значимой, но при этом динамический тариф дает большую прибыль, следовательно, во втором случае стоит выбрать динамический тариф

Ответ: а; 4260000; а; 6; 2.18/2.19