

Задания

1. Про время ожидания транспорта

Вася добирается до школы в несколько этапов. Выходя из дома, он сразу оказывается у автобусной остановки и начинает ждать свой автобус. Зайдя в него, он ровно за 15 минут прибывает на станцию метро. Пересадка у Васи проходит мгновенно: он сразу спускается на перрон и начинает ждать поезд. Поезд везет его ровно 8 минут до станции возле школы. Оттуда Вася идет пешком 3 минуты — и вот он уже у дверей школы.

Вася давно наблюдает за расписанием и заметил закономерность: когда он подходит к остановке автобуса, то автобус приезжает равновероятно через 1, 2, 3, 4, 5, 6 или 7 минут. А когда Вася подходит к остановке поезда, тот приезжает равновероятно через 1, 2 или 3 минуты.

- Найдите вероятность, что, подойдя к остановке автобуса сегодня, Вася будет ждать автобус не меньше 5 минут.
- Найдите вероятность, что суммарное время ожидания транспорта у Васи сегодня будет не больше 4 минут.
- Друг Васи Саша имеет очень точные часы, и он поделился с Васей, что на самом деле и автобус, и поезд приезжают не в целые минуты от ожидания, как посчитал Вася, а равновероятно в течение этого времени (то есть непрерывно в течение диапазона от 0 до 7 и от 0 до 3 минут соответственно). При этом остальные подсчеты Васи (про время пешком и на транспорте) оказались верны. Сегодня Вася дошел до автобусной остановки чуть позже обычного. Посчитайте, какова вероятность, что он успеет добраться до школы менее чем за 32 минуты.

Примечание: Для решения этой задачи не используйте интегралы, такие решения не будут засчитаны.

2. Про регрессию и ее свойства

Вы исследуете влияние качества картографической прометки и навигации на рейтинг туристических маршрутов, чтобы подтвердить необходимость развития прометки для улучшения туристического опыта в регионе. Рейтинг туристических маршрутов — это усредненная оценка маршрута среди туристов, оставивших о нем отзыв. Качество картографической прометки и навигации — это обобщенный индекс, включающий наличие точек маршрута на общедоступных онлайн- и офлайн-картах, наличие указателей по пути маршрута и опросов туристов на маршруте об удобстве и достаточности навигации.

В датасете, с которым вы будете работать, есть следующие переменные:

- map_quality — качество картографической прометки и навигации (далее — «Прометка»);
- landmarks_n — количество уникальных природных и искусственных достопримечательностей на маршруте («Достопримечательности»);
- rent_cost — средняя стоимость аренды жилья вдоль маршрута (тыс. руб./ночь) («Стоимость аренды»);
- route_rating — рейтинг туристического маршрута («Рейтинг»).

Выполните следующие задания по данному датасету.

1. Выполните предварительный анализ данных: постройте графики распределения для каждой переменной (прикрепите их к вашему решению), а также проанализируйте попарные связи между переменными (графически и численно, прикрепите результаты к решению). Какие из переменных наиболее вероятно влияют на исследуемую зависимую переменную и из чего следует этот вывод?

2. Оцените последовательно три регрессионные модели:

Модель 1: $\text{Рейтинг} = \beta_0 + \beta_1 \cdot \text{Прометка} + \varepsilon$

Модель 2: $\text{Рейтинг} = \beta_0 + \beta_1 \cdot \text{Прометка} + \beta_2 \cdot \text{Достопримечательности} + \varepsilon$

Модель 3: $\text{Рейтинг} = \beta_0 + \beta_1 \cdot \text{Прометка} + \beta_2 \cdot \text{Достопримечательности} + \beta_3 \cdot \text{Стоимость аренды} + \varepsilon$

Представьте результат в виде перечисления оцененных коэффициентов для каждой из моделей. Приведите также t-статистики и p-value, используемые для проверки значимости коэффициентов.

3. Прокомментируйте изменение коэффициентов b_0 и b_1 в трех моделях последовательно: почему значения этих коэффициентов изменяются именно таким образом?
4. Прокомментируйте изменение значения R^2 в трех последовательных моделях: почему значение R^2 изменяется именно таким образом?
5. Стандартизируйте исследуемые объясняющие переменные (независимые). На стандартизованных данных постройте аналогичные три модели из п. 2. Представьте результат в виде перечисления оцененных коэффициентов для каждой из моделей. Приведите также t-статистики и p-value, используемые для проверки значимости коэффициентов.
6. Как изменилась интерпретация коэффициентов b_0 и b_1 во всех трех моделях по сравнению с предыдущим случаем (до стандартизации)? Поменялась ли динамика этих коэффициентов? Объясните, почему так произошло.
7. Как соотносятся R^2 моделей на исходных и стандартизованных данных? Почему так происходит?
8. Как добавление переменной количества достопримечательностей на маршруте изменило оценку влияния качества прометки на рейтинг маршрута? Какой теоретический механизм взаимодействия может объяснять это изменение?

9. Сгенерируйте 100 дополнительных колонок со случайными данными в вашем датасете. Оцените модель с 103 независимыми переменными на рейтинг маршрута. Улучшились ли метрики качества модели? Означает ли это, что модель правда стала лучше?

3. Про фонтаны

В далеком городе Энске есть три района (Левобережный, Правобережный и Нагорье) и 500 человек жителей: в файле [fountains.xlsx](#) вам доступны координаты их домов (переменные x и y) и район, к которому относится дом (переменная $location$).

Сейчас в Энске активно обсуждается вопрос строительства фонтанов. Известно, что полезность любого жителя города от фонтана рядом с его домом рассчитывается по формуле $U_i = \frac{10}{2 + dist_i}$, где $dist_i$ — дистанция до ближайшего фонтана. Для расчета этой дистанции используется обычное евклидово расстояние: расстояние между точками (x_1, y_1) и (x_2, y_2) равняется $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

Известно также, что жители готовы платить за такое удовольствие, как фонтан рядом с домом! Как и в случае с полезностью, чем ближе фонтан к дому, тем больше готовность платить: сумма P_i , которую готов заплатить за фонтан i -й житель, рассчитывается по формуле $P_i = \frac{A_i}{2 + dist_i}$, где A_i — параметр, $dist_i$ — расстояние до ближайшего фонтана.

- 1) Мэрия Энска рассматривает три схемы установки фонтанов, все координаты здесь и далее записаны как (x, y) :

Один фонтан с координатами (5; 3,5)

Два фонтана с координатами (2; 2); (8; 2)

Три фонтана с координатами (2; 2); (8; 2); (5; 5)

Рассчитайте среднюю полезность жителей города от каждой из трех схем. Какая из этих схем максимизирует среднюю полезность?

- 2) Мэрия не может решить, какую схему выбрать, поэтому выносит этот вопрос на голосование. На голосование приходят все жители города, при этом каждый из них голосует за тот вариант, который принесет ему больше всего полезности. Если два варианта эквивалентны — он голосует за тот, при котором фонтанов будет меньше (потому что установка фонтанов стоит денег и при прочих равных жители предпочтут сэкономить общественные деньги). Сколько голосов наберёт каждая из схем? Какой вариант победит?
- 3) Энск-2 из параллельной вселенной ничем не отличается от предыдущего Энска, но мэрия там решает отдать установку фонтанов во власть рынка! Специально аккредитованная компания выбирает схему (те же варианты, что в п. 1), собирает со всех жителей города деньги в соответствии с их готовностью платить и устанавливает фонтаны! Какую схему выберет эта компания, если $A_i = 20$ для всех жителей города, себестоимость установки одного фонтана равна 1500, а компания (что логично) максимизирует свою прибыль? Рассчитывайте прибыль для каждой схемы и выберите оптимальную.

- 4) Энск-3 из еще одной параллельной вселенной тоже очень похож на предыдущие два Энска, в нем фонтаны тоже устанавливает коммерческая компания, но мэрия строго велела установить ровно три фонтана! Второе отличие: в Энске-3 в Нагорье живёт местная элита и их параметр A_i выше, чем у других жителей города. У всех жителей Левобережного и Правобережного $A_i = 20$, у всех жителей Нагорья A_i одинаковый, но его размер они могут выбрать сами.

Базовая схема установки фонтанов в Энске-3 — это «Три фонтана с координатами (2; 2); (8; 2); (5; 5)», но жители Нагорья очень хотят, чтобы все фонтаны были в их районе! Им нравится схема «три фонтана с координатами (5; 5,5); (4,5; 4,5); (5,5; 4,5)».

Компания по установке фонтанов выберет тот вариант (из этих двух), который обеспечит ей максимальную прибыль. Все жители города по-прежнему платят по формуле $P_i = \frac{A_i}{2 + \text{dist}_i}$. Какое значение параметра A_i должны выбрать жители Нагорья, чтобы победила та схема, которая им нравится?

4. Про географию и цены на недвижимость

Датасет с некоторыми промежуточными столбцами

Датасет в том виде, который даем участникам

Гендальф планирует возвращение на родину и изучает рынок недвижимости Валинора. На листе «Квартиры по городам» представлены данные о стоимости и площади квартир в различных городах Валинора.

apartment_id — идентификатор каждой отдельной квартиры

city_id — идентификатор каждого отдельного города

city_name — название каждого отдельного города

area — площадь квартиры в квадратных метрах

price — стоимость квартиры в золотых монетах

На листе «Города» приведены координаты расположения городов на карте Валинора. Валинор — относительно небольшая страна, поэтому можно считать, что все города расположены в одной плоскости $X-Y$.

city_id — идентификатор каждого отдельного города

city_name — название каждого отдельного города

x — координата города по оси X

y — координата города по оси Y

Помогите Гендальфу, выполнив следующие пункты задачи.

- 1) Приведите график, который мог бы визуализировать расположение всех городов Валинора на плоскости $X-Y$ и отражал бы среднюю стоимость квартир в этих городах.
- 2) Гендальф заработал довольно много золотых за время своих приключений в Средиземье, в связи с чем рассматривает покупку недвижимости

в Старом городе (это название города из датасета). Рассчитайте корреляцию между евклидовым расстоянием от Старого города до всех остальных городов Валинора и средней стоимостью квартир в этих городах.

- 3) Гендальфу известно, что посреди Валинора ровно на прямой $X = 0$ возвышается длинный горный хребет, разделяющий страну на левую и правую части. В таком случае ему будет сложно навещать своих друзей, проживающих в правой части Валинора. Рассчитайте корреляцию между расстоянием от Старого города до городов Валинора, находящихся в левой части страны ($X < 0$), и средней стоимостью квартир в этих городах.
- 4) Во время посиделок в таверне Элронд напомнил Гендальфу, что горы не представляют большой проблемы, так как Гендальф может всегда перемещаться между любыми городами напрямую по воздуху на орлах. Кроме того, Старый город и Серебряный Лист являются самыми развитыми городами Валинора и в них установлен телепорт, связывающий их напрямую. То есть между этими городами всегда можно перемещаться напрямую, затрачивая на это 0 минут!

Для выполнения данного пункта вам понадобится третий лист датасета «Матрица со временем пути». Каждый элемент этой матрицы a_{ij} показывает время, затрачиваемое на самую короткую возможную дорогу между городами i и j в минутах.

Используя эту матрицу, рассчитайте корреляцию между временем пути от Старого города до всех остальных городов Валинора и средней стоимостью квартир в этих городах.

Внимательно посмотрите на свои результаты из пунктов 1—4 и ответьте на следующие вопросы:

- A) Какая закономерность для средней стоимости квартир наблюдается в городах Валинора?
- B) Из-за чего может наблюдаться подобная закономерность? Можете считать, что с точки зрения социально-экономических процессов города Валинора функционируют так же, как и города какой-то страны в реальной жизни.
- C) Почему в пунктах 2—4 могли получиться разные результаты?
- D) В процессе решения этой задачи мы использовали несколько разных способов измерения «дистанции» (силы связи из-за особенностей географического положения) между городами: просто расстояние по карте и время в пути. Предложите еще какой-нибудь разумный способ определения «дистанции» между городами, который мог бы описывать наблюдаемую зависимость.

5. Про парадокс Симпсона

У Даши есть набор из 100 наблюдений по некоторым показателям X и Y , по которым она оценивает модель линейной регрессии. Ответьте на следующие вопросы.

1. Даша оценила три модели, используя разные подвыборки и полную выборку по переменным X и Y :

$\begin{aligned}$

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, 50 \tag{1} \\ \hat{Y}_i &= \hat{\alpha}_0 + \hat{\alpha}_1 X_i, \quad i = 51, \dots, 100 \tag{2} \\ \hat{Y}_i &= \hat{\gamma}_0 + \hat{\gamma}_1 X_i, \quad i = 1, \dots, 100 \tag{3} \end{aligned}$$

При этом оказалось, что $\hat{\beta}_1 > 0$, $\hat{\alpha}_1 > 0$, а $\hat{\gamma}_1 < 0$. Приведите диаграмму рассеяния по переменным X и Y , отражающую набор наблюдений, по которым Даша могла получить такие оценки коэффициентов.

- Предложите реальную пару показателей X и Y , для которых Даша могла бы получить подобную зависимость в пункте 1. Что в таком случае могут представлять собой полная выборка и подвыборки, чем эти подвыборки могут отличаться друг от друга? Обоснуйте, почему для этих показателей могла бы наблюдаться такая зависимость.
- Даша еще раз внимательно посмотрела на свои результаты в пункте 1 и заметила, что для оценок коэффициентов в моделях (1)-(2) дополнительно выполняется неравенство $\hat{\alpha}_1 > \hat{\beta}_1$. Используя свой пример показателей из пункта 2, объясните, чем может быть обусловлена такая зависимость. Запишите уравнение линейной регрессии, которое бы описывало такую зависимость одновременно для всех наблюдений (X_i, Y_i) , где $i = 1, \dots, 100$.

6. Про RLMS и рисование графиков

В этой задаче вам предстоит немного поработать с [данными обследования РМЭЗ НИУ ВШЭ](#). РМЭЗ — это серия ежегодных общенациональных репрезентативных опросов, проводимая с 1994 года, по которой можно изучать огромный спектр вопросов о жизни людей в России. Вы будете работать с самой свежей волной опроса — за 2024 год.

Для решения этой задачи вам будут нужны два файла: [RLMS_ind.xlsx](#) и [RLMS_HH.xlsx](#).

Вам также пригодятся описания переменных [базы данных по индивидам](#) и [базы данных по домохозяйствам](#).

В файле [RLMS_ind.xlsx](#) содержится небольшая часть базы данных по индивидам:

- idind**, **ccid_i** — идентификационные номера индивида
- ccid_h** — идентификационный номер домохозяйства
- cch4** — номер члена семьи
- wage** — заработная плата человека по основному месту работы (переменная **ssj10** из исходной базы данных)
- educ** — уровень образования человека (переменная **ss_diplom** из исходной базы данных)
- gender** — пол человека (переменная **cch5** из исходной базы данных)
- age** — возраст человека (переменная **ss_age** из исходной базы данных)

В файле [RLMS_HH.xlsx](#) содержится полная база данных по домохозяйствам со всеми переменными, которые в ней есть. Описания переменных, которые будут нужны в процессе решения задачи, следует искать в документации ([описании базы данных по домохозяйствам](#)).

Начнем с того, что удалим из базы по индивидам все наблюдения, для которых нет информации о зарплате (в поле `wage` стоит пустое значение или респондент не ответил по той или иной причине). Обратите внимание на то, как в RLMS кодируются варианты «затрудняюсь ответить», «отказ от ответа» и «нет ответа».

- 1) Рассчитайте, чему равна средняя зарплата респондента, и нарисуйте гистограмму зарплат.
- 2) Рассчитайте среднюю зарплату респондентов в зависимости от уровня образования. Запишите ее в таблицу и визуализируйте. Обратите внимание, что в данных уровень образования кодируется цифрой, но расшифровка доступна в описании базы по индивидам. При построении графика используйте не кодировку уровня образования (число), а текстовое описание (расшифровку).

Для решения следующих пунктов вам будет нужно объединить два файла базы данных. Вы можете сделать это по переменной `ccid_h`.

- 3) Для каждого индивида из вашей базы данных подтяните количество членов его семьи (переменная `cc_nfm`). Каков средний размер семьи в нашей выборке? Визуализируйте распределение размера семьи.
- 4) Визуализируйте связь дохода и размера семьи при помощи диаграммы рассеяния. Рассчитайте корреляцию этих двух переменных и проверьте ее статистическую значимость. Какой вывод о наличии связи дохода человека и размера его семьи можно сделать на основе этих расчетов? Поясните.
- 5) Изменится ли ответ на предыдущий вопрос, если рассматривать отдельно мужчин, отдельно женщин? Если рассматривать только людей в возрасте от 40 лет (40 и старше)?

Для решения следующего пункта вам надо будет изучать родственные связи респондентов в нашей выборке. В этом вам помогут переменные из файла по домохозяйствам вида `ccbX.9.Y` (где `X` и `Y` — номера членов семьи из переменной `sch4`) и [кодификатор родственных связей](#).

К примеру: у вас есть два индивида из одного домохозяйства (с одним значением `ccid_h`): один с номером 3, другой — с номером 5 (это значения переменной `sch4`). Значит, чтобы выяснить их родственные связи, нужно смотреть на переменную `ccb5.9.3`: «кем 5-й член семьи приходится 3-му члену семьи». Пусть там стоит 9 — это значит (в соответствии с [кодификатором](#)), что 5-й член семьи — внук или внучка 3-го члена семьи! Чтобы узнать более подробную информацию об этом 5-м члене семьи, достаточно найти его `id` в переменной `ccidind5` и подробнее изучить уже по файлу базы данных по индивидам.

- 6) Найдите индивида с самым высоким доходом. Сколько человек живет с ним в одном домохозяйстве? Какого они пола (переменные вида `ccbX.4`), возраста (года рождения содержатся в `ccbX.5`) и кем они приходятся нашему герою?