DANO

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Т БАНК

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.

а) Найдите вероятность, что, подойдя к остановке автобуса сегодня, Вася будет ждать автобус не меньше 5 минут.

б) Найдите вероятность, что суммарное время ожидания транспорта у Васи сегодня будет не больше 4 минут.

в) 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.

Решение:

1. 3 / 7
2. 6 / 21 = 2 / 7
3. x + 15 + y + 8 + 3 = 32

x+y = 6, следовательно y = 6 – x, рисуем график этой функции

Рисуем графики функций x = 7 и y = 3 — это ограничения по времени ожидания каждого из видов транспорта.

Считаем долю площади трапеции под y = 6 – x при ограничении y <= 3 от всего прямоугольника. Получаем 13,5 / 21 ~ 0,623.

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 оффлайн-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:

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

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

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

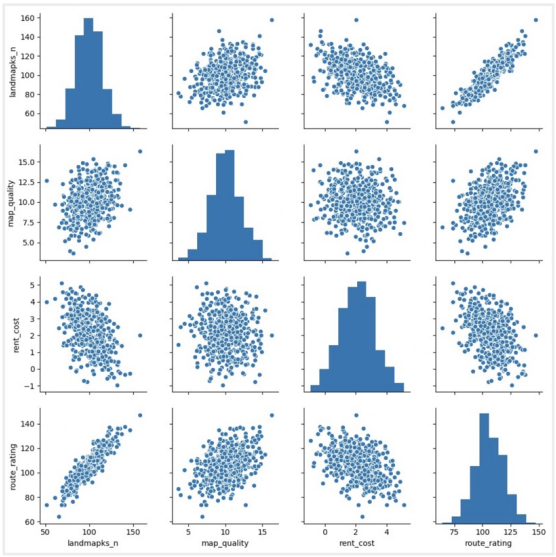
   Представьте результат в виде перечисления оцененных коэффициентов для каждой из моделей. Приведите также t-статистики и p-value, используемые для проверки значимости коэффициентов.
3. Прокомментируйте изменение коэффициентов b_0 и b_1 в трех моделях последовательно: почему значения этих коэффициентов изменяются именно таким образом?
4. Прокомментируйте изменение значения R^2 в трех последовательных моделях: почему значение R^2 изменяется именно таким образом?
5. Стандартизуйте исследуемые объясняющие переменные (независимые). На стандартизированных данных постройте аналогичные три модели из п. 2.

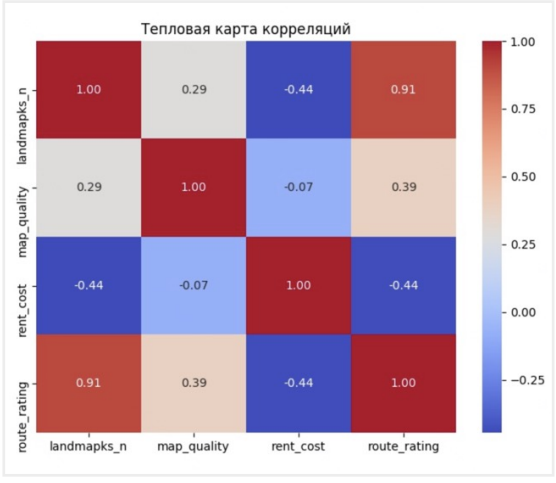
Представьте результат в виде перечисления оцененных коэффициентов для каждой из моделей. Приведите также t-статистики и p-value, используемые для проверки значимости коэффициентов.

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 стандартизированных 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?

Решение:

1.





Наибольшее потенциальное влияние среди представленных переменных на рейтинг маршрута оказывает количество достопримечательностей на маршруте, так как: 1) коэффициент корреляции между этими переменными наибольший, 2) на графиках видно наличие линейной взаимосвязи между этими двумя переменными, и отсутствие других значимых попарных взаимосвязей.

2. Модель 1:

   const — 82.2426

   map_quality — 2.3840

   Модель 2:

   const — 23.6367

   map_quality — 0.8391

   landmarks_n — 0.7409

   Модель 3:

   const — 26.7211

   map_quality — 0.8608

   landmarks_n — 0.7201

   rent_cost — –0.5842

3. tbd
4. tbd
5. Модель 1:

   const — 106.2469

   map_quality — 4.8516

   Модель 2:

   const — 106.2469

   map_quality — 1.7075

   landmarks_n — 10.8936

   Модель 3:

   const — 106.2469

   map_quality — 1.7518

   landmarks_n — 10.5891

   rent_cost — –0.6570

6. tbd
7. tbd

8. tbd
9.

1) видим стат. значимые фичи, например:

   random_feature_68 с p-value 0.028

   random_feature_42 с p-value 0.011

   random_feature_21 с p-value 0.023

   переменные перед нашими фичами почти не изменились:

   const — 106.3819

   map_quality — 1.7663

   landmarks_n — 10.5046

   rent_cost — –0.7293

2) R-squared: 0.880

   Хотя был R-squared: 0.847

   То есть добавление рандомных фичей увеличило R2 почти на 4%.

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 домой 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 $P_i = \frac{A_i}{2 + dist_i}$. Какое значение параметра A_i должны выбрать жители Нагорья, чтобы победила та схема, которая им нравится?

Решение:

Общий комментарий по всем пунктам задачи: ответы принимаются только при наличии решения (кода, эксель-файла с расчётами, какого-то другого описания пути получения решения), цифры без обоснования (даже верные) оцениваются в 0 баллов.

Вопрос 1 — средние полезности

1 фонтан	2.064462
2 фонтана	3.089928
3 фонтана	3.503346

Вопрос 2 — голосование

Схема	Голосов
1	15

2	392
3	93

Побеждает вариант 2.

Вопрос 3 — про прибыли

	Выручка	Прибыль
1 фонтан	2064.462	564.4617
2 фонтана	3089.928	89.92847
3 фонтана	3503.346	-996.654

Побеждает вариант с 1 фонтаном.

Вопрос 4 — элитное Нагорье

Ответ: 409.37 (можно принять на полный балл 409 или 410)

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 эвклидовым 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?

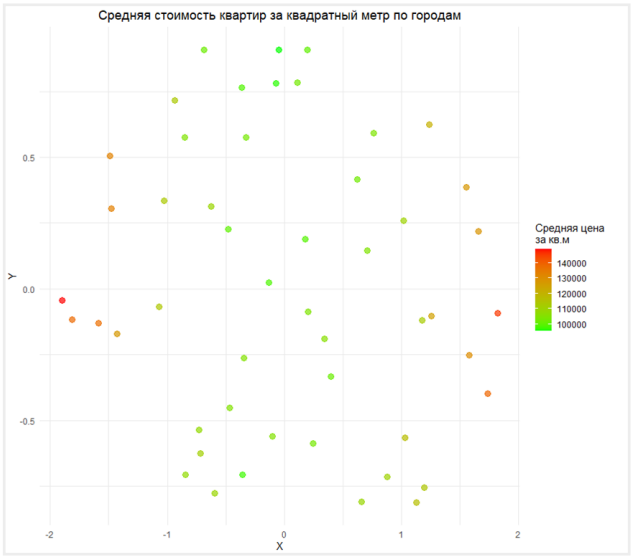
B) Из-за чего может наблюдаться подобная закономерность? Можете считать, что с точки зрения социально-экономических процессов города Валинора функционируют так же, как и города какой-то страны в реальной жизни.

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.

Решение:

Вопрос 1 — карта (график).



Вопрос 2 — корреляция между ценами и расстоянием

Ответ_1: 0.1191375 или округлённые варианты, например 0.119, 0.1191, 0.11914, 0.119147

или

Ответ_2: 0.0877331 или округлённые варианты, например, 0.088, 0.0877, 0.08773, 0.087733

Вопрос 3 — корреляция между ценами и расстоянием при условии X < 0

Ответ_1: -0.8716399 или округлённые варианты, например -0.87, -0.872, -0.8716, -0.871614

или

Ответ_2: -0.9410586 или округлённые варианты, например, -0.94, -0.941, -0.9411, -0.94106, -0.941059

Вопрос 4 — корреляция между ценами и временем пути из матрицы

Ответ_1: -0.8529438 или округлённые варианты, например, -0.853, -0.8529, -0.85294, -0.852944

Ответ_2: -0.9288044 или округлённые варианты, например, -0.93, -0.929, -0.9288, -0.928804

Вопрос А) — про закономерность для стоимости квартир

Очевидное падение стоимости с увеличением расстояния от Старого Города (city_id = 13)

В правой части карты тоже есть кластер с высокими ценами на недвижимость, и для него наблюдается аналогичная ситуация. Соответственно цены на квартиры в рассматриваемом городе выше/ниже, чем ближе/дальше он находится от условных от левого и правого края карты.

Вопрос B) — про закономерность для стоимости квартир

Много рабочих мест в Старом Городе, высокий уровень зарплат, более развитая инфраструктура, лучше социальное обеспечение, всё это стимулирует граждан рассматривать приобретение квартиры непосредственно в нём или в городах, которые лежат неподалёку. Серебряный Лист также является важным экономическим центром.

Вопрос C) — почему получились разные корреляции в разных пунктах

Здесь необходимо прокомментировать особенности расчёта в каждом из трёх пунктов 2–4.

2. Мы смотрим корреляцию расстояния от Старого Города до других городов и цен на квартиры, сначала она отрицательная (больше расстояние – меньше цена). Затем, приближаясь к правому краю, цена вновь начинает расти, хотя расстояние по-прежнему увеличивается, то есть корреляция становится положительной. В связи с этим, итоговая корреляция получилась небольшая. В целом, главное сказать, что для городов в левой части корреляция отрицательная, а для городов в правой части — положительная. Количество городов в левой и правой части почти одинаковое (если быть точным, то в левой части их 26, а в правой 24). Из-за чего слагаемые в выборочной корреляции почти полностью взаимно уничтожаются. Падение цен с расстоянием для крайнего правого города (Серебряного Листа) происходит очень похожим образом (или симметрично) Старому Городу. Таким образом, с учётом того, что характер зависимости аналогичный, слагаемые в выборочной корреляции почти полностью взаимно уничтожаются.

3. Мы смотрим корреляцию расстояния от Старого Города до других городов и цен на квартиры, сначала она отрицательная (больше расстояние — меньше цена). Затем, начиная с центра карты ($X = 0$) цена вновь начинает расти, хотя расстояние по-прежнему увеличивается, то есть корреляция становится положительной. Но мы ограничили выборку только левой частью карты ($X < 0$), в связи с чем итоговая корреляция получилась сильная и отрицательная.

4. Для начала следует сказать, что время пути по смыслу эквивалентно расстоянию из предыдущих ситуаций.

   Добавление телепорта между городами меняет кратчайшее расстояние/время пути между городами по сравнению с предыдущими пунктами задачи (без подробностей как именно или с ошибками в описании того, как именно).

   Матрица показывает кратчайшее время пути из Старого Города в любой другой, следовательно, её элементы, отвечающие за время пути до городов из левой части карты ($X < 0$), отражают время полёта на орлах (эквивалент обычного расстояния из пунктов 1)-2)).

   Матрица показывает кратчайшее время пути из Старого Города в любой другой, следовательно, с учётом телепорта её элементы, отвечающие за время пути до городов из правой части карты ($X > 0$), отражают использование телепорта (0 минут пути) + время полёта на орлах из Серебряного Листа в город назначения.

Корреляция для городов в левой части карты всё так же остаётся сильной и отрицательной по аналогии с ситуациями из предыдущих пунктов. За счёт телепорта расстояние/время пути из Старого Города до городов с дорогой недвижимостью, расположенных близко к Серебряному Листу, но далеко от Старого Города, тоже становится маленьким. Таким образом, для городов из правой части мы тоже будем видеть отрицательную корреляцию между временем пути и ценами на квартиры. Следовательно, итоговая корреляция будет сильная и отрицательная.

Вопрос D) — предложите ещё какой-нибудь способ определения дистанции

В условии мы сказали, что можно считать, что города Валинора функционируют примерно так же, как и города в реальной жизни, поэтому здесь могут быть очень разные варианты ответа. Для референса можно представить ситуацию Москвы — Московской области или любую другую «Региональный центр — его область»

Но какой есть важный момент. На самом деле время пути и расстояние в целом покрывает практически любой чисто «географический» вариант ответа, а мы хотим увидеть что-то альтернативное.

1. Например, это может быть история про миграционные потоки. Сколько человек туда-сюда ежедневно ездят на работу. Чем больше, тем меньше «расстояние» между городами. Эту же идею можно оформить в виде какого-нибудь индекса транспортной доступности, который рассчитывается на основе количества рейсов по всем доступным видам транспорта в день.
2. Ещё можно предложить объём торговли между городами: чем больше, тем меньше расстояние.
3. На самом деле подойдёт разность адекватных социальных/экономических/демографических показателей городов. Например, воображаемый аналог ВВП для города (в России, насколько мне известно, такое не считают, а в США вроде есть), уровень дохода горожан, население/плотность населения и т.д.

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:

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

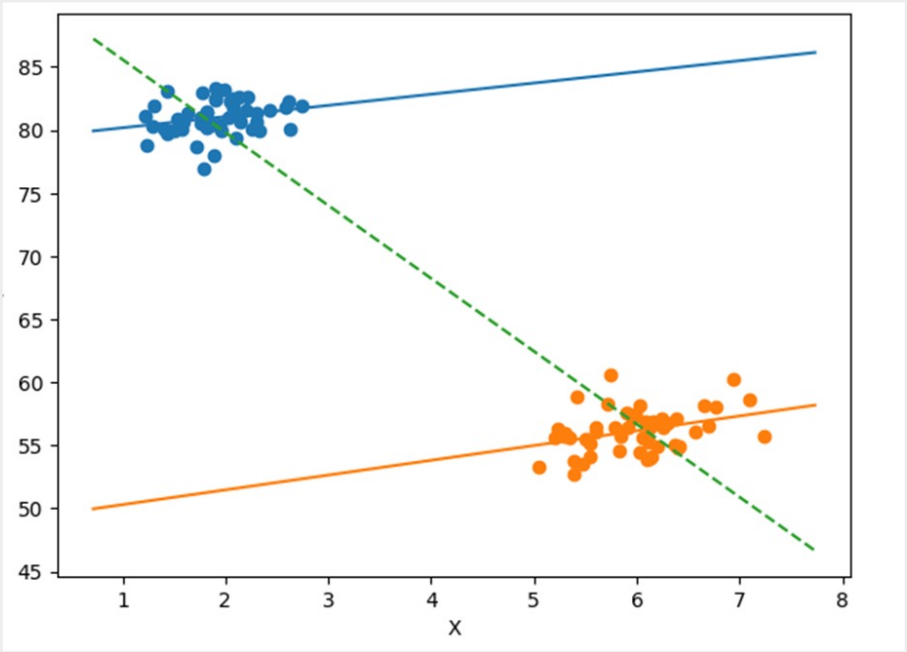
   При этом оказалось, что $\hat{\beta}_1 > 0$, $\hat{\alpha}_1 > 0$, а $\hat{\gamma}_1 < 0$. Приведите диаграмму рассеяния по переменным X и Y, отражающую набор наблюдений, по которым Даша могла получить такие оценки коэффициентов.

2. 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.

3. 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 $i = 1, \ldots , 100$.

Решение:

Вопрос 1 — привести пример диаграммы рассеяния, иллюстрирующей парадокс Симпсона



Вопрос 2 — привести пример показателей X и Y

Например, X — часы подготовки к экзамену за неделю, Y — балл экзамена. Подвыборки — группы разной сложности курса/учителя.

Группа 1 (легкий курс): студенты в среднем тратят меньше часов (X), но имеют высокий базовый уровень Y (экзамен проще). Внутри группы: больше часов — выше балл — положительный наклон.

Группа 2 (сложный курс): студенты в среднем тратят больше часов, но базовый уровень Y ниже (экзамен сложный). Внутри группы: больше часов — выше балл — положительный наклон.

В целом по всем 100 наблюдениям: из-за сдвига средних (сложный курс «живёт» при больших X, но даёт меньшие Y) агрегированная линия может стать отрицательной.

Вопрос 3 — оказалось, что коэффициент наклона у одной из подвыборок больше

На сложном курсе (группа 2) маржинальная отдача часа подготовки выше — каждый дополнительный час заметнее повышает шанс разобрать трудные темы.

В идеале здесь должна была бы быть короткая фраза, поясняющая, почему разбор трудных тем даёт больший прирост к оценке (например, что они весят больше баллов в экзамене).

$Y_i = \delta 0 + \delta 1 * X_i + \delta 2 * D_i + \delta 3 * (X_i * D_i) + \varepsilon_i$

$D_i = 1$ для наблюдений из подвыборки 2, Иначе $D_i = 0$

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 cc|10 из исходной базы данных)
- **educ** — уровень образования человека (переменная **cc_diplom** из исходной базы данных)
- **gender** — пол человека (переменная **cch5** из исходной базы данных)
- **age** — возраст человека (переменная **cc_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 зарплаты.

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)?

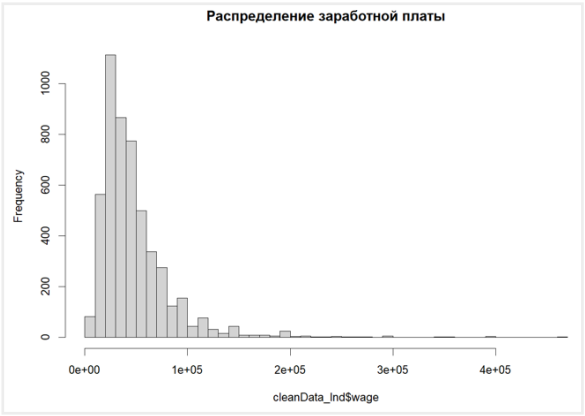
Для решения следующего пункта вам надо будет изучать родственные связи респондентов в нашей выборке. В этом вам помогут переменные из файла по домохозяйствам вида ccbX.9.Y (где X и Y — номера членов семьи из переменной cch4) и кодификатор родственных связей**.**

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 cch4). 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?

Решение:

1)



2) [1] "educ = 4 ; mean wage = 43912.0746164575"

[1] "educ = 6 ; mean wage = 59949.2265676756"

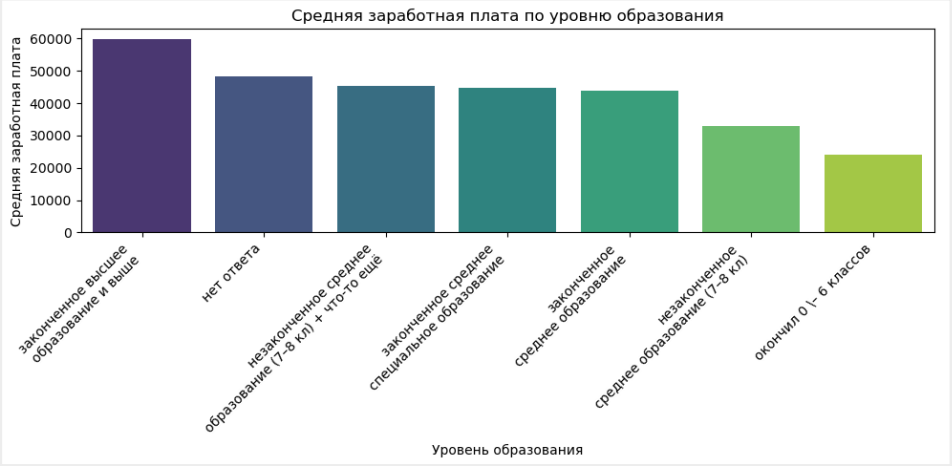
[1] "educ = 5 ; mean wage = 44748.0083333333"

[1] "educ = 2 ; mean wage = 32908"

[1] "educ = 3 ; mean wage = 45497.2146892655"

[1] "educ = 1 ; mean wage = 24214.2857142857"

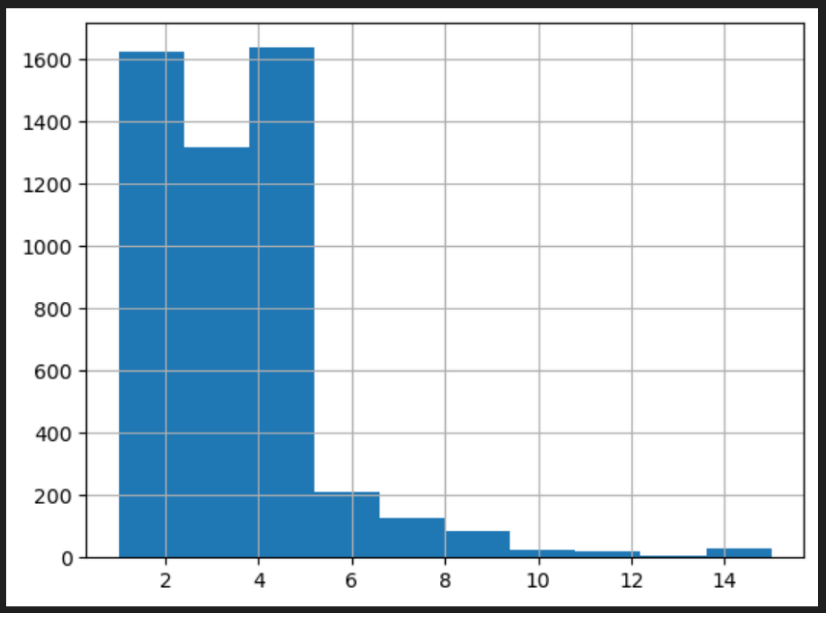
[1] "educ = 99999999 ; mean wage = 48352.9411764706"



3)

```
Среднее: 3.4824734147302085

<Axes: >
```



4)

Pearson's product-moment correlation

data: tmp3$wage and tmp3$cc_nfm

t = 1.1748, df = 5076, p-value = 0.2401

alternative hypothesis: true correlation is not equal to 0

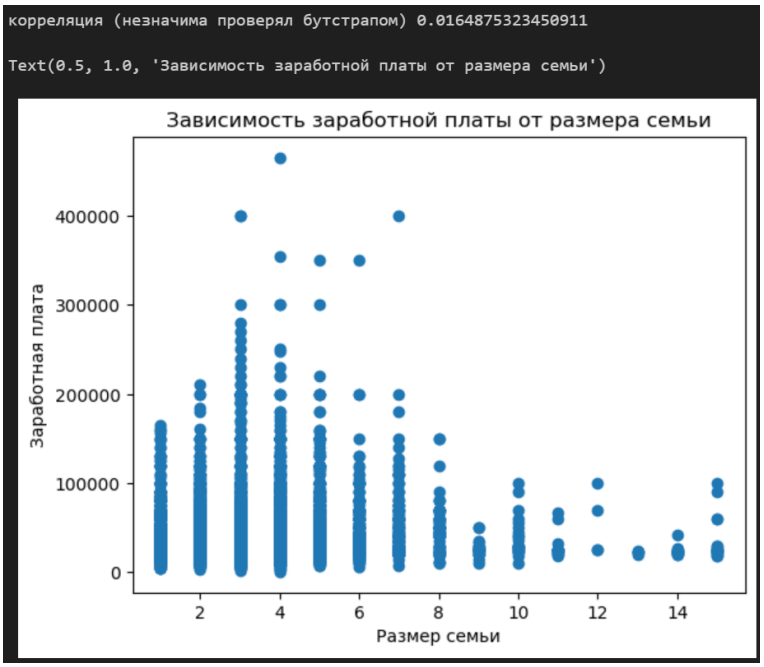
95 percent confidence interval:

-0.01102303 0.04397315

sample estimates:

cor

0.01648753



5)

> cor.test(tmp3$wage[tmp3$gender == 1], tmp3$cc_nfm[tmp3$gender == 1])

Pearson's product-moment correlation

data: tmp3$wage[tmp3$gender == 1] and tmp3$cc_nfm[tmp3$gender == 1]

t = 1.5713, df = 2406, p-value = 0.1162

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.007936831 0.071870845

sample estimates:

cor

0.03201804

```
> cor.test(tmp3$wage[tmp3$gender == 2], tmp3$cc_nfm[tmp3$gender == 2])

    Pearson's product-moment correlation

data:  tmp3$wage[tmp3$gender == 2] and tmp3$cc_nfm[tmp3$gender == 2]

t = -2.1788, df = 2668, p-value = 0.02943

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.079951319 -0.004217971

sample estimates:

    cor

-0.04214518
```

Для людей за 40 тоже значима:

```
> cor.test(tmp3$wage[tmp3$age >= 40], tmp3$cc_nfm[tmp3$age >= 40])

    Pearson's product-moment correlation

data:  tmp3$wage[tmp3$age >= 40] and tmp3$cc_nfm[tmp3$age >= 40]

t = 2.5841, df = 3114, p-value = 0.009808

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.01116199 0.08123984

sample estimates:

    cor

0.04625783
```

6)

```
> max(cleanData_Ind$wage)

[1] 465000

> highestWage_n <- which(cleanData_Ind$wage == max(cleanData_Ind$wage))

> hisHH_id <- cleanData_Ind[highestWage_n, ]$ccid_h

> hisHH_id

[1] 13811087
```

Женщина, 1989 г.р. — возраст ≈ 35–36 лет — вероятно супруга (код родства ожидается 1 — муж/жена, в выгрузке код не указан).

Женщина, 1956 г.р. — возраст ≈ 68–69 лет — тёща (код 11).

Женщина, 2014 г.р. — возраст ≈ 10–11 лет — дочь (код 4).