DANO

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Т БАНК

1. (2 балла) Вы работаете помощником следователя в небольшом городе. Недавно вам передали ряд дел об ограблениях. Вы составили статистику по этим делам и выяснили, что в половине ограбленных домов имеется домашнее животное. Также вы знаете, что в вашем городе домашние питомцы имеются в 80% домов. Какой вывод вы можете сделать исходя из этой информации?

   а) преступник чаще грабит дома, в которых есть домашние животные

   б) преступник реже грабит дома, в которых есть домашние животные (+)

   в) преступник одинаково часто грабит дома, где есть домашние животные и где их нет

   г) предоставленной информации недостаточно, чтобы сделать любой из перечисленных выше ответов

   Решение. В данной формулировке условия возможно две интерпретации, соответственно и два решения засчитывались за верные. Разница состоит в том, относительно чего мы смотрим частоту.

   Подход 1. Если бы преступник одинаково часто выбирал как цели для ограбления дома с питомцами и без них, то в итоговой статистике ограбленных домов соотношение было бы таким же. Однако домов без питомцев оказалось больше. Это значит, дома, где есть домашние животные, грабили реже.

   Подход 2. По условию половина ограбленных домов была с питомцами, половина без них. В итоге получается, что преступник одинаково часто грабил и дома, где животные есть, и дома, где их нет.

2. (4 балла) Офлайн-магазин электроники продает телефоны и аксессуары для них. В магазине можно купить любое количество телефонов и/или аксессуаров. Известно, что за сентябрь аксессуаров (в штуках) было продано больше, чем телефонов. При этом 1% от всех товаров вернули. Выберите одно или несколько верных утверждений, которые следуют из этой информации:

   а) В сентябре вернули не более 1% аксессуаров

   б) В сентябре был хотя бы один покупатель, который приобрел в магазине и телефон, и аксессуар

   в) В сентябре был хотя бы один покупатель, который приобрел в магазине телефон без аксессуара

   г) Аксессуаров без телефонов купили больше, чем телефонов без аксессуаров

   д) Среди других утверждений нет таких, которые следуют из приведенной в условии информации (+)

Решение.

а) Не обязательно. Например, в сентябре могли купить 99 телефонов и 101 аксессуар, при этом вернуть из товаров 2 аксессуара.

б) Не обязательно. Каждый клиент мог купить любое количество телефонов и/или аксессуаров. А значит, все чеки в сентябре могли содержать только телефоны или только аксессуары.

в) Не обязательно. Каждый клиент мог купить любое количество телефонов и/или аксессуаров. А значит, все покупки, содержащие телефоны, могли содержать и хотя бы один аксессуар.

г) Не обязательно. Каждый клиент мог купить любое количество телефонов и/или аксессуаров. А значит, могло не быть ни одной покупки, состоящей только из аксессуаров, и одновременно с этим могла быть хотя бы одна покупка, состоящая только из телефона.

д) Так как другие утверждения не обязательно следуют из приведенной информации, то это единственно верное утверждение.

3. (2 балла) Вам стало интересно узнать, сколько читают ученики старших классов в вашей школе. Вы создали небольшой опрос и планируете свои дальнейшие действия. Для этих целей вы решили в течение недели общаться со всеми старшеклассниками, которые ходят на литературный кружок. В результате полученная вами оценка количества часов чтения, скорее всего:

а) будет завышена (+)

б) будет занижена

в) будет близка к истинному значению

г) может быть как завышена, так и занижена

Решение.

Основная идея решения состоит в том, что у людей есть ограничение по времени, и они распределяют свое свободное время исходя из предпочтений. Это распределение в среднем одинаковое у разных людей. Однако, если говорится о наличии некоторого «хобби», которым занимается человек, это уже не свободное время, в это время он занимается чем-то конкретным. Если говорится о том, что человек участвует в литературном кружке, в этот час он читает больше, чем в «обычное» время. Если говорится о любом другом кружке, не связанном с чтением (например, спортивная секция, решение задач по экономике, рукоделие и т.д.), конкретный промежуток времени чтению уделяться не может. В итоге получается, что если 23 часа в сутки люди распределяют «как все», а 1 час на соответствующий кружок, то если этот кружок связан с чтением, данные люди будут читать больше (и оценка времени чтения будет завышена), чем в среднем все остальные, если не связан с чтением — то на чтение будет уделяться чуть меньше времени (и оценка будет занижена). Отдельные люди за пределами данных кружков могут вести себя по-разному (один из посещающих литературный кружок может за его пределами не трогать книги, а один из спортсменов все оставшееся время читать), но также могут быть и их противоположности, что в среднем будет давать некоторого среднего человека.

4. (4 балла) В городе Монетск каждый житель является клиентом по крайней мере одного из двух банков-конкурентов: банка А или банка Б. Аналитикам банка А Пете и Свете поручили посчитать, какой процент жителей города Монетск одновременно является и клиентами банка А, и клиентами банка Б.

   Петя запустил в приложении банка А опрос, доступный клиентам-жителям Монетска. Света опрашивала людей, выходящих из разных отделений банка Б в городе Монетск. И Петя, и Света задавали один и тот же вопрос: «Являетесь ли вы клиентом банка-конкурента?» (т.е. банка Б для клиентов банка А — и наоборот). Оба опроса проводились одновременно и длились неделю. В Петином опросе приняли участие 12095 человек, из них 9434 являются, по их словам, клиентами банка Б. В опросе Светы приняли участие 2274 человека, из них 978 являются, по их словам, клиентами банка А.

   В результате Петя сделал вывод, что 78% жителей является клиентами обоих банков, а Света — что таких жителей 43%.

   Выберите одно или несколько верных утверждений, которые следуют из этой информации:

   а) Оценка Пети ближе к истинному значению, чем оценка Светы

   б) Оценка Светы ближе к истинному значению, чем оценка Пети

   в) Истинное значение больше, чем оценки Светы и Пети

   г) Истинное значение находится между оценками Светы и Пети

   д) Недостаточно данных, чтобы сделать выводы (+)

   Решение.

   Из приведенных данных не ясно ни сколько всего клиентов в банках А и Б, ни какая доля клиентов банка А пользуется приложением (и что это за клиенты), ни какая доля клиентов банка Б посещает отделения (и что это за клиенты). В зависимости от тех или иных значений этих фактических данных, может быть верно каждое из утверждений а)–г). Поэтому нельзя оценить репрезентативность проведенных Петей и Светой опросов — как и сделать какие-либо выводы о взаимосвязи их оценок и истинного значения процента жителей, являющихся клиентами обоих банков. Делать такие выводы, основываясь лишь на соотношении количества опрошенных ими людей и/или общих представлениях о клиентах банковских приложений и людей, выходящих из отделений банка, в данном случае некорректно.

5. (5 баллов) Некоторые пассажиры, купившие авиабилеты, не пользуются ими из-за опоздания на рейс или изменения планов. Понимая это, авиакомпании иногда специально продают больше билетов, чем есть мест в самолете, — это называется овербукинг.

   Компания «Птичка Airlines» продала 190 билетов на рейс со 188 местами. За 15 минут до окончания регистрации трое из 190 пассажиров не зарегистрированы на рейс — Елена С., Владислав П. и Ирина К. По статистике предыдущих перелетов Елена С. придет на регистрацию с вероятностью 90%, Владислав П. — 50%, Ирина К. — 40%. Какова вероятность того, что кому-то из пришедших на регистрацию пассажиров не хватит мест? Дайте ответ в процентах. Если ответ представлен в виде дроби, округлите его до целого числа. Единицы измерения указывать не нужно.

   Ответ: 65

Решение.

Зарегистрировалось 187 пассажиров, а мест всего 188, поэтому кому-то из еще не зарегистрировавшихся пассажиров не хватит мест в случае, если на регистрацию придет хотя бы 2 пассажира. Вероятность того, что на регистрацию придут Елена С. и Владислав П., равна 0,9 · 0,5; Елена С. и Ирина К. — 0,9 · 0,4; Владислав П. и Ирина К. — 0,5 · 0,4.

Каждое из этих трех событий учитывает возможность пройти регистрацию одновременно всем трем пассажирам (вероятность этого 0,9 · 0,5 · 0,4). Мы должны включить в ответ эту вероятность, но не должны ее дублировать, поэтому нужно дважды вычесть ее из ответа.

Итоговая вероятность тогда будет равна:

0,9 · 0,5 + 0,9 · 0,4 + 0,5 · 0,4 – 2 · 0,9 · 0,5 · 0, 4= 0,65 (65%)

6. (5 баллов) Компания, занимающаяся продажей и доставкой цветов, решила подарить пяти своим самым активным пользователям 8 одноразовых промокодов: 7 промокодов на скидку 15% и 1 промокод — на 25%. Каждый промокод должен быть отправлен ровно одному пользователю. При этом один пользователь может как не получить ничего, так и получить несколько промокодов. Также компания приняла решение не отправлять все 7 промокодов на 15% одному пользователю. Сколько существует способов раздать все 8 промокодов? В ответе единицы измерения указывать не нужно.

   Примечание: Промокоды на 15% считайте одинаковыми: формально это разный набор букв и цифр, но важен именно эффект промокода, а не уникальное сочетание символов.

   Ответ: 1625

   Решение.

   1) Для начала разберемся с 7 промокодами на 15%. Их нужно раздать 5 пользователям, и каждый промокод должен достаться кому-то, причем ровно одному из пользователей.

      Пронумеруем пользователей: 1,2,3,4,5 и расположим их номера в одной строке в порядке возрастания. Каждый промокод обозначим за N и также поместим в эту строку на какую-то позицию между цифрами или после цифр. Тогда можно сказать, что комбинация

      1NN2N3NNN45N

      соответствует случаю, когда 1й пользователь получил 2 промокода, 2й — 1, 3й — 3, 4й — 0, и 5й — 1. Взяв все строки такого вида, начинающиеся с 1, где цифры идут в порядке возрастания, получим все возможные способы раздать 7 промокодов.

      Чтобы получить строку такого вида, нужно из 7 + 5 – 1 = 11 позиций (строка должна начинаться с 1) выбрать 4 позиции для номеров пользователей 2, 3, 4, 5 (на всех остальных позициях будут стоять N-ки). Значит, всего таких строк $C_{7+5-1}^{5-1}=C_{11}^{4}$. При этом 5 из них соответствуют случаям, когда все 7 промокодов на 15% получает ровно один пользователь. По условию это невозможно, поэтому остается C_{11}^{4} — 5 способов.

2) Промокод на 25% получит один из 5 пользователей, поэтому способов раздать его 5.

3) Можно считать, что промокоды на 15% и промокод на 25% раздаются независимо друг от друга, поэтому способов раздать все 8 промокодов итого:

$(C_{11}^4 - 5) \cdot 5 = 325 \cdot 5 = 1625$.

Примечание: задача поиска кол-ва способов распределения n промокодов на 15% между k пользователями тождественна задаче поиска кол-ва решений уравнения $x_1 + x_2 + \cdot + x_k = n$ в целых неотрицальных числах — и называется задачей Муавра. А искомое кол-во согласно выведенной нами формуле равно C_{n+k-1}^{k-1}.

7. (5 баллов) Лена работает маркетологом в банке. Один из продуктов банка — платная подписка, дающая различные бонусы клиенту. Лена решила прорекламировать подписку клиентам без подписки, которые больше всего похожи на клиентов с подпиской. Для этого она случайным образом выбрала группу А из 1000 клиентов с подпиской и группу Б из 100 000 клиентов без подписки.

Далее Лена рассчитала значения 8 параметров для каждого из клиентов в каждой из групп: возраст, доход в рублях в месяц и т.д. Все параметры — целые неотрицательные числа. Оказалось, что не существует двух клиентов из выбранных 101 000 с полностью совпадающим набором значений этих 8 параметров.

Дальше Лена объединила выборки А и В и произвела стандартизацию значений каждого из 8 параметров для всех 101 000 клиентов по формуле:

$$x_{new} = \frac{x - m}{\sqrt{D}}$$

Здесь x — значение параметра, m — среднее арифметическое значений параметра, D — дисперсия значений параметра, а x_{new} — стандартизированное значение параметра.

Так Лена получила для каждого клиента X из группы А значения стандартизированных 8 параметров $x_1, x_2, ..., x_8$, а для каждого клиента Y из группы В — значения $y_1, y_2, ..., y_8$. Наконец, для каждого клиента X из группы А Лена решила найти «наиболее похожего» на него клиента Y из группы В. «Наиболее похожий» на X клиент из группы В — такой, для которого принимает наименьшее значение (среди всех клиентов в группе В с их набором y_1, y2, ..., y_8) сумма:

$S = |x_1 - y_1| + |x_2 - y_2| + ... + |x_8 - y_8|$

Выберите одно или несколько верных утверждений:

а) Один и тот же клиент из группы В мог оказаться «наиболее похожим» на нескольких клиентов из группы А. (+)

б) Все значения всех параметров из групп А и В после стандартизации будут лежать в интервале [– 1; 1].

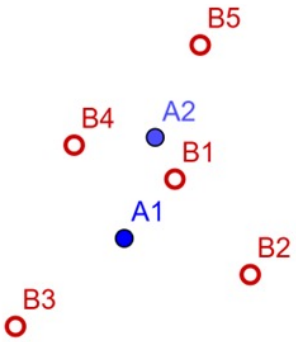
в) Может найтись клиент из группы А, для которого сумма значений всех параметров после стандартизации окажется равна 0. (+)

г) Некоторые клиенты из группы В могли не оказаться «наиболее похожими» ни на кого из клиентов из группы А. (+)

д) Среди других утверждений нет верных.

Решение.

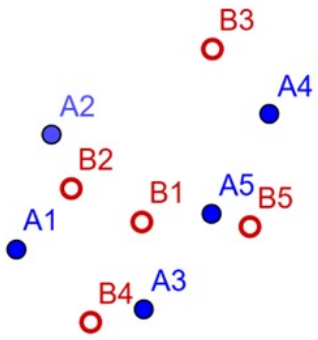
а) Верно. Ничто не запрещает клиенту B1 из группы B оказаться «наиболее похожим» одновременно на клиентов A1 и A2 из группы A. Ниже приведен пример такого случая для упрощенного двумерного варианта (когда параметров 2, а не 8). Здесь каждый клиент отображен точкой на координатой плоскости, а расстояние между каждыми двумя точками равно значению S — степени «похожести» клиентов.



б) Неверно. Стандартизация значений не гарантирует попадание новых значений в интервал [–1; 1]. Хотя и большинство значений действительно будет лежать в этом интервале.

в) Верно. Несмотря на то, что в данном случае сумма значений всех параметров после стандартизации сама по себе ничего не значит, она может оказаться равна 0. Например, после стандартизации в группе A может оказаться клиент со значениями параметров , и как следствие сумма значений параметров для него тоже окажется равной 0.

г) Верно. Клиент B1 из группы B может не оказаться «наиболее похожим» ни на одного клиента из A. Ниже приведен пример такого случая для упрощенного двумерного варианта (когда параметров 2, а не 8). Здесь каждый клиент отображен точкой на координатой плоскости, а расстояние между каждыми двумя точками равно значению S — степени «похожести» клиентов.



д) Так как среди других утверждений есть верные, то это утверждение неверно.

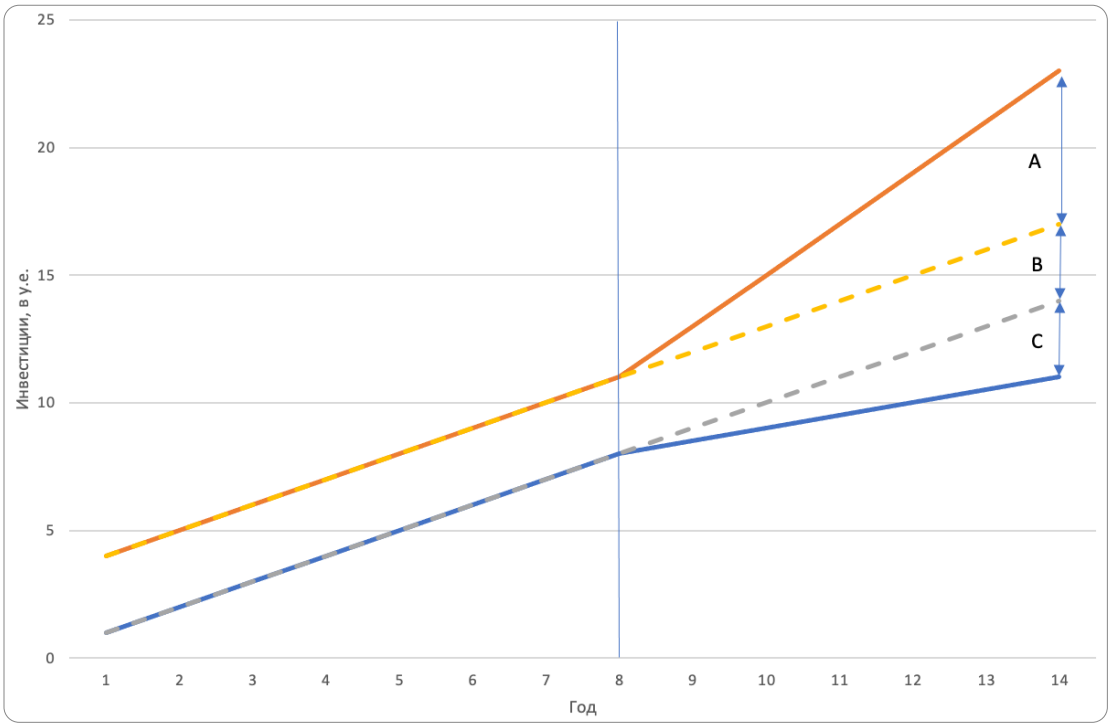
8. (4 балла) Вы являетесь главным экономистом в одной далекой-далекой стране, где есть два очень похожих друг на друга региона — Север и Юг. В 8-й год правительство Севера решает снизить налоги на инвестиции, на Юге этого не происходит. В это же время и Север, и Юг сталкиваются с резким ужесточением

требований для выдачи кредитов на инвестиции, что оказывает сдерживающий эффект на оба региона.

В 14-м году вас просят оценить эффект от снижения налогов. Собранные данные показывают, что инвестиции ведут себя, как показано на рисунке ниже. Оранжевая линия — фактическая динамика инвестиций на Севере, синяя — на Юге. Желтая и серая пунктирные линии показывают, как вели бы себя инвестиции, если бы в 8-м году ничего не менялось.

Величина отрезка A составляет 6 условных единиц (у.е.), отрезка B — 3 у.е., отрезка C — 3 у.е. Чему будет равен эффект от снижения налогов в условных единицах?

Запишите ответ в виде числа. Если ответ представлен в виде дроби, округлите его до целого. Единицы измерения указывать не нужно.

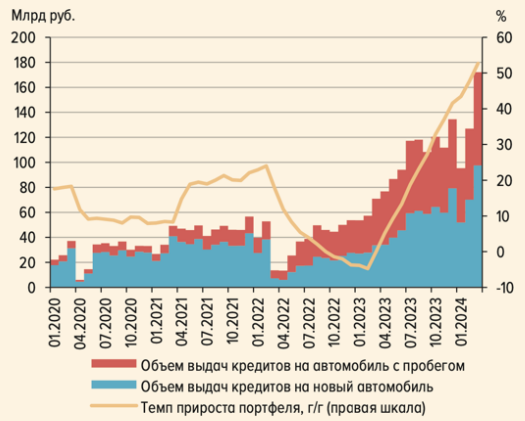


Ответ: 9

Решение: В условии сказано о том, что на Севере происходит снижение налогов, а также на Севере и на Юге ужесточаются требования к инвестиционным кредитам. Таким образом, если бы не изменение налогов, то линии продолжили бы идти параллельно (и оранжевая линия шла бы параллельно синей), и эффект от ужесточения требований составляет 3 у.е. (отрезок C). Однако на Севере не происходит снижения на 3 у.е., а происходит рост на 6 у.е. (отрезок A). Таким образом, эффект от снижения налогов на Севере составляет 9 у.е.: 6 у.е. (на которые инвестиции выросли относительно состояния без каких-либо изменений в налогах и требованиях к кредитам) плюс 3 у.е. (на которые инвестиции должны снизиться из-за эффекта увеличения требования по кредитам).

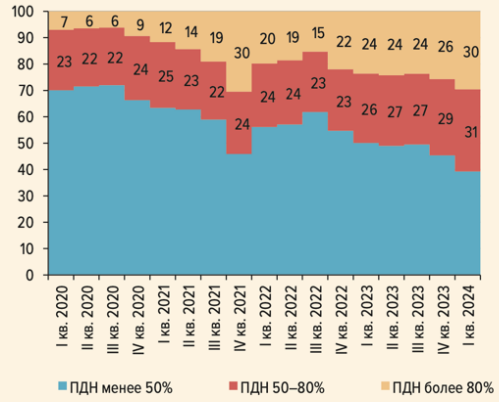
9. (5 баллов)

ДИНАМИКА ВЫДАЧ И ТЕМПЫ РОСТА ПОРТФЕЛЯ АВТОКРЕДИТОВ *Рис. 17*



Источники: формы отчетности 0409115 и 0409704.

РАСПРЕДЕЛЕНИЕ ВЫДАННЫХ АВТОКРЕДИТОВ ПО ПДН *Рис. 18*



Источник: форма отчетности 0409704.

На рисунке представлены два графика из Обзора финансовой стабильности ЦБ РФ за 4-й квартал 2023 и 1-й квартал 2024 года (cbr.ru/Collection/Collection/File/49151/4_q_2023_1_q_2024.pdf).

На левом графике представлены объемы выдач кредитов на автомобили с 2020 года, а также темп прироста общего портфеля автокредитов. На правом графике представлено распределение кредитов по величине показателя долговой нагрузки (ПДН) — это отношение между суммой ежемесячных платежей человека по всем кредитам к ежемесячному доходу. Чем выше данный показатель, тем больше шанс, что заемщик не сможет вернуть взятый кредит.

Выберите одно или несколько верных утверждений, которые можно сделать на основе данных графиков

а) В первом квартале 2023 года портфель автокредитов снижался (+)

б) За 2022-2023 гг. рискованность выданных автокредитов выросла (+)

в) В январе 2021 года доля кредитов, выданных на автомобили с пробегом, в общем объеме выдач была выше, чем доля этих кредитов в январе 2023-го

г) Объем выдач кредитов на новые автомобили вырос с 2022 года (+)

д) Среди других утверждений нет верных.

Решение. Портфель автокредитов характеризуется желтой линией на левом графике (Рис. 17). Важно понимать, что там отражаются темпы прироста, а значит портфель растет, если эта линия находится выше нуля, и падает, если линия находится ниже. Первый квартал 2023 года находится между значениями 01.2023 и 04.2023, и весь этот период на графике значения темпов прироста не превышают нуля.

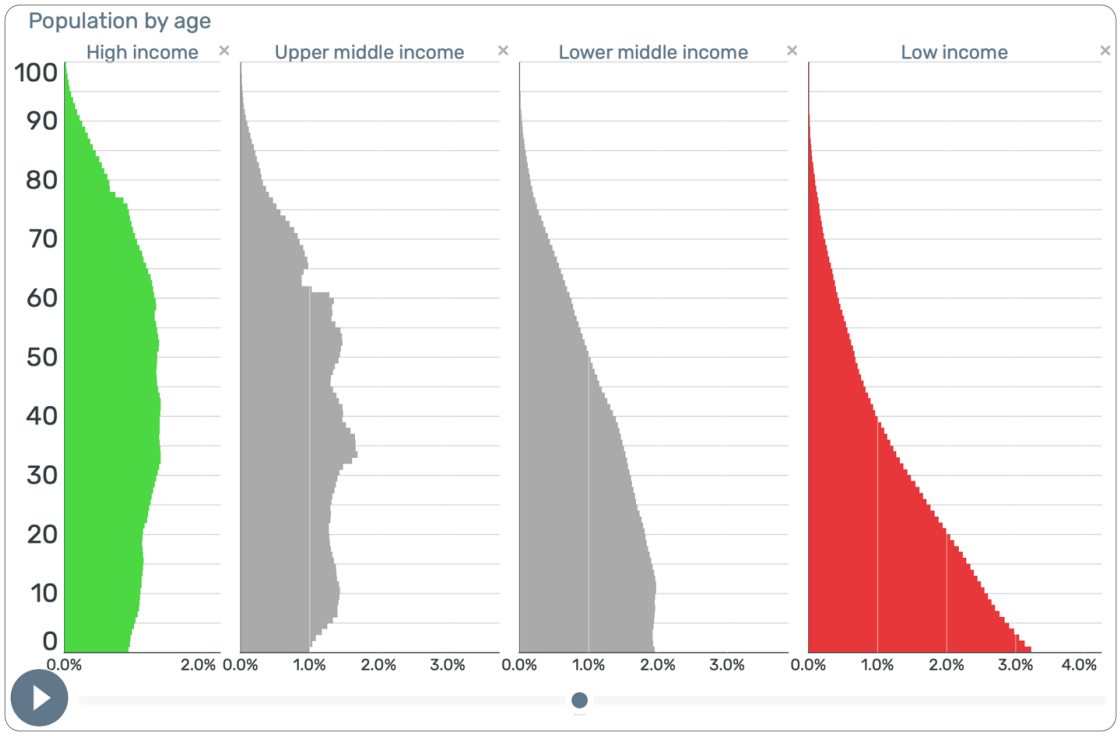
Рискованность портфеля оценивается по доле кредитов с высоким ПДН. За 2022 и 2023 год выросла и доля кредитов с ПДН более 50%.

Доли выданных кредитов на автомобили с пробегом и на новые оцениваются по соотношению частей столбцов на левом графике. В данном случае в январе 2021 года доля кредитов на автомобили с пробегом (доля красного в общей высоте

соответствующего столбца не больше 25%) сильно меньше, чем в январе 2023 года (красного около половины).

Объем выдач кредитов с 2022 года вырос: около 40 млрд рублей в январе 2022 по сравнению со 170 млрд рублей в марте 2024 (последние доступные данные). Кредиты действительно падали в марте и апреле 2022, однако здесь вопрос о том, выросли ли они в целом с 2022 года по последний доступный нам момент времени.

10. (5 баллов) На графиках представлены возрастные пирамиды (гистограммы, отражающие процент населения в каждой возрастной группе; в данном случае их 100 — одна группа соответствует каждому возможному значению возраста от 0 до 100) для четырех групп стран (на основе классификации Всемирного банка blogs.worldbank.org/en/opendata/new-world-bank-group-country-classifications-income-level-fy24), слева для самых богатых стран, справа — для самых бедных. Данные представлены на 2023 год.



Выберите одно или несколько верных утверждений на основе данных графиков:

а) В самых богатых странах происходит снижение отношения числа детей до 10 лет к численности взрослых в возрасте 30-40 лет в течение последних 10 лет (+/-)

б) В странах с самым высоким уровнем дохода в среднем самое молодое население

в) В странах с высоким доходом в ближайшие годы необходимо уделить большое внимание постройке новых школ для обеспечения всех детей возможностью для обучения

г) В странах с доходом ниже среднего люди в среднем моложе, чем в странах с низким доходом.

д) Среди других утверждений нет верных.

Решение. Пункт а) в оценивании не учитывался, поскольку действительно корректно оценить изменение соотношение за 10 лет по статическому графику без дополнительных предположений нельзя.

Соотношение средних возрастов можно оценить по долям людей каждого возраста и их соотношению. Так, в богатых странах людей в возрасте до 20 лет точно меньше 2%, в то время как в бедных странах (самый правый график) их точно больше. При этом обратная картина наблюдается для более старших возрастов (старше 40 лет). Из этой логики в среднем самое молодое население в самых бедных странах, а значит оно моложе, чем в странах с доходом ниже среднего.

В странах с высоким доходом очевидный тренд на снижение количества детей (по сути на снижение рождаемости). Поэтому старых школ достаточно для обеспечения меньшего количества детей местами и вопрос создания дополнительных мест не стоит. В самых же бедных странах этот вопрос стоит острее. Даже несмотря на более высокий уровень детской смертности (да и в целом смертности из-за проблем с доступом к медицине), непопулярность образования среди бедных семей (дети рано начинают работать и помогать с хозяйством родителям), если сравнить число детей, которые в текущем году выпускаются из школы (группа около 18 лет) и которые придут в школу (около 6 лет), численность различается в 1,5 раза. Сложно представить себе ситуацию, в которой такое увеличение числа учеников исправляется естественным образом. Таким образом, если не уделять внимание этому вопросу, возникающий дефицит учебных мест будет вести к снижению уровня образованности населения, а значит и к снижению качества жизни. При этом важно понимать, что «уделить внимание» не равно «строить обязательно». Возможно в стране и так уделяется внимание данному вопросу, и поэтому где-то дефицита и не возникает. Больше необходимого строить также никто не будет, потому что это экономически неэффективно, и подобный контраргумент не валиден.

11. (5 баллов) Компании часто берут кредиты в виде кредитных линий. В этом случае по договору с банком они могут в любой момент получить деньги, но только в определенных пределах. Обычно эти пределы восстанавливаются, если компания вовремя возвращает деньги. Исследования показывают, что в кризисные времена компании начинают чаще пользоваться этим видом кредитования.

    Группа ученых решила выяснить, как активно российские компании использовали кредитные линии во время пандемии. Для этого они собрали данные о всех кредитах, которые компании получили в период пандемии, а также за несколько лет до этого для сравнения. Также кредиты разделили на две группы: старые кредитные линии и новые кредиты. Затем рассчитали долю кредитов первой группы в общем объеме займов. Эти доли в зависимости от финансового состояния компании (платит ли она зарплату, расплачивается ли по старым займам вовремя, получает ли прибыль и т. д.), был ли этот период пандемийным или нет, а также относится ли компания к отрасли, сильно пострадавшей от кризиса (например, туризм, общественное питание и т. д.), показали на графиках.

Рис. 3. Доля ранее одобренных кредитных линий в общем объеме займов для компаний из не подверженных пандемии отраслей, в разбивке по децилям

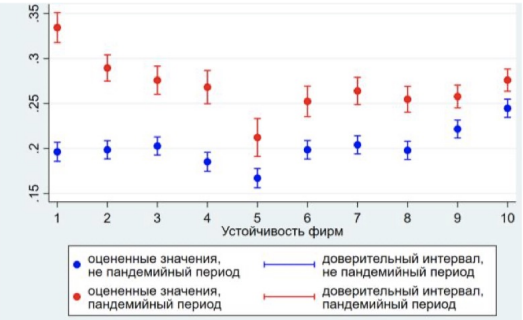
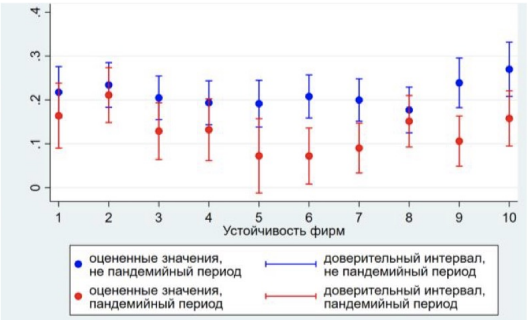


Рис. 4. Доля ранее одобренных кредитных линий в общем объеме займов для компаний из подверженных пандемии отраслей, в разбивке по децилям



Примечание. Оценки и доверительные интервалы для доли ранее одобренных кредитных линий в общем объеме заимствований компаний. Синие точки показывают оцененные значения для допандемийного периода (2017–2019 гг.), красные точки – для периода пандемии (2020 г.). Более высокие значения переменной «Устойчивость фирм» по шкале X соответствуют более высоким оценкам вероятности дефолта заемщика (то есть по более низким финансовым показателям компании).

Выберите одно или несколько верных утверждений, которые можно сделать на основании анализа указанных графиков

Обратите внимание на подпись к графикам: чем больше показатель «Устойчивость фирм», тем хуже финансовое состояние компании.

а) В допандемийный период менее устойчивые компании среди не подверженных пандемийному шоку активнее использовали имеющиеся у них кредитные линии по сравнению с более устойчивыми (+)

б) В пандемийный период компании из отраслей, подверженных пандемии, использовали кредитные линии с такой же интенсивностью или меньше по сравнению с допандемийным периодом (+)

в) В период до пандемии компании из отраслей, подверженных пандемии, использовали кредитные линии более интенсивно, чем компании из не подверженных шоку отраслей

г) В пандемийный период среди компаний, подверженных влиянию шока, самые финансово устойчивые (1-3 группы) и самые финансово неустойчивые (8-10 группы) компании пользовались кредитными линиями значительно активнее, чем средние фирмы (5 группа).

д) Среди других утверждений нет верных.

Решение. В п. а) необходимо посмотреть на положение синих точек (допандемийный период) на левом графике (неподверженные компании). Мы видим повышающийся тренд, при этом правый хвост находится статистически значимо выше левого. Отсюда можем сделать вывод, что менее устойчивые компании (которые находятся правее) использовали кредитные линии больше, чем более финансово устойчивые.

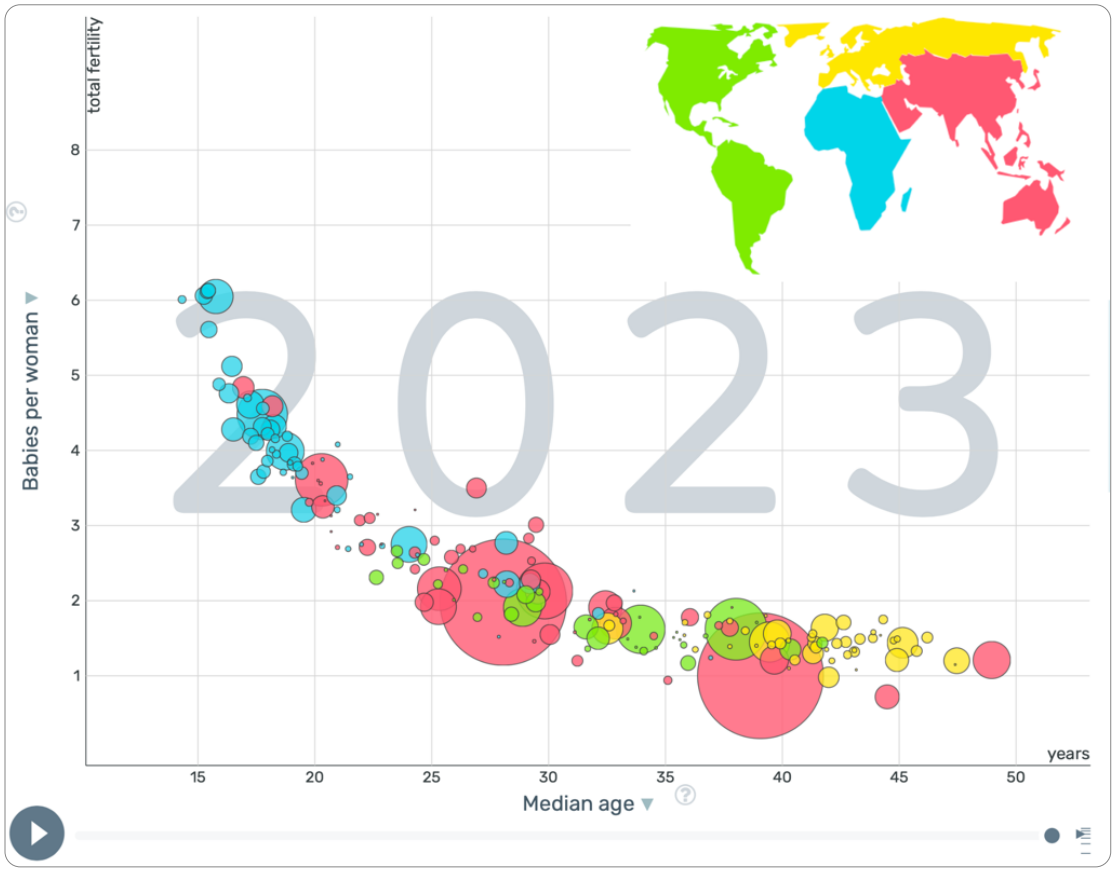
В п. б) необходимо сравнить синие (допандемийный период) и красные (пандемийный период) точки на правом графике (подверженные компании). Мы

видим, что красные точки с учетом доверительных интервалов в большинстве не отличаются от синих, а для некоторых групп фирм (например, 6 и 9 по шкале «Устойчивость фирм») лежат ниже.

В п. в) необходимо сравнить синие точки (допандемийный период) на левом графике (неподверженные компании) и на правом графике (подверженные компании). На первый взгляд может показаться, что справа точки выше, однако важно ориентироваться на шкалу и довольно широкие доверительные интервалы на правом графике, можно увидеть, что в обоих случаях значения находятся около 0,2 с подъемом до 0,25 в самом конце.

В п. г) смотрим на правый график (подверженные компании), красные точки (пандемийный период), и сравниваем три самые левые группы, три самые правые, и пятую в середине. С учетом доверительных интервалов, даже если найдется различие для одной группы из трех, нельзя сказать что все три наиболее устойчивые и три наименее устойчивые группы пользовались кредитными линиями активнее.

12. (5 баллов) На графике представлено распределение стран в зависимости от медианного возраста жителей (по горизонтальной оси), количества детей, приходящихся на одну женщину (по вертикальной оси) и размера населения (размер кругов).



Цвет кругов обозначает разные регионы мира в соответствии с картой.

Выберите одно или несколько верных утверждений, которые следуют из приведенного графика:

а) В среднем медианный возраст людей в странах Африки ниже, чем в Азии (включая Австралию, Океанию и Новую Зеландию) (+)

б) Среднее количество детей на одну женщину в Америке ниже, чем в Европе (включая Россию)

в) Чем больше численность населения в стране, тем больше медианный возраст людей в стране

г) Среднее количество детей на женщину в среднем отрицательно взаимосвязано с медианным возрастом (+)

д) Среди других утверждений нет верных.

Решение.

а) Африка отражается на графике синим цветом, Азия — красным. Большая часть стран Африки располагается в медианном возрасте 15-30 лет (большинство сконцентрировано в районе 15-20 лет), в то время как в странах Азии медианный возраст доходит до 50 лет (большая концентрация наблюдений в районе 25-30 лет). Поэтому в среднем в Африканских странах медианный возраст ниже, чем в Азии.

б) Америка отражена на карте зеленым цветом, Европа — желтым. Большая часть наблюдений в этих странах лежит между 1 и 2 детьми, однако в случае Америки есть ряд стран с количеством детей 2-3 на одну женщину. Соответственно, в Америке не может быть в среднем меньше детей, чем в Европе.

в) Два больших круга на графике — это Китай и Индия. Они находятся довольно близко к середине по медианному возрасту. Остальные круги так не выделяются и расположены более-менее равномерно по всем медианным возрастам, так что нельзя говорит о наличии положительной взаимосвязи между медианным возрастом и численностью населения.

г) На графике очевидна отрицательная взаимосвязь этих двух показателей. При этом важно отметить, что мы говорим именно о взаимосвязи, выводах о причинности в данном случае мы бы не смогли сделать.

13. (2 балла) Если бы вы хотели оценить взаимосвязь среднего количества детей на одну женщину и медианного возраста, то какое из приведенных ниже уравнений подошло бы больше всего для данных на рисунке выше? Выберите один верный вариант ответа. Все коэффициенты положительные.

а) b0-b1*Median Age

б) b0+b1/Median Age (+)

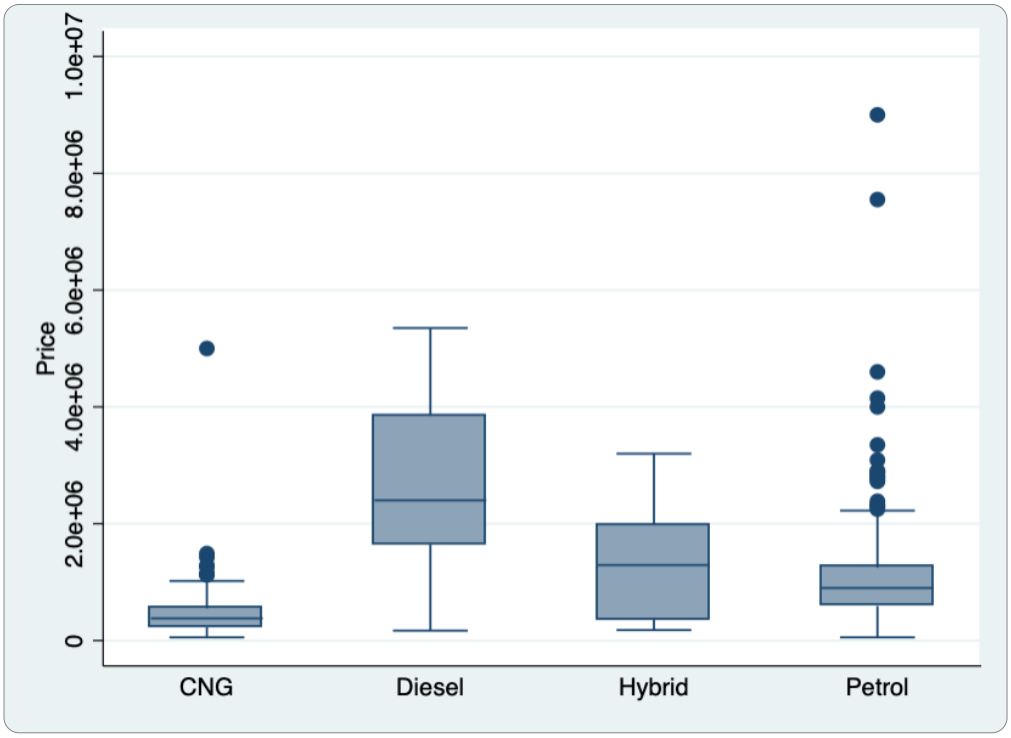
в) b0-b1*(Median Age)^2

г) b0+b1*(Median Age)^0,5

Решение. По графику мы видим убывающую взаимосвязь, больше всего похожую на гиперболу. Соответственно, из перечисленных наиболее подходящее уравнение регрессии имело бы вид

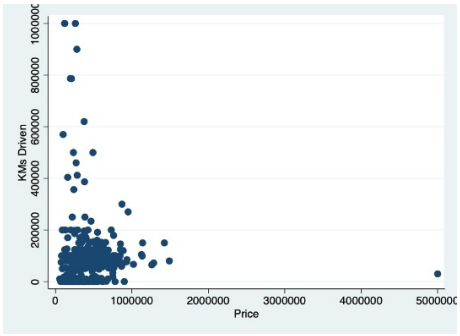
Babies per woman=b0+b1/Median age. Однако похожую форму может иметь график корня, если взять его с минусом и сдвинуть выше (то есть b0-b1*(Median Age)^0,5), либо функция логарифма, взятая с минусом и также сдвинутая выше (b0-b1*ln (Median age)). Важно, что функция должна отражать отрицательную взаимосвязь и иметь определенную форму (сначала сокращаться довольно быстро, а затем все медленнее, стремясь к горизонтали).

14. (8 баллов, по 2 за каждое верное соотнесение) На графике представлены «ящики с усами», отражающие распределение цены автомобиля в зависимости от типа топлива.

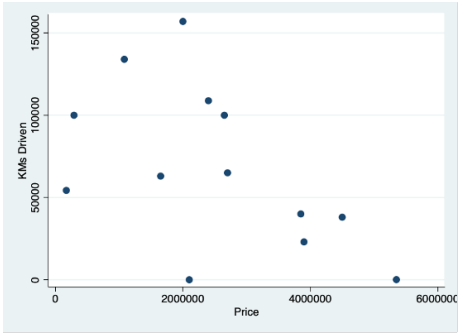


Также имеется 4 диаграммы рассеяния для изучения зависимости цены автомобиля от пробега по подгруппам в зависимости от топлива. Соотнесите диаграммы рассеивания и тип топлива автомобилей.

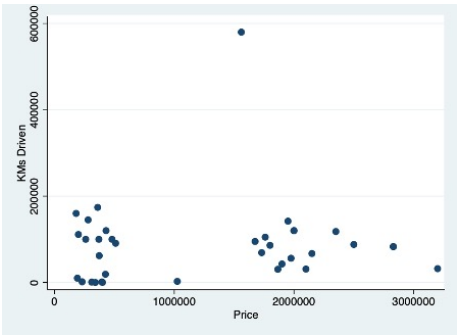
а)



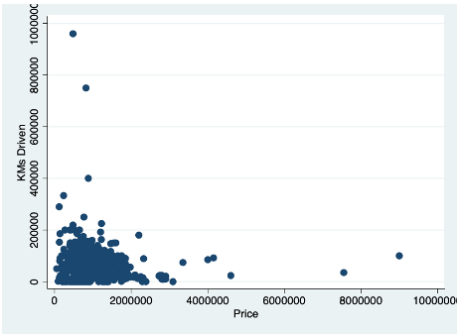
б)



в)



г)



Решение:

а) CNG (выявляется по явному одиночному выбросу Price=5 000 000)

б) Diesel (равномерно распределены наблюдения, почти от 0 до 6 000 000)

в) Hybrid (две ценовые группы на «ящике с усами» выделяться существенно не будут, однако здесь максимальная цена ниже, чем у предыдущего типа топлива)

г) Petrol (в основном отличается сильной концентрацией наблюдений от 0 до примерно 2 00 0000, и далее два сильно выделяющихся наблюдения справа: близкое к 8 000 000 и около 9 000 000)

15. (4 балла)

С использованием данных из предыдущей задачи была оценена регрессионная модель зависимости цены автомобиля от года выпуска, пробега, типа топлива и новизны. Получилось следующее уравнение (все перечисленные коэффициенты значимы на 5% уровне значимости):

Price= -62376.67 — 0.137*KMs + 31.52*Year — 277.11*D_used + 1863.70*D_Diesel

KMs — пробег в тысячах километров, Year — год выпуска, D_used — переменная, равная 1, если автомобиль подержанный («с пробегом»), и 0, если новый. Последняя переменная — индикатор для типа топлива (то есть D_Diesel=1, если это двигатель работает на дизельном топливе, и 0, если тип топлива другой).

Возьмем два автомобиля: оба 2000 года выпуска, подержанные, на дизельном топливе. Пробег одного из них составляет 60 тыс. километров, а второго — 120 тыс. километров. На сколько процентов дороже будет первый автомобиль по сравнению со вторым, исходя из полученных оценок модели? Укажите ответ в процентах от меньшей цены. Если ответ представлен в виде дроби, округлите его до двух знаков после запятой. Единицы измерения указывать не нужно.

Ответ: 0,37%

Решение: Из условия в уравнение необходимо подставить значение для соответствующих переменных (год выпуска 2000; индикатор автомобиля с пробегом равен 1, поскольку обе машины в примере с пробегом; аналогично индикатор автомобиля на дизельном топливе будет в обоих случаях равен 1; пробег у двух автомобилей одинаковый, в уравнение вносится без дополнительных нулей, поскольку было сказано, что KMs уже в тысячах). Разница в процентах указана в таблице ниже.

	коэффициенты	**Авто 1**	**Авто 2**
const	–62 376,67		
KMs	–0,14	60	120
Year	31,52	2000	2 000
D_used	–277,11	1	1
D_Diesel	1863,7	1	1
Оценка цены		2241,7	2 233,48
			0,3680%

16. (3 балла)

Для более полного анализа факторов, влияющих на цену, было решено учесть влияние марки автомобиля, и в модель добавили индикатор на одну из наиболее часто встречающихся марок, а именно на марку Suzuki. Коэффициент для данной переменной и 95% доверительный интервал для него получились следующие:

Переменная	**Коэффициент**	**[доверительный интервал]**	
Suzuki	–443.64	–514.86	–372.42

Какой вывод о взаимосвязи марки автомобиля на цену вы можете сделать исходя из данной информации? Выберите верный вариант ответа:

а) При прочих равных автомобили марки Suzuki стоят значимо дешевле, чем средний автомобиль другой марки (+)

б) При прочих равных автомобили марки Suzuki стоят значимо дороже, чем средний автомобиль другой марки

в) При прочих равных автомобили марки Suzuki значимо не отличаются по стоимости от среднего автомобиля другой марки

г) Приведенных данных недостаточно для того, чтобы сделать вывод о значимости различия цены автомобиля Suzuki от среднего автомобиля другой марки

Решение: Проверка гипотезы о значимости отличия коэффициента от 0 с помощью доверительного интервала проводится следующим образом: если 0 принадлежит доверительному интервалу, то мы не может сказать, что коэффициент значимо от него отличается, если 0 не принадлежит доверительному интервалу, то может сказать, что коэффициент значимо отличен от нуля. В последнем случае можем интерпретировать знак коэффициента. В данном случае индикатор ставился только на автомобили марки Suzuki, чтобы сравнить их цену со средней ценой машин остальных марок. В данном случае видим, что коэффициент отрицательный, а ноль не принадлежит доверительному интервалу. Значит машины марки Suzuki стоят значимо дешевле автомобиля другой марки.

17. (8 баллов) (см. файл 17_Распределение средств клиентов банков)

В таблице приведены данные о распределении средств различных клиентов банков (физических лиц, юридических лиц и ИП) в зависимости от размера вкладов: например, группа «<100» означает, что в этой группе собраны все вклады размером 100 тыс. рублей и меньше. Для каждого такого диапазона размера вклада и типа вкладчика указана общая сумма всех депозитов в млрд. руб. Вклады до 1,4 млн рублей являются застрахованными. Если размер вклада превышает эту сумму, то он является частично застрахованным.

Выберите одно или несколько верных утверждений, которые можно сделать на основе анализа данной таблицы.

а) На 1 июля 2024 г. в общем объеме вкладов наибольшую долю составляют депозиты размером более 20 млн. рублей (+)

б) За 1 квартал 2024 года увеличилась общая сумма вкладов физических лиц на депозиты в каждом диапазоне объемов

в) Темп прироста полностью застрахованных депозитов в июне 2024 года превышает среднемесячный темп прироста таких депозитов в 2023 году (+)

г) Все группы депозитов юридических лиц объемом свыше 700 тыс. рублей в апреле и мае 2024 года росли в среднем темпом не менее 1% в месяц (+)

д) Среди других утверждений нет верных.

Решение:

а) Для расчетов используем столбец R. Сумма вкладов на депозитах размеров более 20 млн. рублей составляет около 32% от всех. Эта доля наибольшая из всех групп (следующая по объему группа депозитов размером от 1,4 млн до 3 млн рублей.

б) В данном случае используем столбец K, и считаем темп прироста относительно 1 января 2024 года (т.е. столбца G). Можем заметить, что для самой первой группы происходит снижение объемов депозитов в объемах до 100 тыс. рублей. Соответственно приведенное в этом пункте утверждение неверное.

в) Темп прироста в июне считается исходя из отношения данных в столбце R к данным в столбце N. Темп прироста в 2023 году считается из отношения столбца F к столбцу B. Получаем 1,40%. Для получения среднемесячного темпа необходимо воспользоваться формулой сложных процентов, то есть из темпа роста (единица плюс темп прироста) взять корень 12 степени и вычесть 1. Получаем 0,93%.

г) Изменение депозитов юридических лиц за апрель и май оценивается по сравнению столбцов Q и M. Начиная с группы в 700 тыс. рублей и по возрастанию размера депозита, среднемесячные темпы прироста составили 1,57%, 2,39%, 2,45%, 2,99%, 2,59%, 2,62% и 4,32%. Все они больше 1%.

18. (8 баллов) 18_Hit.csv и 18_Hits_description

В файле 1 собраны данные о более чем 14 тысячах популярных песен с 1899 года. Описание характеристик можно найти в файле 2. Проанализируйте представленные данные и выберите одно или несколько верных утверждений.

а) Более энергичные песни в среднем более громкие (+)

б) Средняя продолжительность песни в жанре Folk составляет не более 3 минут

в) усредненное по песням жанра значение переменной Acoustic является наибольшим для жанра Jazz (+)

г) В жанре Country более продолжительные композиции являются более популярными, в то время как в жанре EDM напротив более популярны менее продолжительные композиции (+)

д) Среди других утверждений нет верных.

Решение.

а) Корреляция loudness и energy составляет примерно 0,75, она статистически значима. Значит более энергичные песни в среднем более громкие.

б) Находим среднее значение только среди песен, где значение genge=»-Folk». Можно предварительно перевести продолжительность каждой песни

из миллисекунд в минуты, можно сделать уже после расчета среднего. Получаем примерно 3,75 минуты, что больше 3 минут.

в) Находим среднее значение переменной Acousticness по каждому жанру. Далее выбираем из них максимальное. Оно действительно относится к жанру Jazz и составляет примерно 0,73.

г) Здесь необходимо сравнить корреляции двух переменных, которые мы находим для песен двух жанров: Country и EDM. В первом случае она составляет 0.29 и статистически значимая. Во втором случае она составляет -0,20 и она тоже статистически значимая. Таким образом, в жанре Country более популярны более продолжительные композиции, а в жанре EDM более популярны менее продолжительные композиции.

19. (4 балла, по 1 баллу за каждый верно заполненный пропуск) 19_stock_price_1.xlsx

Ваши коллеги из другого отдела занимаются учетом стоимости имеющихся у компании акций разных компаний. Данные за июль 2024 года представлены в файле stock_price. Там отражена цена акций четырех компаний: A, B, C и D. Вашим коллегам важно, чтобы цена у каждой бумаги была каждый день, и они прибегают к следующему способу вести свой учет. В дни, когда торги проводились (т.е. в рабочие дни), указывается цена закрытия этого дня. В другие дни (выходные или праздники) указывается последняя известная цена. Они просят вас отранжировать бумаги по их риску и выявить наиболее и наименее рискованные акции.

В качестве меры риска предполагается использование коэффициента вариации. Вы также знаете, что при анализе рисков следует ориентироваться только на дни, когда проводятся торги. Исходя из сделанных вами расчетов, запишите в соответствующие поля название наиболее и наименее рискованной акции, а также укажите коэффициент вариации для них. В ответе необходимо указать название компании и коэффициент вариации в долях. Если ответ представлен в виде дроби, округлите его до трех знаков после запятой. Единицы измерения указывать не нужно.

Наиболее рискованная акция компании « », коэффициент вариации « »

Наименее рискованная акция компании « », коэффициент вариации « »

Решение:

Для расчетов необходимо было удалить наблюдения за выходные дни (остается 23 наблюдения). Далее рассчитывается коэффициент вариации как отношение стандартного отклонения к среднему. Акции компания, где коэффициент вариации максимальный из четырех, являются наиболее рискованными, а где коэффициент минимальный — наименее рискованными. Поскольку не было сказано, выборка представлена или генеральная совокупность, засчитывались оба варианта подсчета стандартного отклонения.

Наиболее рискованная акция компании «B», коэффициент вариации «0,059»(0,058)

Наименее рискованная акция компании «C», коэффициент вариации «0,033»(0,032)

20. (8 баллов + 4 балла) 20_employee_v1.csv и 20_employee_description

Вам необходимо проанализировать ряд показателей, связанных с исполнением трудовых функций работников некоторой компании. Данные представленны в файле employee.csv, их описание — в файле employee_description.

В частности, руководство компании интересуется, какое влияние коллеги оказывают друг на друга (например, более производительные коллеги мотивируют работать лучше, или напротив, мешают выполнению обязанностей). В данном случае коллегами будут считаться люди, выполняющие одну и ту же роль (jobrole) в одном и том же департаменте (department) в одном регионе (state) в один и тот же год проведения оценивания. Оно проводится раз в год, дата проведения указана в переменной reviewdate.

Взаимосвязи между количественными переменными рассматриваются на основе статистической значимости соответствующей корреляции, количественной переменной с категориальной — на основе статистически значимого различия в средних на подвыборках.

Часть 1. Расчеты (8 баллов, по одному за каждое верно заполненное число)

Заполните пропуски необходимыми значениями. Если ответ представлен в виде дроби, округлите его до двух знаков после запятой. Единицы измерения указывать не нужно.

Для исследования влияния опыта работы сотрудника на оценку его деятельности менеджером будем ориентироваться на два показателя: время работы в компании и время работы на последней должности. В среднем продолжительность работы в компании в годах составляет приблизительно 5,70 (5,26), а продолжительность работы на последней должности в годах — 2,81 (2,59).

При анализе влияния коллег мы фокусируем внимание на 6015 (1154, 6013, 1153) работниках из общего количества в 6059, у которых есть хотя бы один коллега. Среднее значение оценок менеджера, полученных коллегами, составляет 3,47. В среднем у таких работников имеется 50.54 (50,56) коллег.

Коэффициент корреляции времени работы в компании с оценкой менеджера составляет 0.02. Коэффициент корреляции времени работы на последней должности с оценкой менеджера составляет 0.03. Коэффициент корреляции оценки менеджера для работника с оценкой его коллег составляет 0,00.

Комментарии по решению:

Количество коллег и среднее по оценкам, которые им выставил менеджер, не дано. Исходя из описания, коллеги определялись по нескольким факторам: роль, департамент, регион и год. После группировки по указанным характеристикам необходимо было посчитать количество людей в группе, за исключением самого человека, для которого считаем число, и среднюю оценку людей в группе, опять же за исключением самого себя. Эта средняя оценка коллег и есть та переменная, с которой необходимо было посчитать последнюю корреляцию.

Первые два числа (продолжительность работы в компании и на последней должности) можно рассчитывать из двух соображений. Во-первых, мы характеризуем целиком наши данные, которые являются несбалансированными панельными данными (когда много людей наблюдаются в течение нескольких периодов времени и имеют много характеристик, при этом некоторые люди уходят

из выборки, а некоторые появляются). Тогда мы можем просто посчитать среднее значение длительности работы в компании по всем наблюдениям. У этого подхода есть ограничение в виде того, что люди, работающие несколько лет, будут создавать искажение. Поэтому можно использовать второй подход — взять срез на последний год. В данной выборке последний год 2022, поэтому можно было посчитать среднее значение показателей только по этому году. Во всех остальных подходах (когда оставляют одно какое-то наблюдение или только последнее), результат не может быть никак проинтерпретирован. Других оснований для исключения наблюдений из подсчета среднего нет (в том числе, не подходит индикатор увольнения сотрудника, если по нему есть данные на 2022 год, потому что человек уволился после проведения опроса).

Переходим к подсчету количества работников. Да, действительно, если считать по уникальным идентификаторам работников, людей будет сильно меньше, чем 6059 даже в общем наборе. Однако в одном из подходов к работе с панельными данными наблюдения за одним человеком в разные периоды времени может приниматься за два разных наблюдения за разными людьми (так называемая сквозная регрессия). В силу того, что формулировку можно интерпретировать по-разному, засчитывалось как подсчитанное количество строк, которое остается без учета людей без коллег, так и количество уникальных идентификаторов. Более того, в выборке имеются люди, которые дважды проходили опрос в течение одного года (1 января и 31 декабря). Один из них из-за этого становился коллегой сам себе, и отсюда могло браться еще одно расхождение, однако если эта проблема с базой не была замечена, ошибкой это не считалось. На дальнейшие расчеты этот нюанс не оказывал влияния (кроме количества коллег, что также отражено в наличии второго возможного ответа).

Корреляции времени работы в компании и на должности необходимо было считать по всем людям, а не только по тем, у кого есть коллеги: в тексте задания, когда надо вписывать, уточняется, что при анализе влияния коллег мы ограничиваем выборку, но не при анализе влияния своего стажа. Статистическая значимость корреляции проверяется отдельным тестом (как проверяется значимость коэффициентов в уравнении регрессии, или значимость различия средних). Граничные значения, которые приводятся в некоторых учебниках (0,2; 0,3; 0,5 и пр.) — не определяются статистической значимостью. Они могут говорить о силе взаимосвязи, о близости точек к некоторой прямой, о некоторой «экономической значимости», но они не говорят о статистической значимости. В данном случае из трех регрессий статистически значимой на 5% уровне значимости оказывается лишь одна — корреляция времени работы на последней должности с оценкой менеджера. Можно ли при таком уровне говорить о том, что на scatter plot точки будут лежать близко к одной линии или вообще будут похожи на вытянутое в нужном направлении облако? Нет, не будут. Грубо говоря, из всего разброса одной переменной мы с помощью другой можем объяснить лишь 3%. Но этого достаточно, чтобы корреляция была статистически значимой.

Часть 2. Интерпретация. (4 балла, по 1 баллу за каждый верно заполненный пропуск)

Выберите наиболее подходящее заполнение пропуска из предложенных вариантов исходя из анализа, проведенного в Части 1.

Переходя к анализу влияния опыта работы, время работы в компании (связано положительно/связано отрицательно/не связано) с оценкой менеджера. Время работы на последней должности с оценкой менеджера (связано положительно/связано отрицательно/не связано).

Корреляция средней оценки менеджера, полученной коллегами, и показателя с оценкой самих работников является (положительной/отрицательной/нулевой), что говорит о (наличии положительной/наличии отрицательной/отсутствии) взаимосвязи между оценкой работника и средней оценкой его коллег.

Комментарии к решению: из трех корреляций статистически значимой и положительной оказалась лишь одна — корреляция оценки менеджера с временем работы на последней должности. Поэтому второй пропуск заполняется как «связано положительно». Более развернутый комментарий по поводу статистической значимости приведет выше. Две остальные корреляции статистически не отличимы от нуля, а значит взаимосвязи там нет, а про корреляцию мы можем сказать, что она нулевая, вне зависимости от полученного знака.