

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 у половины пострадавших людей имеется домашнее животное. Также вы знаете, что в вашем городе домашние питомцы имеются в 50% 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
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 ноябрь телефонов (в штуках) было продано меньше, чем аксессуаров. При этом 3% от всех товаров вернули. Выберите одно или несколько верных утверждений, которые следуют из этой информации:
 - а) В ноябре вернули не более 3% аксессуаров
 - б) В ноябре был хотя бы один покупатель, который приобрел в магазине аксессуар без телефона
 - в) В ноябре был хотя бы один покупатель, который приобрел в магазине телефон без аксессуара
 - г) Телефонов без аксессуаров купили меньше, чем аксессуаров без телефонов
 - д) Среди других утверждений нет таких, которые следуют из приведенной в условии информации (+)
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 выходят после занятий из школы. В результате полученная вами оценка количества часов чтения:
 - а) Будет завышена
 - б) Будет занижена
 - в) Будет близка к истинному значению (+)
 - г) Может быть как завышена, так и занижена

4. (4 балла) В городе Банкнотск 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 Банкнотск одновременно является и клиентами банка А, и клиентами банка Б.

Петя запустил в приложении банка А опрос, доступный клиентам-жителям Банкнотска. Света опрашивала людей, выходящих из разных отделений банка Б в городе Банкнотск. 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 Петинем опросе приняли участие 1504 человека, из них 512 являются по их словам клиентами банка Б. В опросе Светы приняли участие 2235 человек, из них 1096 являются по их словам клиентами банка А.

В результате Петя сделал вывод, что 34% жителей является клиентами обоих банков, а Света что таких жителей 49%. 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 меньше, чем оценки Светы и Пети
 - д) Недостаточно данных, чтобы сделать выводы (+)
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 192 билета на рейс со 190 местами. За 15 минут до окончания регистрации трое из 192 пассажиров не зарегистрированы на рейс — Аркадий А., Ирина К. и Никита П. По статистике предыдущих перелетов Аркадий А. придет на регистрацию с вероятностью 80%, Ирина К. — 40%, Никита П. — 20%. 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: 43

6. (5 баллов) Компания, занимающаяся доставкой продуктов, решила подарить пяти своим самым активным пользователям 6 одноразовых промокодов: 5 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 5 промокодов на 15% одному пользователю. Сколько существует способов раздать все 6 промокодов?

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

Ответ: 605

7. (5 баллов) Лиза работает маркетологом в банке. Один из продуктов банка — платная подписка, дающая различные бонусы клиенту. Лиза 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 В из 100 000 клиентов без подписки. Далее Лиза 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 Лиза 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 Лиза получила для каждого клиента X из группы А значения стандартизированных 8 параметров $x_1, x_2, \dots, x_8$, а для каждого клиента Y из группы В — значения $y_1, y_2, \dots, y_8$. Наконец, для каждого клиента X из группы А Лиза 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 $y_1, y_2, \dots, y_8$) сумма:

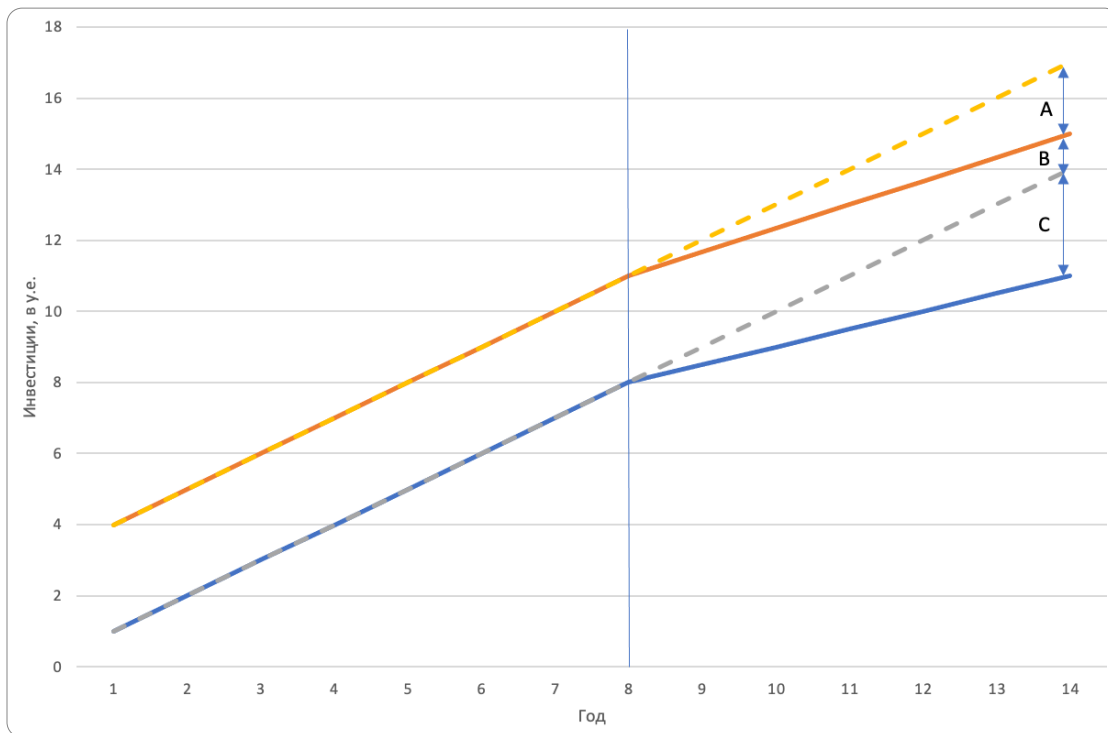
$$S = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_8 - y_8|$$

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 группы А после стандартизации окажутся не больше 1.
 - в) Может найтись клиент из группы В, для которого сумма значений всех параметров после стандартизации окажется равна 0. (+)
 - г) Каждый клиент из группы В оказался «наиболее похожим» хотя бы на одного клиента из группы А.
 - д) Среди других утверждений нет верных.
8. (4 балла) Вы являетесь главным экономистом в одной далекой-далекой стране, где Север и Юг отделены друг от друга, но при этом они очень похожи друг на друга. В 8-й год правительство Севера решает снизить налоги на инвестиции. Однако 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.

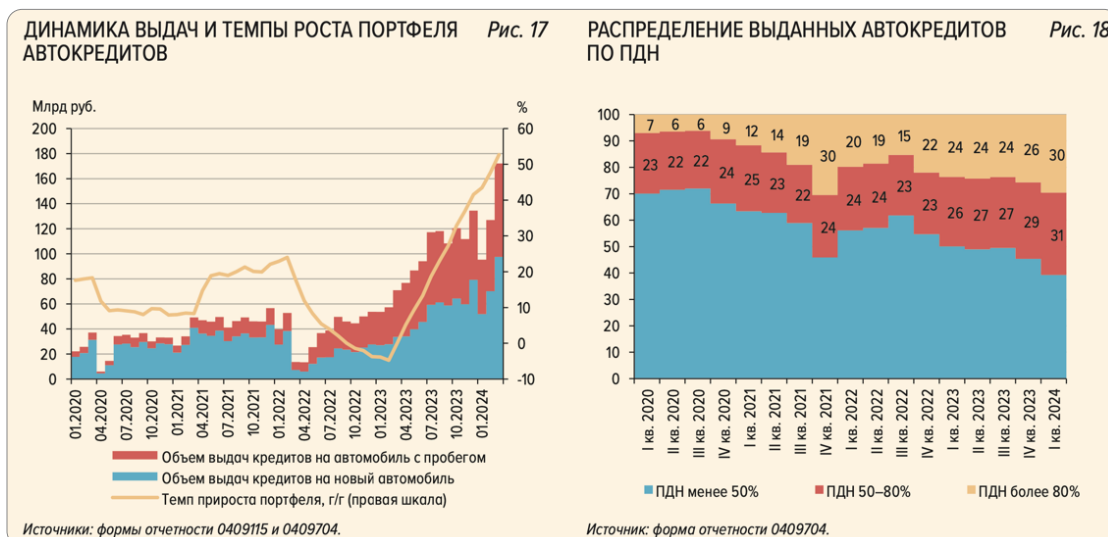
В 14-м году вас просят оценить эффект от снижения налогов. Собранные данные показывают, что инвестиции ведут себя, как показано на рисунке ниже. Оранжевая

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 А составляет 2 условных единицы (у.е.), отрезка В — 1 у.е., отрезка С — 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: 1

9. (5 баллов)

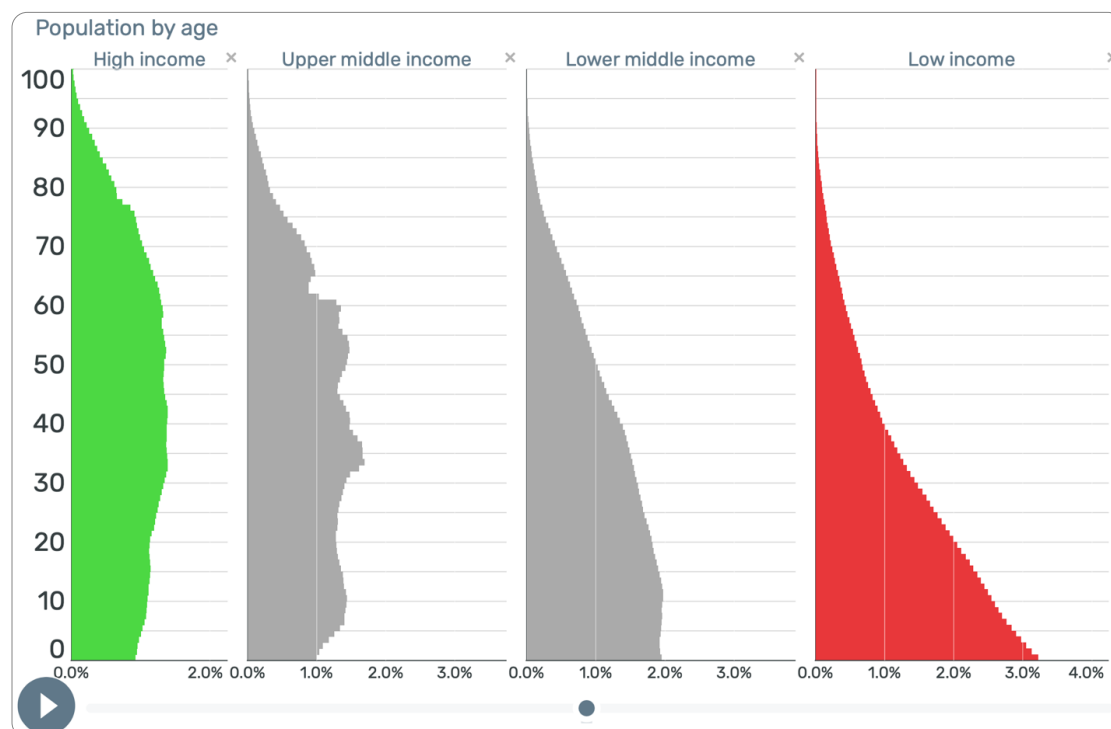


На рисунке представлены два графика из Обзора финансовой стабильности ЦБ РФ за 4-й квартал 2023 и 1-й квартал 2024 года (cbr.ru/Collection/Collection/File/49151/4_q_2023_1_q_2024.pdf).

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 третьем квартале 2023 года портфель автокредитов снижался
 - б) За 2020г. рискованность выданных автокредитов выросла (+)
 - в) В январе 2021 года доля кредитов, выданных на новые автомобили, в общем объеме выдач ниже, чем в январе 2024
 - г) Объем выдач кредитов на автомобили с пробегом в январе 2022 года вырос по сравнению с январем 2021 года. (+)
 - д) Среди перечисленного выше нет верных ответов
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 бедных странах происходит рост отношения числа детей до 10 лет к численности взрослых в возрасте 30-40 лет в течение последних 10 лет (+)

- б) В странах с самым низким уровнем дохода в среднем самое молодое население (+)
- в) В странах с доходом выше среднего 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 доходом выше среднего. (+)
- д) Среди других утверждений нет верных.

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

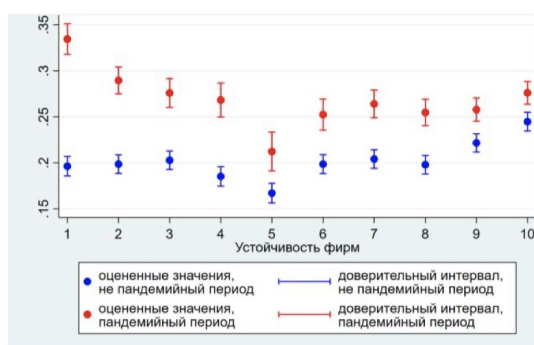
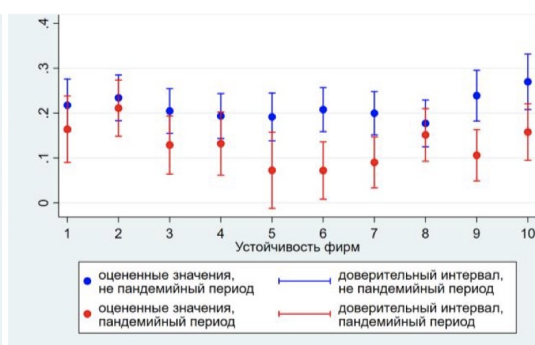


Рис. 4. Доля ранее одобренных кредитных линий в общем объеме займов для компаний из подверженных пандемии отраслей, в разбивке по децилям

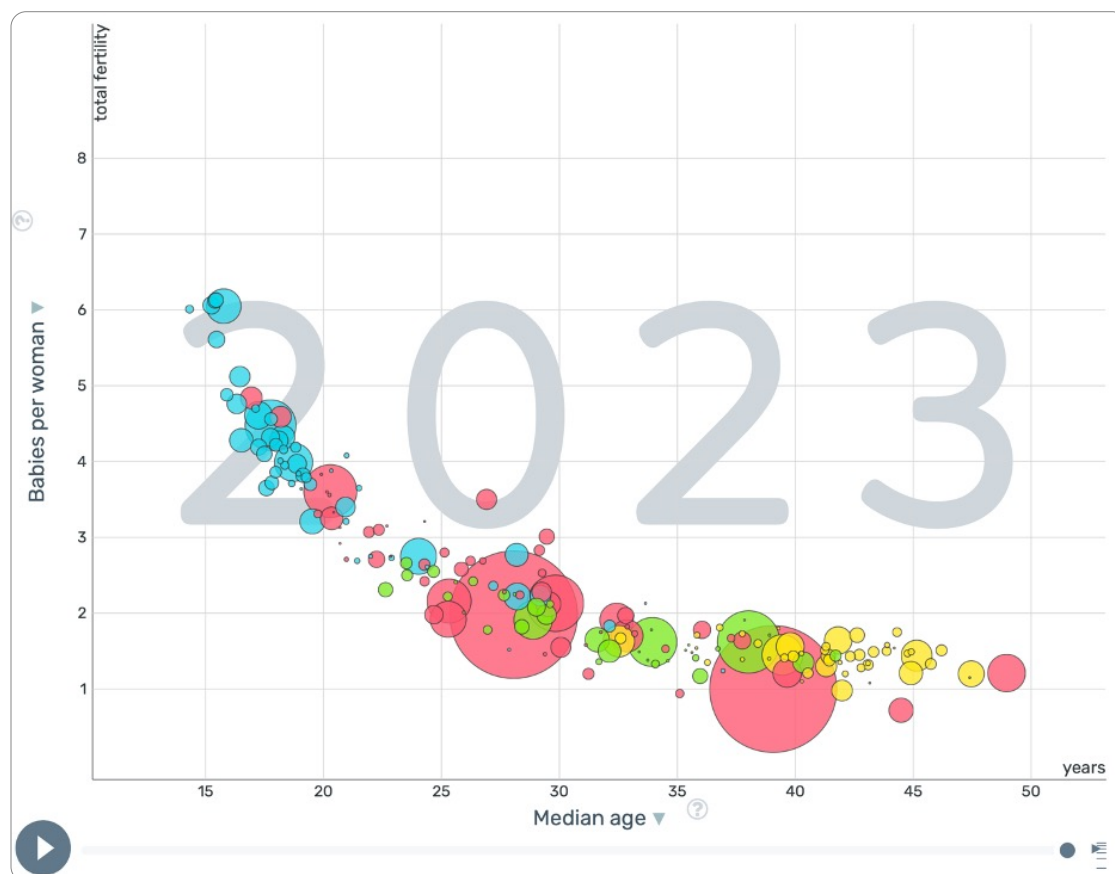


Примечание. Оценки и доверительные интервалы для доли ранее одобренных кредитных линий в общем объеме заимствований компаний. Синие точки показывают оценочные значения для допандемийного периода (2017–2019 гг.), красные точки – для периода пандемии (2020 г.). Более высокие значения переменной «Устойчивость 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?

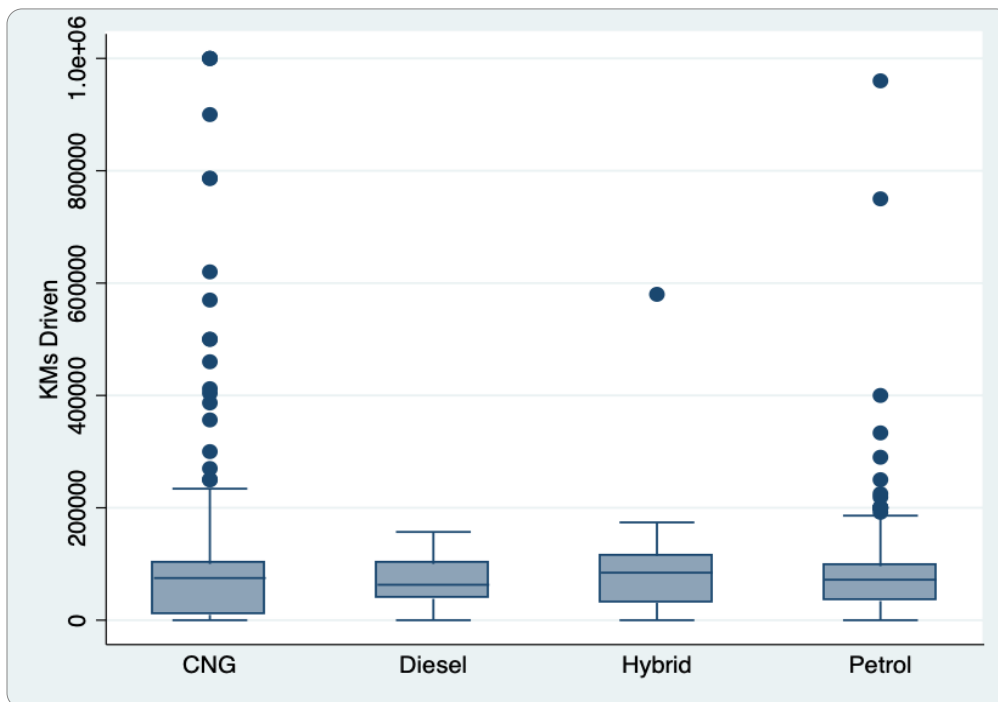
Обратите внимание на подпись к графикам: чем больше показатель «Устойчивость фирм», тем хуже финансовое состояние компании

- а) В пандемийный 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, не подверженных пандемии, использовали кредитные линии с такой же интенсивностью или больше, по сравнению с допандемийным периодом (+)
 - в) В период до пандемии компании из отраслей, подверженных пандемии, использовали кредитные линии менее интенсивно, чем компании из не подверженных шоку отраслей
 - г) В допандемийный период среди компаний, подверженных влиянию шока, самые финансово устойчивые (1-3 группы) и самые финансово неустойчивые (8-10 группы) компании пользовались кредитными линиями так же активно, как средние фирмы (5 группа). (+)
 - д) Среди других утверждений нет верных.
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: голубой — Африка, красный — Азия (включая Австралию и Новую Зеландию), зеленый — Америка (Северная и Южная вместе), желтый — Европа.

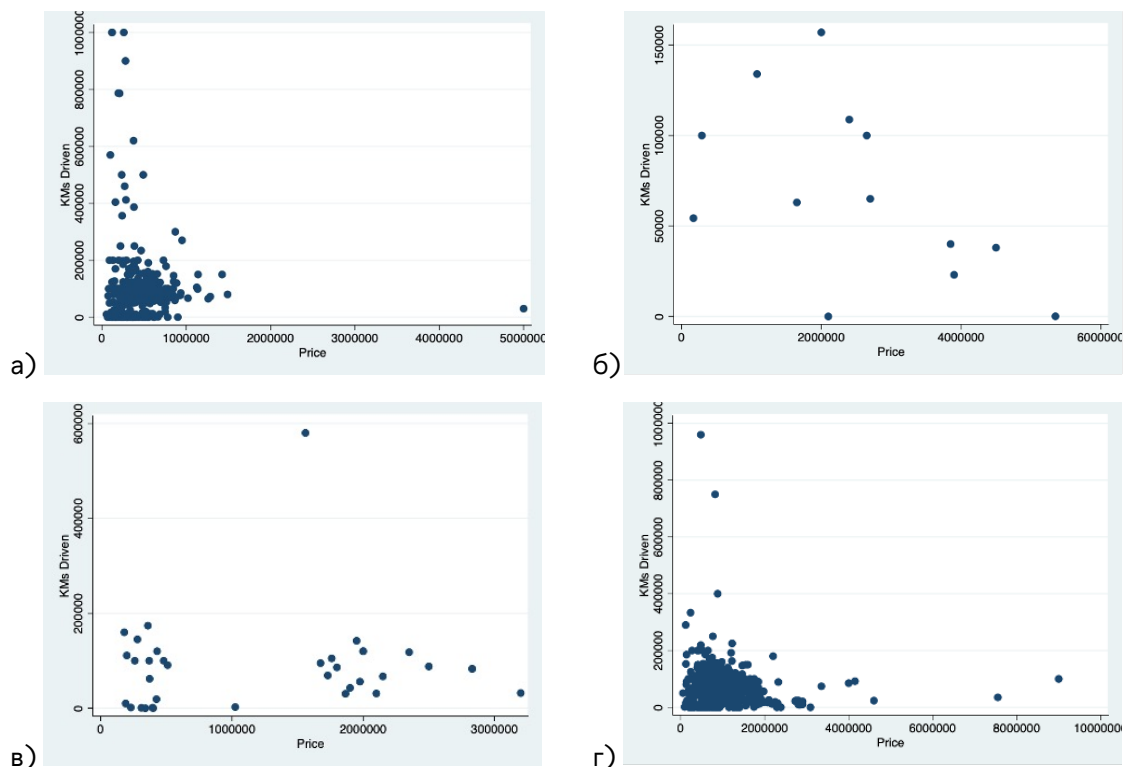


Выберите одно или несколько верных утверждений, следующие из приведенного графика:

- а) В среднем медиана медианный возраст в странах Азии (включая Австралию, Океанию и Новую Зеландию) ниже, чем в Европе (включая Россию) (+)
 - б) Среднее количество детей на одну женщину в Африке выше, чем в Америке (+)
 - в) Чем больше численность населения в стране, тем больше медианный возраст людей в стране
 - г) Рост среднего количества детей на женщину ведет к снижению медианного возраста
 - д) Среди других утверждений нет верных.
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) $b_0 - b_1 * \text{Median Age}$
 - б) $b_0 - b_1 * \ln(\text{Median age})$ (+)
 - в) $b_0 + b_1 * (\text{Median Age})^2$
 - г) $b_0 + b_1 * (\text{Median Age})^{0,5}$
14. (8 баллов, по 2 за каждое верное соотнесение) На графике представлены ящики с усами, отражающие распределение пробега 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



- a) CNG
- б) Diesel
- в) Hybrid
- г) Petrol

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):

$$\text{Price} = -62376.67 - 0.137 \cdot \text{KMs} + 31.52 \cdot \text{Year} - 277.11 \cdot \text{D_used} + 1863.70 \cdot \text{D_Diesel}$$

KMs — пробег в тысячах километров, Year — год выпуска, D_used — переменная, равная 1, если автомобиль использованный, 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 30 тыс. километров, а второго — 60 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: 0,183%)

16. (3 балла) Для более полного анализа факторов, влияющих на цену, было решено учесть влияние марки автомобиля, и добавили индикатор на одну из наиболее часто встречающихся марок, а именно на марку Honda. Коэффициент для данной переменной, и 95% доверительный интервал для него получились следующие:

Переменная	Коэффициент	[доверительный интервал]	
Honda	145.42	45.39	245.45

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 Honda стоят значительно дешевле, чем средний автомобиль другой марки
 - б) При прочих равных, автомобили марки Honda стоят значительно дороже, чем средний автомобиль другой марки (+)
 - в) При прочих равных, автомобили марки Honda значительно не отличаются по стоимости от среднего автомобиля другой марки
 - г) При прочих равных, приведенных данных недостаточно для того, чтобы сделать вывод о значимости различия цены автомобиля Honda от среднего автомобиля другой марки
17. (8 баллов) (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. более трети всех вкладов ИП приходилось на депозиты размером более 20 млн. рублей
 - б) За 1 квартал 2024 года увеличилась общая сумма вкладов юридических лиц на депозиты в каждом диапазоне объемов (+)
 - в) Среднемесячный темп прироста полностью застрахованных депозитов в 1 квартале 2024 года превышает среднемесячный темп прироста таких депозитов в 2023 году
 - г) Все группы депозитов ИП объемом свыше 1400 тыс. рублей в апреле и мае 2024 года росли в среднем темпом не менее 1% в месяц
 - д) Среди других утверждений нет верных.
18. (8 баллов) (Файлы 18_Hit.csv и 18_Hits_description)

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

- а) Средняя продолжительность песни в жанре Jazz составляет не более 5 минут
- б) Со временем акустические песни встречаются реже (+)
- в) Усредненное по песням жанра значение переменной Liveness является наибольшим для жанра Gospel (+)
- г) В жанре R&B более продолжительные композиции являются более популярными, в то время как в жанре Pop напротив более популярны менее продолжительные композиции
- д) Среди других утверждений нет верных.

19. (4 балла, по 1 баллу за каждый верно заполненный пропуск)
(Файл 19_stock_price_3.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.

Наиболее рискованная акция компании «А», коэффициент вариации «0,045»(0,044)

Наименее рискованная акция компании «В», коэффициент вариации «0,025»

20. (8 баллов + 4 балла) (Файлы 20_employee_v3.csv и 20_employee_description)

Вам необходимо проанализировать ряд показателей, связанных с исполнением трудовых функций работников некоторой компании. 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. Она 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)

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 (5,27), а продолжительность работы на последней должности в годах — 2,81 (2,60).

При анализе влияния коллег мы фокусируем внимание на 5741 (1151, 5739, 1150) работников из общего количества в 5785, у которых есть хотя бы один коллега. Среднее значение оценок менеджера, полученных коллегами, составляет 3,48. В среднем у таких работников имеется 48,13 (48,15) 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 0.01.

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.