Всем привет!

В этом файле вы можете ознакомиться с решениями заданий одного из вариантов второго (отборочного) этапа в 2022/2023 учебном году. Задания других вариантов решаются аналогично.

1. Средняя оценка пользователей (2 балл)

На одном из сайтов с оценкой недвижимости рейтинги жилых комплексов формируются как средняя оценка пользователей по пятибалльной шкале без округления. Жилой комплекс «Сердце столицы» компании «Донстрой» имеет рейтинг 4,7 по оценкам 820 пользователей. Какое наименьшее количество пользователей должно еще выставить свои оценки, чтобы рейтинг этого жилого комплекса увеличился хотя бы на 0,1 без округления?

а) 420

б) 220

в) 410 (+)

г) 210

д) Среди остальных ответов нет ни одного верного

Решение и комментарии. Чтобы рейтинг жилого комплекса увеличился при наименьшем количестве новых пользователей N, все эти пользователи должны выставить оценки «5». Тогда пользователей станет всего $820 + N$ человек, и сумма их оценок будет равна $4{,}7 \cdot 820 + 5 \cdot N$. А так как итоговый рейтинг должен стать равным 4,8, то получаем уравнение:

$$\frac{4{,}7 \cdot 820 + 5 \cdot N}{820 + N} = 4{,}8$$

Откуда $N = 410$.

2. Зоопарк (3 балла)

Владельцы одного зоопарка заметили, что с наступлением холодов 80% млекопитающих перестают покидать свои места ночлега, в то время как из северных животных таких всего 20%. Еще владельцы зоопарка выяснили, что животные, покидающие свои места ночлега с наступлением холодов, — это в точности все северные млекопитающие и только они. Известно, что число животных, не относящихся ни к северным, ни к млекопитающим, составляет 80% от общего числа животных в зоопарке.

Сколько процентов от общего числа животных в зоопарке составляют северные млекопитающие? Округлите ответ до двух знаков после запятой.

(3,81)

Решение и комментарии. Пусть x млекопитающих с наступлением холодов покидают свои места ночлега. Тогда не покидают места ночлега $4x$ млекопитающих – все эти млекопитающие являются северными животными. Северных животных, покидающих

свои места ночлега, тогда всего 16x. Всех вышеперечисленных животных в зоопарке суммарно 21x, что составляет 20%, т.е. 1/5 от всех животных зоопарка. Таким образом, в зоопарке находится 105x животных, 4x из которых составляют северные млекопитающие. Значит, доля северных млекопитающих составляет от всех животных зоопарка 4/105 ≈3,81%.

3. Задача про торговую сеть (5 балла)

Менеджеры одной сети розничных магазинов электроники решили изучить объемы продаж товаров в одном из магазинов сети за февраль 2022 года. Они выделили две категории покупателей:

- покупатели типа А — те, кто приобрели ровно одну единицу любого товара за месяц;
- покупатели типа Б — все остальные.

Менеджеры выяснили, что за февраль на каждых четырех покупателей типа А приходится ровно один покупатель типа Б. Известно, что за месяц данным магазином было продано не более 1020 единиц любых товаров. Какие выводы могут сделать менеджеры на основе имеющейся информации? Выберите все верные ответы.

а) Количество покупателей типа Б составляет 25% от количества всех покупателей магазина в этом месяце
б) За каждый день магазином было продано в среднем 34 единицы товаров
в) В этом месяце количество единиц товаров, проданных покупателям типа Б, не превышает количество единиц товаров, проданных покупателям типа А
г) В этом месяце покупателей типа Б было не более 170 (+)
д) Среди остальных ответов нет ни одного верного

Решение и комментарии.
а) Если на каждых четырех покупателей типа А приходится ровно один покупатель типа Б, то их соотношение равно 4:1. То есть, покупателей типа А ровно 4/5 = 80%, а покупателей типа Б ровно 1/5 = 20% от всех покупателей магазина в этом месяце. Пункт а) неверен.
б) В феврале 2022 было 28 дней, поэтому за каждый день магазином было продано в среднем 1020/28 ≈ 36,42 единиц товара. Пункт б) неверен.
в) Пусть в феврале в магазине было 120 покупателей типа А, каждый из которых приобрел одну единицу товара, и 30 покупателей типа Б, каждый из которых приобрел 30 единиц любых товаров. Тогда количество единиц товаров, проданных покупателям типа Б (900), превышает количество единиц товаров, проданных покупателям типа А (120). Пункт в) неверен.
г) Если за февраль покупателей типа А было 4N человек, то они приобрели суммарно 4N единиц товаров. Тогда покупателей типа Б по условию было N человек, и они приобрели суммарно не менее 2N единиц товаров (так как каждый из них приобрел хотя бы 2 единицы любых товаров). Значит, за месяц в магазине было приобретено не менее 6N единиц товаров. С другой стороны, по условию за месяц магазином было продано не более 1020 единиц товаров. Таким образом, должно выполняться неравенство 6N <= 1020, откуда N <= 170. При этом возможен крайний случай N = 170

(680 покупателей типа А и 170 покупателей типа Б, каждый из которых приобрел по 2 единицы товаров). Пункт г) верен.

4. Задача про настольные игры (5 балла)

Компания, занимающаяся производством и продажей настольных игр, решила изучить, как COVID-19 повлиял на ее выручку. Все свои игры эта компания делит на несколько сегментов, в каждом из которых игры стоят одинаково. Для этого аналитики компании строят одни и те же графики за период с сентября по ноябрь 2019 года, когда пандемия еще не наступила, и за период с сентября по ноябрь 2020 года, после начала пандемии, и сравнивают их. Какие из перечисленных графиков помогут компании проанализировать влияние COVID-19 на выручку компании? Выберите все подходящие варианты ответа.

а) столбчатая диаграмма, показывающая среднее количество покупателей магазина за каждый день недели
б) столбчатая диаграмма, показывающая количество проданных игр по каждому ценовому сегменту (+)
в) гистограмма распределения расстояния от места жительства покупателей до магазина
г) круговая диаграмма с процентным соотношением суммарной стоимости проданных игр по дням недели
д) Среди остальных ответов нет ни одного верного

Решение и комментарии.

а) Столбчатая диаграмма, показывающая среднее количество покупателей магазина за каждый день недели, не отражает выручку компании, так как не показывает суммарную стоимость игр, приобретенных каждым из этих покупателей. Например, возможна ситуация, когда количество покупателей магазина возросло в определенные дни, но при этом сумма, затрачиваемая ими на покупки, уменьшилась (из-за большей экономии – и, как следствие, приобретения меньшего количества игр и/или игр из более дешевых сегментов) – в таком случае данный график не отразит реальные изменения в размере выручки. Пункт а) не подходит.

б) Столбчатая диаграмма, показывающая количество проданных игр по каждому ценовому сегменту, позволит оценить сумму выручки магазина в оба периода и сравнить полученные суммы. Пункт б) подходит.

в) Гистограмма распределения расстояния от места жительства покупателей до магазина никак не отражает выручку компании. Пункт в) не подходит.

г) Круговая диаграмма с процентным соотношением суммарной стоимости проданных игр по дням недели не отражает абсолютного значения выручки – поэтому при изменении распределения выручки по дням недели невозможно будет сказать, выросла, уменьшилась или не изменилась выручка в абсолютном значении как в общем, так и в каждый день недели в отдельности. Пункт г) не подходит.

5. Selection bias

Вы консультант министерства здравоохранения и хотите выяснить, насколько хорошо работают больницы, а именно улучшают ли они состояние здоровья людей. Вам

доступны результаты опрос населения, в ходе которого у людей узнавали, был ли человек хотя бы раз госпитализирован за последний год (то есть оставался ли он в больнице более чем на сутки), а также попросили оценить общее состояние своего здоровья по шкале от 1 до 5 (где 1 — очень плохо, а 5 — прекрасно). Эти результаты представлены в таблице ниже. Разница между двумя средними показателями значимая.

Группа	Размер выборки	Среднее состояние здоровья	Стандартная ошибка
Был госпитализирован	7 774	3,21	0,014
Не был госпитализирован	90 049	3,93	0,003

Вам необходимо проинтерпретировать полученные в ходе опроса результаты. Какие из приведенных ниже утверждений являются верными?

а) Так как средняя оценка здоровья госпитализированных ниже, мы можем сделать вывод, что больницы ухудшают состояние здоровья пациентов
б) Люди с плохим состоянием здоровья скорее окажутся госпитализированными, поэтому среднее состояние в группе госпитализированных могло оказаться ниже. (+)
в) Люди врут в опросах, и если бы состояние их здоровья было оценено объективно, то соотношение было бы другим
г) Среди остальных ответов нет ни одного верного

Решение и комментарии. В данном случае по проведенному опросу некорректно делать выводы относительно того, улучшается или ухудшается состояние здоровья людей после госпитализации (пункт а) неверен). Основная причина этого, что на итоговую разницу в средних величинах влияет сразу несколько факторов: сам эффект госпитализации и различие в исходном состоянии здоровья у людей из разных групп. Госпитализируют обычно людей в очень тяжелом состоянии, либо у которых много разных заболеваний и им необходим контроль врачей при лечении. Скорее всего именно этот эффект мы и видим, когда анализируем среднюю оценку самочувствия госпитализированных и не госпитализированных (пункт б) верен). При этом важно отметить, что негативный эффект госпитализации (связанный с повышенной вероятностью столкнуться с заболеваниями в больнице) также может существовать, однако из-за особенностей опроса мы не сможем отделить его от описанного выше эффекта.
Как можно заметить, эффект смещения оценки из-за того, что негоспитализированные люди в среднем более здоровые не связан никак с тем, что люди склонны искажать информацию о себе. То есть, даже если бы мы взяли для всех людей в опросе их реальное состояние здоровья, группа госпитализированных имела бы более низкие оценки (пункт в) неверен).

6. **Gapminder (5 балла)**

На рисунке изображена диаграмма рассеяния, показывающая взаимосвязь между экономическим развитием и продолжительностью жизни в 2018 году.

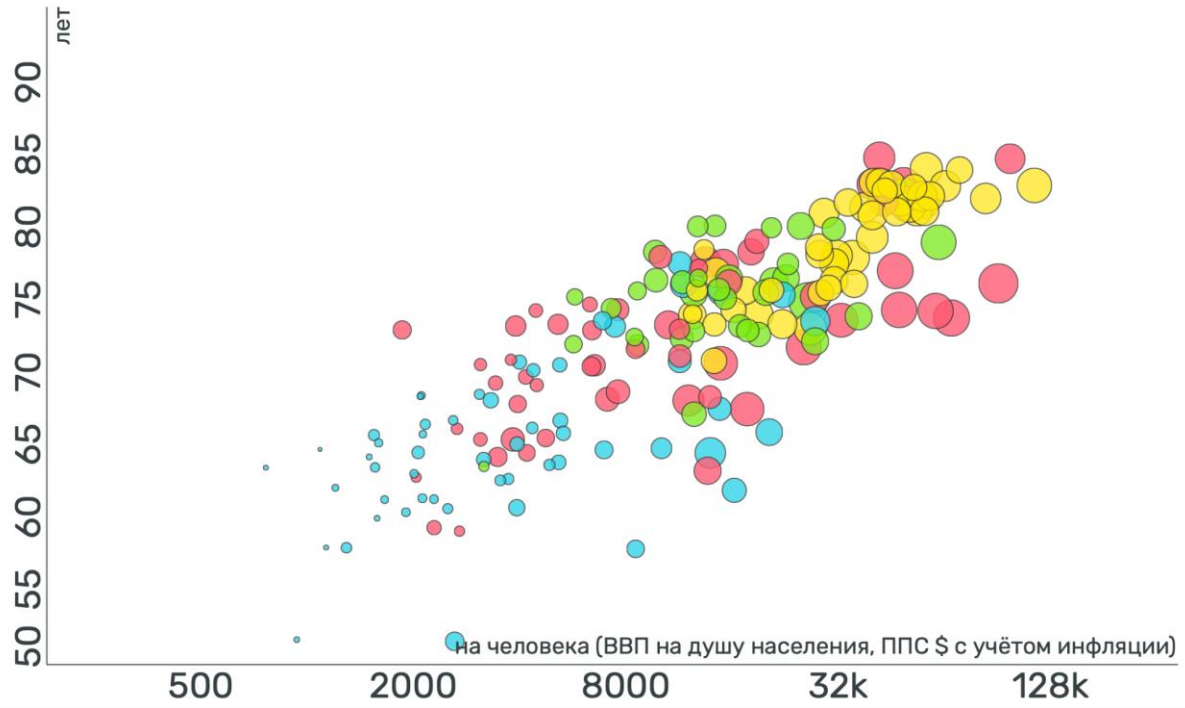
По вертикальной оси показана средняя продолжительность жизни в годах.

По горизонтальной оси отражен реальный ВВП на душу населения в долларах (показатель скорректирован на инфляцию и пересчитан по паритету покупательной способности для сопоставимости между странами). Для наглядности графика значения ВВП преобразованы специальным образом, который при этом качественно не влияет на выводы.

Размер кругов отражает третий показатель — выбросы CO_2 в атмосферу — и измеряется в метрических тоннах выбросов на человека. Чем больше размер круга, тем больше выбросы.

Цвет кругов обозначает разные регионы мира: голубой — Африка, красный — Азия (включая Австралию и Новую зеландию), зеленый — Америка (Северная и Южная вместе), желтый — Европа.

Источник: Gapminder, URL: https://www.gapminder.org



Выберите все верные утверждения, следующие из графика.

а) Существует положительная корреляция между экономическим развитием и средней продолжительностью жизни населения (+)

б) Выбросы CO_2 влияют на продолжительность жизни населения

в) Средняя продолжительность жизни в странах Африки в среднем выше, чем в странах Европы

г) Среди остальных ответов нет ни одного верного

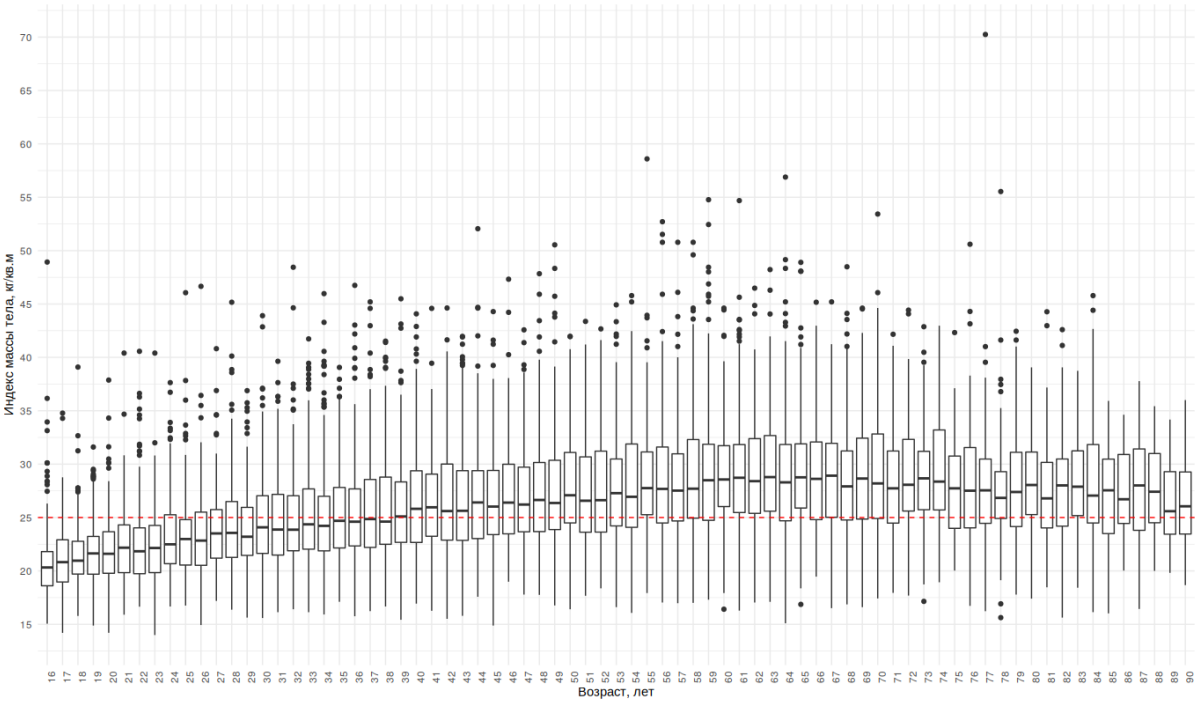
Решение и комментарии. В данном задании по графику мы можем выявлять только наличие взаимосвязей (корреляции), но никак не причинно-следственную связь. Более того, положительная связь выбросов с продолжительностью жизни будет означать очень контринтуитивный вывод (раз люди дольше живут из-за загрязнения воздуха, надо не бороться с загрязнением окружающей среды, а наоборот развивать его, чтобы увеличить продолжительность жизни).
В целом можно видеть положительный наклон у данного scatter plot-а, что говорит о наличии положительной корреляции между доходами на душу населения и продолжительностью жизни. Также можно заметить, что размер кружочков, которые отражают объем выбросов CO_2 также растет по мере движения вправо-вверх вдоль облака точек, что говорит о положительной корреляции выбросов как с доходами, так и с возрастом.

Также видно существенные отличия стран Африки (облако голубых точек), стран Азии (красные точки) и Европы (желтые точки). В данном случае очевидно, что практически все облако голубых точек лежит ниже облака желтых точек, что означает, что средняя продолжительность жизни в Африке ниже, чем в Европе.

7. Индекс массы тела (5 балла)

Один из основных показателей здоровья — *индекс массы тела (ИМТ)*. Показатель рассчитывается путем деления веса в килограммах на квадрат роста в метрах. В своих исследованиях ученые используют различные уровни (или интервалы) ИМТ в качестве некоторой базы для сравнения. Здесь в качестве такого значения будет использован ИМТ, равный 25 (отмечен на графике красным пунктиром).

На рисунке представлен график распределения ИМТ среди россиян разного возраста, а именно: для каждого возраста от 16 до 90 лет построен отдельный «ящик с усами» (boxplot).



Источник: РМЭЗ, 28-я волна, данные за 2019 год

Выберите все утверждения, следующие из графика.

а) Разброс ИМТ у людей в возрасте 50 лет выше, чем у людей в возрасте 90 лет (+)

б) Медианный уровень ИМТ превышает 25 для людей в возрасте от 39 до 90 лет (+)

в) Первый квартиль (25-й перцентиль) ИМТ ниже 25 в любом возрасте

г) Среди остальных ответов нет ни одного верного

Решение и комментарии. Данный тип графика строится следующим образом. Датасет сортируется в порядке возрастания переменной (в данном случае, ИМТ). Размер самого ящика определяется 25 и 75 перцентилем распределения, черная линия в середине – медиана распределения. Усы определяются прибавлением (или вычитанием) к 75 (или 25) перцентилю 1,5 величин «межквартильного размаха» (расстояние между 25 и 75 перцентилем, которые являются 1 и 3 квартилем соответственно). Все, что не попадает в этот размах отмечается точками, которые являются выбросами.
Можем видеть, что начиная с 39 лет медиана распределения стабильно выше 25, 25 перцентиль оказывается выше 25 для ряда возрастных групп (например, для 60-63 лет). Межквартильный разброс напрямую связан с любыми метриками разброса (например, с дисперсией или стандартным отклонением), так что по графику можно сделать однозначный вывод, что в 50 лет разброс оказывается больше, чем в 90 (нижний ус для 50 начинается на уровне ниже, чем для 90 лет, а верхний заканчивается на уровне выше).

8. Задача про нормирование (6 баллов, по 3 на каждый способ, можно частичные баллы (если к одному способу ответ верный, а к другому нет))

Андрей слышал, что для решения задач по анализу данных бывает полезным нормировать данные. У него был набор данных, состоящий из 100 значений, — целых чисел от −1000 до 1000 (известно, что не все числа одинаковые). Он решил нормировать каждое значение из набора двумя способами, о которых недавно узнал.

Способ 1

$$x_{new} = (x - x_{min})/(x_{max} - x_{min})$$

Из каждого значения вычитается минимальное значение из набора данных x_{min} и делится на разность максимального и минимального значений из набора данных $x_{max} - x_{min}$.

Способ 2

$$x_{new} = (x - \bar{x})/\sqrt{D}$$

Из каждого значения вычитается среднее арифметическое всех значений набора данных $\bar{x}$ и делится на корень из дисперсии набора данных $\sqrt{D}$ (дисперсия равна D).

Укажите по очереди для каждого способа в алфавитном порядке все пункты, которые выполняются для этого способа (например, 1бг2аг).

Обратите внимание: не обязательно должны быть использованы все пункты, при этом некоторые свойства могут соответствовать и первому, и второму способу:

а) новое значение может оказаться меньше нуля;
б) новое значение может оказаться больше 1;
в) новое значение может оказаться больше старого значения;
г) одно из новых значений обязательно окажется равным 0.

Ответ: 1вг; 2абв

Решение и комментарии.

1.

а), г) Для любого x из набора выполняется неравенство $x \geq x_{min}$. Тогда

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \geq \frac{x_{min} - x_{min}}{x_{max} - x_{min}} = 0.$$

Таким образом, любое значение $x_{new} \geq 0$. Отметим также, что всегда найдётся $x_{new} = 0$ (так как всегда существует значение $x = x_{min}$).

б) Для любого x из набора выполняется неравенство $x \leq x_{max}$. Тогда

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \leq \frac{x_{max} - x_{min}}{x_{max} - x_{min}} = 1.$$

Таким образом, любое значение $x_{new} \leq 1$. Отметим также, что всегда найдётся $x_{new} = 1$ (так как всегда существует значение $x = x_{max}$).

в) Допустим, в наборе было несколько отрицательных значений x. Тогда все соответствующие им новые значения x_{new} согласно пункту а) станут неотрицательными, т.е. каждое из них будет больше соответствующего ему старого значения.

2.

а) Допустим, все значения в наборе были положительные. Тогда после вычитания из каждого из них среднего арифметического $\bar{x}$ найдется хотя бы одно отрицательное (так как найдется хотя бы одно значение меньше среднего арифметического). Таким образом, найдутся значения меньше нуля.

б) Рассмотрим набор данных из 99 чисел «0» и 1 числа «100». В этом случае среднее арифметическое чисел $\bar{x} = 1$, $D = 99$. Таким образом, для старого значения $x = 100$ из этого набора новое значение $x_{new} = (100 - 1)/\sqrt{99} \approx 9{,}95 > 1$.

в) Допустим, все значения в наборе были отрицательные. Тогда после вычитания из каждого из них среднего арифметического $\bar{x}$ найдется хотя бы одно положительное (так как найдется хотя бы одно значение больше среднего арифметического). Таким образом, найдутся новые значения большие нуля, а значит и большие старых.

г) В случае из пункта б) для старых значений $x = 0$ из набора новые значения

$$x_{new} = (0 - 1)/\sqrt{99} \approx -0{,}1$$

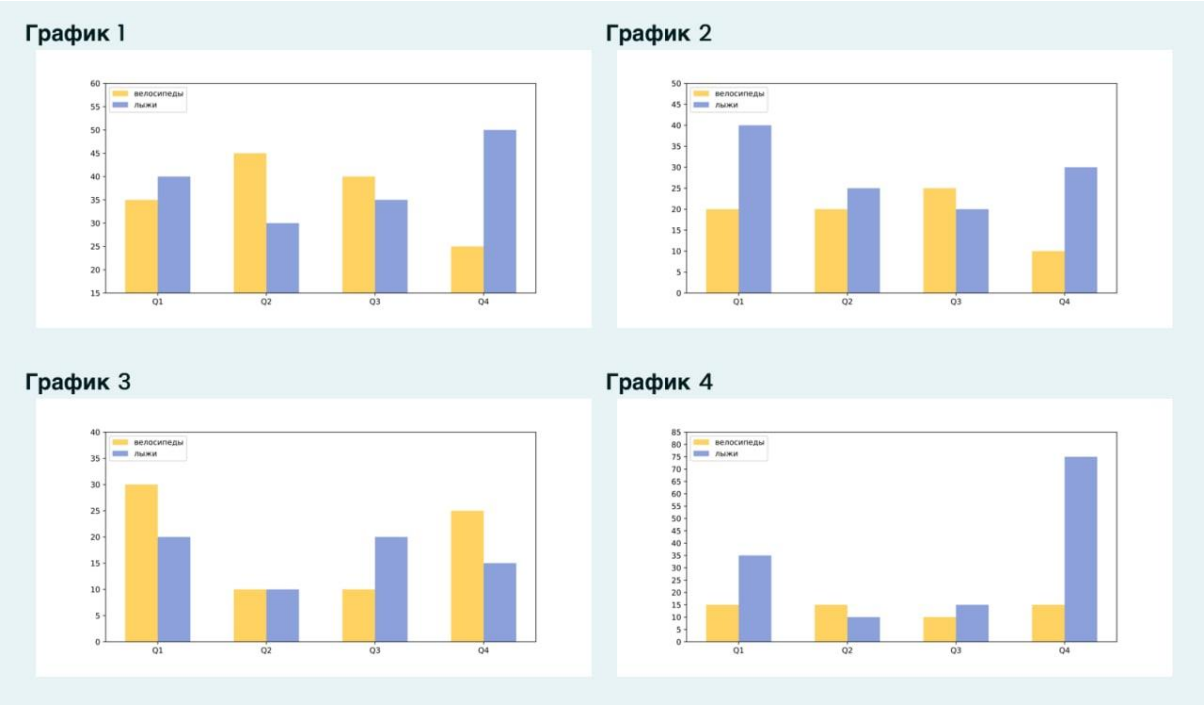
Таким образом, в данном случае ни одно из новых значений не равно 0.

9. Задача на графики (8 баллов, по 2 за каждое верное соотнесение, можно частичные баллы)

Аналитики международной компании, занимающейся продажей спортивного инвентаря, решили проанализировать продажи основных товаров, распространяемых компанией, — велосипедов и лыж — за каждый квартал 2021 года. Они построили графики и диаграммы продаж в четырех магазинах компании, находящихся в разных городах мира. Каждому магазину соответствует один график и одна диаграмма. Каждый график показывает продажи велосипедов и лыж по отдельности, за каждый

квартал по очереди. Каждая диаграмма отображает общую долю продаж (и велосипедов, и лыж) за каждый квартал относительно суммы всех продаж (и велосипедов, и лыж) в этом магазине за год. Для каждого графика укажите диаграмму, построенную для того же магазина, что и этот график.

Примечание: секторы одного и того же цвета на разных диаграммах могут соответствовать разным кварталам.



Решение и комментарии.

Посчитаем по каждому графику суммарные продажи велосипедов и лыж за каждый квартал, а затем переведем их в проценты от всех продаж магазина за год. Для удобства округлим проценты до целого числа с поправкой на то, что значения по одному магазину должны складываться в 100% (где-то возможно отклонение от реального значения менее чем на 1%, на ответ это не повлияет).

График 1) 75, 75, 75, 75 -> 25%, 25%, 25%, 25% (все доли равны)

График 2) 60, 45, 45, 40 -> 31%, 24%, 24%, 21% (равны только две средние доли)

График 3) 50, 20, 30, 40 -> 36%, 14%, 21%, 29% (все доли различны)

График 4) 50, 25, 25, 90 -> 26%, 13%, 13%, 48% (равны только две меньшие доли)

Соотнесем графики с диаграммами; для диаграммы 4, в которой равны две большие и две меньшие доли, соответствующий ей график отсутствует.

Ответ: График 1 – Диаграмма 3, График 2 – Диаграмма 1, График 3 – Диаграмма 5, График 4 – Диаграмма 2.

10. Платежеспособность заемщика (8 баллов)

Банк заинтересован в анализе платежеспособности своих заемщиков. В идеале он хотел бы точно знать, кто вернет кредит, а кто нет, и давать кредит только первым. На практике банки пытаются оценить вероятность невозврата кредита и используют для этого различные способы. Используя эти оценки вероятности, банк принимает решение о выдаче кредита. При этом важным оказывается выбор порогового значения вероятности. Тогда, если предсказанная вероятность больше порогового значения, будет предсказана невыплата (и банк откажет такому клиенту в выдаче средств), а если меньше — будет предсказано, что заемщик вернет кредит полностью (и такому клиенту банк выдаст кредит). Если поставить пороговое значение слишком низко (скажем, выдавать кредиты только людям, у которых вероятность невыплат составляет 0,01), то таких людей может быть слишком мало и банк не сможет получить прибыль, а если слишком высоко, то число неплатежей будет большим и банк будет терпеть убытки.

В файле ‘default.csv’ содержатся данные о 100 фактических заемщиках банка. Переменная DEFAULT принимает значение 1, если заемщик по факту не вернул кредит полностью или частично (и такого заемщика мы будем называть недобросовестным), и принимает значение 0, если заемщик полностью вернул кредит (такого заемщика мы будем называть добросовестным). Переменная PD принимает значения от 0 до 1 и показывает вероятность невозврата кредита для этих заемщиков, оцененную банком.

Сделав необходимые расчеты, выберите все верные утверждения.

а) При пороговом значении 0,45 модель верно предсказывает добросовестного заемщика в 67 случаях (+)

б) При пороговом значении в 0,25 модель не может предсказать дефолт у 5 заемщиков (+)

в) Чем ниже выбирается пороговое значение, тем в большем числе случаев правильно предсказываются добросовестные заемщики

г) Среди остальных ответов нет ни одного верного

Решение и комментарии.

а) необходимо на основании имеющейся вероятности невозврата кредита (переменная PD) и данного граничного значения в 0,45 приписать каждому заемщику, предсказывает ли модель дефолт. Если значение превышает граничное, значит модель предсказывает дефолт, и такому человеку в *предсказание дефолта* надо поставить 1. Если же PD ниже, то такому человеку приписывается значение 0 в *прогнозируемый дефолт*. Если прогноз совпадает с фактическим событием (которое отражено в переменной DEFAULT), значит модель предсказывает конкретно этого человека верно (если у него в прогнозе дефолта 1 и фактически был дефолт, значит модель верно предсказала недобросовестного заемщика). В случае порогового значения 0,45 прогнозируется, что 79 человек будут добросовестными заемщиками, однако 12 человек в итоге допускают дефолт и для них модель ошибается, а 67 человек оказываются добросовестными заемщиками, для которых модель предсказывала отсутствие дефолта.

б) По аналогии с предыдущим пунктом строим переменную, предсказывается заемщику дефолт или нет, но с другим граничным значением. Теперь нам необходимо посчитать, сколько было людей, которые фактически допустили дефолт (DEFAULT=1), но модель прогнозировала, что эти заемщики добросовестные (то есть прогноз равен 0). Таких наблюдений оказывается 5.

в) Общая логика состоит в том, что чем ниже выбирается пороговое значение, тем меньше заемщиков будут прогнозироваться как добросовестные, а значит меньше количество тех, кто фактически является добросовестным. Значит случаев верного прогноза также становится меньше. Это можно проверить на конкретных примерах, взяв различные пороговые значения.

11. Характеристики людей (8 баллов)

В файле gender.xlsx представлены данные, касающиеся различных физиологических параметров лица человека. Описание всех переменных приведено в файле на листе «Описание». На основании предложенных данных и ваших расчетов выберите все верные варианты ответа.

а) Стандартное отклонение в ширине лба среди мужчин больше, чем среди женщин (+)

б) У женщин лоб в среднем выше, чем у мужчин

в) У большинства мужчин длинные волосы (+)

г) Среди остальных ответов нет ни одного верного

Решение и комментарии.

Необходимо посчитать выбранные показатели для каждой из групп (мужчин и женщин).

Стандартное отклонение – это общепринятый термин, который равен квадратному корню из дисперсии, которая в свою очередь рассчитывается как сумма квадратов отклонения наблюдений от среднего значения, деленное на количество наблюдений. В Excel, например, для расчета стандартного отклонения, используется функция =СТАНДОТКЛОН(). В данном случае, стандартное отклонение ширины лба для мужчин составляет 1,19, а у женщин – 0,88.

Если показатель принимает значения 0 или 1, то долю единиц можно посчитать через среднее значение. Используя этот факт, можно посчитать долю мужчин с длинными волосами, она равна 0,87, то есть больше половины.

Среднее значение высоты лба у женщин составляет 5,80, а у мужчин 6,10.

12. Фильмы (8 баллов)

Вам представлены два файла. Один из них (movies_ratings.csv) содержит информацию о названиях фильмов, их популярности, бюджете, кассовых сборах, а также оценках, выставленных отдельными пользователями каждому из фильмов. Описание всех переменных находится на листе «Описание» в файле movies.xlsx. Посчитайте корреляции необходимых переменных и выберите верные утверждения из приведенных ниже.

а) чем новее фильм, тем более высокую оценку пользователей в среднем он имеет

б) более продолжительные фильмы приносят больше кассовых сборов (+)

в) Фильм, получивший наибольшую среднюю оценку пользователей, был выпущен в 21-м веке

г) Среди остальных ответов нет ни одного верного

Решение и комментарии.

Исходный файл с данными был представлен в csv формате, соответственно, чтобы конвертировать его в Excel, необходимо было правильно его открыть. 1 способ: импортировать файл, указав в соответствующем разделе в качестве разделителя столбцов запятую. 2 способ: во вкладке Данные использовать функцию «Текст по столбцам», также указав запятую в качестве разделителя. В некоторых случаях могло сложится так, что название фильма “20,000 Leagues Under the Sea” делилось и данные съезжали на 1 столбец. В этом случае было необходимо сделать замену «20,000» на «20000».

Следующий этап анализа этого набора данных – агрегирование имеющихся показателей. В исходной версии таблицы имеется порядка 900 тыс. строк, которые отражают оценки отдельных пользователей к каждому фильму. Однако, чтобы считать взаимосвязь кассовых сборов с рейтингом и т.п. по фильмам, необходимо сделать так, чтобы одно наблюдение соответствовало одному фильму и больше не встречалось. Для этого можно было сделать сводную таблицу в Excel, или группировку, или любые аналогичные функции. В результате должна получиться таблица в 546 наблюдений (по количеству фильмов), где самое важное – посчитать показатель rating как среднее из оценок пользователей. Переменная с номером пользователя нам не нужна, поэтому ее можно было не сохранять.

Далее идет расчет корреляций. Ниже представлена таблица с корреляциями, где звездочкой отмечена статистически значимая корреляция. Да, во многих источниках говорят, что корреляция ниже примерно 30% считается отсутствующей, однако это очень упрощенный анализ. Корреляция показывает меру линейной взаимосвязи, но при этом важно отметить, что она показывает и долю разброса, которую одна переменная объясняет в другой. И говорить о том, что корреляция в 20% несущественна означает говорить, что 20% разброса можно просто пренебречь, что не совсем корректно. Верный ответ всегда имеет нужный знак и значимый уровень корреляции.

Ниже по порядку представлены корреляционные матрицы для всего набора данных, для 20 века и для 21 века. Если около числа есть звездочка, то данная корреляция считается значимой.

Таким образом, в а) смотрим корреляцию года и оценки. Она отрицательная (-0,1512), то есть у новых фильмов средний рейтинг ниже. В б) продолжительность и сборы связаны положительно (0,2165). Для в) необходимо было найти тот фильм, у которого максимальная средняя оценка (это “The Sixth Sense” 1999 года). Аналогично в других вариантах надо было рассмотреть корреляции года и оценки в 21 веке (третья таблица), которая составляет 0,12 и незначима; года и продолжительности (-0,25, значимая); продолжительности и бюджета (0,17, значимая); популярность и год по векам (-0,04, незначимая в XX веке и 0,15, значимая в XXI); бюджет и популярность (0,51, значимая); бюджет и год (-0,03, незначимая).

```
. pwcorr rating_m budget popularity year revenue runtime, st(0.05)
```

	rating_m	budget	popula~y	year	revenue	runtime
rating_m	1.0000					
budget	0.1054*	1.0000				
popularity	0.2796*	0.5125*	1.0000			
year	-0.1512*	-0.0274	0.0002	1.0000		
revenue	0.2002*	0.6369*	0.6438*	-0.0368	1.0000	
runtime	0.2382*	0.1704*	0.1948*	-0.2504*	0.2165*	1.0000

Figure 1. Корреляционная таблица переменных, объединенная выборка

```
. pwcorr rating_m budget popularity year revenue runtime if year<=2000, st(0.05)
```

	rating_m	budget	popula~y	year	revenue	runtime
rating_m	1.0000					
budget	0.0273	1.0000				
popularity	0.2202*	0.5118*	1.0000			
year	-0.1591*	-0.0429	-0.0412	1.0000		
revenue	0.1185*	0.6042*	0.6677*	-0.0568	1.0000	
runtime	0.1679*	0.1329*	0.1548*	-0.2325*	0.2118*	1.0000

Figure 2. Корреляционная таблица переменных, XX век

	rating_m	budget	popula~y	year	revenue	runtime
rating_m	1.0000					
budget	0.2563*	1.0000				
popularity	0.3787*	0.5339*	1.0000			
year	0.1180	-0.0154	0.1471*	1.0000		
revenue	0.3949*	0.7205*	0.6687*	0.0618	1.0000	
runtime	0.4099*	0.2956*	0.3335*	0.0673	0.2530*	1.0000

Figure 3. Корреляционная таблица переменных, XXI век

13. Рейтинг команд (15 баллов, по 5 за каждый верный ряд, частичные баллы (если 1 ряд верный, а 2 и 3 нет, получают 5 из 15))

В файле «Оценки.xlsx» вам представлены результаты команд, которые принимали участие в некотором соревновании.

Итоговый балл (ОБЩИЙ ИТОГ) рассчитывался как сумма трех итоговых оценок по теоретическому блоку, практическому блоку и за презентацию. Каждая из итоговых оценок складывалась из взвешенной суммы оценок по отдельным критериям по формулам, приведенным ниже. Каждый критерий оценивался по шкале от 0 до 3.

Итог за теоретический блок = 2*«Гипотеза и механизм» + 1*«Корректность и оригинальность»
Итог за практический блок = 2*«Интерпретация статанализа» + 1*«Перспективы и применимость» + 2*«Дан ответ на поставленный вопрос» + 1*«Визуализация результата»
Итого за презентацию = 1*«Командная работа» + 1*«Логика, связность и читаемость слайдов»

У жюри есть несколько вариантов подведения итогов и сложения всех оценок:

- по указанной формуле посчитать итоговую оценку каждого члена жюри, далее найти средний балл по всем членам жюри и выбрать победителем команду с наибольшим баллом;
- по указанной формуле посчитать итоговую оценку для каждого члена жюри, далее найти медианный балл и выбрать победителем команду с наибольшим баллом;
- найти медианный балл по отдельным критериям, а затем получить из них итоговую оценку по приведенным формулам.

Каждый из этих подходов имеет свои плюсы и минусы. В данном задании вам необходимо построить рейтинг команд по каждому из указанных принципов. При вводе ответа укажите номера команд через запятую, начиная с команды, набравшей наибольшее количество баллов, и далее по убыванию баллов (то есть команда, занявшая первое место, будет идти в списке первой, второе место — второй и так далее). Если у команд получается одинаковый итоговый балл, ранжируйте их по возрастанию номеров (например, если команда 1 и команда 2 набрали по 29 баллов, то в ответе необходимо указать “1, 2”).

При выборе победителя по принципу нахождения среднего итогового балла из оценок всех членов жюри порядок команд будет следующим: (9, 17, 12, 1, 14)

Код команды	итоговый балл
9	29
17	28
12	25
1	24,67
14	18

При выборе победителя по принципу нахождения медианного итогового балла порядок команд будет следующим: (9, 17, 1, 12, 14)

Код команды	итоговый балл
9	31
17	27
1	25
12	24
14	18

При выборе победителя по принципу подсчета итогового балла из медианных баллов жюри по отдельным критериям порядок команд будет следующим: (9, 17, 1, 12, 14)

Код команды	итоговый балл
9	29
17	28
1	24
12	24
14	18

Решение и комментарии.
При первом подходе необходимо сначала по формулам из задания получить итоговый балл, который каждый член жюри выставил каждой команде, а затем посчитать среднее значение из этих полученных баллов. Отдельным критерием считается каждый из 8 пунктов оценивания, которые были приведены в таблице. Столбец «№п/п» означал номер строки по порядку, а не номер команды.

14. Задача про машины-1 (14 баллов, по 2 за каждый пропуск)

В файле Auto.xlsx вам показаны данные по некоторым характеристикам подержанных автомобилей, представленных на рынке некой страны. Подробное описание переменных дано в файле на листе «Описание». Сами данные находятся на листе «Данные». Проведите небольшой анализ представленных там показателей и заполните пропуски в тексте ниже.
Если ваш ответ представлен дробным числом, запишите его через запятую с округлением до 2 знаков после запятой (например, «0,33»).
Обратите внимание: запомните полученные вами результаты — они будут нужны в следующем задании.

При анализе факторов, влияющих на ценообразование подержанных автомобилей, необходимо рассчитать корреляцию с основными факторами. Так, корреляция цены с возрастом автомобиля (по состоянию на 2022 год) составляет XXXX (–0,58), цены с пробегом автомобиля — XXXX (–0,57), а цены с объемом двигателя — XXXX (0,33).

Отметим также, что наибольший средний возраст автомобилей наблюдается в группе с двигателями типа XXXX (LPG) и составляет XXXX (12,82) лет, наибольший средний пробег — в группе с двигателями XXXX (LPG) и составляет XXXX (198027,12).

Решение и комментарии. В данном случае с данными никаких манипуляций проводить было не надо. Тип двигателя также засчитывался при указании, что это двигатель на сжиженном природном газе и другие альтернативы, эквивалентные LPG.

15. Задача про машины-2 (3 балла, за все верные, иначе 0)
(продолжение задачи 14)
В любом анализе важно указать возможный механизм влияния одного фактора на другой. Ниже предлагаем вам заполнить пропуски в описании взаимосвязи цены с возрастом и пробегом автомобиля, которую вы получили в предыдущем задании, чтобы получилась верная логическая цепочка.

Наличие такой взаимосвязи цены автомобиля с его возрастом и пробегом объясняется тем, что чем больше возраст автомобиля, тем XXXX (больше/меньше) на нем могли проехать, и все это обуславливает XXXX (больший/меньший) износ автомобиля, что будет действовать на цену в сторону ее XXXX (увеличения/уменьшения).

Решение и комментарии. В итоге корреляция возраста и цены автомобиля получилась отрицательная, поэтому в ходе объяснения надо было прийти к тому, что с ростом возраста автомобиля ценя его должна снижаться.