Содержание

Вариант 1 (с ответами)

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

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)

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

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

5. Selection bias (5 балла)

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

6. Gapminder (5 балла)

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

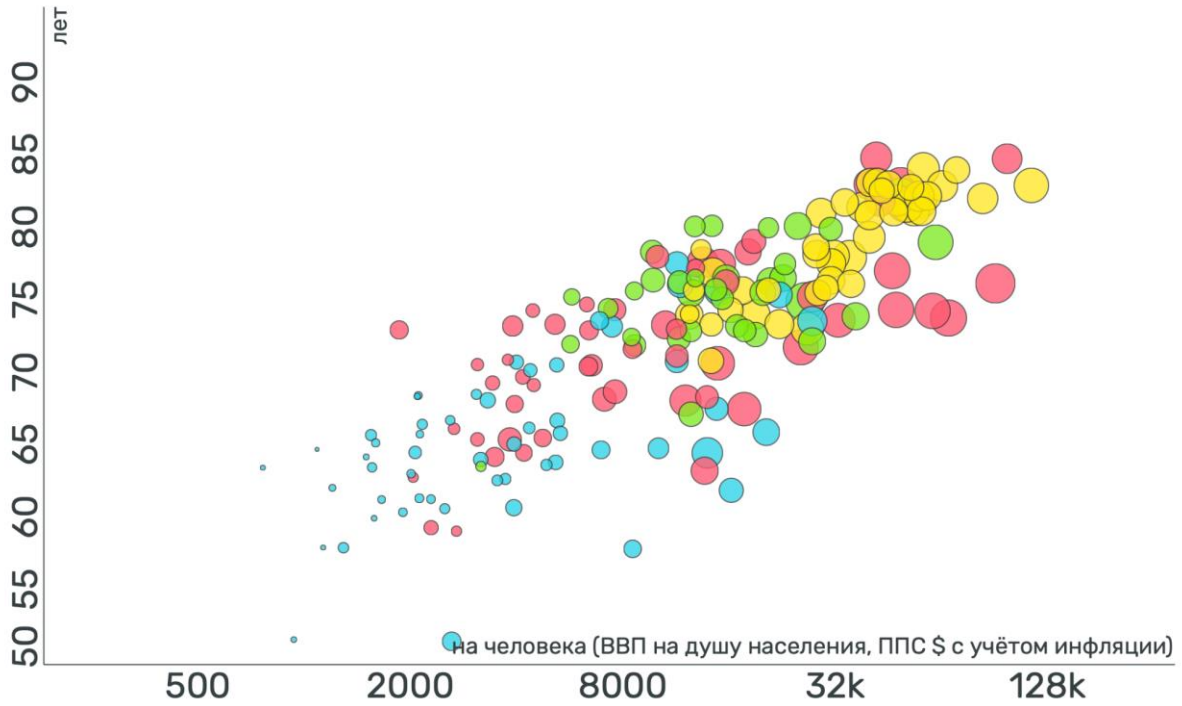
По вертикальной оси показана средняя продолжительность жизни в годах.

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

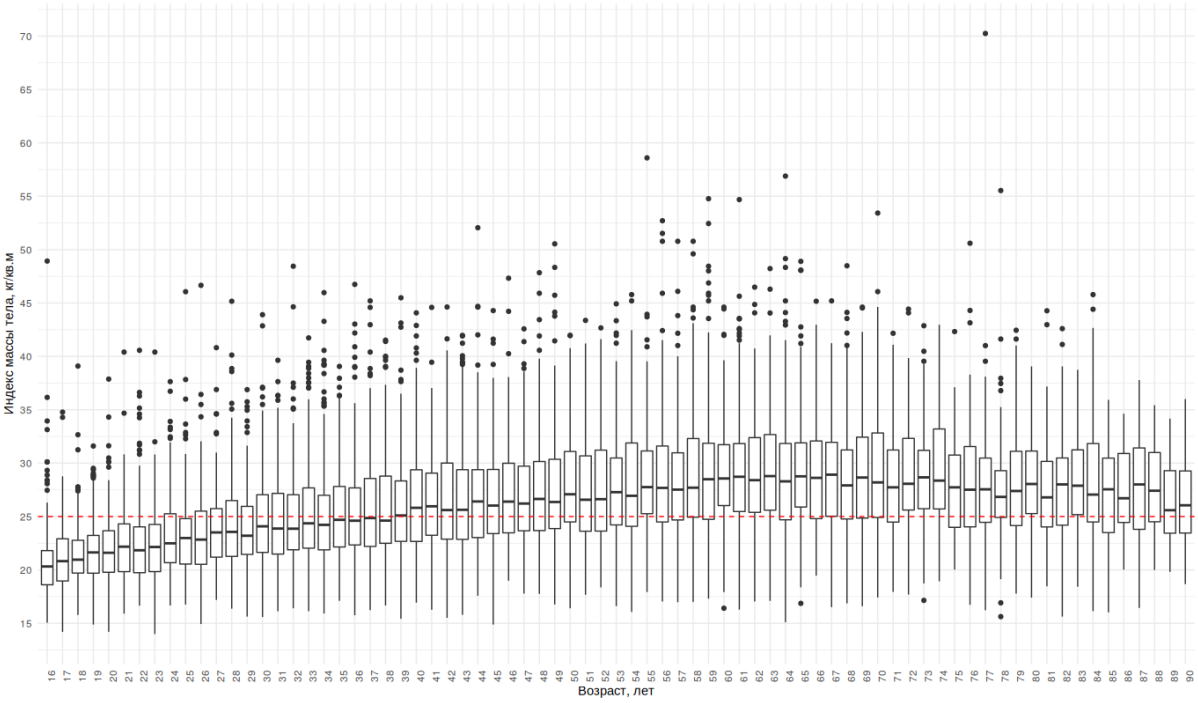
в) Средняя продолжительность жизни в странах Африки в среднем выше, чем в странах Европы

г) Среди остальных ответов нет ни одного верного

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

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

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.

(Условие общее для всех 3 вариантов, картинки лежат в отдельных папках, у парных совпадают номер)

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

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

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

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

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).

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).

Вариант 1 (без ответов)

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

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. **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

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

5. Selection bias (5 балла)

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

6. Gapminder (5 балла)

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

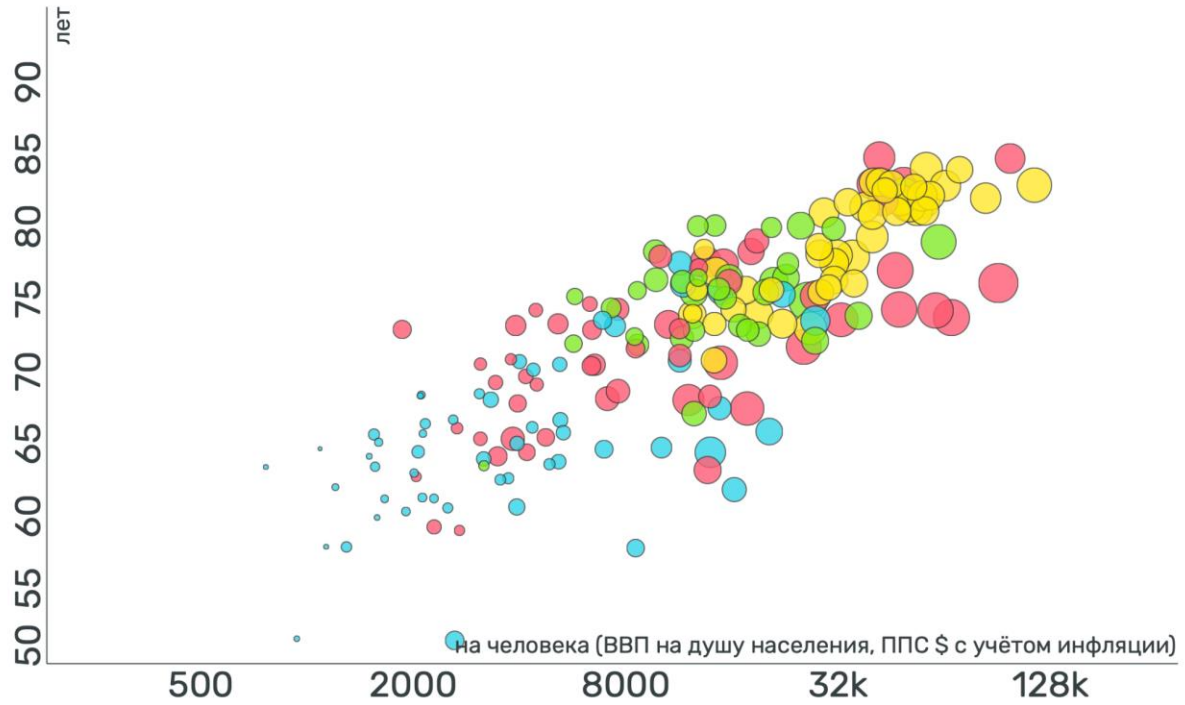
По вертикальной оси показана средняя продолжительность жизни в годах.

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

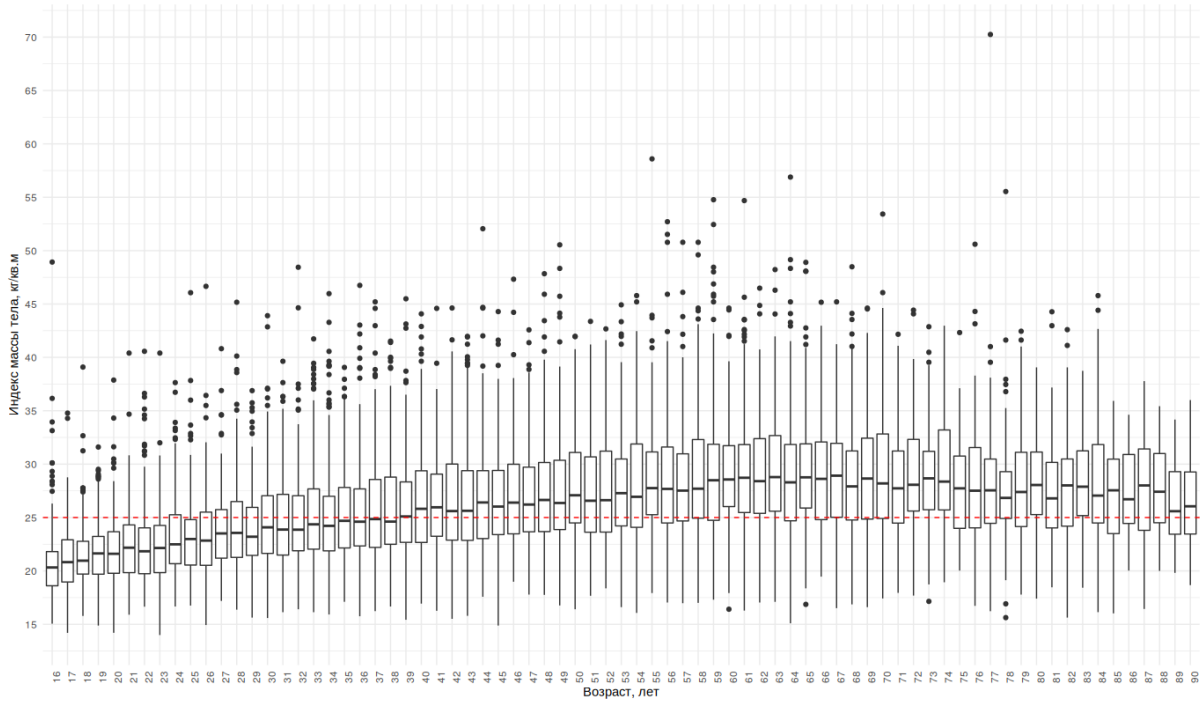
в) Средняя продолжительность жизни в странах Африки в среднем выше, чем в странах Европы

г) Среди остальных ответов нет ни одного верного

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

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.

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.

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

10. 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

11. 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

12. 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”).

13. 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, цены с пробегом автомобиля — XXXX, а цены с объемом двигателя — XXXX.

Отметим также, что наибольший средний возраст автомобилей наблюдается в группе с двигателями типа XXXX и составляет XXXX лет, наибольший средний пробег — в группе с двигателями XXXX и составляет XXXX.

14. 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).