Демонстрационный вариант 2022/2023. Задачный тур заключительного этапа

Время выполнения – 240 минут

Максимальное количество баллов – 100

Задание 1. – 25 баллов

Перед вами текст презентации одной сети пиццерий для потенциальных инвесторов. Вы являетесь аналитиком инвестиционной компании и хотите произвести оценку данной компании по информации ниже. Аналитик, который готовил презентацию, использовал различные ухищрения, чтобы выставить компанию в более выгодном свете. Вам нужно их найти. Выпишите все проблемы и обоснуйте их, при надобности приведите поясняющие примеры.

Решение и критерии:

«Фэмили Пицца» является современной сетью пиццерий семейного формата с широким ассортиментом. Пока сеть состоит из двух пиццерий в районе Девятково, а в прошлом месяце открылась еще одна в районе Десятково.

На рис. 1 продемонстрирован стремительный рост заказов в нашей сети за последние полгода. **Видно, что за полгода количество заказов выросло всего на 10 – это не говорит о стремительном росте. (3 балла) Здесь аналитик убрал ноль на оси OY, чтобы создать видимость резкого роста на графике. (2 балла)**

Рисунок 1

А при анализе рынков в этих двух районах мы выяснили, что за прошлый месяц доли рынков в каждом из этих районов выросли. **В новом районе не было пиццерии в прошлом месяце, поэтому «рост доли рынка в районе Десятково» значит только открытие пиццерии. (2 балла)**

Главным образом пиццерия ориентирована на семейных клиентов с детьми дошкольного возраста – специально для этого разработано детское меню, а в каждой пиццерии есть зона для детей с различными играми. Ниже на рис. 2 Вы можете заметить стабильность количества посетителей с детьми возраста 0-7 лет. **На графике сложно проследить динамику изменения количества посетителей с детьми, так как вторая отображаемая величина имеет бОльшие абсолютные значения, из-за чего меньшая величина теряется. (2 балла) На самом же деле видно, что оранжевая величина не растет (и даже падает), в отличие от синей, то есть доля интересующей нас аудитории уменьшается. (3 балла)**

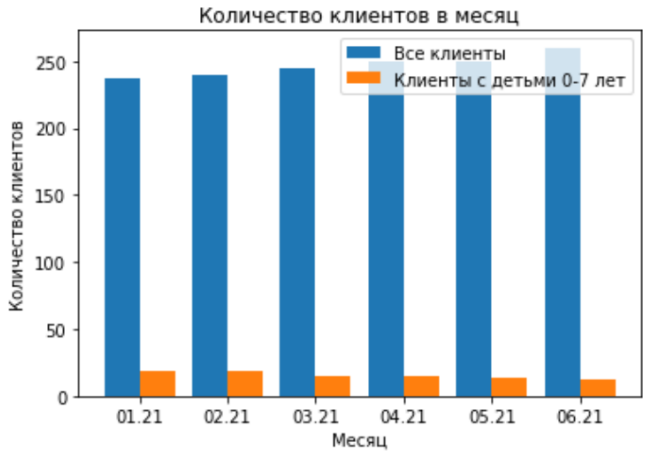

Рисунок 2

Наша компания заботится и о будущем подрастающего поколения, поэтому мы пытаемся свести количество отходов к минимуму и используем вторсырье по возможности. На рис. 3 вы можете заметить, что в нашей сети используется минимальное количество неперерабатываемого пластика на одного посетителя. **На графике изображено слишком мало конкурентов, чтобы делать такие выводы. Более того, по шкале видно, что разница в этом показателе между пиццериями достаточно маленькая, что даже может входить в погрешности измерения этой величины. (3 балла)** Более того, наша компания с гордостью носит титул ТОП-1 среди экологичных пиццерий района. **Не хватает определения того, какая пиццерия называется «экологичной». Если данная пиццерия единственная позиционирует себя как «экологичная», то титул «ТОП-1» не имеет большого веса. (2 балла)**

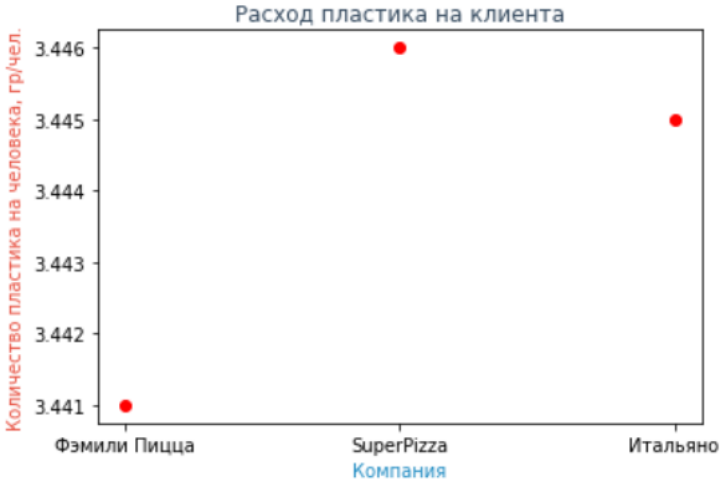

Рисунок 3

Посетители высоко оценивают качество блюд и обслуживания: на платформе Яндекс.Карты средняя оценка нашей сети 4.9 (у главных конкурентов 4.7). **Не хватает информации о количестве оценок. Если у данной пиццерии средняя оценка определялась при 10 оценках клиентов, а у конкурентов на основе 1000 оценок, то сложно делать выводы о преимуществе «Фэмили Пицца». (3 балла)** В условиях пандемии мы так же разработали интернет-платформу для заказов. После каждого заказа мы проводим опрос по удобству сайта, средняя оценка пользователей 4.96, что выше средних оценок аналогичных сервисов конкурентов в Google Play / AppStore. **Неправильно сравнивать оценки из разных источников. (2 балла) Более того, пользователи сайта пиццерии могут поставить оценку только по завершении заказа – у сайта могут быть**

существенный проблемы, из-за которых часть клиентов не могут сделать заказ, но на оценке сервиса это не отразится, так как эти клиенты не дойдут до этапа обратной связи. (3 балла)

Исходя из данных выше, мы имеем стабильное конкурирующее положение по рынку, а по некоторым показателям уже превосходим конкурентов.

Задание 2. – 25 баллов

Петя работает в маркетинговом отделе и занимается подготовкой данных для рубрики «интересные факты о вас», которая отображается клиентам. За прошлую неделю Петя подготовил много выборок, но у одной из потерялось описание. Эта выборка дана в файле target.csv. Однако Петя помнит, что для расчетов ему понадобился в том числе граф связей клиентов друг с другом. Этот граф лежит в файле graph_example.csv

Помогите Пете – объясните, как строилась выборка и по какому принципу проставлялся target. Вам не даны все данные, которые использовал Петя, поэтому у вас может не получиться повторить target идеально. Однако вы можете искать и подтверждать зависимости, которые Петя использовал. Воссоздайте искомую величину и проверьте ее корреляцию с исходным target. Заведомо известно, что можно получить корреляцию > 0.99.

Описание полей graph_example.csv:

- v1 – ID первого клиента в связи
- v2 – ID второго клиента в связи
- weight – «вес» связи, ее значимость

Описание полей target.csv:

- vertex – ID клиента
- target – некоторая величина, характеризующая клиента, описание которой требуется восстановить

Решение:

Задача заключается в том, чтобы по связям клиентов построить некоторое число, характеризующее их. При просмотре данных замечаем, что веса всех связей в графе равны 1, поэтому пока что все связи клиентов можно считать равнозначными. В первую очередь хочется посмотреть на связь target и суммы весов всех связей для каждого клиента (так как все веса связей = 1, то по сути это количество связей для одного клиента) – назовем эту величину X. Посмотрев на диаграмму рассеяния этих двух величин, замечаем некоторую линейную зависимость с некоторыми отклонениями (рисунок 4).

```
Зависимость: target = 0.351 * x + 2.385
Корреляция исходного target и рассчитанной величины x: 0.916
```

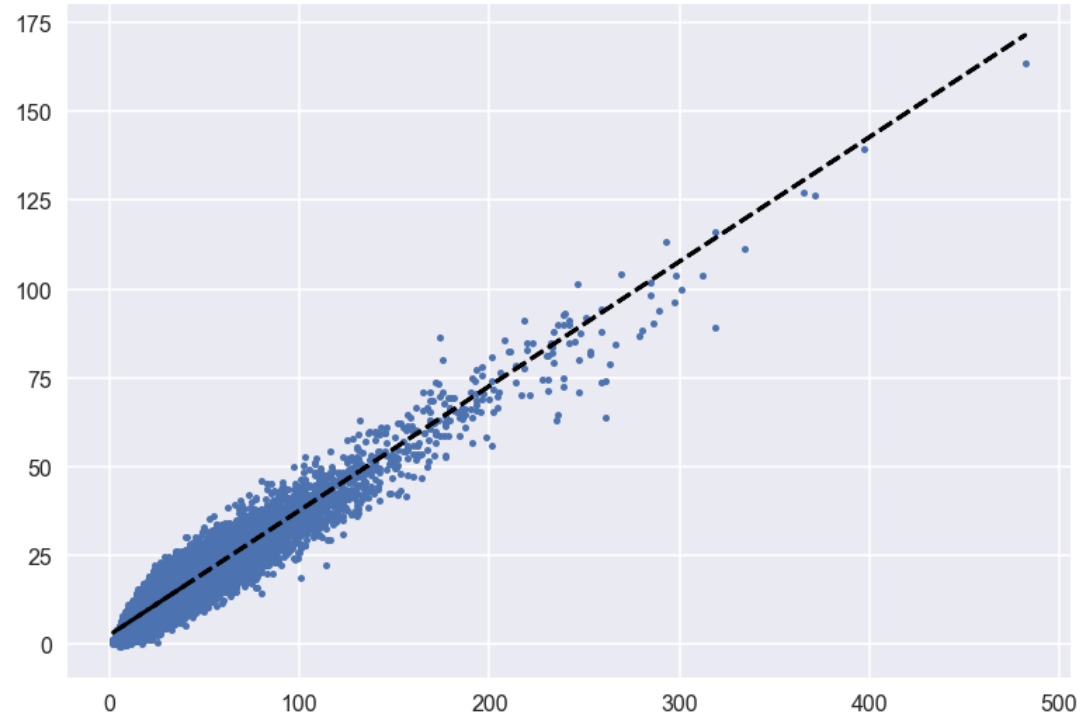

Рисунок 4

Получаем высокий коэффициент корреляции, но все же пока довольно большой разброс. Попробуем улучшить результат.

Сейчас мы учитывали веса связей, но не учли, что люди, с которыми связан наш клиент, тоже могут иметь разный приоритет. По имеющимся данным не так много вариантов ранжирования клиентов – попробуем построить величину X таким образом: сначала для каждого клиента посчитаем количество его связей и назовем эту величину весом клиента, а затем для каждого клиента посчитаем сумму весов его «друзей». Посмотрим, как эта величина кореллирует с target (рисунок 5).

```
Зависимость: target = 0.011 * x + 1.095
Корреляция исходного target и рассчитанной величины x: 0.943
```

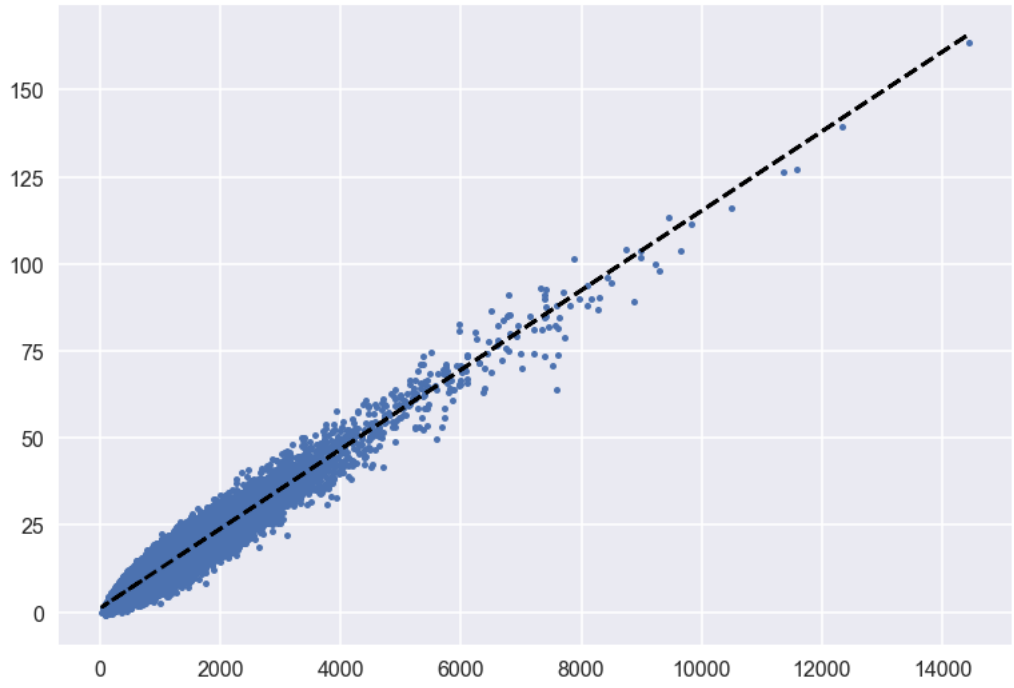

Рисунок 5

Мы видим увеличение корреляции, что указывает на зависимость между весами клиентов-друзей и target. Посмотрим, можем ли мы еще улучшить X.

Заметим, что для каждого клиента значение target меньше, чем количество его связей. Это наталкивает на гипотезу, что не все связи учитываются в target. Попробуем произвести выбор нужных связей в соответствии с ранжированием клиентов-друзей по их весу. Есть два варианта: мы учитываем связи с клиентами «маленького» веса (их вес меньше некоторого фиксированного числа), или же учитываются связи с клиентами «большого» веса (аналогично, их вес больше некоторого числа). Предположим, что эта фиксированная граница выбора связи будет одинаковой для всех клиентов (можно экспериментировать и выбирать для каждого клиента свою границу, но для начала мы ограничимся единой границей для всех). Для двух выделенных вариантов попробуем найти оптимальную границу перебором (для каждой границы не учитываем связи, которые меньше/больше текущей границы, в итоге X – количество связей с теми клиентами, чей вес больше/меньше этой границы). Оптимум выбираем через максимизацию корреляции.

Такой перебор находит оптимальный вариант, в котором для каждого клиента выбираются связи с клиентами веса >= 28. Для этого варианта корреляция target и X составляет 0.992 (рисунок 6).

Критерии (X – воспроизведенная участником величина):

- **При корреляции X и target > 0.90 – 10 баллов**
- **При корреляции X и target > 0.94 – 20 баллов**
- **При корреляции X и target > 0.99 – 25 баллов**

```
Зависимость: target = 0.987 * x + 0.185
Корреляция исходного target и рассчитанной величины x: 0.993
```

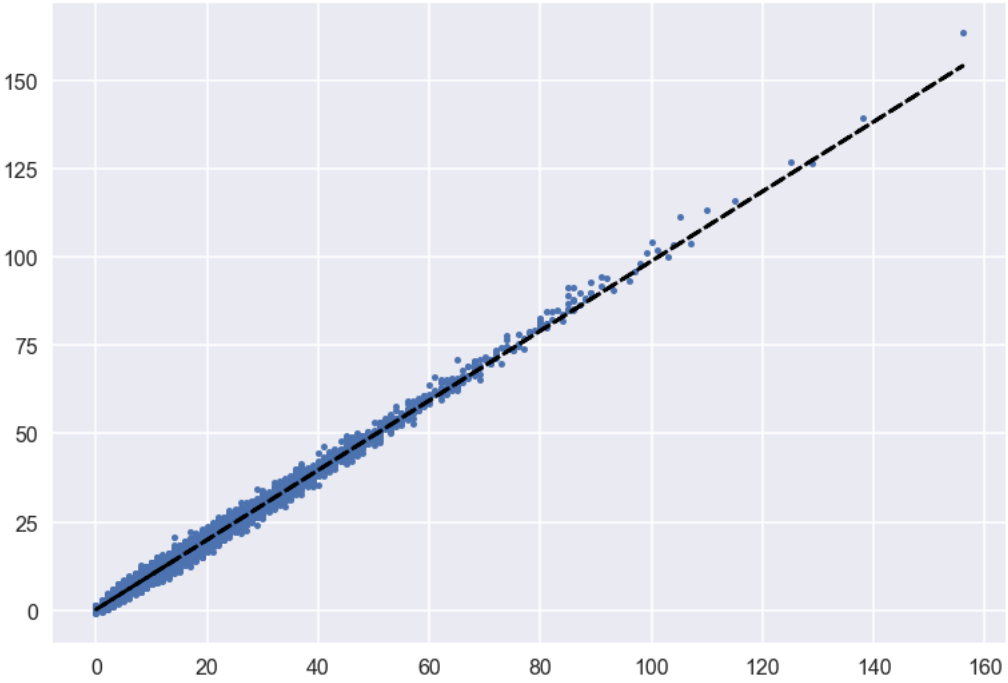

Рисунок 6

Задание 3. – 25 баллов

Петя решил исследовать статистику трат клиентов банка в зависимости от региона. Для этого он собрал выборку, содержащуюся в файле task4_data.csv. Данные описываются следующими полями:

- party_rk – идентификатор клиента
- purchase_sum – сумма трат за определенный период
- russian_region_nm – название региона проживания клиента
- russian_region_cd – код региона проживания клиента
- timediff_to_msk_hour_cnt – разница в часах между часовым поясом проживания клиента и МСК
- russian_federal_district_cd – код федерального округа проживания клиента

Петю интересуют следующие вопросы о выборке:

1. Как различаются траты населения в разных частях страны?
2. Какие регионы самые «тратящие»? Какие самые богатые?
3. Где наибольший разрыв в доходах?
4. Влияет ли часовой пояс на то, сколько тратят люди?

Попробуйте ответить на эти вопросы с помощью визуализации. Будьте лаконичны – максимум 3 графика. Не используйте в визуализации карту, это не требуется и не приведет к получению дополнительных баллов. Прокомментируйте графики и сделайте выводы, где возможно. Если вы делали предобработку данных, то объясните ее и приложите файл.

Решение:

При первичной обработке данных видим наличие нескольких дублей среди party_rk – требуется понять, что с ними сделать. Две пары дублей абсолютно идентичные (party_rk = 148830292 и party_rk = 115257196), поэтому достаточно выкинуть одну из этих строк. В других двух парах регионы различаются (party_rk = 122480825 и party_rk = 145632455), – можно убрать их из рассмотрения, так как однозначно определить их принадлежность к конкретному региону не получается.

Для первого вопроса рассмотрим распределения трат по федеральным округам. Данных для каждого округа достаточно, чтобы рассматривать различные статистики, характеризующие распределения. На графике (рисунок 7) мы видим значения среднего, медианы, 25%-перцентиль и 75%-перцентиль. По этим данным видно, что ФО 001 (Центральный ФО) можно назвать самым тратящим, а ФО 004 (Северо-Кавказский ФО) тратит меньше всех. Так же среди оставшихся ФО выделяется ФО 002 (Северо-Западный ФО), остальные же имеют примерно одинаковые по порядку характеристики, но со своими особенностями.

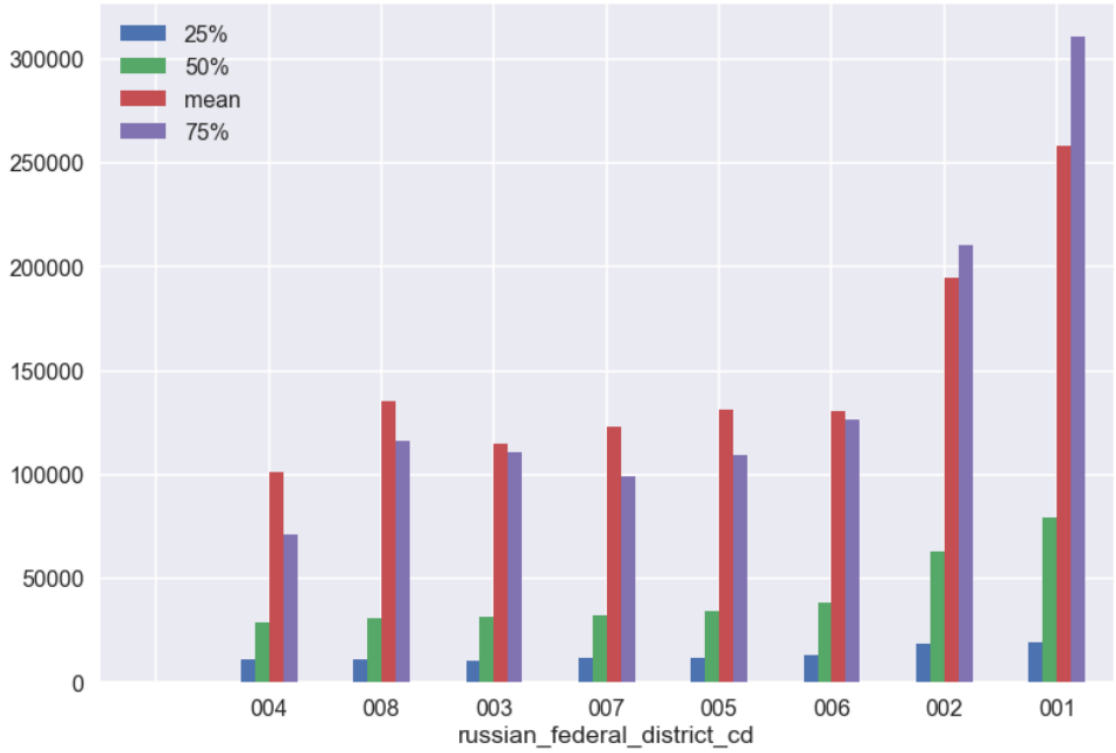

Рисунок 7

Чтобы ответить на следующие два вопроса, мы будем сравнивать максимальные значения трат, но так как в данных у каждого региона могут быть свои выбросы, то предварительно требуется их убрать из рассмотрения, так как они не отражают истинное распределение трат в каждом регионе, но могут сильно сказаться на ответе. Важно это сделать внутри каждого региона! Более того, хочется убрать из рассмотрения регионы, в которых слишком мало транзакций – про распределение трат в таких регионах не получится сделать точные выводы.

Разрыв в доходах мы будем оценивать через разницу между максимальными тратами и медианой по региону – минимальные траты в этом случае непоказательны. На рисунке 8 показаны распределения медианы, 75%-перцентиля, 95%-перцентиля и максимумов трат по регионам. Регионы справа на лево отсортированы по разрыву в доходах (доходы оцениваются через траты). Для удобства здесь приведены ТОП-10 регионов по разрыву, они же получаются самыми тратящими среди остальных.

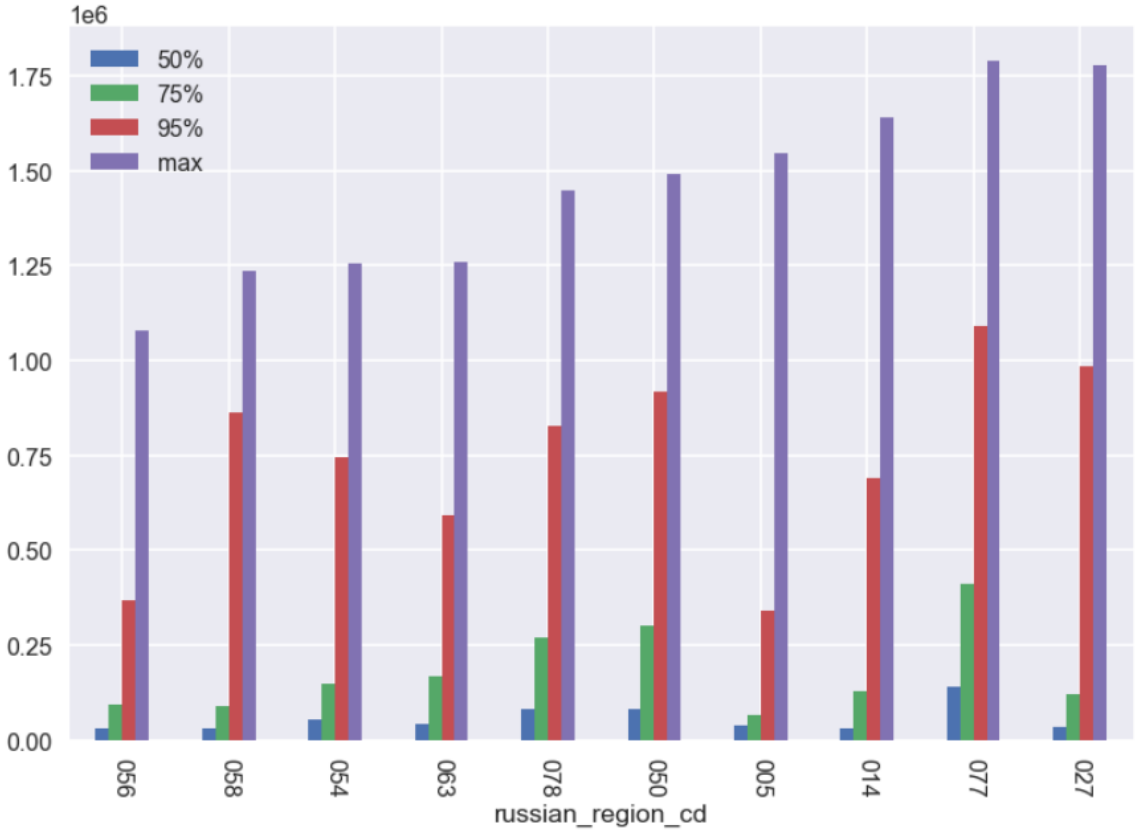

Рисунок 8

С наибольшим разрывом в доходах выделяются регионы 027, 077 и 014 (Хабаровский край, г. Москва и Республика Саха). Самые «богатые» регионы выделяются по приближенности 95%-перцентиля и 75%-перцентиля к максимумам трат (по сути равномерность больших трат): это 077, 050, 078, 063 (г. Москва, Московская область, г. Санкт-Петербург и Самарская область соответственно).

На рисунке 9 показаны распределения трат в зависимости от часового пояса (графики боксплот, где прямоугольник показывает границы 25%-перцентиля и 75%-перцентиля, а линия внутри прямоугольника показывает медиану). Изначально были обработаны выбросы (внутри каждой таймзоны), и убраны территории, данных по которым слишком мало, чтобы делать выводы о характере распределения трат (это территории с разницей в часах с МСК равной -1, 8 и 9 часов).

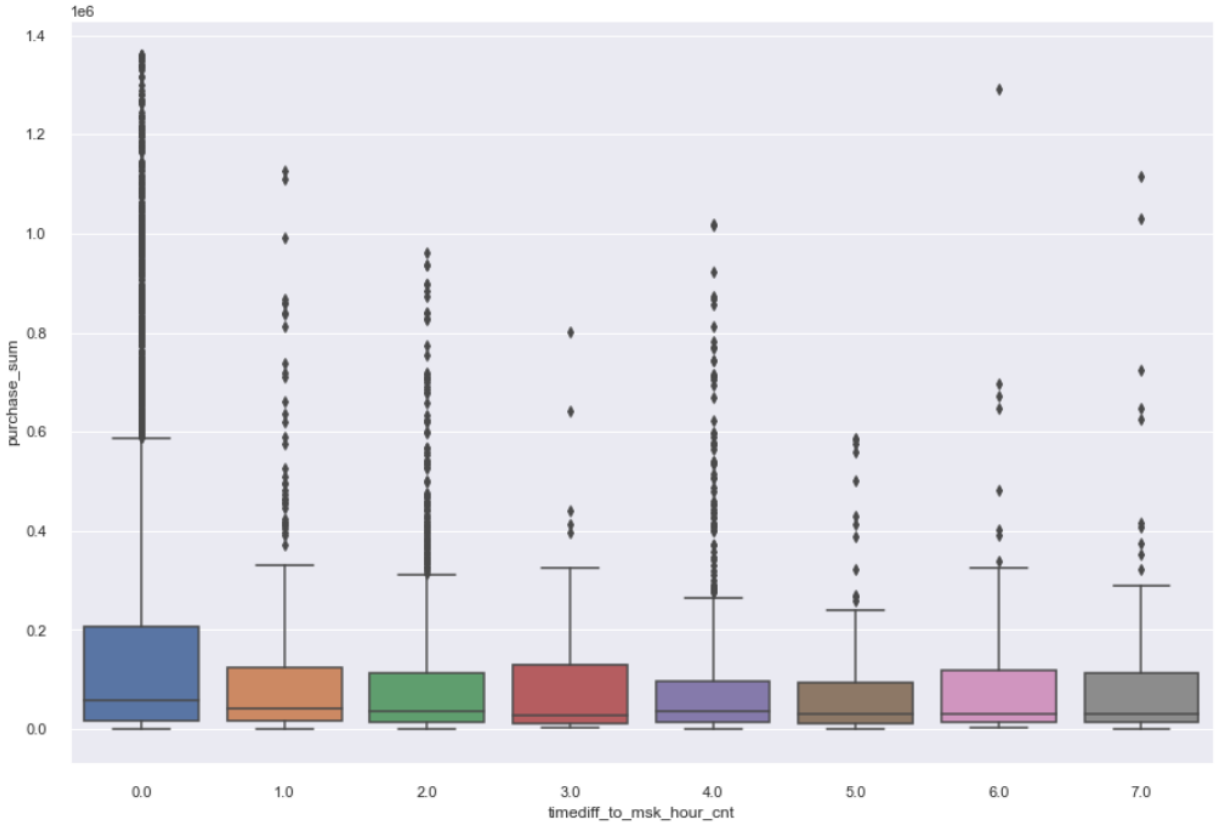

Рисунок 9

Главным образом выделяется московский часовой пояс, что объясняется тем, что в него входит большинство самых тратящих регионов. Распределения трат в других часовых поясах больше похожи между собой. Видимые различия в распределениях сложно назвать значимыми, так как у части таймзон все еще недостаточно данных для того, чтобы делать точные выводы о характере распределений. Дополнительно можно построить графики плотности распределений с разбивкой по часовым поясам (гистограммы или violinplot – аналог boxplot, но точнее видно характер распределения), в данной задаче они не выявят различий в характере распределений, только подтвердят одинаковый характер распределений и наличие тяжелых «хвостов» из больших трат.

Критерии:

- Проделана обработка и изучение данных – 2+2+2 баллов (обработка выбросов максимальных и минимальных значений + обработка регионов с маленьким количеством данных + обработка дублей в данных)
- Графики читаемые и верно размечены – 3 баллов
- За ответ на каждый вопрос – 4 балла (график действительно отвечает на вопрос и правильно интерпретирован)

Задание 4. – 25 баллов

Вам дан датасет с 6 признаками. 4 из этих признака были получены синтетически из других двух базовых с помощью различных математических операций с добавлением некоторого шума. Найдите базовые признаки и алгоритм получения синтетических признаков. Опишите математические преобразования, с помощью которых получились сгенерированные признаки. Главное описать тип преобразований, не указывая точные коэффициенты. Необходимые данные находятся в файле task5_data.csv

Решение:

Построим попарные диаграммы рассеяния для каждой пары столбцов (рисунок 10).

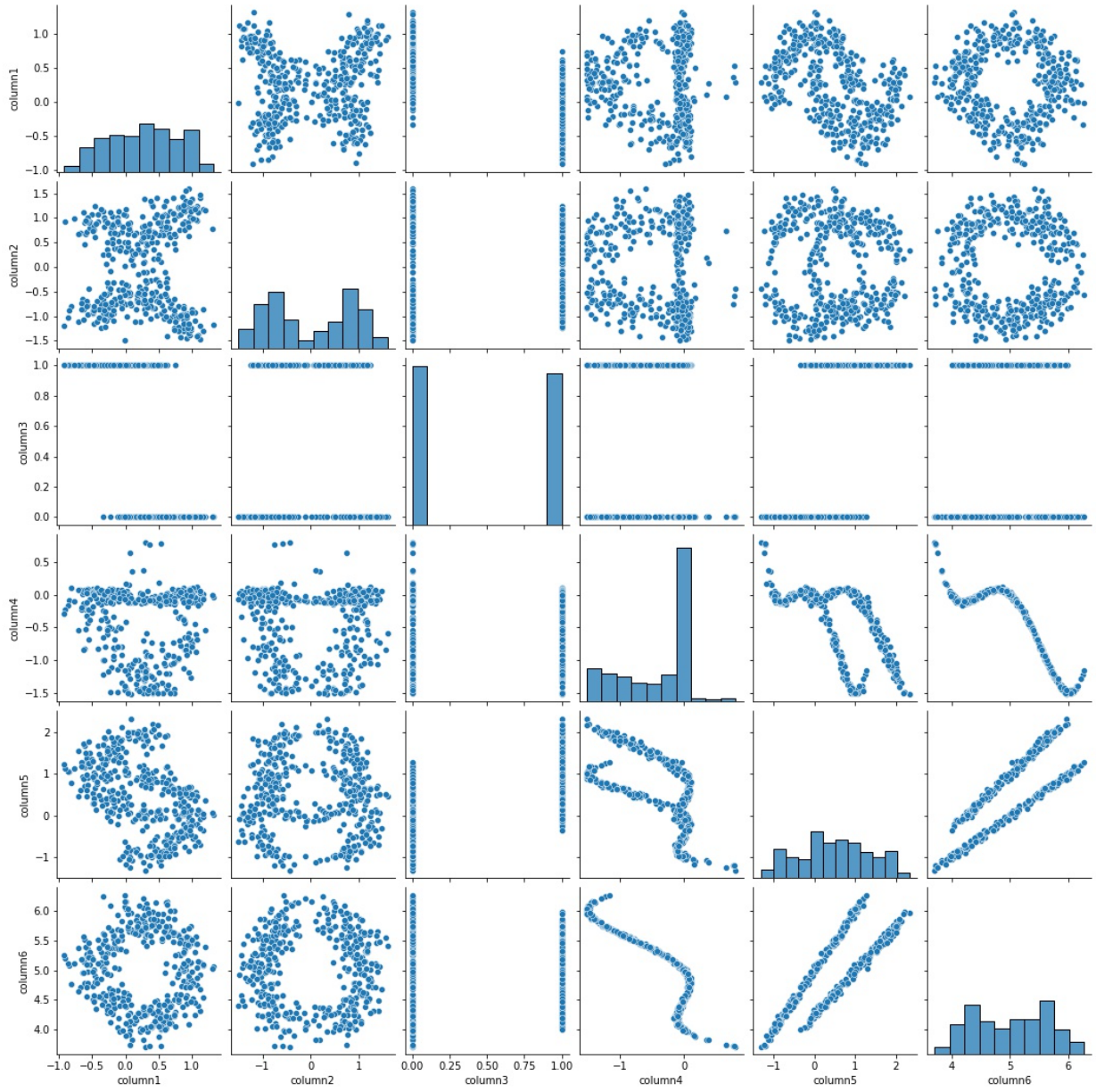

Рисунок 10

Сразу выделяется column3 – это категориальный признак всего с двумя значениями. Так же заметно, что на многих графиках точки визуально разделяются на два множества – вероятно из-за разбиения на 2 класса в соответствии со значениями в column3. Добавим это разделение на диаграммы рассеяния (рисунок 11).

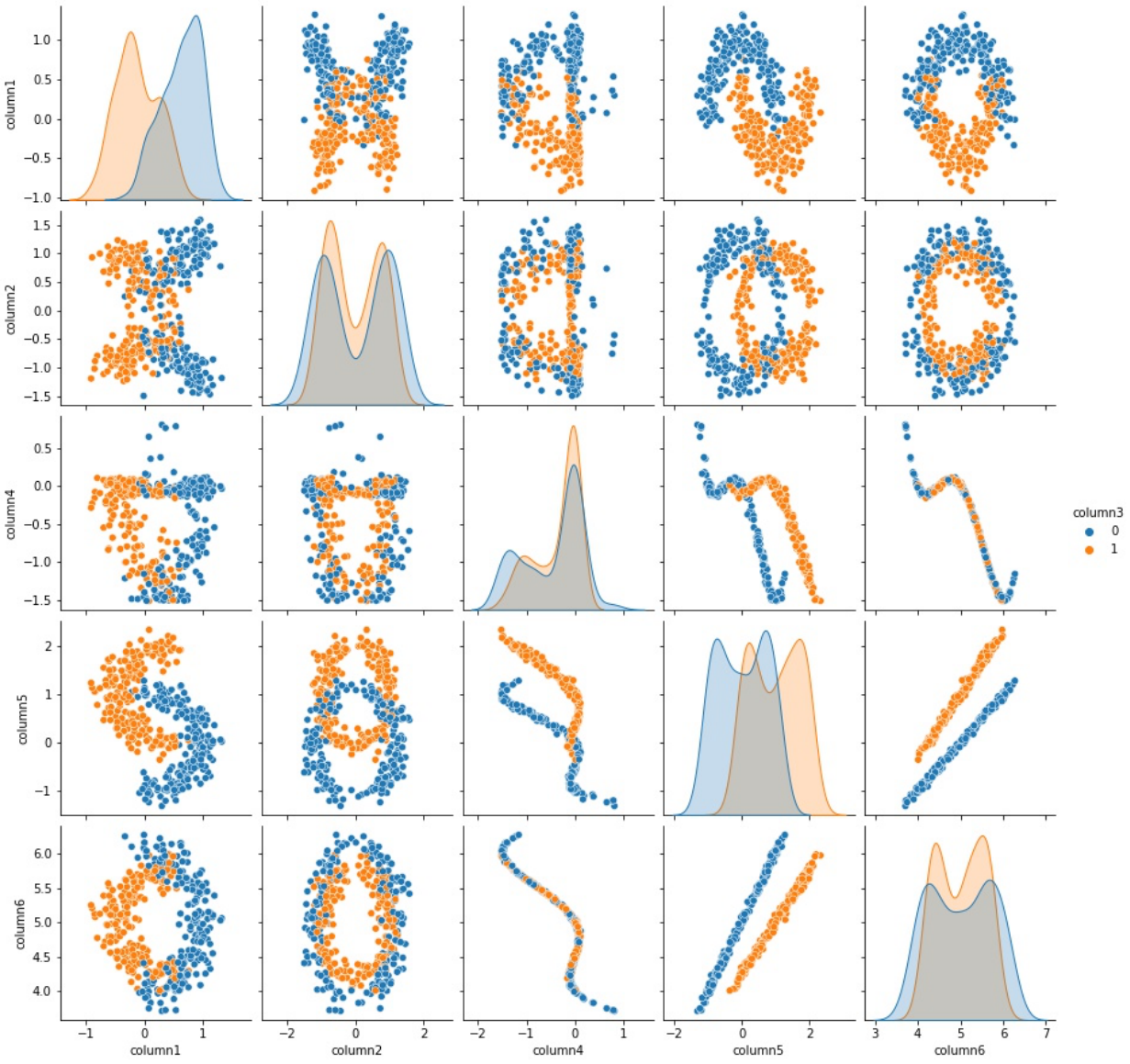

Рисунок 11

Теперь это разбиение стало более заметным. Среди оставшихся 5 признаков больше всех выделяется column6 – в нем зависимости с другими признаками наиболее выражены и проще всего описываются. Опишем зависимости:

1. column4 полиномиально зависит от column6 и не имеет явной зависимости от column3. По наличию трех локальных экстремумов можно предположить, что это зависимость выражается многочленом 4ой степени, по значениям column6, в которых значения column4 находятся в окрестности 0, можно восстановить корни многочлена. При рассмотрении многочлена (x - 4)*(x - 4.5)*(x - 5)*(x - 6.5) видим достаточно хорошую аппроксимацию.
2. column5 с разбиением на классы линейно зависит от column6 (с добавлением некторого шума).
3. column2 получается из column6 постоением двух окружностей с центром в (5, 0) разных радиусов для разных классов, определяемых значениями column3 (но значения column2, изображаемые на оси OY распределены равновероятно относительно оси OX).

4. column1 получается из column6 похожим образом как и column2 – строится окружность, где значения column1 относительно оси OX распределяются в соответствии с классами из признака column3.

Критерии:

- Учтено разбиение на классы – 5 баллов
- Расписаны зависимости – по 4 балла за зависимость
- Верно расписана логика, по которой генерируются переменные – 4 балла