

Исследование компании «Мегафон»

Анализ зависимости технических
показателей качества связи на
удовлетворённость клиентов



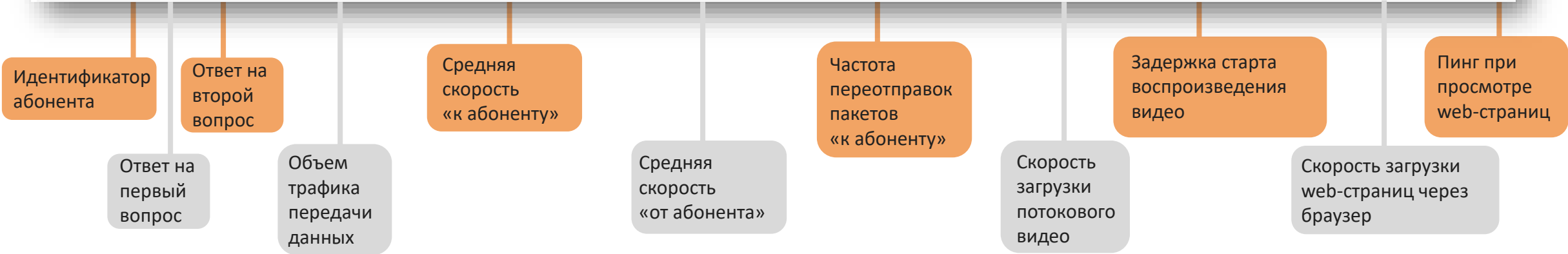
Описание задачи




Подготовка данных


```
#Загружаем библиотеки
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
...
#Загружаем датасет
df_main = pd.read_csv('megafon.csv')
df_main.head()
```

	user_id			Q1	Q2	Total Traffic(MB)	Downlink Throughput (Kbps)	Uplink Throughput (Kbps)	Downlink TCP Retransmission Rate(%)	Video Streaming Download Throughput (Kbps)	Video Streaming xKB Start Delay(ms)	Web Page Download Throughput (Kbps)	Web Average TCP RTT (ms)
0	1	5	NaN			775.48846	360.13	86.56	3.93	1859.15	2309	1007.82	83
1	2	5	4			861.96324	3023.54	411.18	1.27	667.47	2080	255.36	425
2	3	1	4			261.11860	790.96	34.20	1.79	1079.60	6367	535.85	485
3	4	8	3			179.18564	2590.97	325.88	0.80	7053.81	3218	1221.02	51
4	5	2	2, 3, 4			351.99208	731.61	223.54	1.15	4550.38	1767	2336.56	68



```
#Смотрим описание признаков
df_main.info()
```

user_id	3112	non-null	int64		3110	non-null	object
Q1	3110	non-null	object		1315	non-null	object
Q2	1315	non-null	object				
Total Traffic(MB)	3112	non-null	float64				
Downlink Throughput(Kbps)	3112	non-null	float64				
Uplink Throughput(Kbps)	3112	non-null	float64				
Downlink TCP Retransmission Rate(%)	3112	non-null	float64				
Video Streaming Download Throughput(Kbps)	3112	non-null	float64				
Video Streaming xKB Start Delay(ms)	3112	non-null	int64				
Web Page Download Throughput(Kbps)	3112	non-null	float64				
Web Average TCP RTT(ms)	3112	non-null	int64				



Информация

Первый технический показатель представлен как сумма за период в одну неделю перед участием в опросе. Остальные технические показатели отображают среднее значение по данному признаку за период в одну неделю перед участием в опросе

Предобработка данных

```
#Смотрим наличие пропущенных значений  
df_main.null().sum()
```

user_id	0
Q1	2
Q2	1797
Total Traffic(MB)	0
Downlink Throughput(Kbps)	0
Uplink Throughput(Kbps)	0
Downlink TCP Retransmission Rate(%)	0
Video Streaming Download Throughput(Kbps)	0
Video Streaming xKB Start Delay(ms)	0
Web Page Download Throughput(Kbps)	0
Web Average TCP RTT(ms)	0



Вывод

Имеются пропущенные значения по первому вопросу (2 значения) и по второму вопросу (1797 значений).

Если на первый вопрос отвечали 9-10, тогда ответ на второй вопрос отсутствует.

Предобработка данных

Вопрос Q1

NaN

В вопросе не должно быть пропущенных значений

int

Значение в признаке должно быть целочисленным (тип int64)

```
#Смотрим все уникальные значения Q1
pd.unique(df_main.Q1)
```

```
['5', '1', '8', '2', '3', '9', '10', '7', '4', '11', '6', '2, 9',  
'0', '1, 3', '19', '15', nan, '1, 6', '***** ** **',  
'3 - дер.Ширяево Волоколамского района, 9 - в Москве', '10, 9',  
'Чем даль ше,тем лучше.Спасибо за ваш труд.Оценка 10 !',  
'ОЦЕНКА-3/НЕВАЖНО/', 'Отвратительно',  
'Я ценой услуг не удовлетворен', 'Пока не понял', '3, 9', '5, 6',  
'0, 1, 5', '5, 7', 'hi',  
'4. Тульская область Заокский район. Романовские дачи связи почти нет',  
'Немагу дать атценку денги незашто снимаеть скоро выключаю',  
'10, 50',  
'Очень хорошо. Обслуживания я довольно. Спасибо вам.555', '?',  
'Поохое',  
'Когда в Москве-10 а когда в калужской области в деревне Бели-1',  
'Нет', 'Да', 'Ужасно',  
'3 тройка, связь отвратительная, жалко платить за такой тарив',  
'Чдтгдтгчдтгчч', '3, 7', '20, 89031081392', '1, 8', 'Без з',  
'10, 5', '2, 5',  
'Я в Смол. Области живу сейчас, не пользуюсь телефоном совсем'],
```

!

Информация

Имеется множество неформатных значений. Таким образом, все значения не подходящие под нужный формат не имеют вариантов ответа на второй вопрос.

Предобработка данных

Вопрос Q1

NaN

Пропущенные значения

Так как имеется всего два пропущенных значения, то было принято решение о заполнении этих объектов значением 10.

```
#Заполним пропущенные значения
df_main.Q1 = df_main.Q1.fillna(10)
```

###Скучный код###

```
#Заменим значения в соответствии с условиями при
помощи регулярных выражений
for i in range(df_main.shape[0]):
    ...

#Переведём столбец в числовой тип
df_main.Q1 = df_main.Q1.astype(int)

#Проверим
df_main.info()
```

int

Целочисленное значение

Условия форматирования для имеющихся неформатных ответов:

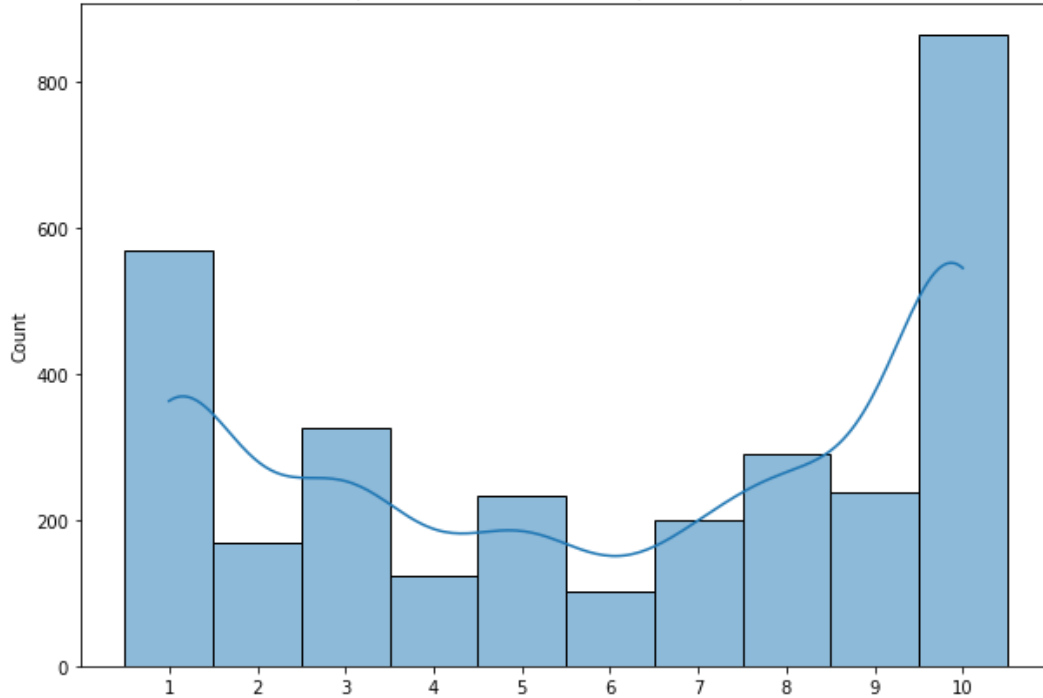
- Если число в ответе больше 9 (а также ответ "Да"), то значение в признаке Q1 меняется на 10;
- Если число в ответе меньше 9 (а также ответ "Нет" и ответы без чисел), то значение в признаке Q1 меняется на 1.

user_id	3112	non-null	int64
Q1	3112	non-null	int64
Q2	1352	non-null	object

Предобработка данных

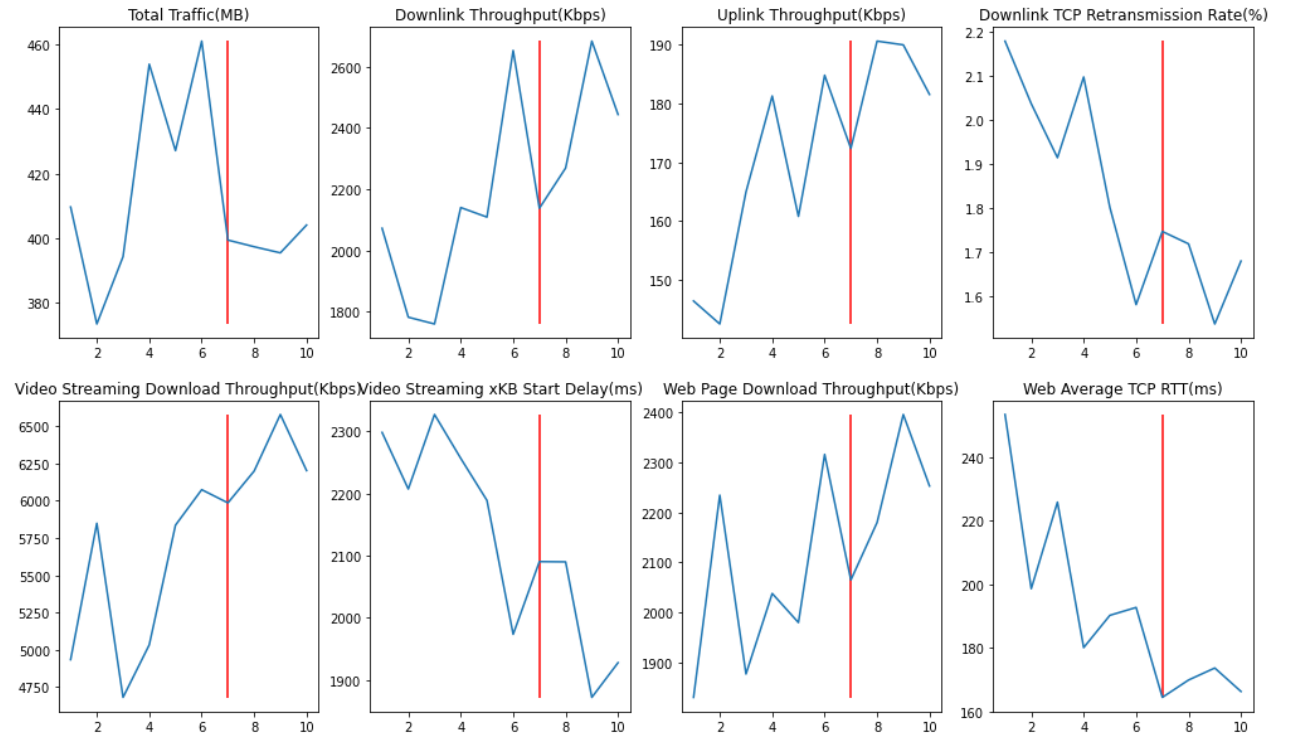
Вопрос Q1

Распределение ответов на первый вопрос (Q1)



Частотное распределение признака Q1

Преобладают предельные значения
(абоненты либо резко недовольны связью,
либо наоборот)



Динамика в изменениях значений

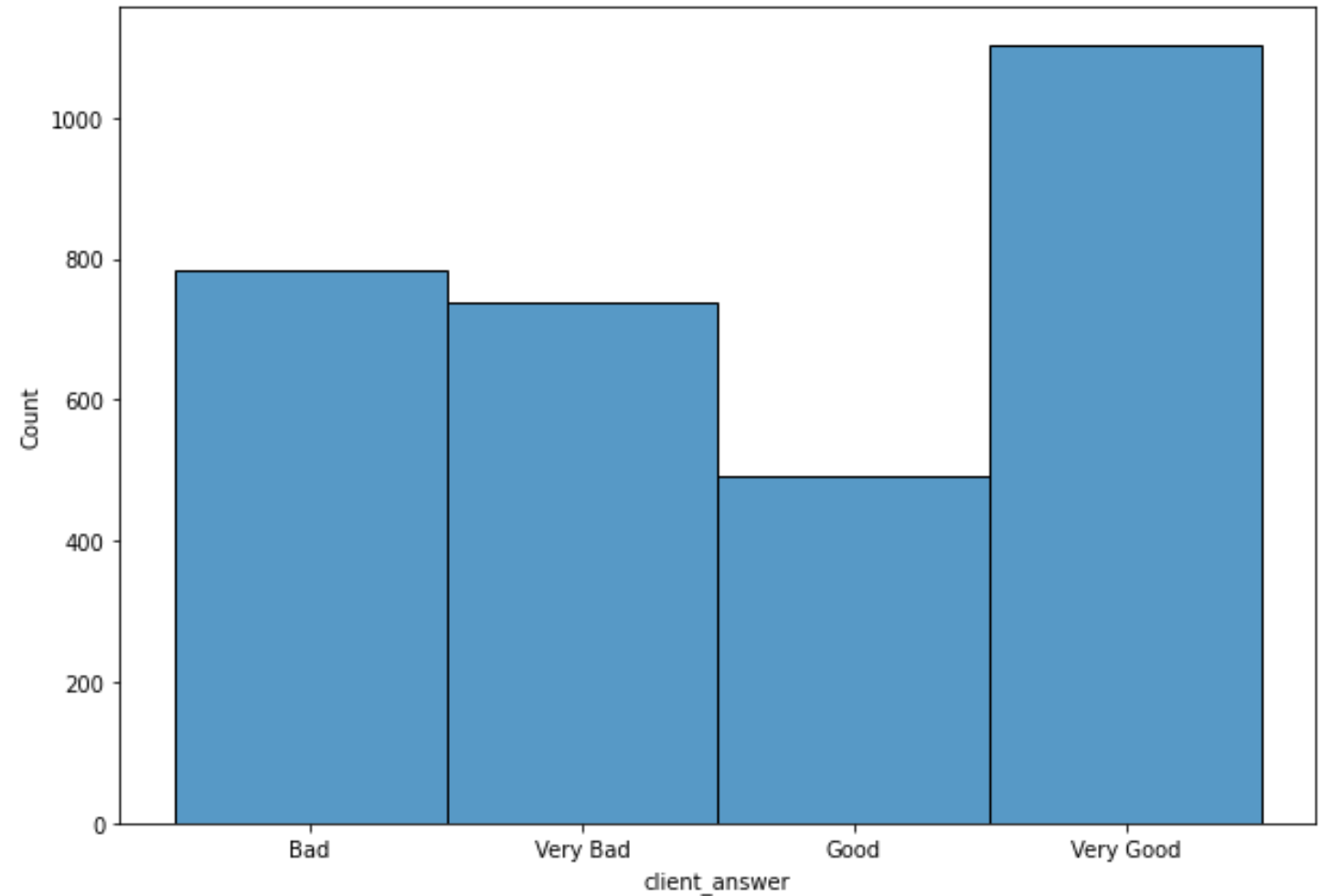
Виден рост, либо снижение темпа падения
значений в технических характеристиках при
оценке от 7 до 10

Предобработка данных Вопрос Q1



Более общие оценки

1-2: Very Bad
3-6: Bad
7-8: Good
9-10: Very Good



Предобработка данных

Вопрос Q2

NaN

В вопросе не должно быть пропущенных значений

int

Значение в признаке должно быть целочисленным (тип int64)

NaN

Пропущенные значения

Если в ответе на Q1 будет 9 или 10, то в ответе на Q2 зададим значение 0.

```
for i in range(df_main.shape[0]):  
    if df_main.iloc[i, 11] == 'Very Good':  
        df_main.iloc[i, 2] = '0'
```

В остальных случаях удалим наблюдения с пропущенным значением по Q2, так как из этих данных сложно будет сделать какой либо вывод.

```
#Удалим пропущенные значения по признаку Q2  
df_main.dropna(subset=['Q2'], inplace=True)
```

int

Целочисленное значение

Сделаем декодирование One-Hot-Encoding.

Создадим следующие признаки:

- A1 (Answer 1) – Нодозвоны, обрывы при звонках;
- A2 – Время ожидания гудков при звонке;
- A3 – Плохое качество связи в зданиях, ТЦ и тд;
- A4 – Медленный мобильный Интернет;
- A5 – Медленная загрузка видео;
- A6 – Затрудняюсь ответить;
- A7 – Свой вариант.



Итоговый вид

Выводы после предварительной обработки

Интерпретация ответов
на вопрос Q2 (One-Hot-
Encoding)

Video Streaming xKB Start Delay(ms)	Web Page Download Throughput(Kbps)	Web Average TCP RTT(ms)	client_answer	A1	A2	A3	A4	A5	A6	A7
2080	255.36	425	Bad	0	0	0	1	0	0	0
6367	535.85	485	Very Bad	0	0	0	1	0	0	0
3218	1221.02	51	Good	0	0	1	0	0	0	0
1767	2336.56	68	Very Bad	0	1	1	1	0	0	0
4223	856.05	220	Very Bad	0	0	0	1	1	0	0

Интерпретация ответов
на вопрос Q1



Удалим ненужные признаки

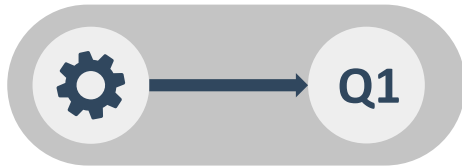
Нет необходимости в признаках Q1 и Q2, так как мы интерпретировали их в новые признаки.

Уберём графы в которых нет конкретного ответа на вопрос Q2. Эти наблюдения не информативны.

Основное исследование

ГИПОТЕЗА №1





Гипотеза №1



Нулевая гипотеза

Технические показатели с высокой оценкой в первом вопросе (Q1) не отличаются по качеству от технических показателей с более низкой оценкой по первому вопросу (Q1)



Альтернативная гипотеза

Существует связь между техническими показателями и оценкой по первому вопросу (Q1)

?

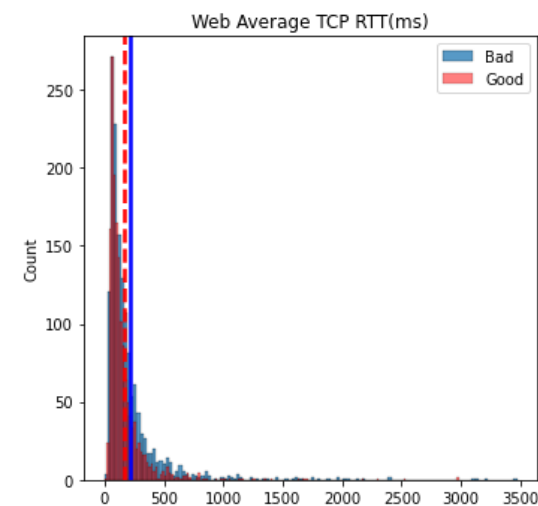
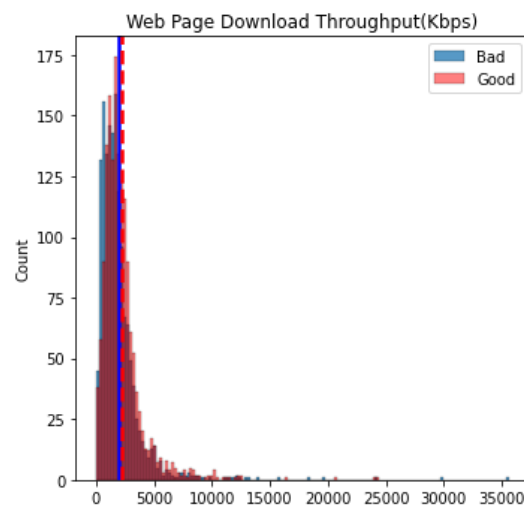
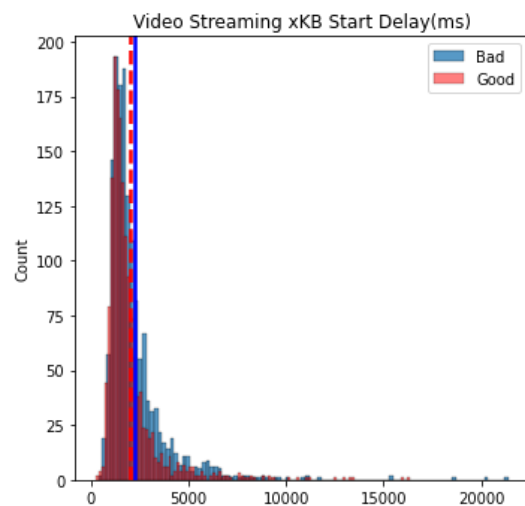
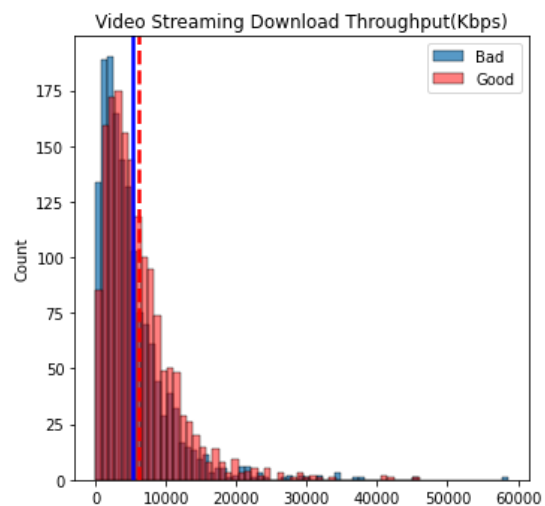
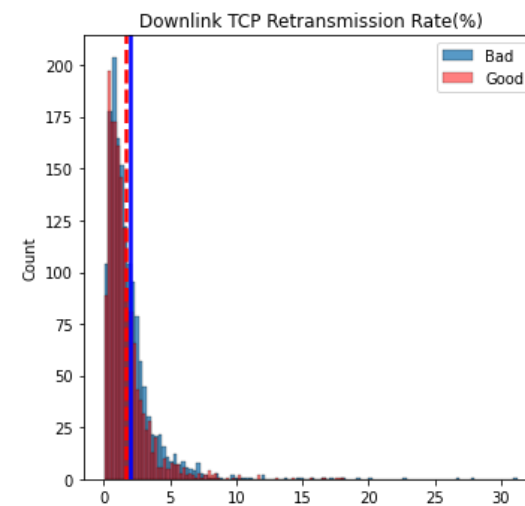
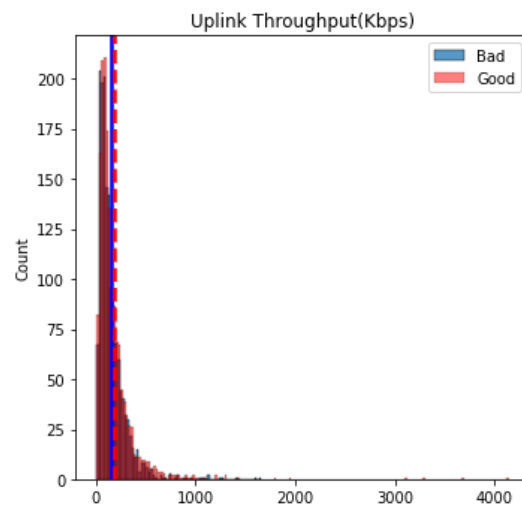
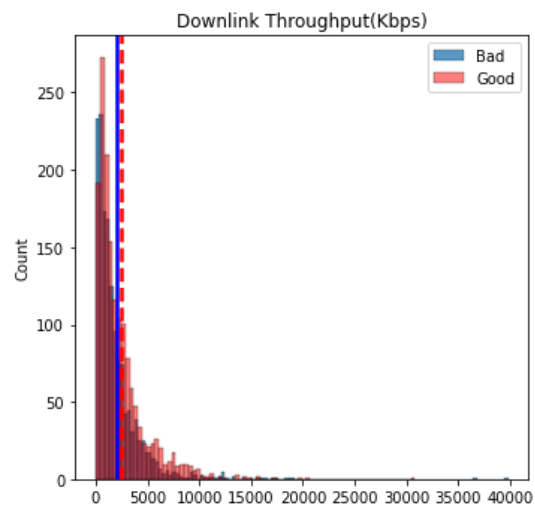
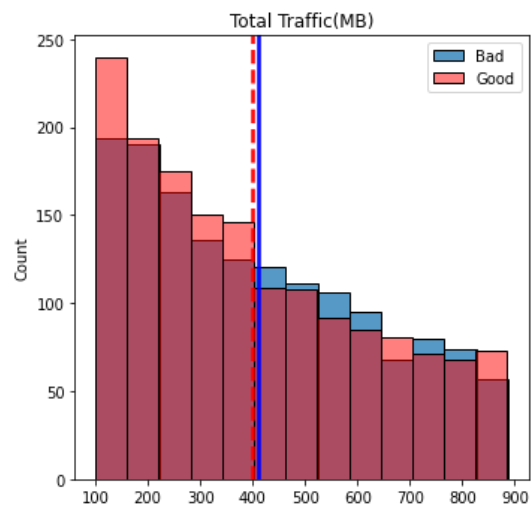
Цель

Определить целесообразность проведения такого рода опросов.



Разведочный анализ

Гипотеза №1

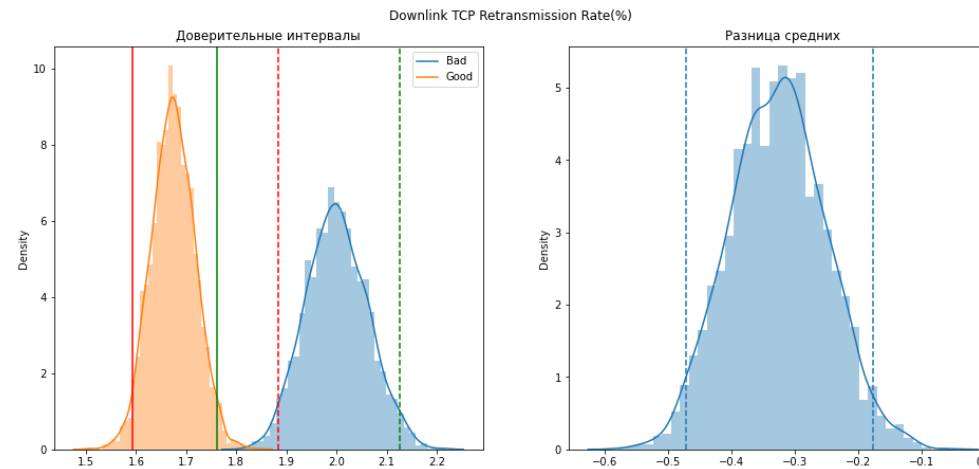
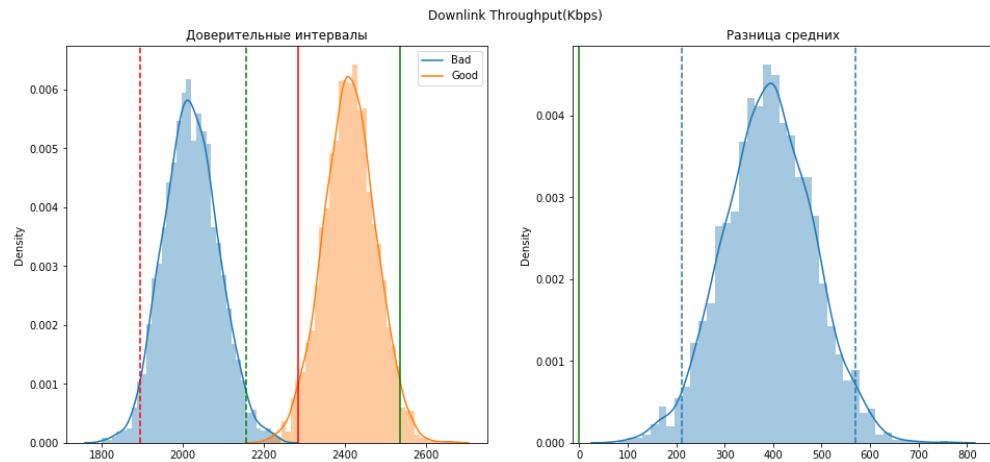
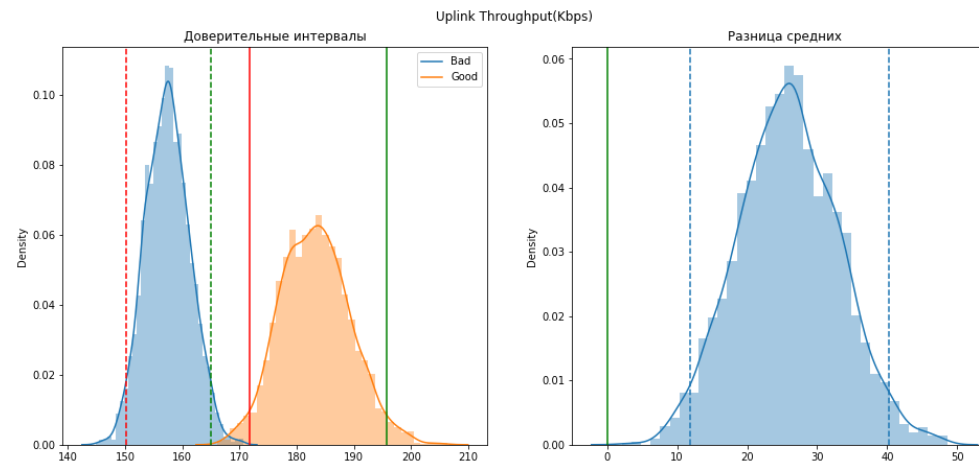
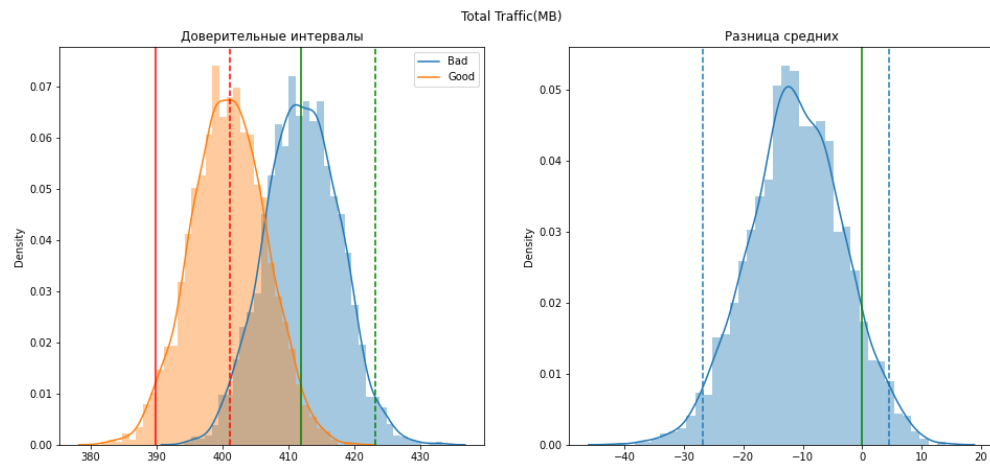




Bootstrap

Гипотеза №1

N = 3000





Выводы

Гипотеза №1



Нулевая гипотеза

Технические показатели с высокой оценкой в первом вопросе (Q1) не отличаются по качеству от технических показателей с более низкой оценкой по первому вопросу (Q1)



Альтернативная гипотеза

Существует связь между техническими показателями и оценкой по первому вопросу (Q1)



Имеется зависимость от технических показателей качества связи абонентов и их оценкой. Качество связи у людей поставивших более низкую оценку действительно хуже. Проведение данного опроса (и подобных ему в дальнейшем) является целесообразным.

Основное
исследование

ГИПОТЕЗА №2





Редактируем

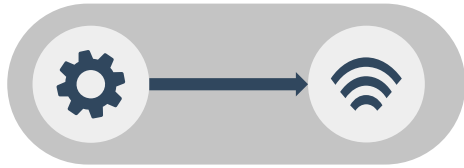
Гипотеза №2

A3 – Плохое качество связи в зданиях, ТЦ и т.п.	705
A1 – Недозвоны, обрывы при звонках	647
A4 – Медленный мобильный интернет	625
A5 – Медленная загрузка видео	222
A2 – Время ожидания гудков при звонке	184
A7 – Свой вариант	86
A6 – Затрудняюсь ответить	50



problem	
Internet and Mobile (A3)	824
Internet (A4, A5)	211
Mobile (A1, A2)	204
Other (A6, A7)	111

A1	A2	A3	A4	A5	A6	A7	problem
0	0	0	1	0	0	0	internet
0	0	0	1	0	0	0	internet
0	0	1	0	0	0	0	internet and mobile
0	1	1	1	0	0	0	internet and mobile
0	0	0	1	1	0	0	internet
1	0	1	1	0	0	0	internet and mobile



Гипотеза №2



Нулевая гипотеза

Метрики для характеристики качества интернет соединения со значением «Internet» в признаке problem не отличаются по качеству от метрик со значением «Mobile» в признаке problem



Альтернативная гипотеза

Существует связь между метриками для характеристики качества интернет соединения и признаком problem (сформированным на основе ответов на второй вопрос Q2)



Цель

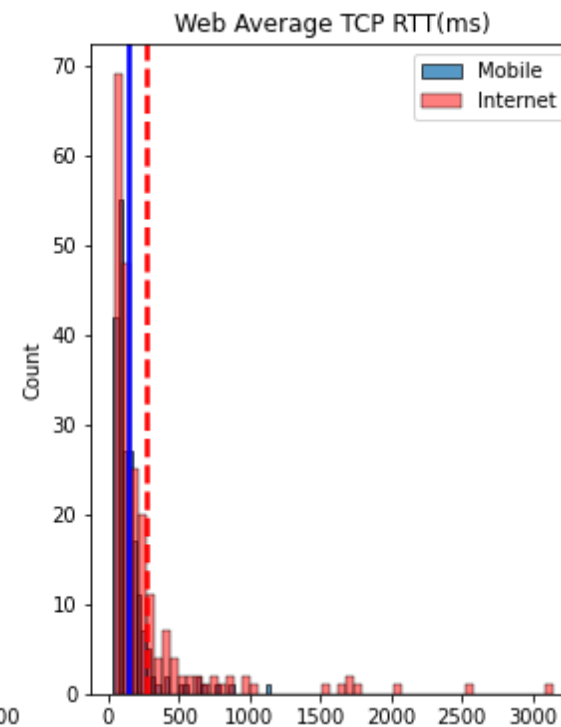
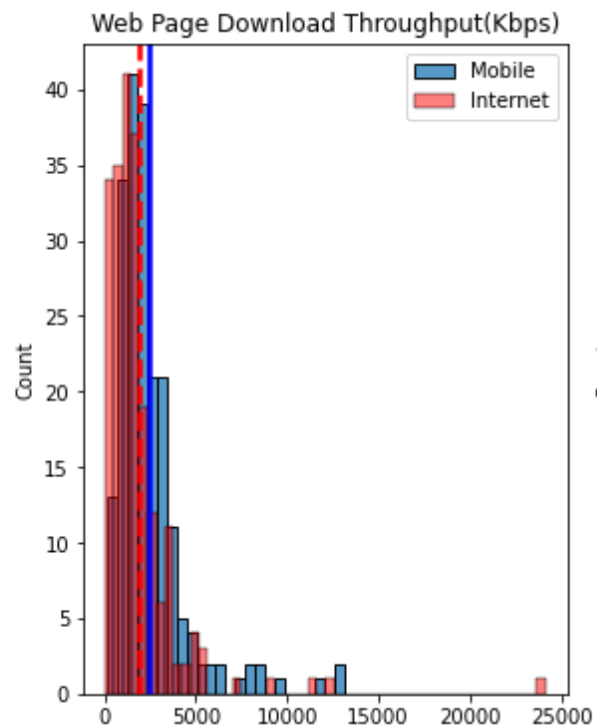
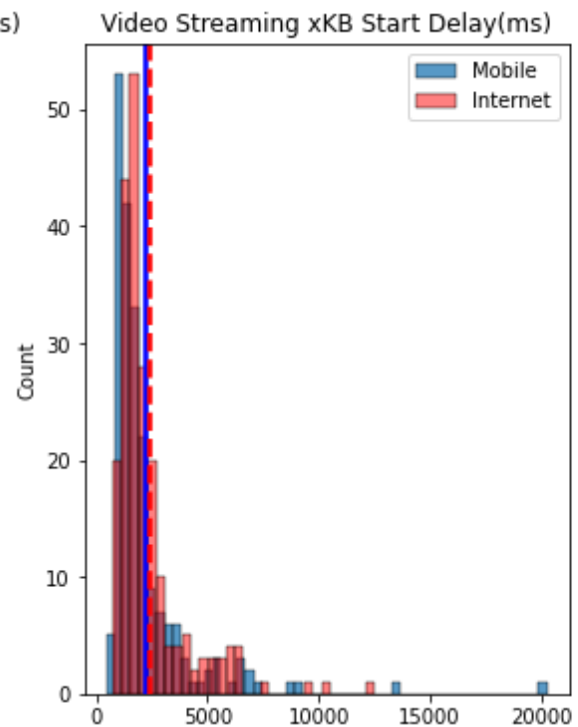
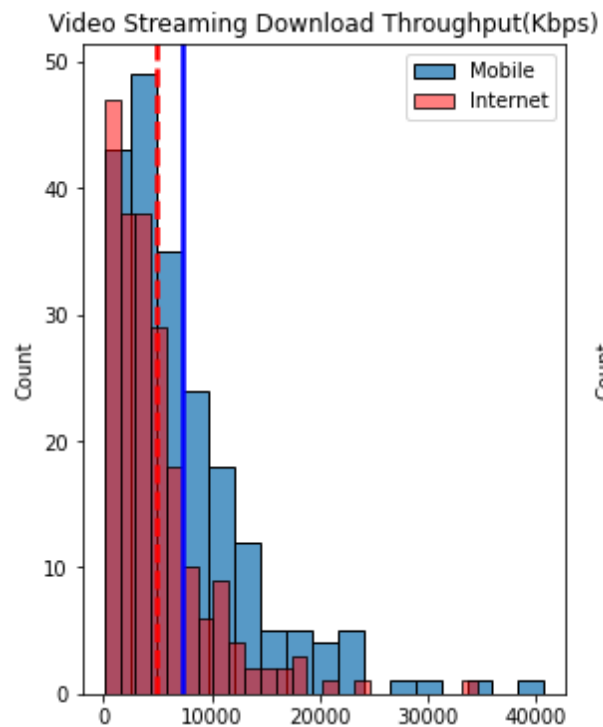
Выявить абонентов с низким качеством интернет соединения.



Разведочный анализ

Гипотеза №2

Video Streaming Download Throughput(Kbps)	Video Streaming xKB Start Delay(ms)	Web Page Download Throughput(Kbps)	Web Average TCP RTT(ms)	problem
667.47	2080	255.36	425	internet
1079.60	6367	535.85	485	internet
1699.64	4223	856.05	220	internet
8878.63	1678	2358.38	94	mobile
23392.06	757	2109.31	60	mobile





Анализ

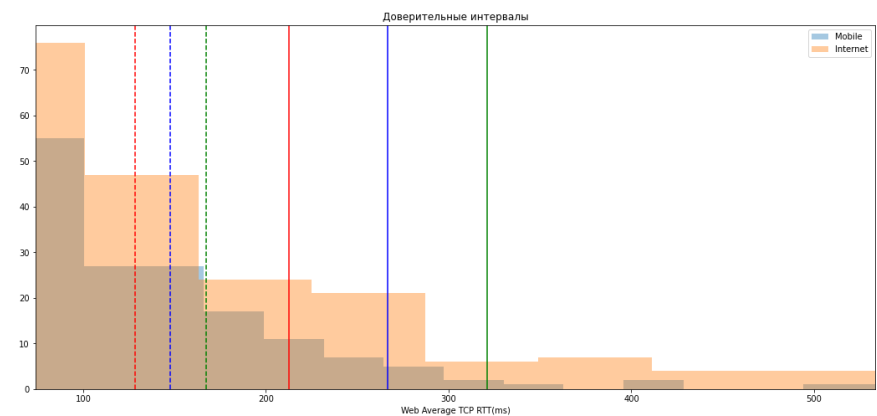
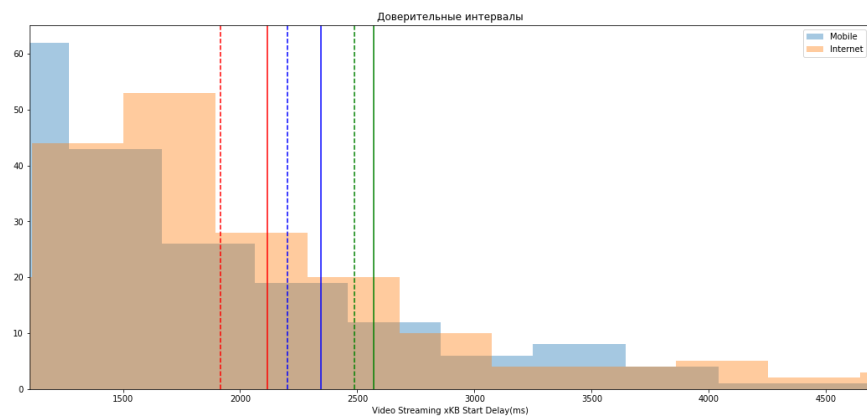
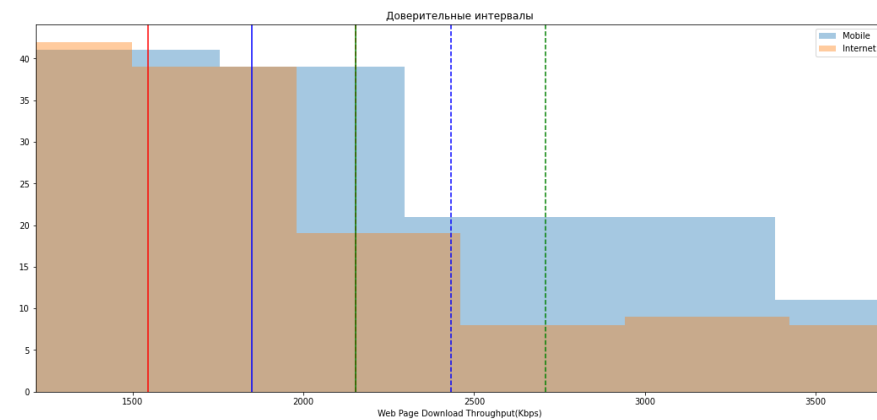
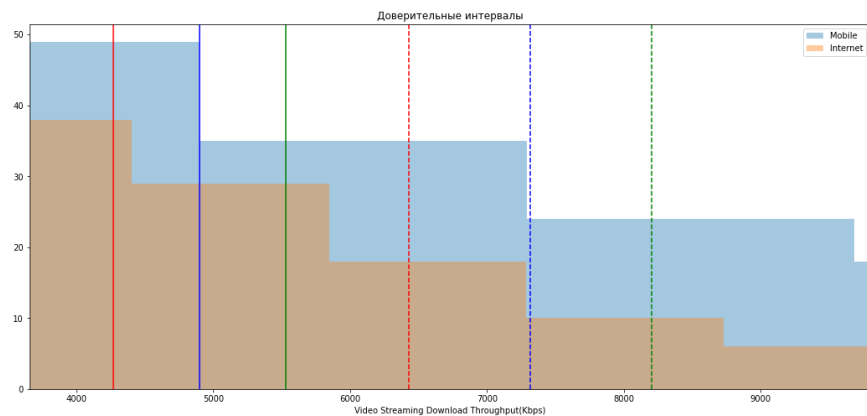


Немного теории

Если данные распределены ненормально и/или неизвестна дисперсия в популяции, выборочное среднее подчиняется t-распределению Стьюдента.

$$x \pm t_{0,05} * \frac{\sigma}{\sqrt{n}}$$

Гипотеза №2

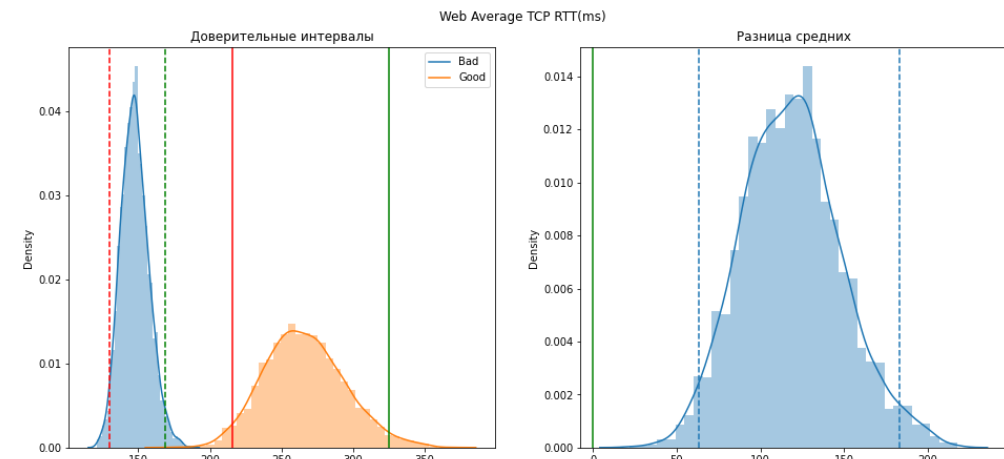
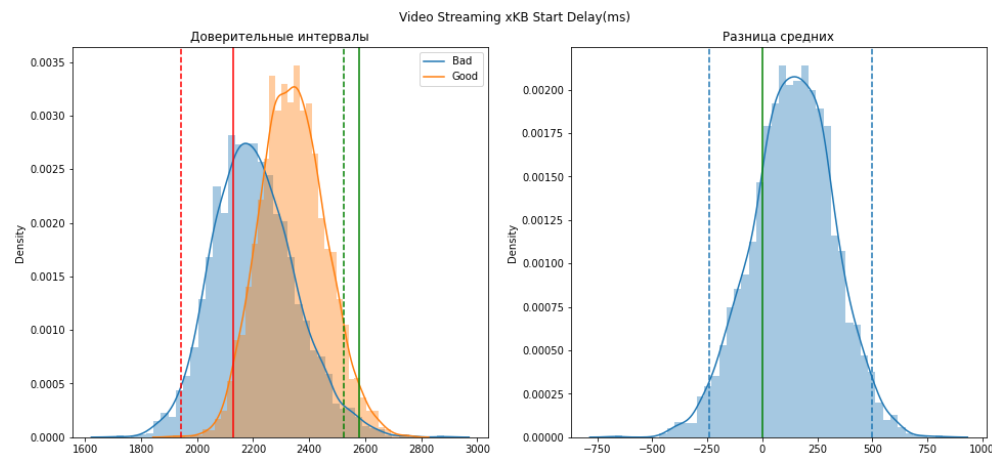
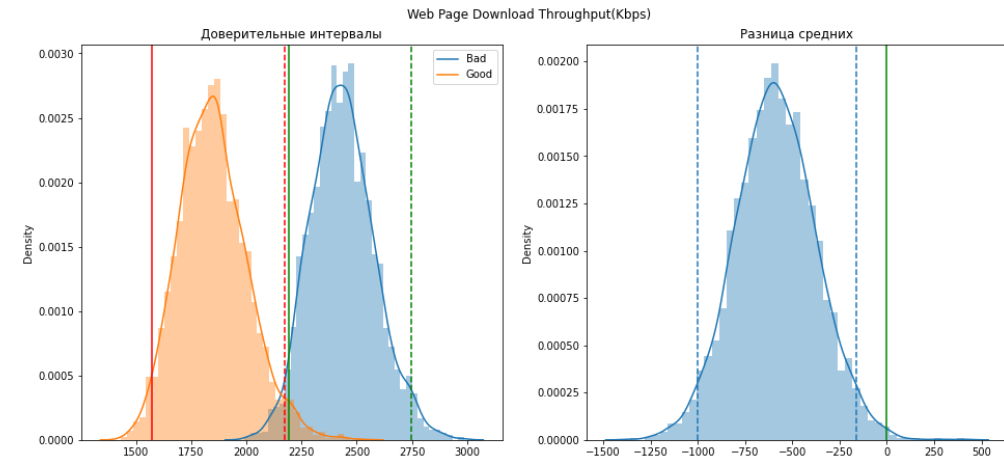
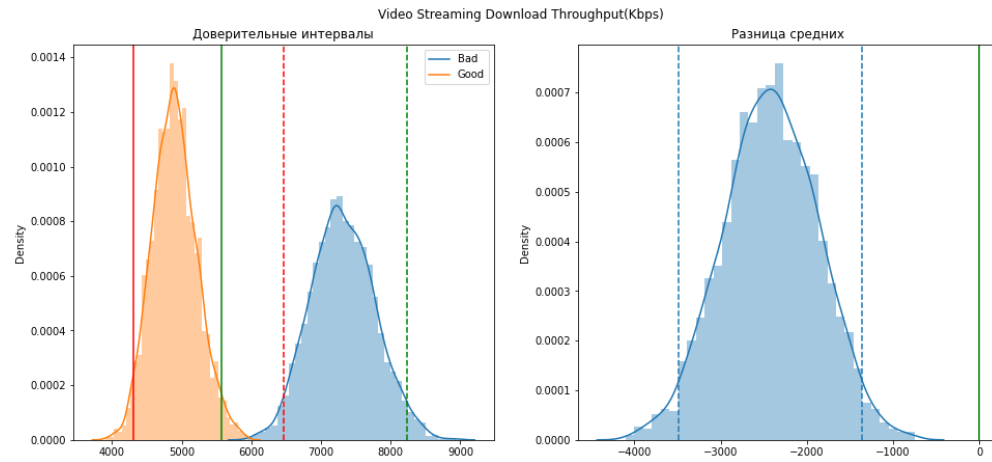




Bootstrap

Гипотеза №2

N = 3000





Выводы

Гипотеза №2



Нулевая гипотеза

Метрики для характеристики качества интернет соединения со значением «Internet» в признаке problem не отличаются по качеству от метрик со значением «Mobile» в признаке problem



Альтернативная гипотеза

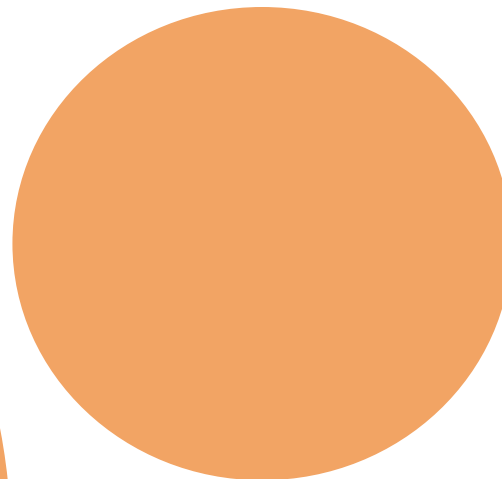
Существует связь между метриками для характеристики качества интернет соединения и признаком problem (сформированным на основе ответов на второй вопрос Q2)

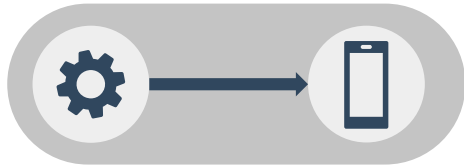


Имеется зависимость технических показателей интернет соединения абонентов и их оценкой. Следует обратить внимание на людей поставивших низкую оценку по причине плохого интернета, возможно стоит улучшить связь конкретных абонентов по соответствующим техническим показателям интернет соединения.

Основное исследование

ГИПОТЕЗА №3





Гипотеза №3



Нулевая гипотеза

Технические показатели для абонентов со значением «Internet and Mobile» в признаке problem не отличаются по качеству от метрик с остальными значениями (Mobile, Internet, Other) в признаке problem



Альтернативная гипотеза

Существует связь между техническими показателями качества и проблемой указанной абонентами как «Internet and Mobile»



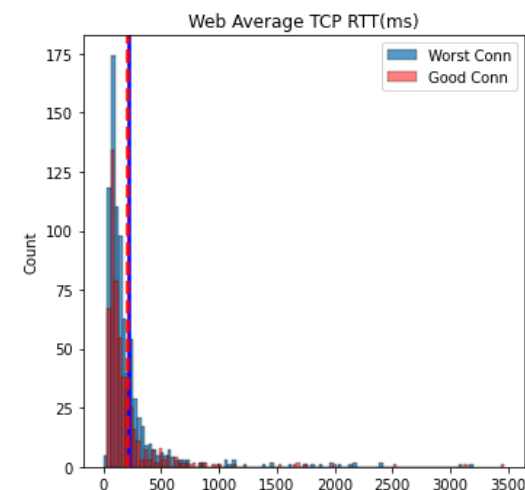
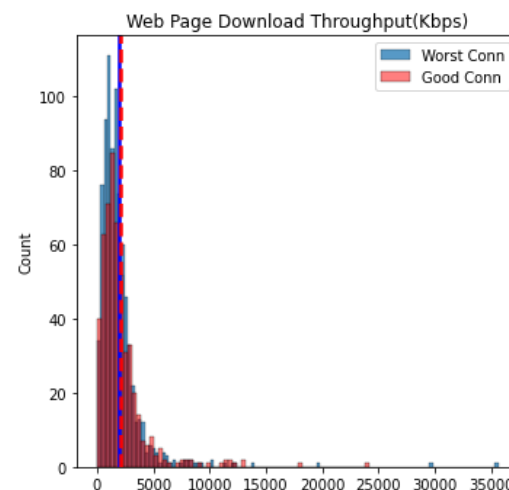
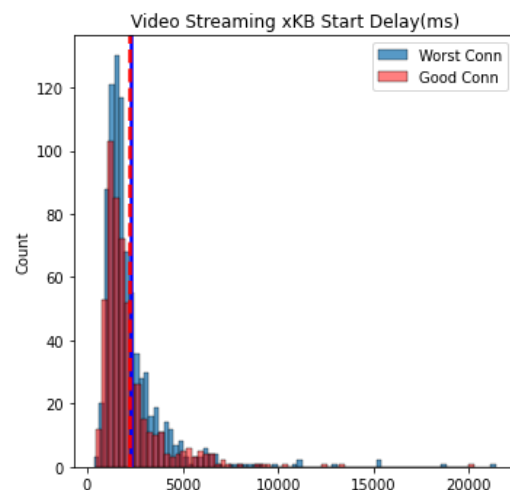
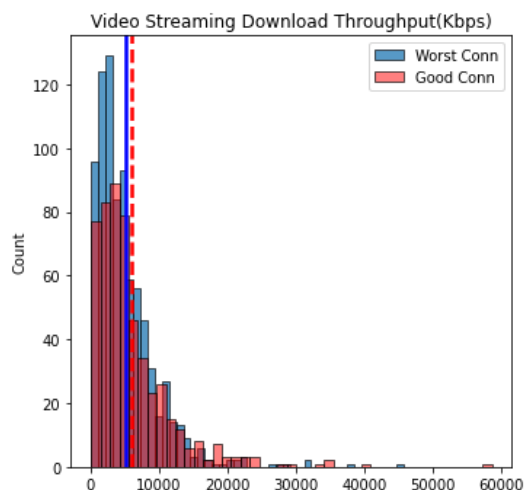
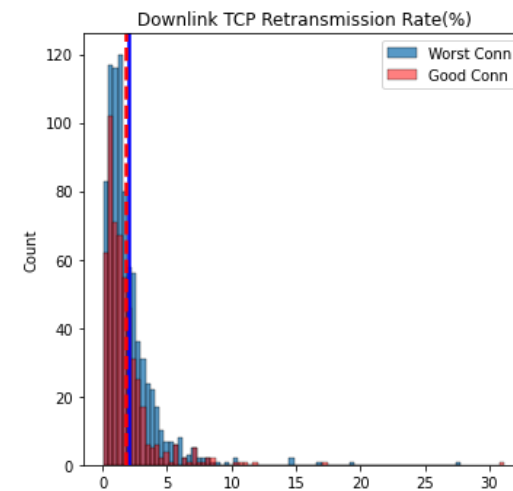
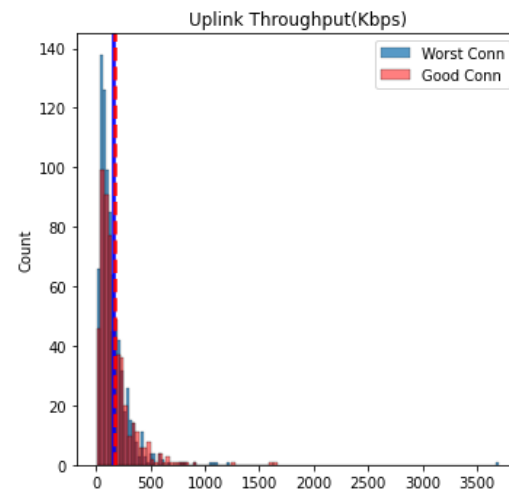
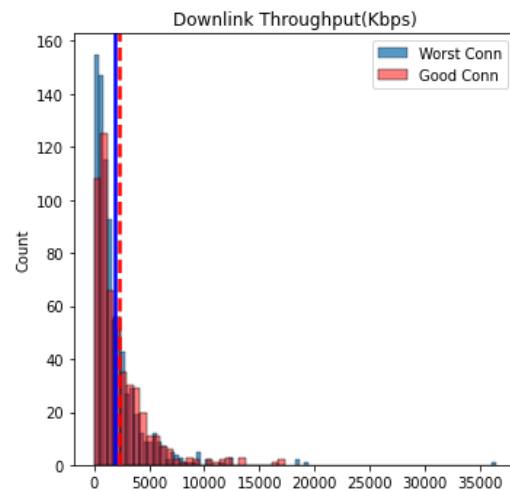
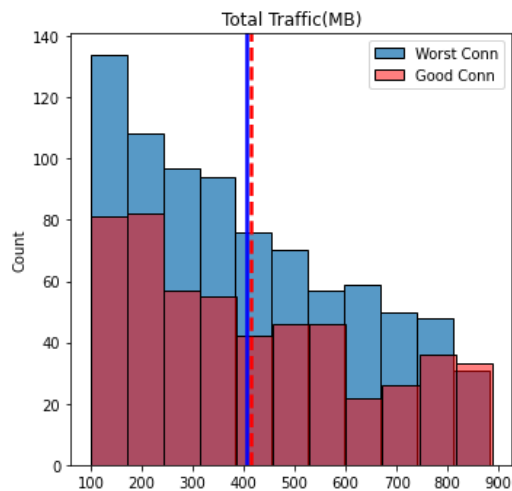
Цель

Определить количество технических показателей необходимое для более точного анализа.



Разведочный анализ

Гипотеза №3

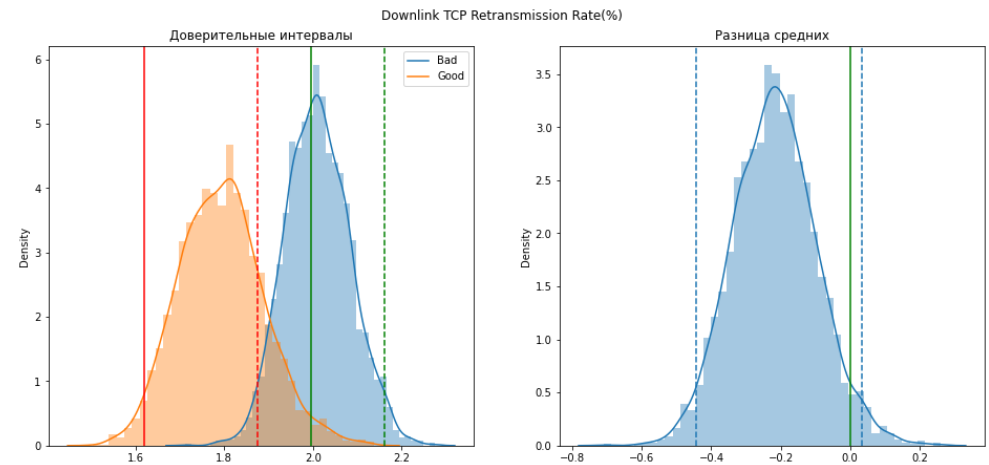
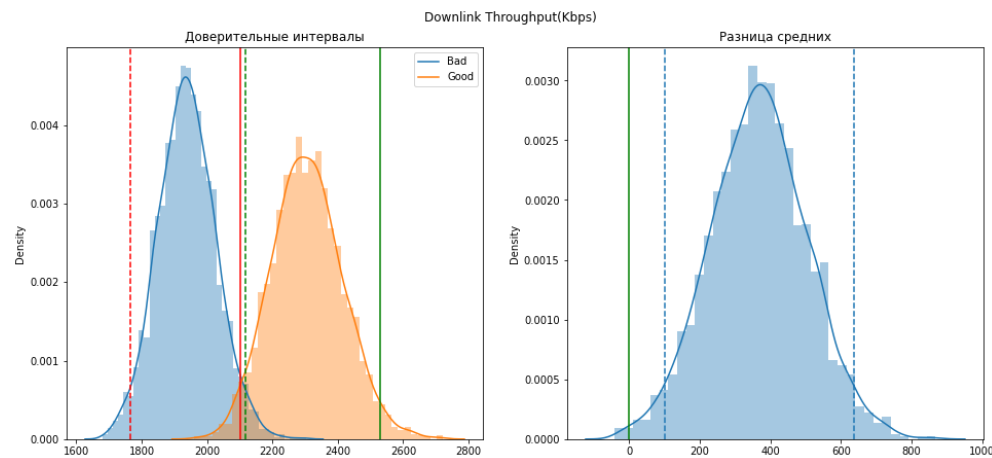
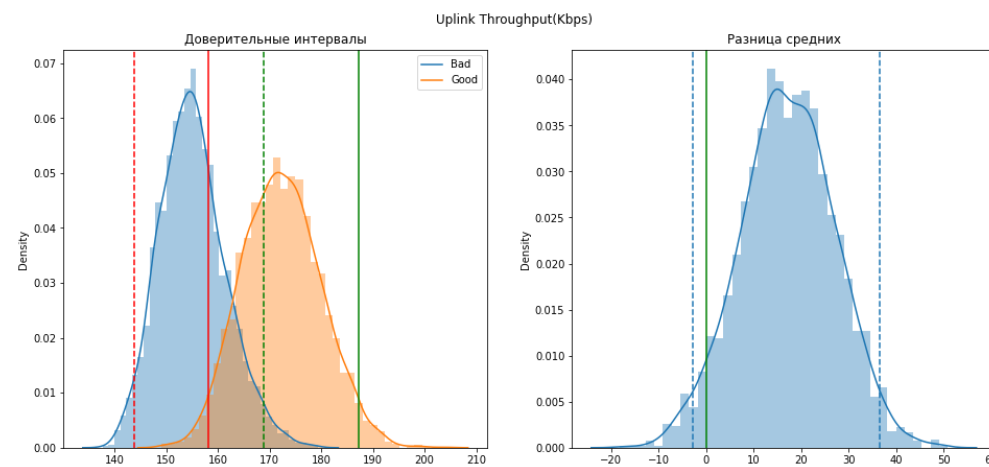
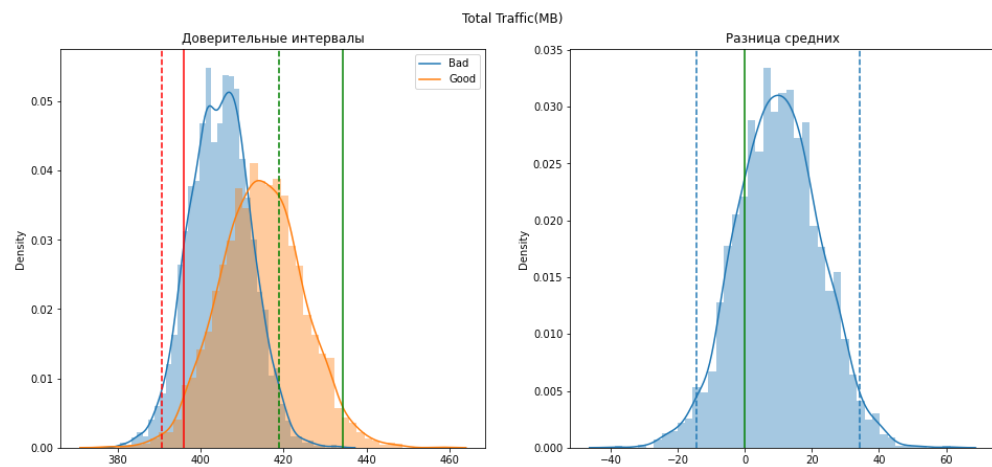




Bootstrap

Гипотеза №3

N = 3000





Выводы

Гипотеза №3



Нулевая гипотеза

Технические показатели для абонентов со значением «Internet and Mobile» в признаке problem не отличаются по качеству от метрик с остальными значениями (Mobile, Internet, Other) в признаке problem



Альтернативная гипотеза

Существует связь между техническими показателями качества и проблемой указанной абонентами как «Internet and Mobile»



Нет оснований полагать, что абоненты с проблемой "Internet and Mobile" имеют более низкие технические показатели, скорее всего данные абоненты более активно используют как мобильные, так и интернет функции оператора.

Необходимо предложить данным абонентам более индивидуальные тарифные планы, основанные на большем времени использования всех функций сети.

Выводы

Гипотеза №1

- Опросы такого рода помогут находить проблемные места в технических характеристиках связи и улучшать их в дальнейшем.

Гипотеза №3

- Группе абонентов с жалобами на «Internet and Mobile» необходимо предложить более индивидуальные тарифные планы, которые основаны на большем времени использования всех функций сети.

Гипотеза №2

- Необходимо провести диагностику интернет соединения у указанных абонентов, так как их качество интернет соединения действительно хуже остальных.

- Технический параметр Video Streaming xKB Start Delay(ms) следует исключить из анализа, так как его показания являются не информативными.

Заключение



Спасибо
за внимание!