МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
ИНСТИТУТ «ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ПРИКЛАДНАЯ МАТЕМАТИКА»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Искусственный интеллект»

ЛАБОРАТОРНАЯ РАБОТА №0

VI семестр

Студент: Калинина А.В.
Группа: М8О-308Б-19
Преподаватель: Самир Ахмед

Подпись: _____
Оценка: _____
Дата сдачи: «___»_____22г.
Дата проверки: «___»_____22г.

Москва, 2022

1. Постановка задачи

Требуется определить задачу и найти под нее необходимые данные. Выбранный датасет проанализировать, визуализировать зависимости, показать распределения данных. Подготовить отчет с результатами лабораторной работы.

2. Описание

Для выполнения лабораторной работы была поставлена задача бинарной классификации оттока клиентов телефонной компании: перестанет абонент пользоваться услугами или нет. Датасет содержит:

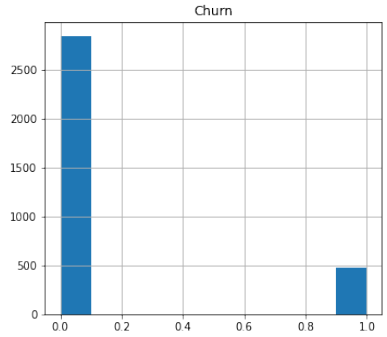
1. Churn - целевая переменная, перестанет клиент пользоваться услугами или нет
2. ID — индекс
3. AccountWeeks - количество недель, при которых у пользователя активный аккаунт
4. ContractRenewal - 1, если клиент продлял недавно договор, иначе — 0
5. DataPlan - 1, если у клиента есть тарифный план, 0, если нет
6. DataUsage - ежемесячное количество гигабайт
7. CustServCalls - количество обращений в службу поддержки
8. DayMins - среднее время в минутах за месяц
9. DayCalls - среднее количество звонков в месяц
10. MonthlyCharge - средний счет за месяц
11. OverageFree - самая большая плата за перерасход за последний год

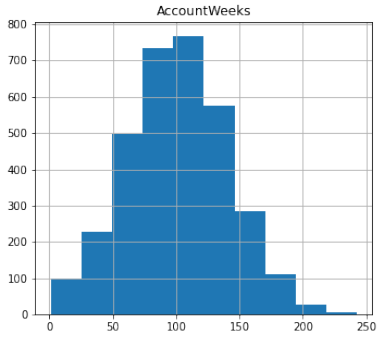
3. Ход работы

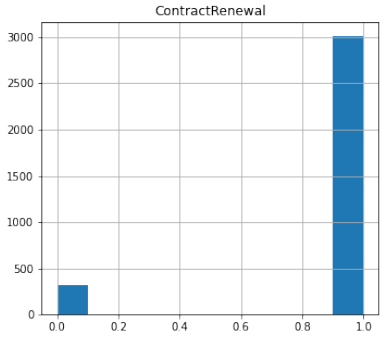
Датасет содержит 11 показателей.

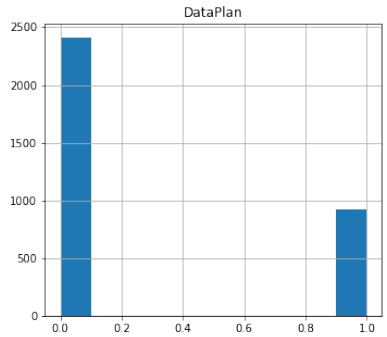
```
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Churn            3333 non-null   int64
 1   AccountWeeks     3333 non-null   int64
 2   ContractRenewal  3333 non-null   int64
 3   DataPlan         3333 non-null   int64
 4   DataUsage        3333 non-null   float64
 5   CustServCalls    3333 non-null   int64
 6   DayMins          3333 non-null   float64
 7   DayCalls         3333 non-null   int64
 8   MonthlyCharge    3333 non-null   float64
 9   OverageFee       3333 non-null   float64
 10  RoamMins         3333 non-null   float64
dtypes: float64(5), int64(6)
```

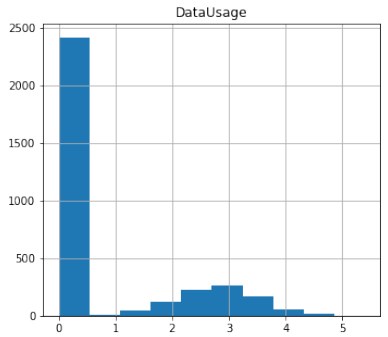
Заметим, что признаки являются числовыми. Датасет не содержит пропущенных значений.

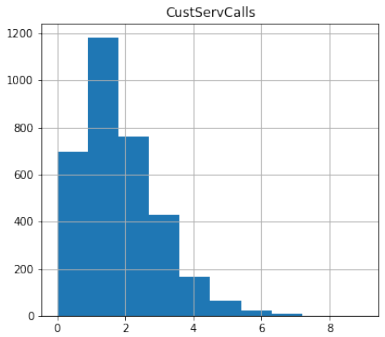


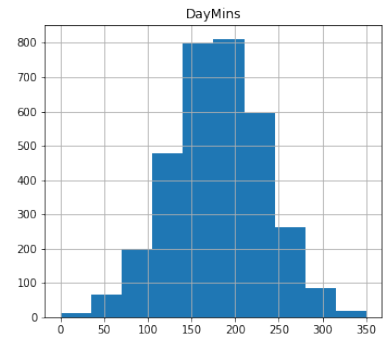


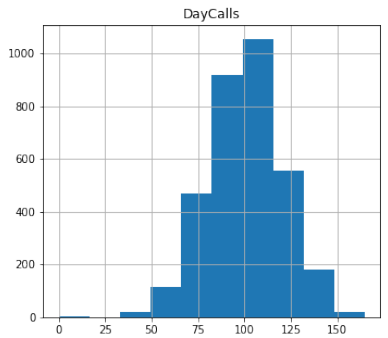


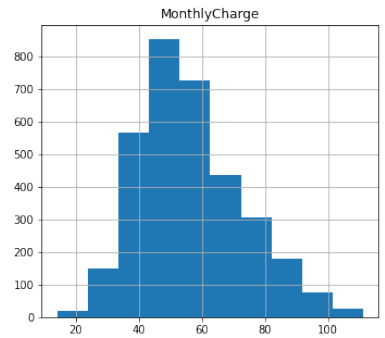


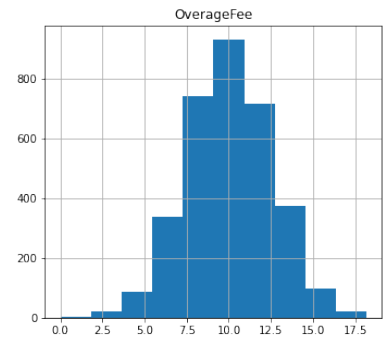


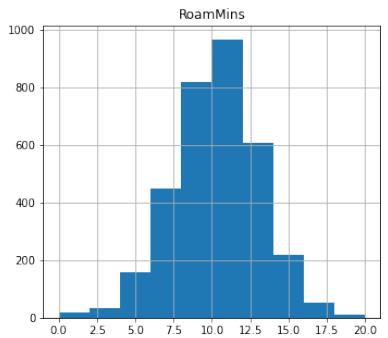












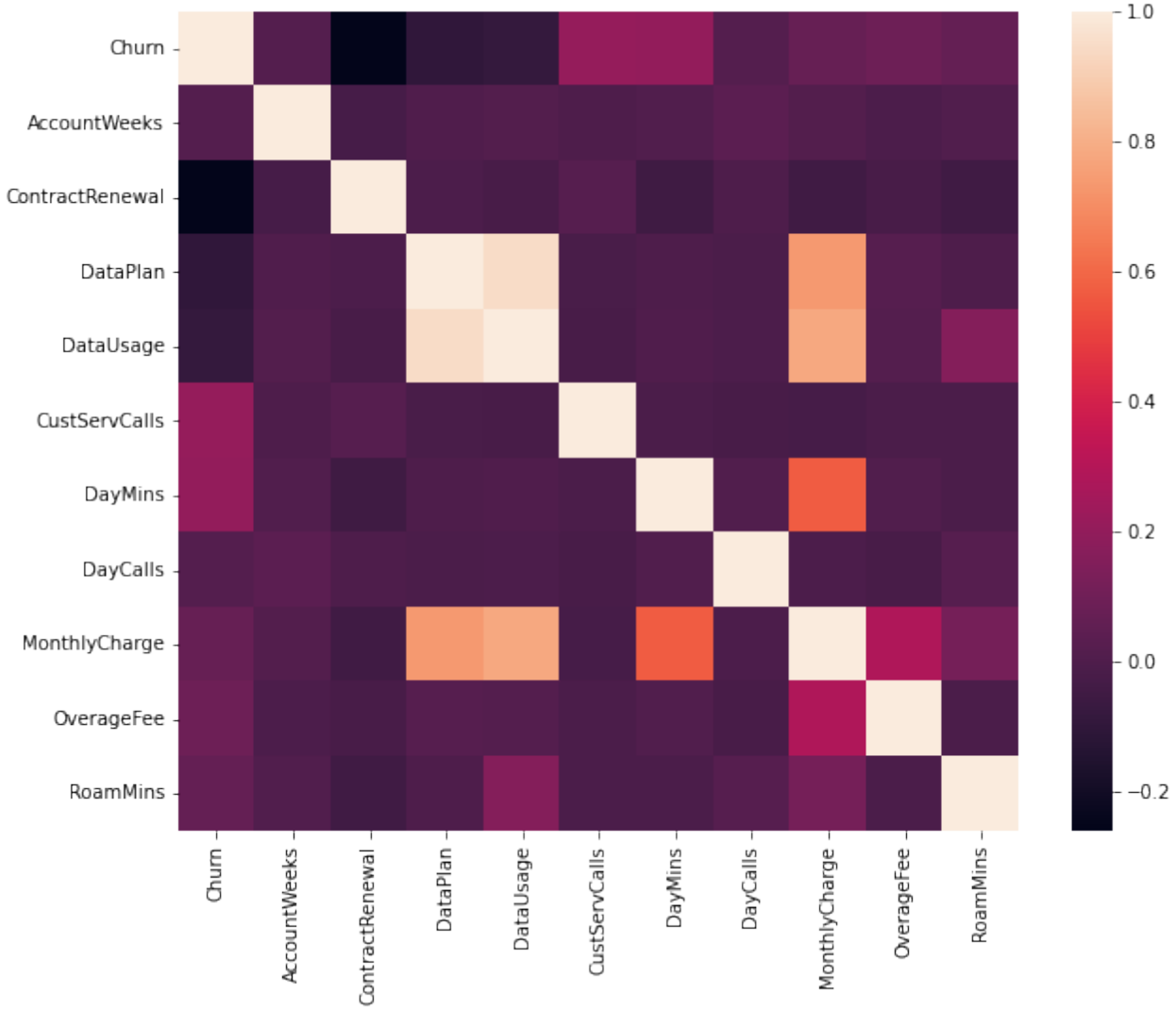
Посмотрим на распределение признаков, визуализировав их в виде гистограммы:

Из анализа гистограмм видно аномальные распределения признаков отсутствуют.

Посчитаем коэффициенты корреляции между целевым признаком, а так же выведем тепловую карту:

```
Churn                1.000000
CustServCalls        0.208750
```

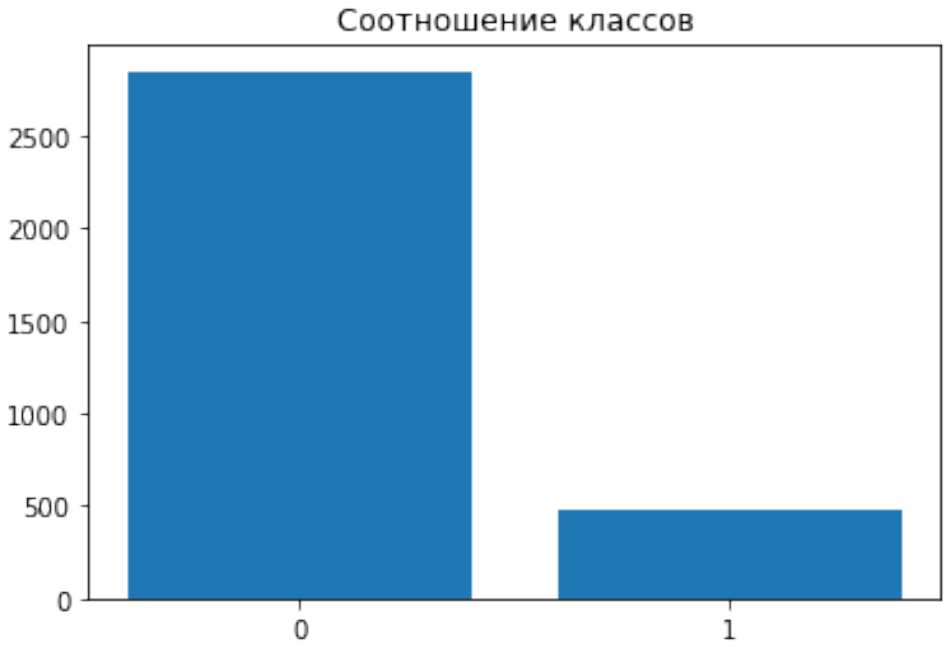
```
DayMins              0.205151
OverageFee           0.092812
MonthlyCharge        0.072313
RoamMins             0.068239
DayCalls             0.018459
AccountWeeks         0.016541
DataUsage           -0.087195
DataPlan            -0.102148
ContractRenewal     -0.259852
```



Сделаем некоторые выводы исходя из анализа матрицы корреляции:

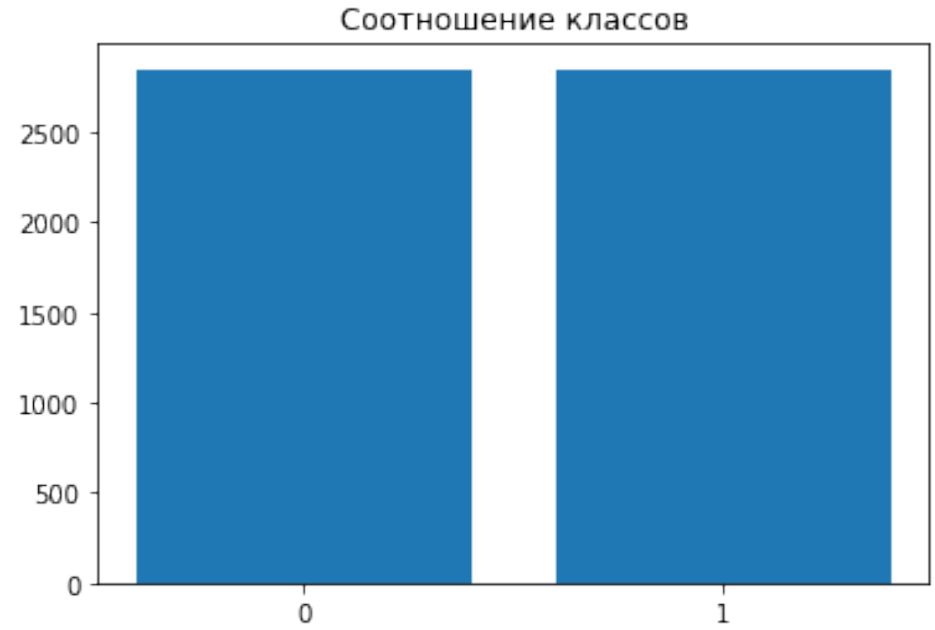
- Наиболее сильная корреляция целевой переменной наблюдается с CustServCall, DayMins и ContractRenewal.
- Меньше целевой признак коррелирует с MonthlyCharge

Проверим соотношение классов



Классы являются несбалансированными. Необходимо изменить соотношение классов для корректного обучения модели в дальнейшем.

Новое соотношение классов:



4. Выводы

В ходе выполнения лабораторной работы был проведен анализ данных, для их дальнейшего использования при обучении линейной модели. Так же произведены визуализация распределения данных и корреляционной матрицы с целью изучения зависимостей между ними. Выяснилось, что целевой признак churn зависит от имеющихся данных, что позволит получить хорошую модель в дальнейшем.