Московский Авиационный Институт

(Национальный Исследовательский Университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

**Лабораторная работа No0 по курсу**

**«Машинное обучение»**

**Data Mining и исследование данных**

|  |  |
| --- | --- |
| Студент | Мохляков П.А. |
| Группа | М80-308Б-19 |
| Дата | 02.05.2022 |
| Оценка |  |
| Подпись |  |

Москва

2022

**Лабораторная работа №0**

**Задача:** определить задачу, которую вы хотите решить и найти под нее соответствующие данные. Проанализировать данные, визуализировать зависимости.

**Описание датасета**

Дан набор данных, который содержит опрос удовлетворенности авиапассажиров. Необходимо предсказать удовлетворенность пассажиров.

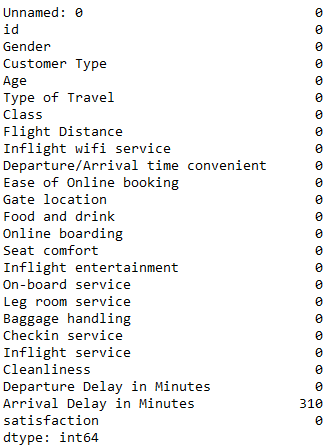
**Ссылка на датасет:** <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?resource=download&select=train.csv>

**Содержание датасета**

* Gender: пол пассажиров (женщина, мужчина)
* Customer Type: тип клиента (постоянный клиент, нелояльный клиент)
* Age: фактический возраст пассажиров
* Type of Travel: цель полета пассажиров (Личная поездка, Деловая поездка)
* Class: класс в самолете пассажиров (Бизнес, Эко, Эко Плюс)
* Flight distance: Расстояние полета этого путешествия
* Inflight wifi service: уровень удовлетворенности услугой Wi-Fi на борту (0: не применимо; 1–5)
* Departure/Arrival time convenient: уровень удовлетворенности удобным временем отправления/прибытия
* Ease of Online booking: уровень удовлетворенности онлайн-бронированием
* Gate location: уровень удовлетворенности расположением ворот
* Food and drink: уровень удовлетворенности едой и напитками
* Online boarding: уровень удовлетворенности онлайн-посадкой
* Seat comfort: уровень удовлетворенности комфортом сидений
* Inflight entertainment: уровень удовлетворенности развлечениями в полете
* On-board service: уровень удовлетворенности обслуживанием на борту
* Leg room service: уровень удовлетворенности обслуживанием в номерах
* Baggage handling: уровень удовлетворенности обработкой багажа
* Check-in service: уровень удовлетворенности сервисом регистрации заезда
* Inflight service: уровень удовлетворенности обслуживанием в полете
* Cleanliness: уровень удовлетворенности чистотой
* Departure Delay in Minutes: минут задержки при отправлении
* Arrival Delay in Minutes: минут задержки при прибытии
* **Satisfaction**: уровень удовлетворенности авиакомпанией (удовлетворенность, нейтральность или неудовлетворенность)

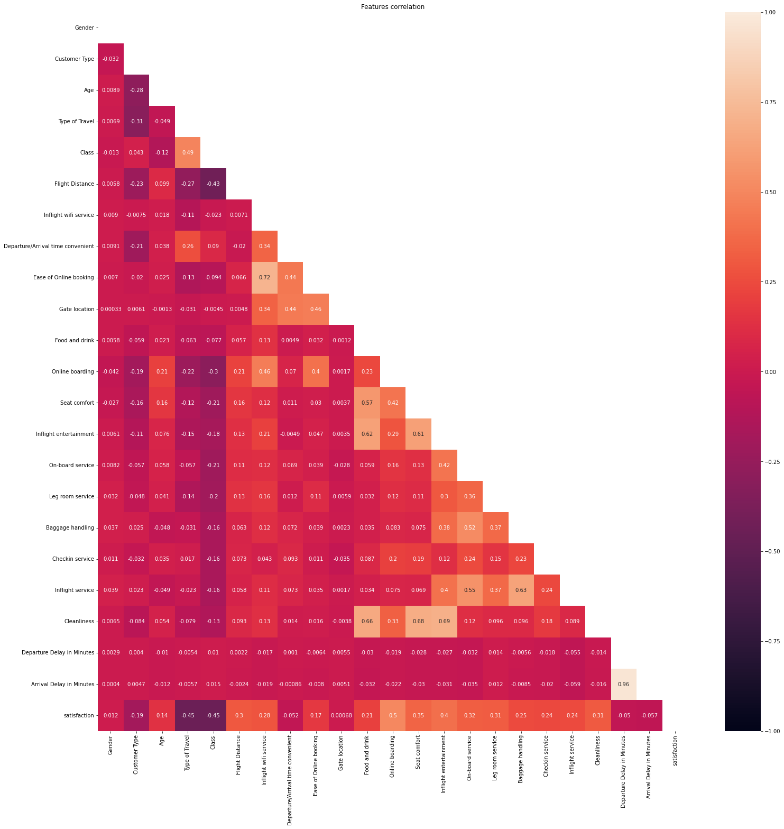
**Анализ датасета на пропуски и корреляции**

Проверим датасет на содержание пропусков.



Во-первых, уберем лишние столбцы с порядковым номером и id, так как они не несут для нас никакой полезной информации.

Во-вторых, мы видим, что есть пропуски в задержках при прибытии.  
Посмотрим корреляции.



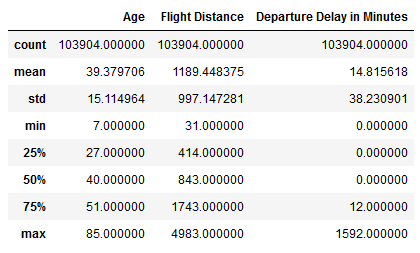
Как мы видим, задержки в прибытии и задержки в отправлении очень сильно коррелируют между собой, что может вызвать проблемы в дальнейшем. Также учитывая тот, факт, что у нас есть пропуски в задержках при прибытии можно удалить данный столбец.

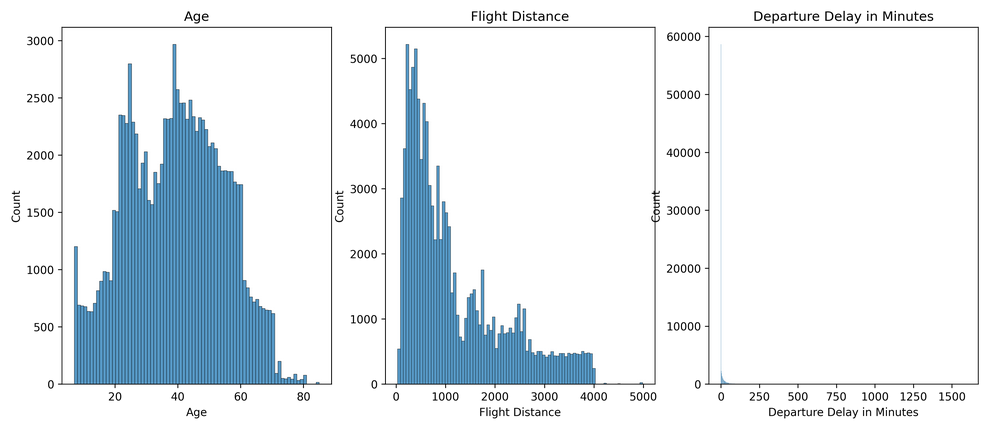
Также по корреляциям можно отметить:

* Корреляция удовлетворенности между чистотой и удовлетворенности едой и напитками, комфортом сидений и развлечениями в полете.
* Корреляция удовлетворенности онлайн-посадкой и удовлетворённостью услугой Wi-Fi на борту
* Обратную корреляцию между уровнем удовлетворенностью авиакомпанией, классом и типом поездки
* Также обратная корреляция между расстояние полета и классом в самолете

**Распределение фичей**

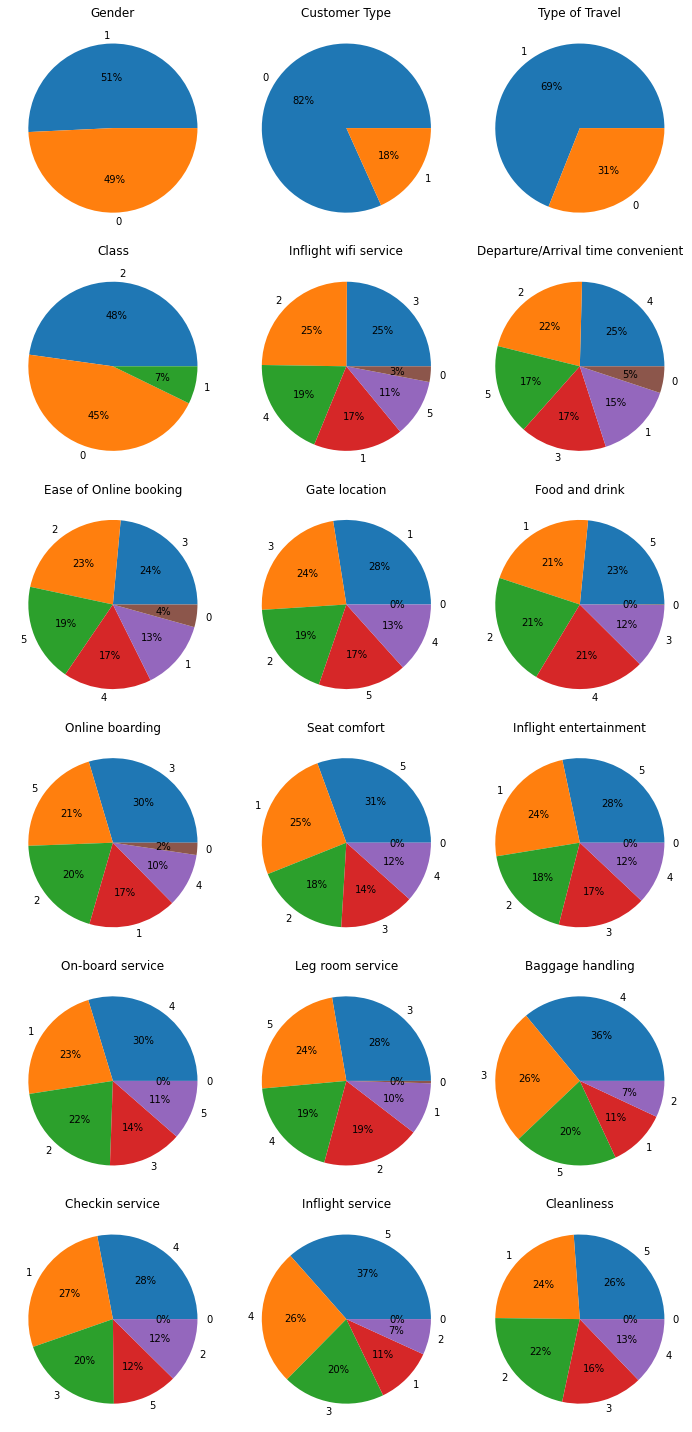
Распределение численных фичей:





Распределение возраста похоже на нормальное, с просадкой в возрасте 30 лет. Расстояние перелета в основном меньше 1500 км, и в основном перелеты без задержек, что логично.

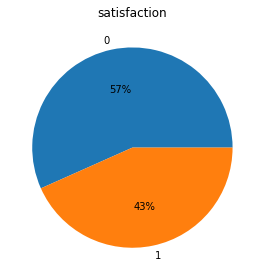
Распределение категориальных фичей:



Все оценки распределены плюс минус равномерно.

**Таргет**

Распределение таргета



Классы выглядят сбалансированными, но при опросе нейтральный уровень удовлетворённости и неудовлетворенность были представлены одним вариантом ответа, что при необходимости нужно учитывать в дальнейшем.

**Вывод**

В данной лабораторной работе я провел исследование датасета. Данный датасет был посвящен удовлетворенностью от авиаперелетов. Я рассмотрел признаки и убрал бесполезные или те, что могли помешать в дальнейшем. Также я убедился, что признаки зависят друг от друга и от них зависит целевой признак. Таким образом мы можем получить работающую модель.