Московский Авиационный Институт

(Национальный Исследовательский Университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

**Лабораторная работа №0 по курсу**

**«Машинное обучение»**

**Сбор и анализ данных**

Студент: Гаптулхаков Руслан Рамилевич

Группа: М80 - 308Б -19

Дата: 09.05.2022

Оценка: \_\_\_\_\_\_\_\_\_\_\_

Подпись: \_\_\_\_\_\_\_\_\_\_\_

Москва, 2022

1. **Постановка задачи**

Собрать или найти готовые данные и провести исследовательский анализ. Написать отчёт по результатам исследования.

1. **Описание датасета**

Представим, что спустя несколько сотен лет мы сможем премещаться в паралельные измерения. Нам нужно предсказать смог ли успешно пассажир переместиться в другое измерение во время столкновения космического корабля с пространственно-временной аномалией. Чтобы помочь вам сделать эти прогнозы, вам дается набор личных записей, извлеченных из поврежденной компьютерной системы корабля. Этот датасет по хож на классический обучающий датасет "Титаник", но в новой тематике.

train.csv — личные записи примерно двух третей (~8700) пассажиров, которые будут использоваться в качестве обучающих данных.

* PassengerId — уникальный идентификатор для каждого пассажира. Каждый идентификатор принимает форму gggg\_pp, где gggg указывает группу, с которой путешествует пассажир, а pp — его номер в группе. Люди в группе часто являются членами семьи, но не всегда.
* HomePlanet — планета, с которой вылетел пассажир, обычно планета его постоянного проживания.
* CryoSleep - указывает, решил ли пассажир быть переведен в режим анабиоза на время рейса. Пассажиры, находящиеся в криосонном состоянии, находятся в своих каютах.
* Cabin — номер каюты, в которой находится пассажир. Принимает форму палуба/число/сторона, где сторона может быть либо P для левого борта, либо S для правого борта.
* Destination — планета, на которую будет высаживаться пассажир.
* Age - возраст пассажира.
* VIP - оплатил ли пассажир специальное VIP-обслуживание во время рейса.
* RoomService, FoodCourt, ShoppingMall, Spa, VRDeck — сумма, которую пассажир выставил в счет за каждое из многочисленных роскошных удобств космического корабля «Титаник».
* Name - имя и фамилия пассажира.
* Transported — был ли пассажир перенесен в другое измерение. Это цель, столбец, который вы пытаетесь предсказать.

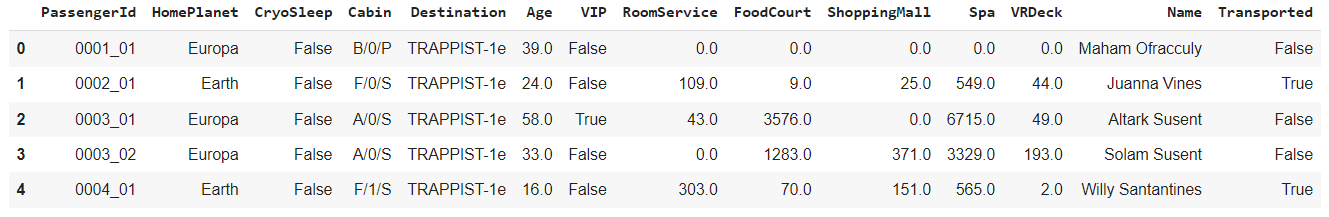
test.csv — личные записи оставшейся трети (~ 4300) пассажиров, которые будут использоваться в качестве тестовых данных. Ваша задача состоит в том, чтобы предсказать значение Перевезено для пассажиров в этом наборе.

Ссылка на датасет: <https://www.kaggle.com/competitions/spaceship-titanic/overview>

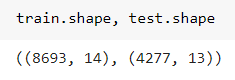
1. **Описание признаков**

Первое, что стоит сделать нужно посмотреть на колличество данных и форму их представления.

*Таблица с данными:*



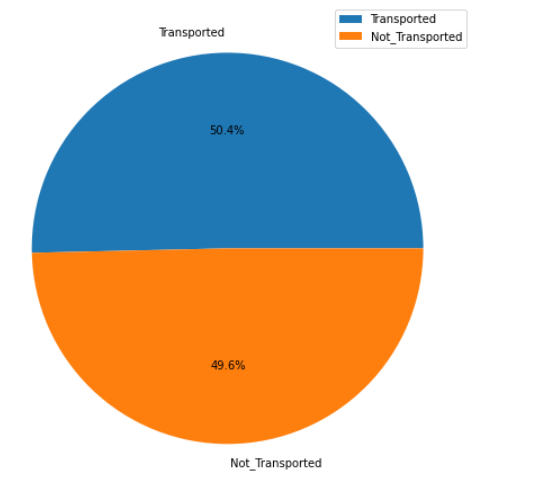
*Колличество данных:*



Данных достаточно много.

Провери есть ли незаполненные поля в таблице и заполнил их ближайшими значениями.

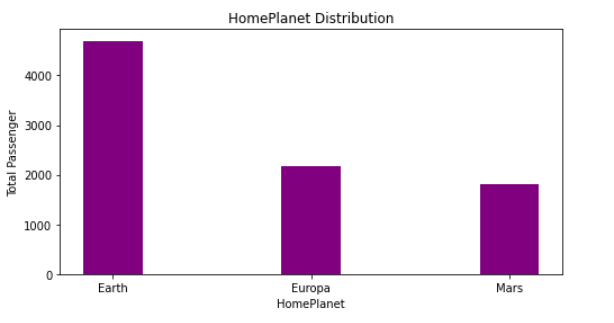
Разделим наши данные по целевому признаку и посмотрим на их сбалансированность. *Распределение целевой пременной:*



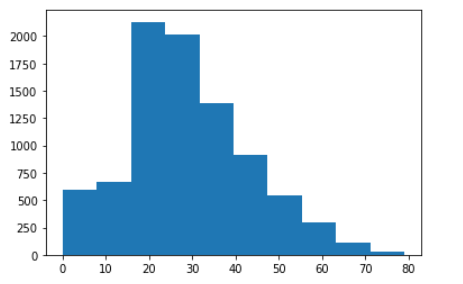
Из pie-диаграмммы видно, целевые признаки сбалансированные. Этот факт нам наруку.

Посмотрим откуда наши пассажиры и в каком количестве.

*На каких планетах живут наши пассажиры:*

Больше всего людей живут на Земле.

*Распределение по возрасту:*

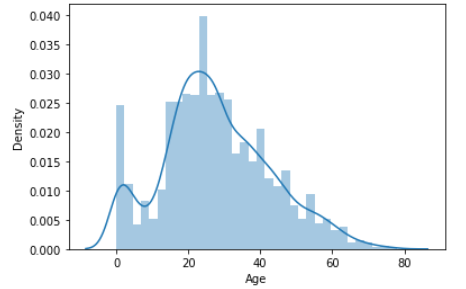


В среднем пассажиры корабля имели 20-30 лет.

*Минимальный, средний, наибольший возраста пассажиров:*



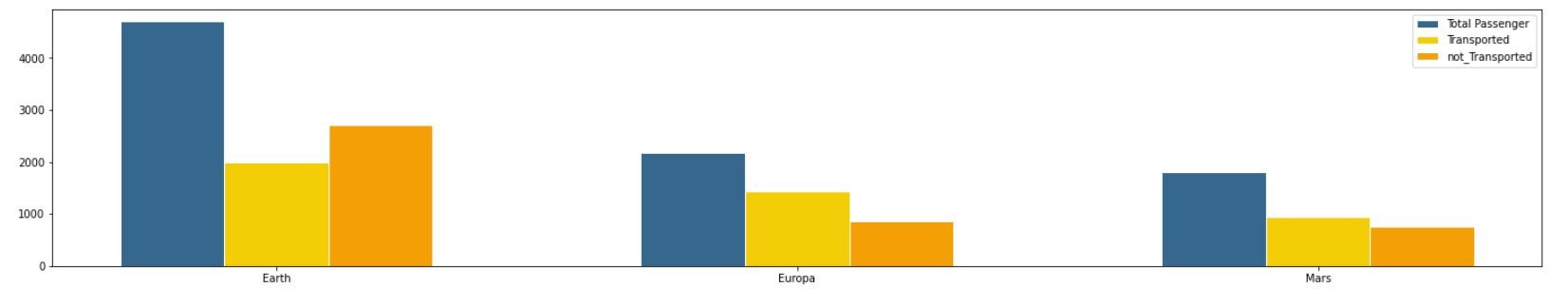
*Распределение по возрасту пассажиров успешно добравшихся:*



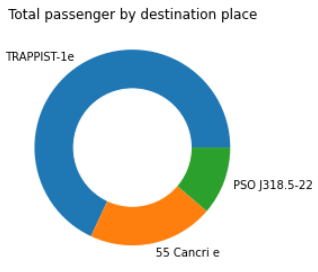
*Минимальный, средний, наибольший возраста пассажиров успешно добравшихся:*



Тут мы видим пропорциональное соотношение по планетам отправки между группами целевого признака. То есть сколько всего отправились с данной планеты, сколько успешно перебралось, сколько не перебралось.

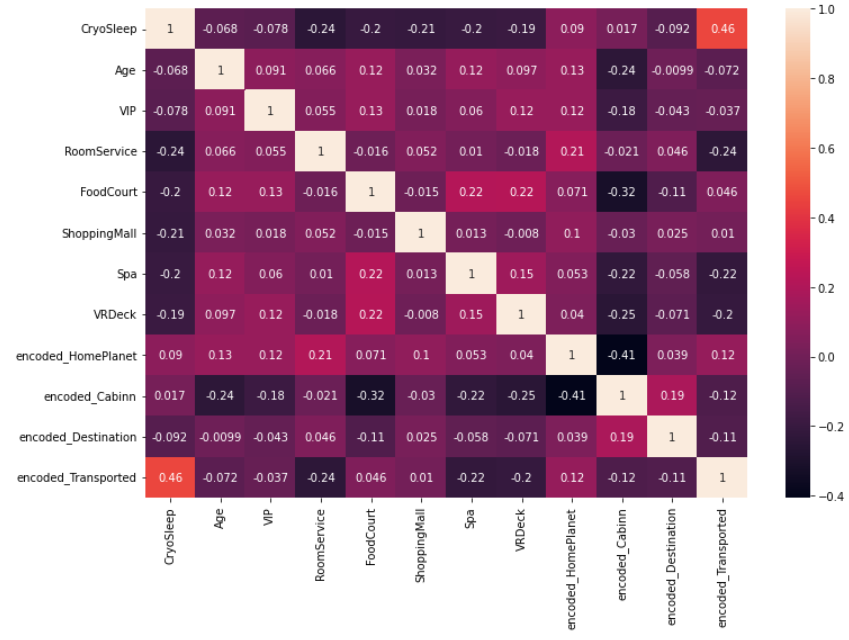


*Посмотрим на распределение пассажиров по конечным пунктам назначения:*

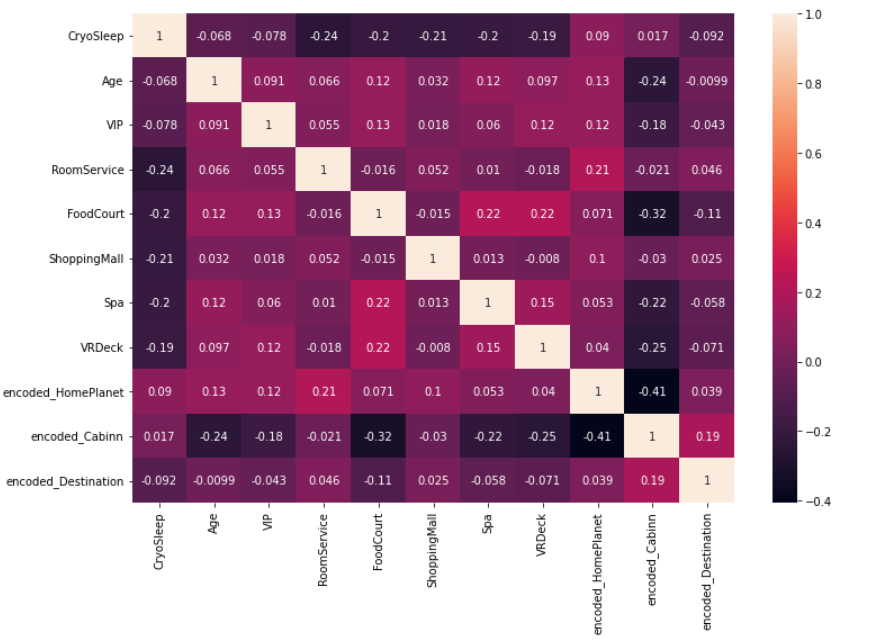


Разобьём номер каюты на составляющие, чтобы больше извлечь оттуда информации. Затем переведём категориальные признаки.

*Посмотрим на корреляционную матрицу с целевой переменной:*



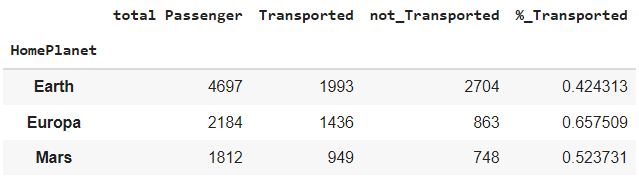
*Посмотрим на корреляционную матрицу без целевой переменной:*



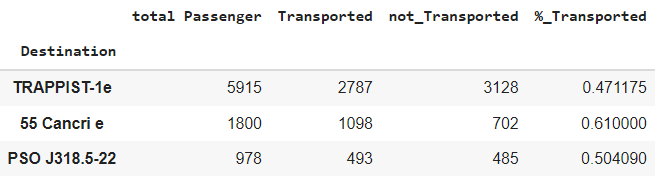
Выкинеем столбец *encoded\_Cabin*, так как он сильно коррелирует с encoded\_HomePlanet.

1. **Основные выводы**

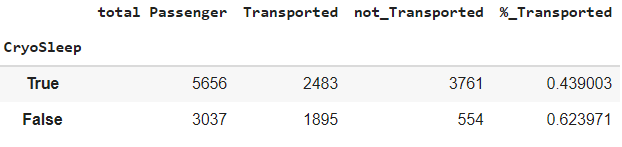
* Средний возраст "выживших" чуть меньше, чем средний возраст всех пассажиров.
* Самое большое число пассажиров отправились с Earth. С самым большим процентом успешно перебравшихся оказались пассажиры, отправлявшиеся с Europa.



* Самое большое число пассажиров добирались до *TRAPPIST-1e*. С самым большим процентом успешно перебравшихся оказались пассажиры, добиравшиеся до *55 Cancri e.*



* У пасажиров находящихся в криосонном состоянии было меньше шансов на успех, чем у бодрствующих.



1. **Вывод**

В данной лабораторной работе я исследовал выбранный датасет. В процессе работы были интуитивно заполнены недостающие данные, проанализированны распределения признаков с визуализацией, извлечены дополнительные признаки из уже имеющихся, переведены все признаки в числа. Посморели зависимсть признаков от целевой переменной и зависимость признаков между собой. Мы удалили один из признаков, так как он достаточно сильно зависил от другого. А также убедились в том, что наши признаки зависят от целевой переменной - это повышает шансы получить хорошую классификационную модель.