МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ им. Р.Е.АЛЕКСЕЕВА

Институт радиоэлектроники и информационных технологий

Кафедра «Вычислительные системы и технологии»

Методы машинного обучения

Отчет

по лабораторной работе №1

Предварительная обработка данных

Вариант 2

ПРОВЕРИЛ:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Новиков И.А.

СТУДЕНТ:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Мусатов М. В.

23-ИВТ-5

Нижний Новгород

2025 г.

**Цель:** получение навыков работы с библиотеками анализа данных, предварительной обработки данных, средствами визуализации.

Описание данных

**YEAR, MONTH, DAY** – Дата рейса

**DAY\_OF\_WEEK** – День недели

**AIRLINE** – Код авиакомпании

**FLIGHT\_NUMBER** – Номер рейса

**TAIL\_NUMBER** – Бортовой номер самолёта

**ORIGIN\_AIRPORT, DESTINATION\_AIRPORT** – Аэропорты вылета и назначения

**SCHEDULED\_DEPARTURE, DEPARTURE\_TIME** – Плановое и фактическое время вылета

**DEPARTURE\_DELAY** – Задержка вылета

**TAXI\_OUT, WHEELS\_OFF** – Время руления и момент отрыва от ВПП

**SCHEDULED\_TIME, ELAPSED\_TIME, AIR\_TIME** – Плановое, общее и чистое время полёта

**DISTANCE** – Расстояние между аэропортами

**WHEELS\_ON, TAXI\_IN** – Момент посадки и время руления до гейта

**SCHEDULED\_ARRIVAL, ARRIVAL\_TIME, ARRIVAL\_DELAY** – Плановое, фактическое время прилёта и задержка

**DIVERTED, CANCELLED** – Признак перенаправления/отмены рейса (0/1)

**CANCELLATION\_REASON** – Причина отмены (A-D: авиакомпания, погода, система, безопасность)

**Типы задержек (DELAYS):**

**AIR\_SYSTEM\_DELAY** – Задержка из-за системы

**SECURITY\_DELAY** – Задержка из-за безопасности

**AIRLINE\_DELAY** – Задержка по вине авиакомпании

**LATE\_AIRCRAFT\_DELAY** – Задержка из-за позднего прибытия самолёта

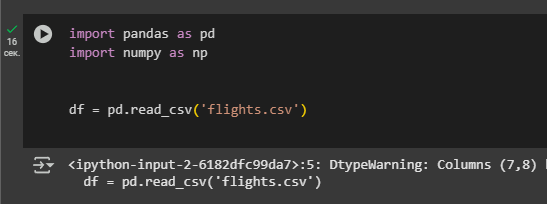
**WEATHER\_DELAY** – Задержка из-за погоды

Работа была выполнена в google colab

**Общие задания:**

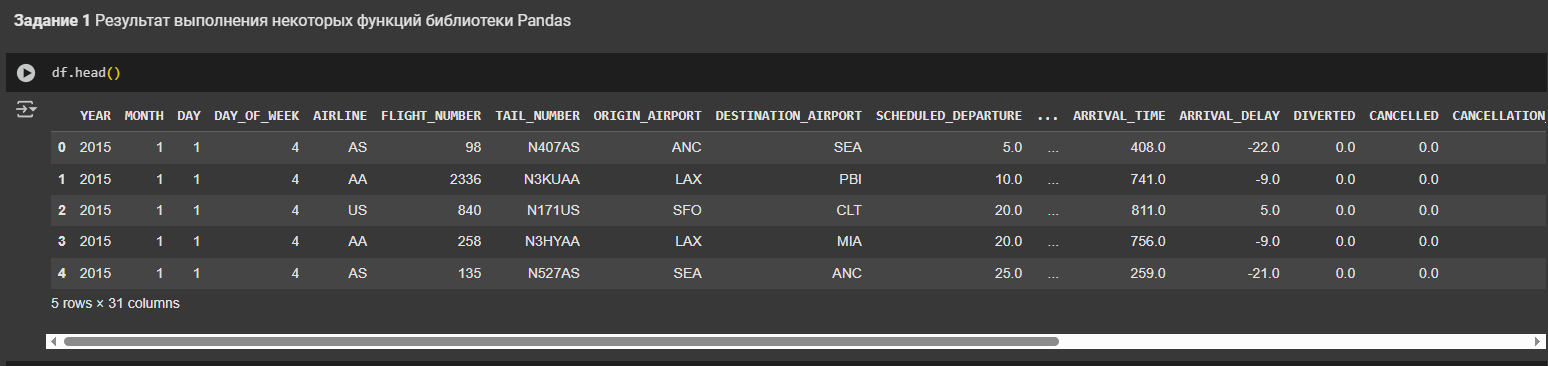
1. **Результат выполнения функций Pandas**

Сначала импортирую библиотеки и загружу файл

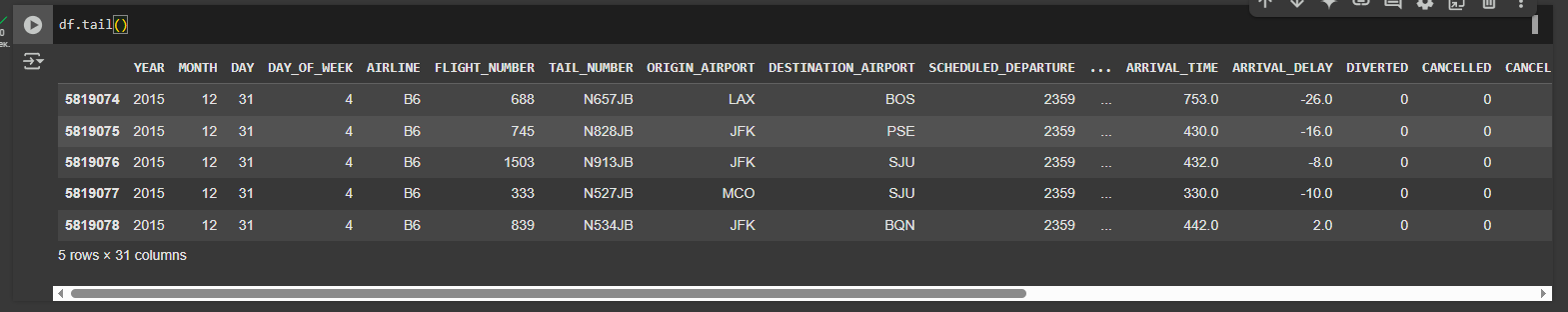


Теперь пробую функции

**Head -** выводит первые 5 строк DataFrame (таблицы данных) для быстрого просмотра содержимого:



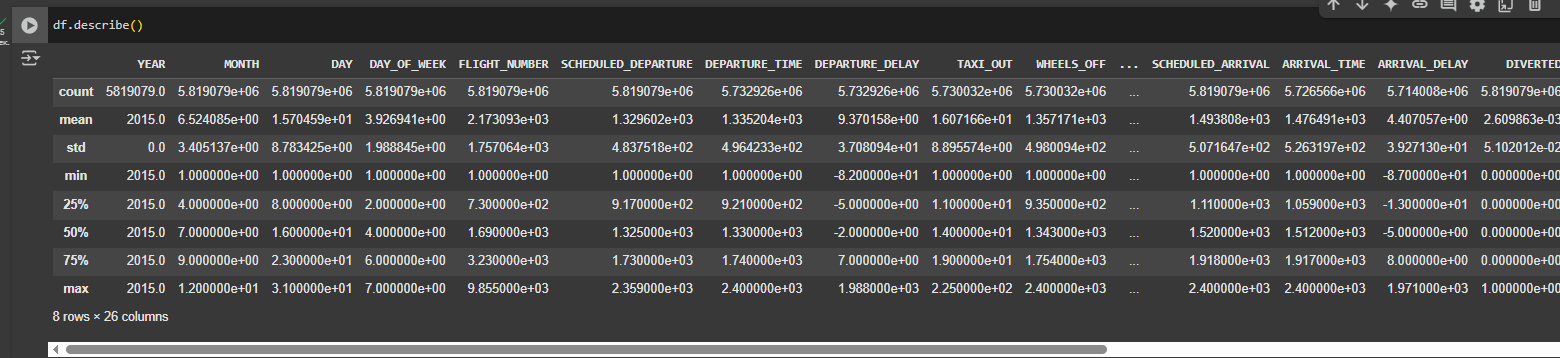
**Tail -** выводит последние 5 строк:



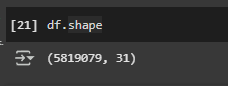
**Info -** предоставляет краткую сводку о DataFrame, включая структуру данных, типы столбцов и использование памяти. Это один из самых полезных методов для первичного анализа данных:



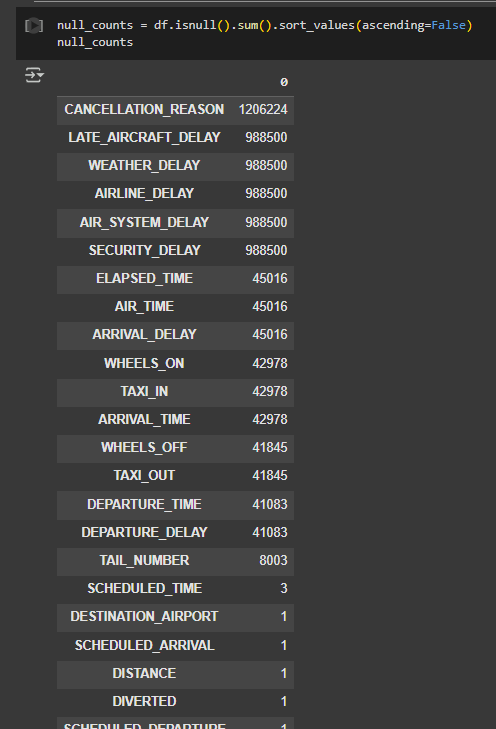
**Describe** - генерирует статистическую сводку для числовых столбцов DataFrame:



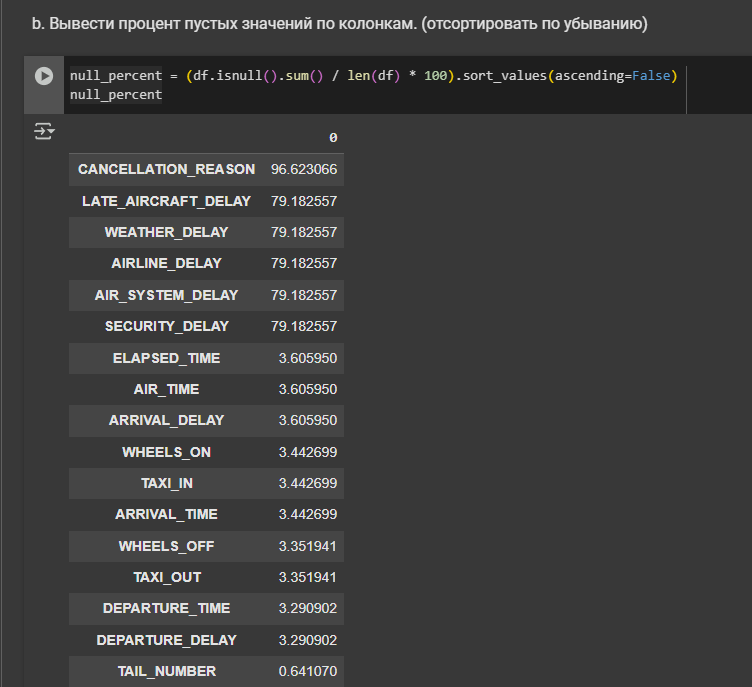
**Shape** - возвращает кортеж с размерностью DataFrame в формате (количество строк, количество столбцов):



1. **Пустые значения.**
   1. Вывести количество пустых значений по колонкам (отсортировать по убыванию)



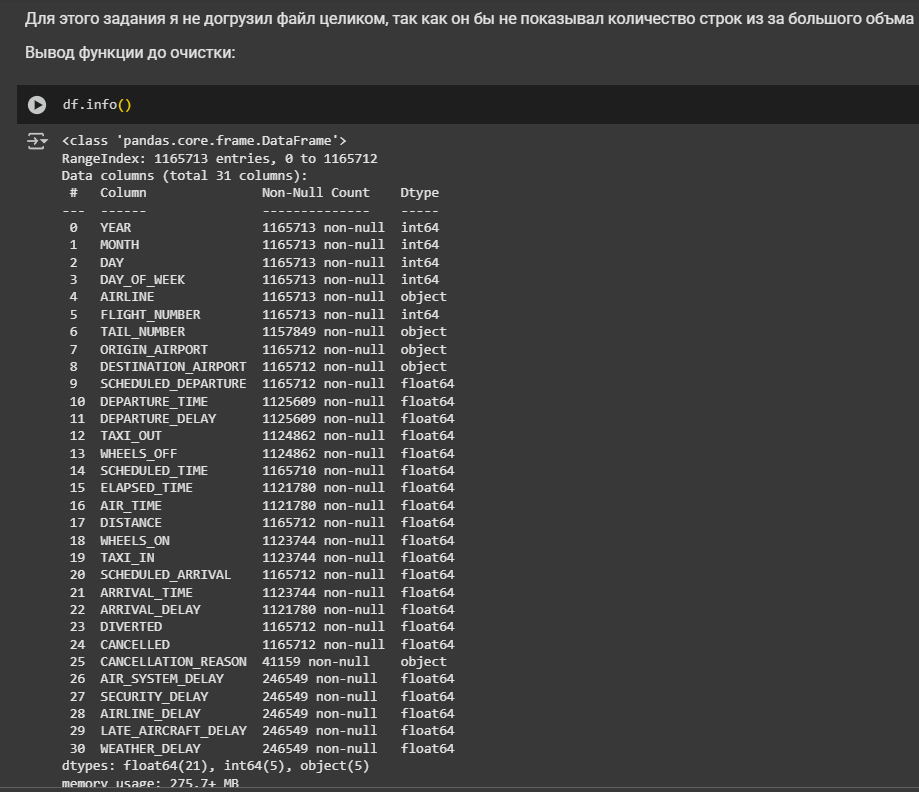
* 1. Вывести процент пустых значений по колонкам. (отсортировать по убыванию)



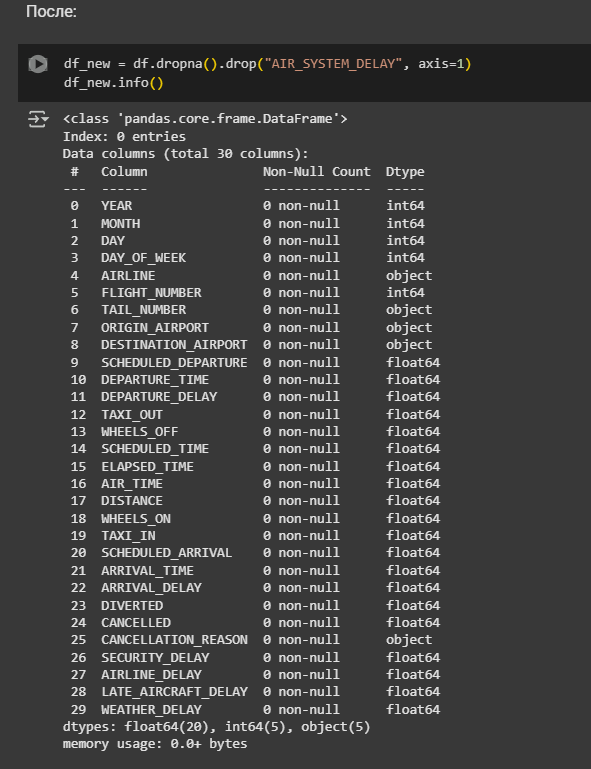
* 1. Показать выводы функции info до и после очистки.

Для этого задания я не догрузил файл целиком, так как он бы не показывал количество строк из за большого объма

Вывод функции до очистки:

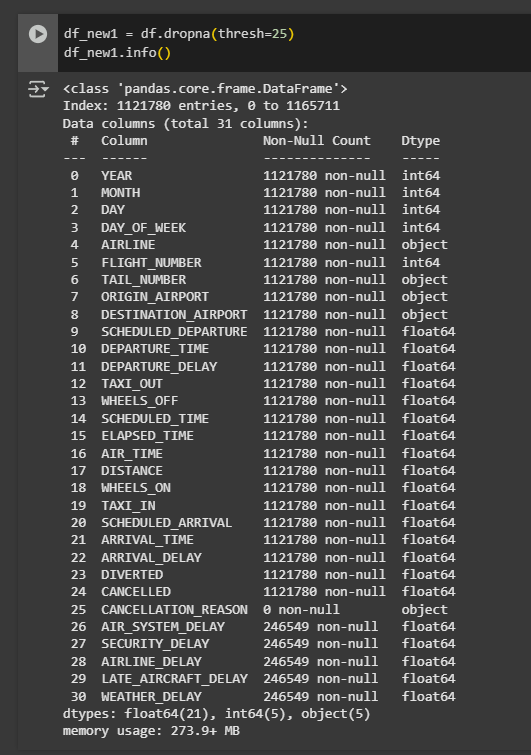


После:

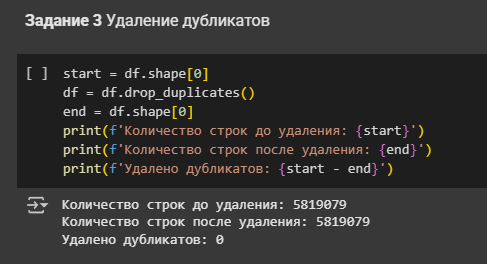


Удалились все строки так как последние 6 столбцов всегда содержат NaN

Поэтому добавляю параметр thresh=25, чтобы удалялись строчки, где NaN больше 7 раз

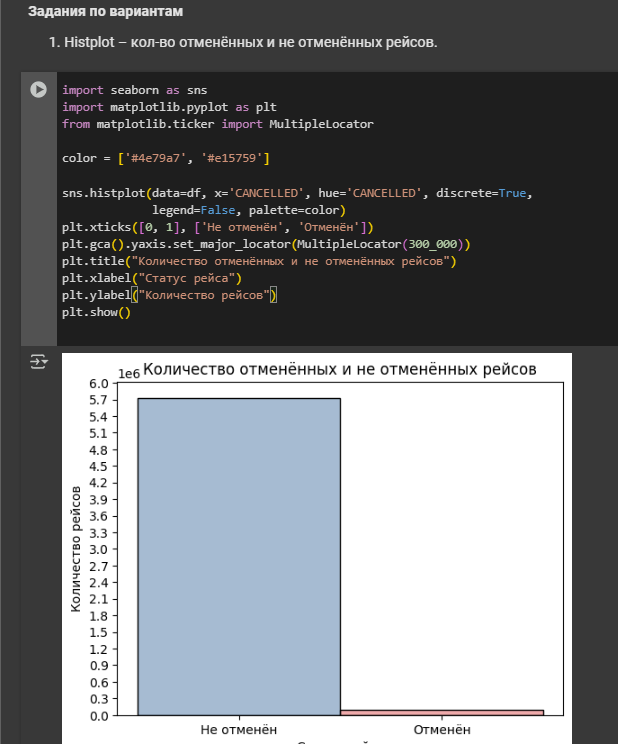


1. **Удаление дубликатов. (drop\_duplicates) Вывести количество строк до удаления и после. Так же вывести количество найденных дубликатов**

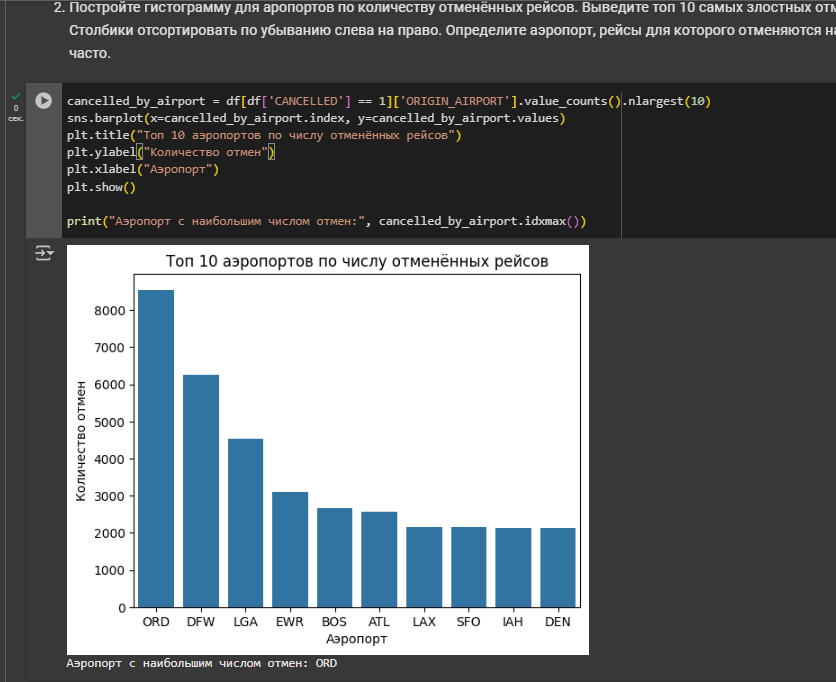


**Задания по вариантам:**

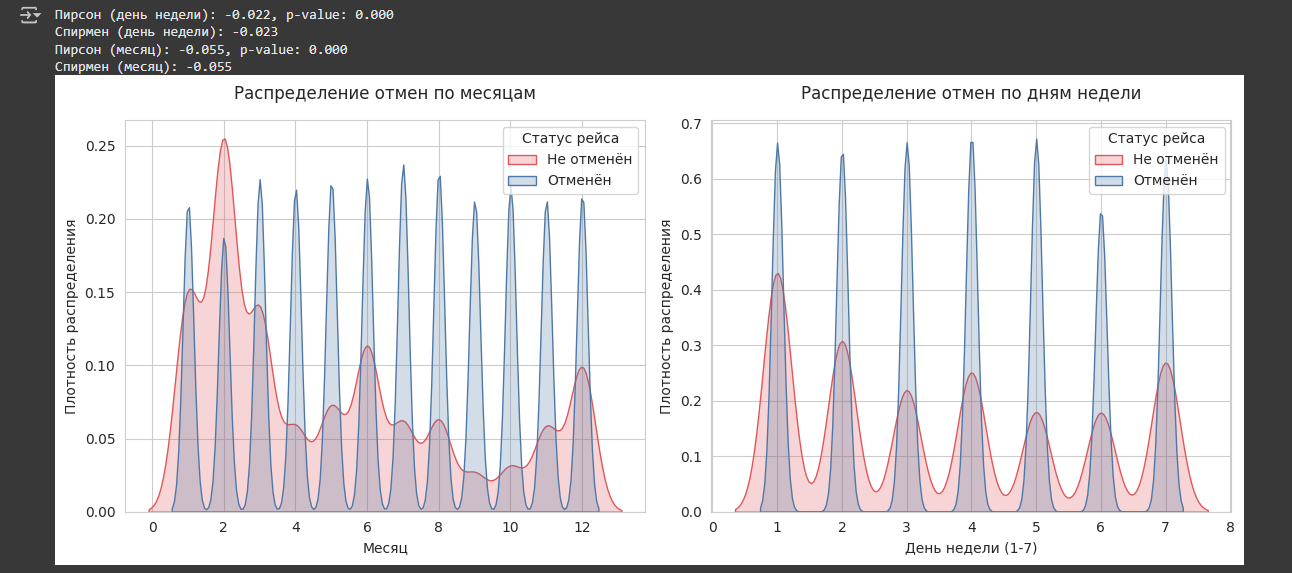
1. Histplot – кол-во отменённых и не отменённых рейсов.



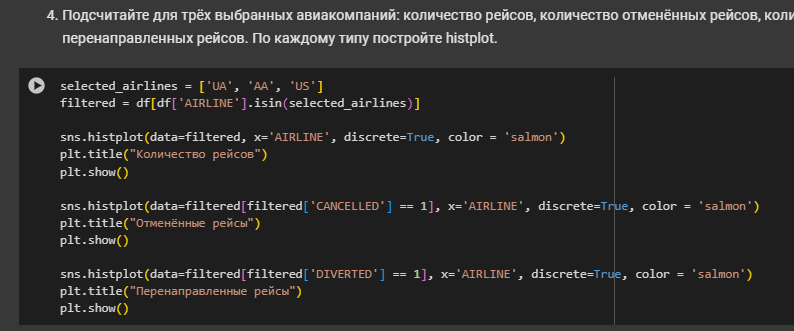
1. Постройте гистограмму для аропортов по количеству отменённых рейсов. Выведите топ 10 самых злостных отменщиков. Столбики отсортировать по убыванию слева на право. Определите аэропорт, рейсы для которого отменяются наиболее часто.

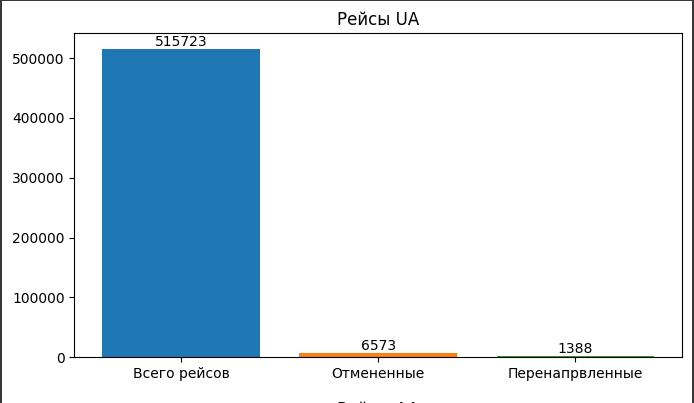


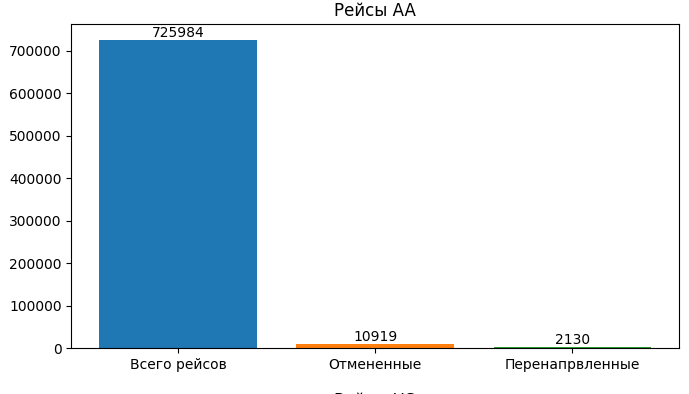
1. Определите коэффициент корреляции Пирсона и Спирмена между отменой рейса и днём недели, месяцем. Оцените значение p-value. Постройте плотность распределения количества отмен по месяцам и по дням недели (используйте kdeplot или displot).

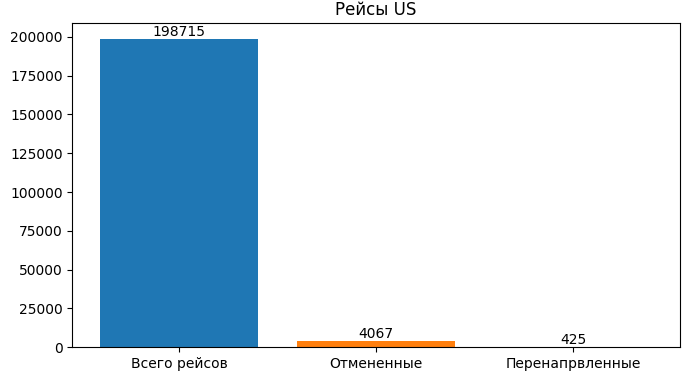


1. Подсчитайте для трёх выбранных авиакомпаний: количество рейсов, количество отменённых рейсов, количество перенаправленных рейсов. По каждому типу постройте histplot.

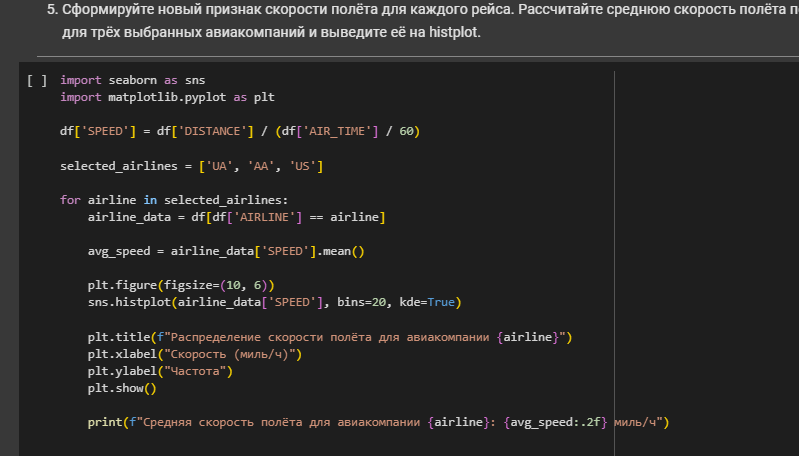




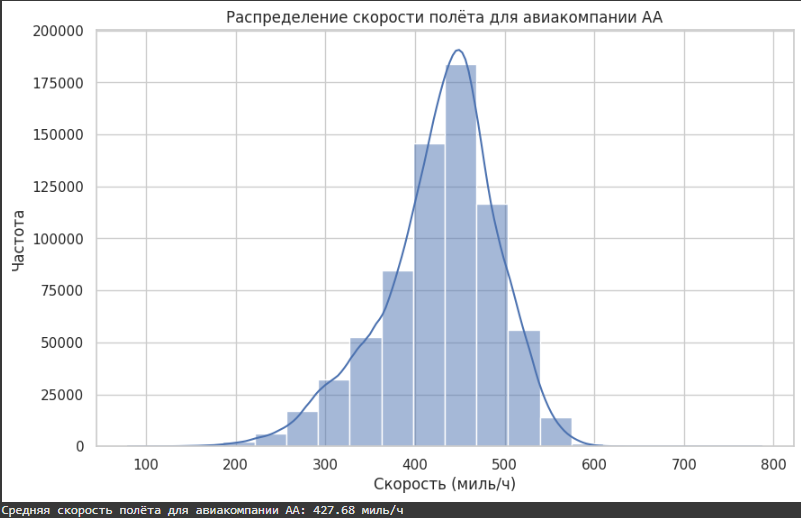


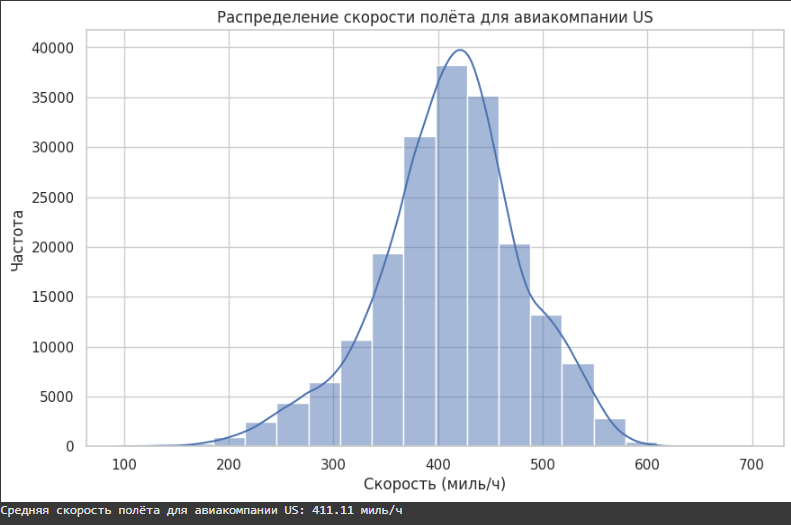


1. Сформируйте новый признак скорости полёта для каждого рейса. Рассчитайте среднюю скорость полёта по всем рейсам для трёх выбранных авиакомпаний и выведите её на histplot

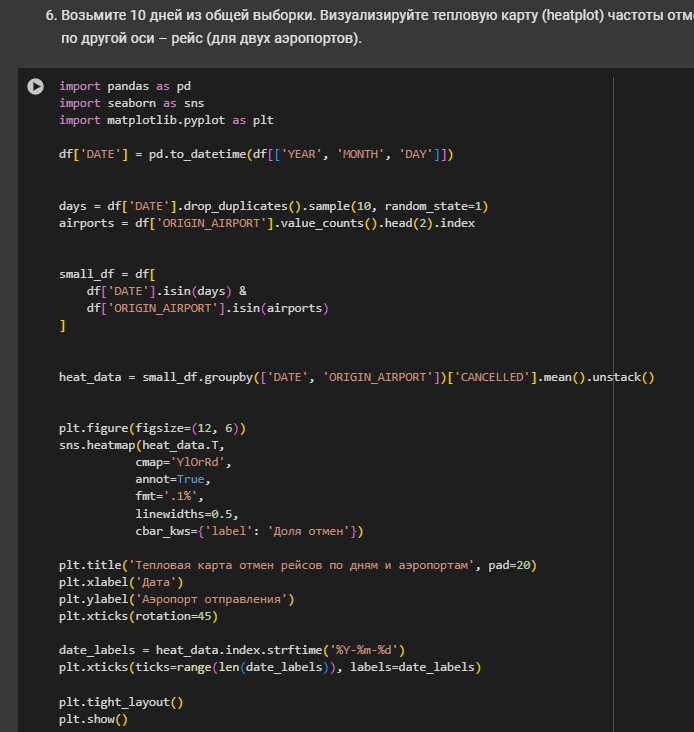


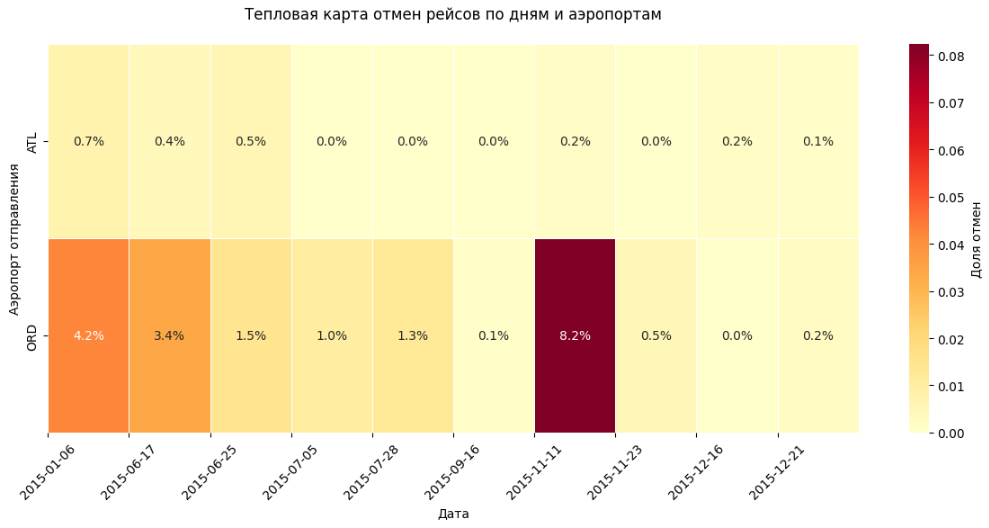




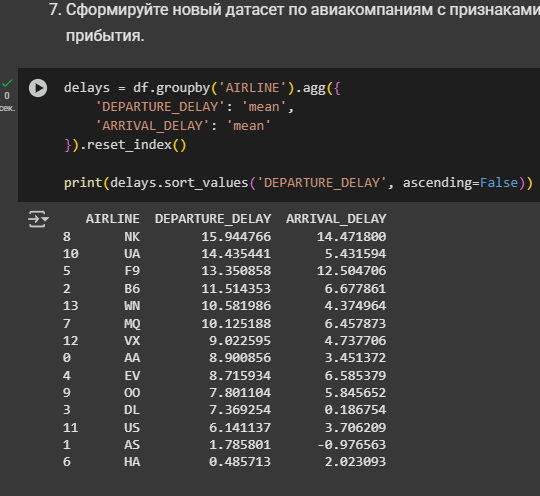


1. Возьмите 10 дней из общей выборки. Визуализируйте тепловую карту (heatplot) частоты отмены рейсов. По одной оси – дни, по другой оси – рейс (для двух аэропортов).

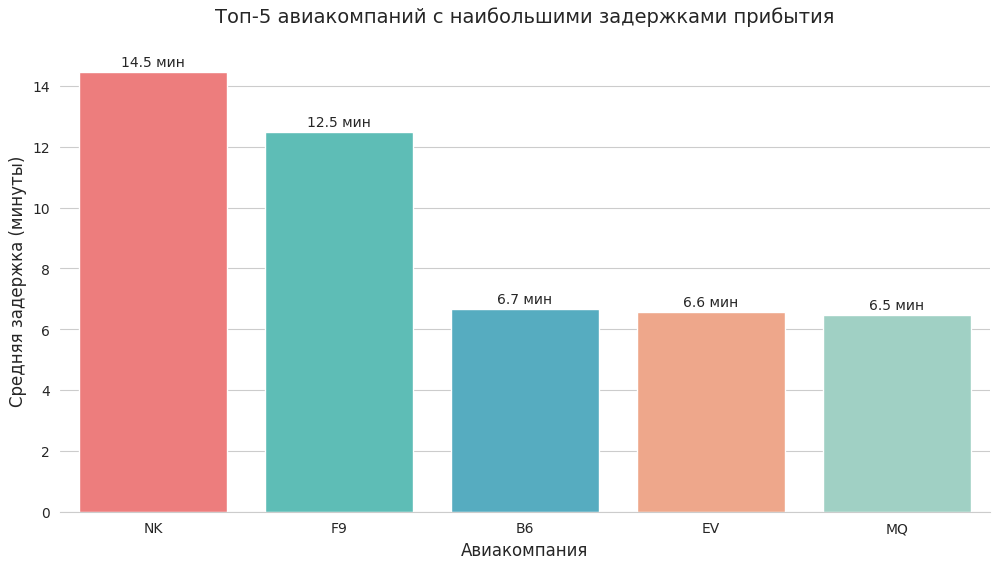


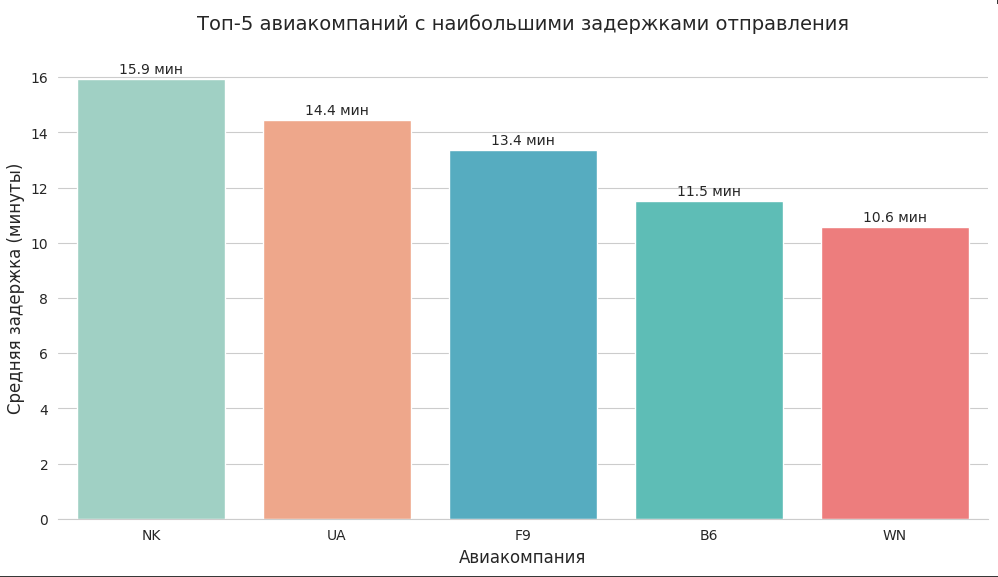


1. Сформируйте новый датасет по авиакомпаниям с признаками: среднее время задержки отправки и среднее время прибытия.

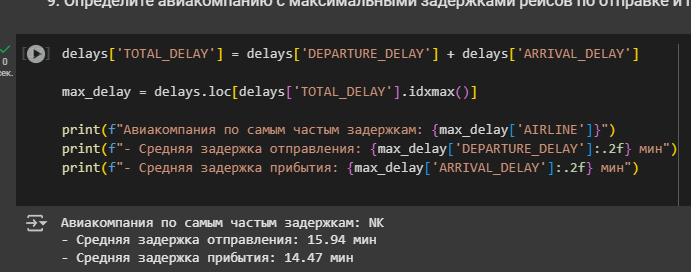


1. Постройте два histplot по задержкам отправления и прибытия для топа 5 компаний.





1. Определите авиакомпанию с максимальными задержками рейсов по отправке и прибытию (самый ярый опаздун).



**Вывод:**

В ходе выполнения лабораторной работы были успешно приобретены навыки работы с библиотеками анализа данных (такими как pandas, numpy), методами предварительной обработки данных (очистка, заполнение пропусков, нормализация) и средствами визуализации (matplotlib, seaborn). Освоены основные этапы обработки данных: загрузка, исследование структуры, преобразование и визуальное представление результатов. Полученные умения позволяют эффективно анализировать данные, выявлять закономерности и представлять их в наглядной форме, что является важным для дальнейшей работы в области анализа данных и машинного обучения.