**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**

**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**

**(РУТ (МИИТ)**

Институт/факультет «Академия “Высшая Инженерная Школа”, АВИШ»

Специальность/Направление подготовки Информатика и вычислительная техника

**КУРСОВАЯ РАБОТА**

**по дисциплине:** Технологии хранения больших данных

|  |  |
| --- | --- |
| **на тему:** | «Аналитика внутренних авиарейсов с помощью методов Big Data» |

|  |  |
| --- | --- |
| **Студент группы ШАД 212** | **(Орлов В.М )** |
| **Научный руководитель** | **(Зуев И.А. )** |

**Москва 2022 г.**

**Оглавление**

1. Основные термины / сокращения
2. Введение
3. Основная часть - описательная
4. Выводы по работе - отчет
5. Заключение
6. Список использованных источников и программ
7. Приложения – код

**Основные термины / сокращения**

Python – ЯП, использованный при выполнении КР.

Библиотека – подключаемый модуль, расширяющий возможности разработки.

Pandas – библиотека python, использованная для работы с данными.

Датасет – набор данных, считанный из использованных файлов.

Датафрейм – тип данных, предоставляемый pandas.

Функция – блок программы, которому присваивается некоторое имя и который пригоден к повторному использованию.

Индекс – в программировании означает положение элемента в некоем массиве. В контексте курсовой работы – столбец данных, уникальные значения которого используются в качестве индекса в его классическом понимании для поиска записей не по их порядковому номеру в датафрейме, а по конкретному интересующему значению индекса.

**Введение**

Цели курсовой работы представлены следующими пунктами:

1. Построение графика количества рейсов по месяцам
2. Построение графика количества рейсов по дням недели
3. Определение 10 городов с наибольшим количеством рейсов
4. Определение 10 авиакомпаний с наибольшим количеством рейсов.
5. Поиск аэропорта с минимальной задержкой вылета
6. Выявление самой пунктуальной авиакомпании на прилет
7. Поиск аэропорта, в котором самолеты проводят в среднем больше всего времени на рулении

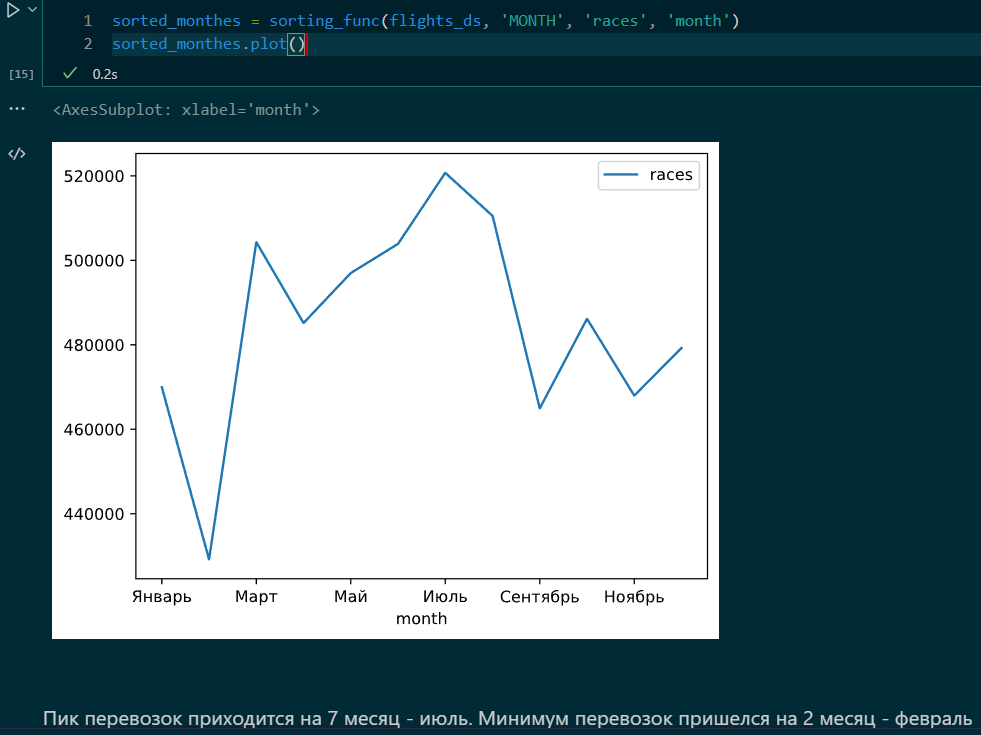
**Основная часть – описательная**

Для выполнения курсовой работы было использовано три файла формата csv – flights, airports, airlines.

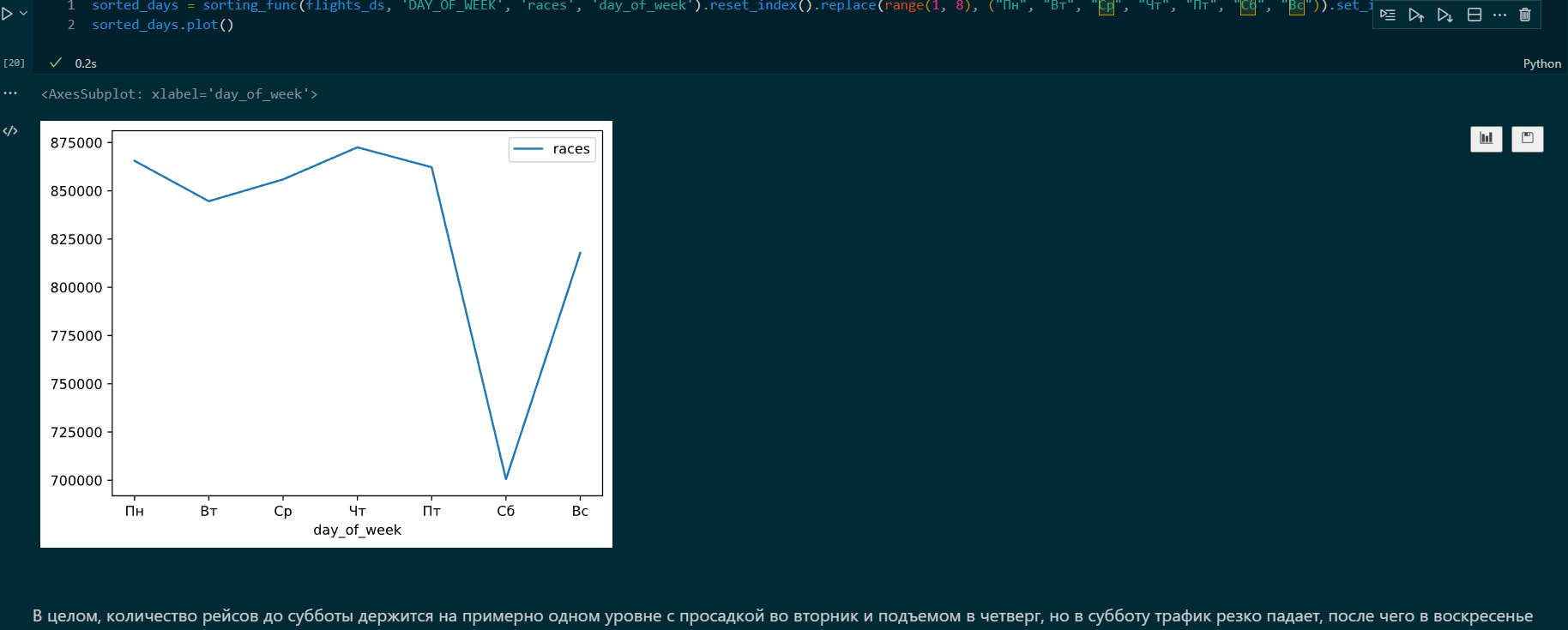
Для решения поставленных задач в программу были загружены все три файла.

Для построения графика количества рейсов, данные о перелетах были сгруппированы по месяцам, когда были проведены перелеты. Для упрощения работы была объявлена функция “sorting\_func”, которая проводит вышеперечисленные действия.

После группировки и обработки данных, был построен график, отображающий количество перелетов ( ось у ) и название месяца ( ось х ).



Для построения графика количества рейсов по дням были использованы те же исходные данные, что и раньше, а для их обработки была применена та же функция «sorting\_func». Единственным изменением стало то, по какому столбцу данных строились графики – в первом случае этим столбцом были

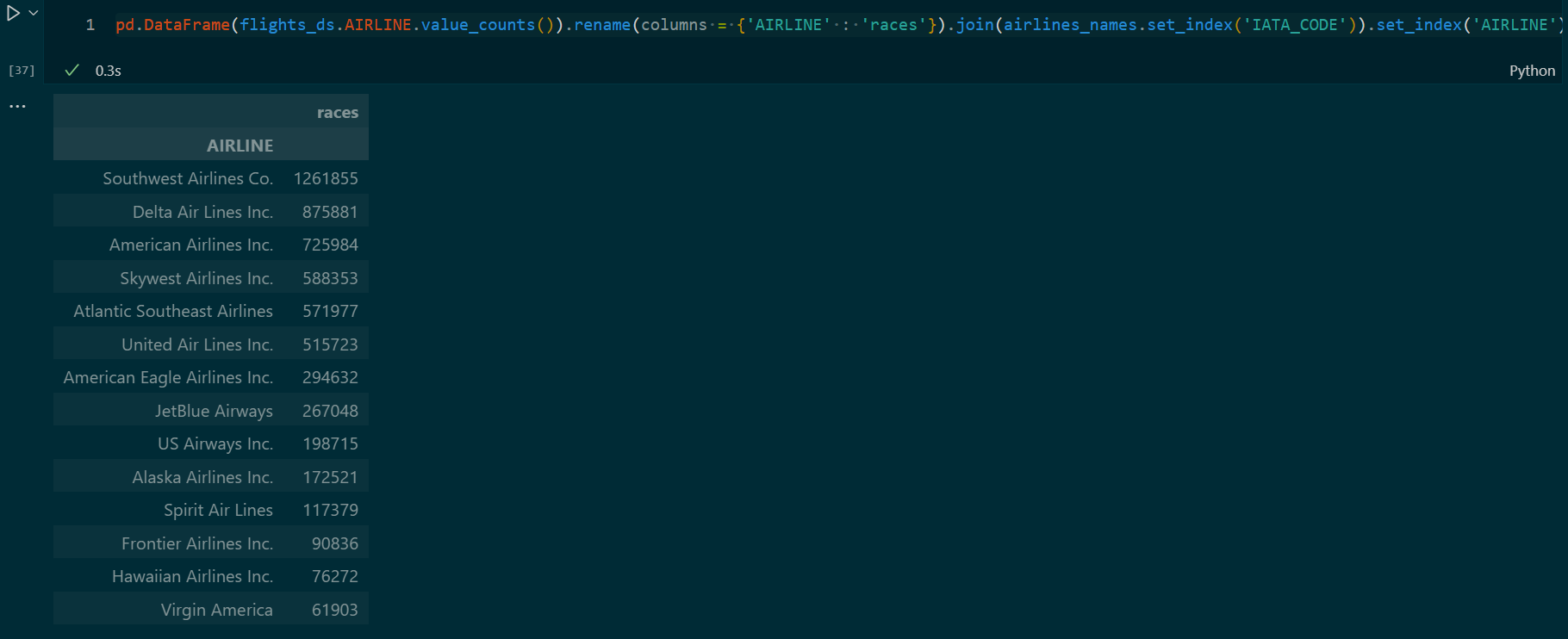
месяца, во втором – дни недели. Для большей понятности и наглядности, цифровые представления месяцев и дней недели были заменены на их названия на русском языке.

При определении 10 городов с наибольшим количеством авиаперевозок для упрощения поиска данных, была создана копия исходного датафрейма перелетов с присвоением ей индекса, отражающего IATA-код аэропорта **отправления**. После создания описанной копии, было произведено ее слияние с другим датафреймом, хранящим данные об аэропортах США. Слияние было произведено по значению короткого кода аэропорта IATA-коду. Итоговые значения были отсортированы по кодам аэропортов в порядке возрастания – от наименьшего к большему. Те же самые действия были проведены повторно, но теперь в качестве индексов были установлены коды аэропортов прибытия, которые так же были объединены с датафреймом аэропортов США и отсортированы по возрастанию. После всех произведенных операций, результаты по количеству перелетов для сгруппированных по аэропортам отправления и прибытия были сложены с установкой соответствия по индексам – кодам аэропортов. Таким образом была получена информация о количестве прибывавших и отбывавших рейсах с каждого аэропорта. Дополнительно итоговые суммы были отсортированы по убыванию и из них были выбраны первые 10 городов, в которых было произведено больше всего перелетов.

Изображение выглядит как текст

Автоматически созданное описание

При определении 10 авиакомпаний с наибольшим количеством рейсов все данные о рейсах были сгруппированы по авиалиниям их выполнявшим и сразу же было посчитано их количество, после чего полученный датафрейм был объединен с датафреймом, содержащим информацию об аэропортах США, где в качестве индекса был установлен код аэропорта.



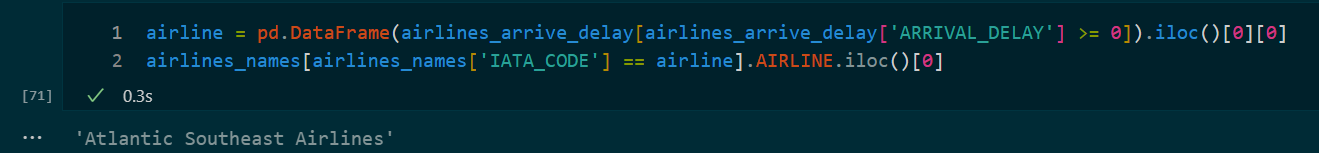
При определении же аэропорта с минимальными задержками вылетов было принято решение о том, что за минимальная задержка – это нулевая. То есть, если в датафрейме будет указана отрицательная задержка, это будет говорить о том, что самолет прилетел с опережением графика, а положительная – о том, что самолет задержался. Таким образом, за минимальную задержку было принято число 0. Используя выбранный подход, делаем выборку данных, содержащую лишь те записи из оригинального датафрейма с перелетами, в которых задержка вылета была равна нулю. После создания выборки, все полученные записи были сгруппированы по кодам аэропортов вылета и было подсчитано количество рейсов для каждого из них. В результате была получена сводка по количеству рейсов с нулевой задержкой у каждого из аэропортов, у которых такая задержка имелась.

Изображение выглядит как текст

Автоматически созданное описание

Для выявления самой пунктуальной авиакомпании был выбран аэропорт PHF.

Из исходного датафрейма были выбраны записи, в которых в столбце аэропорта прибытия фигурирует аэропорт PHF. После получения среза данных значения были отсортированы по величине задержки прибытия по возрастанию, после чего была выбрана авиакомпания, у которой по полученной сводке задержка прибытия минимальна.

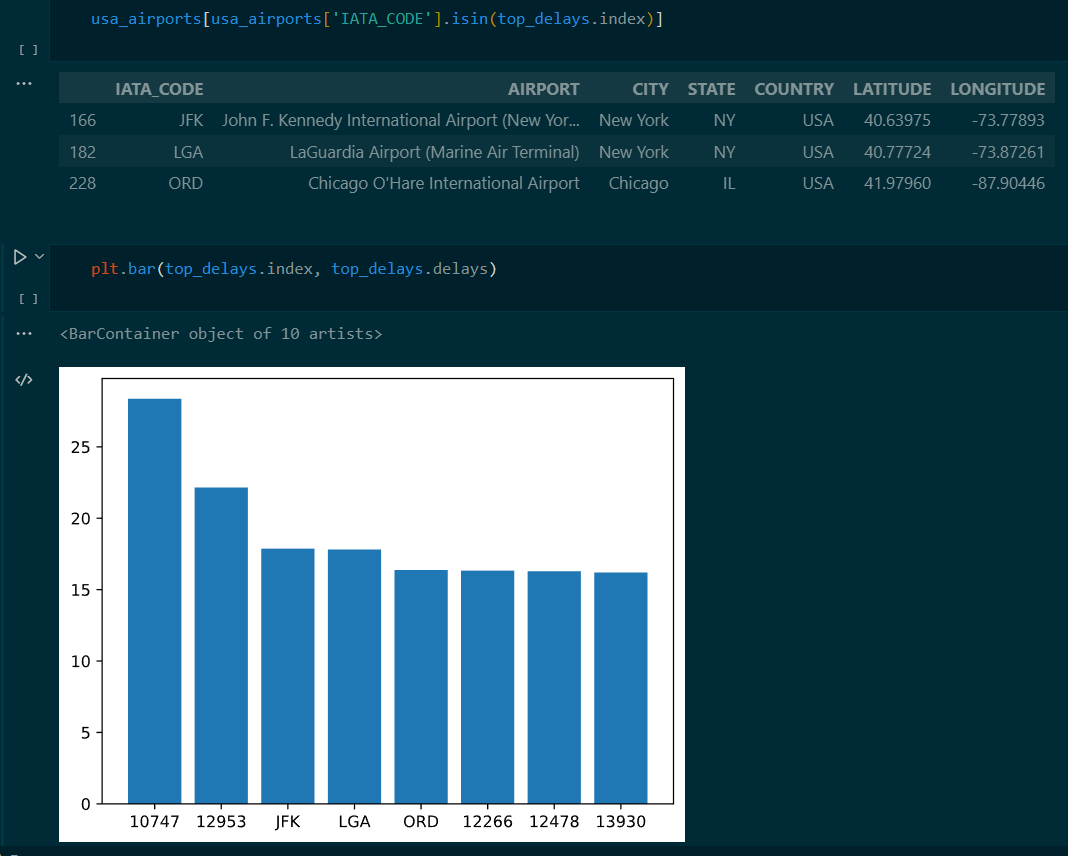


Для определения аэропорта с наиболее долгими рулежками было создано две выборки, в которых данные о перелетах были сгруппированы следующим образом:

в переменной origin\_delay записана выборка, содержащая данные о рулежке по прилете;

в переменной destination\_delay записана выборка, содержащая ланные о рулежке на вылет.

В обеих выборках данные сгруппированы по длительности задержки по убыванию. После получения выборок, их значения задержек были сложены соответственно каждому аэропорту, сгруппированы, были посчитаны средние значения задержек для каждого аэропорта и из них были выбраны 10 аэропортов с самыми большими задержками рулежки. Результаты действий были записаны в переменную top\_delays, которая позже применялась для определения полного названия аэропортов и для построения столбчатой диаграммы для большей наглядности разности задержек между аэропортами.



**Заключение**

В качестве результата проведенной работы были получены данные о работе многих аэропортов США, на основе которых можно сделать выводы о том, как именно работают аэропорты, какой у них пассажиропоток, какое время года наиболее востребовано и в какие дни недели люди предпочитают совершать перелеты. Кроме того, на основе проведенного анализа можно сделать выводы о том, что есть очень много аэропортов, задержки в которых достаточно велики, что, безусловно, необходимо решать.

**Список использованных источников и программ**

Visual Studio Code 2022

Python 3.10.7

Pandas

matplotlib

Windows 10

**Приложения**

Ссылка на выполненную курсовую работу: [ссылка](https://github.com/Meksis/Courseworks/blob/main/coursework_BD.ipynb)