**Flight Delay Analysis and Prediction**

**Problem statement:**

Flight delay analysis is a field of study that involves analyzing historical flight data to identify patterns and trends in flight delays. This analysis can help airlines, airports, and travelers to understand the causes of flight delays and to take steps to reduce them.

Flight delays can be caused by a variety of factors, such as weather conditions, air traffic congestion, mechanical issues, crew scheduling, and airport operations. By analyzing historical flight data, we can identify which airports in future may cause delays.

**Implementation:**

* We used PySpark and MongoDB to process and analyze flight delay data.
* Data was preprocessed to clean and prepare it for analysis.
* Exploratory data analysis was performed to gain insights into the data.
* Machine learning models were trained to predict flight delays based on various factors such as airline, origin, destination, time of day, and weather conditions.
* We created visualizations to communicate the results of our analysis.

**Establishing connection:**

We have initially established a connection to MongoDB Atlas using the MongoDB Spark connector and created a SparkSession named "AirLineTraffic".



**Preprocessing:**

We have selected specific columns from the data and filtered the data based on selected origin and destination airports.

We have created a new DataFrame called "arrivalDelay" by selecting and renaming relevant columns from "airlineDelay" and adding a "delay\_type" column with a constant value of "arrival".

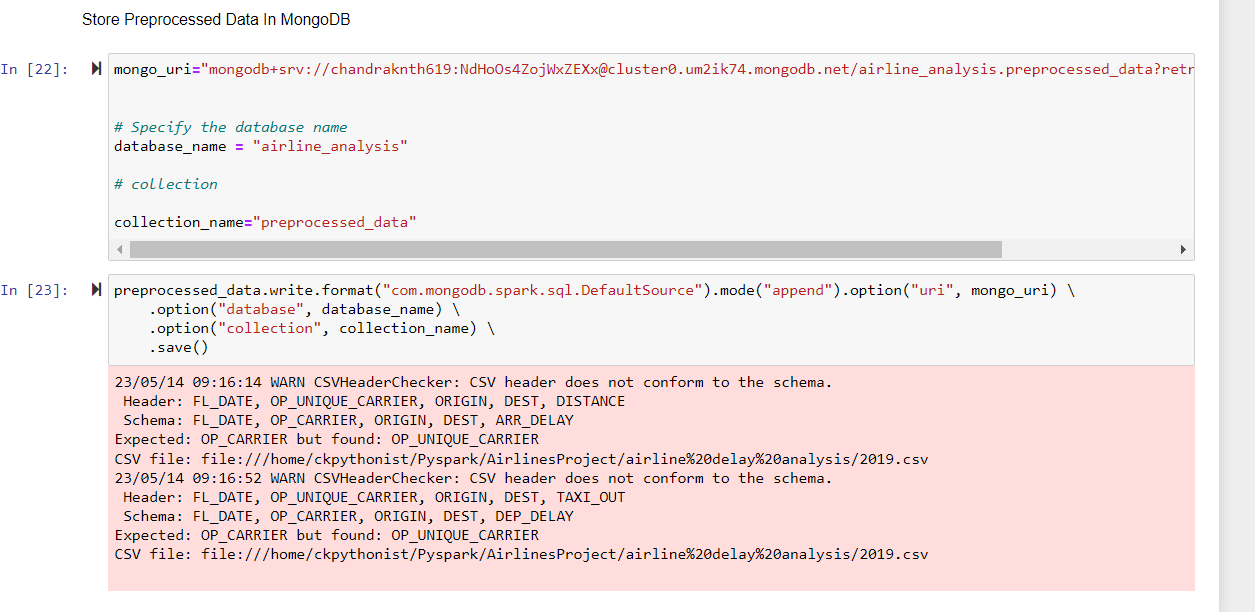
Similarly, we have created another DataFrame called "departDelay" by selecting and renaming columns from "airlineDelay" and adding a "delay\_type" column with a constant value of "departure".

The "arrivalDelay" and "departDelay" DataFrames are then grouped by various columns and aggregated using the average function to calculate the average delay for each combination of airport, carrier, year, month, day, day of the week, and delay type.

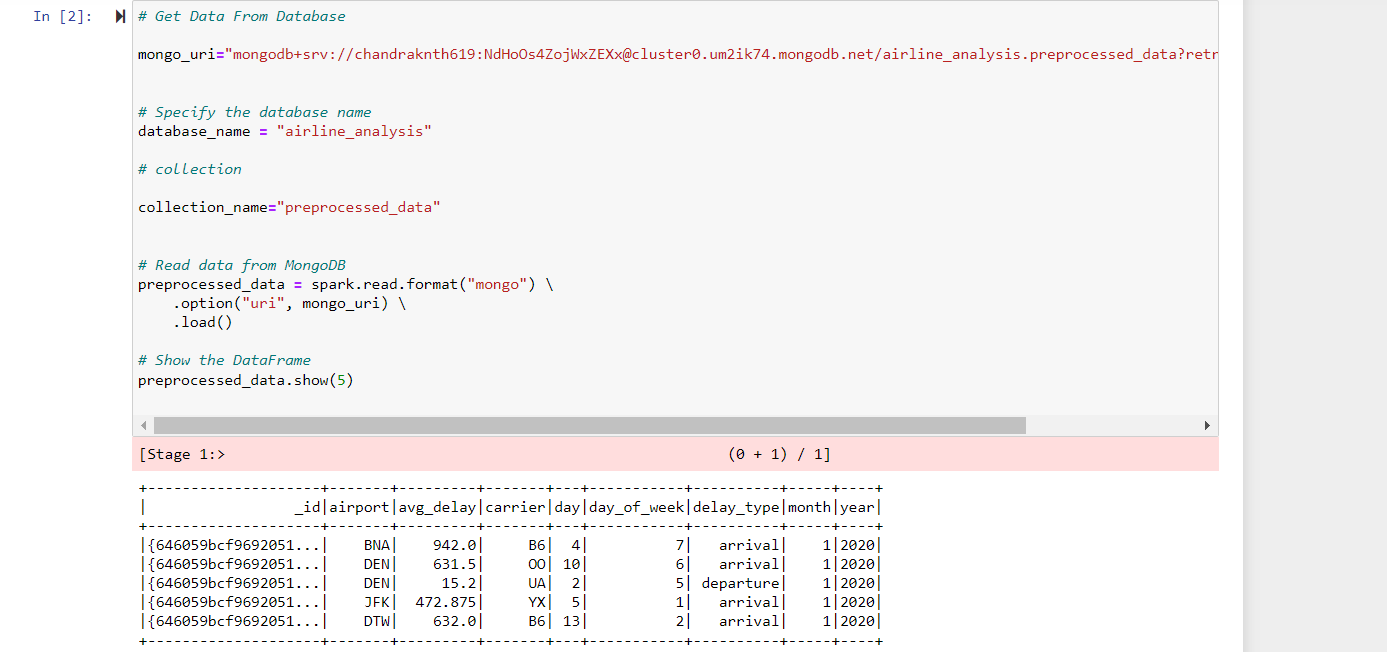
We then pushed this preprocessed data back to MongoDB.

**Loading the preprocessed data into MongoDB:**

Here, the preprocessed data has been loaded into MongoDB under specified database and collection.



**Retrieving the data from MongoDB:**



**Delay Analysis:**

With this data, we have analyzed many business use cases like percentage contribution of arrival delay of each airport, top 10 airports with most departures, top 10 airports with most arrivals, top 10 airports with the highest average of flights handled, etc.

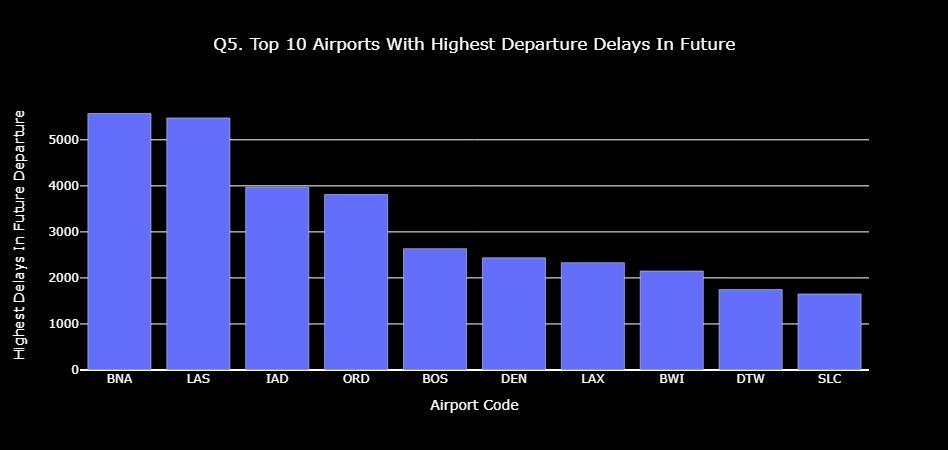
**Delay** **Prediction**:

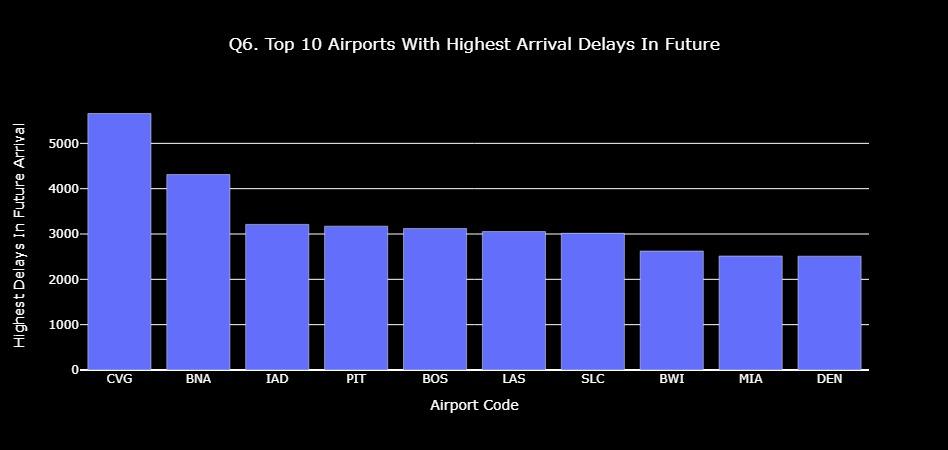
We have applied linear regression machine learning model to predict the future delays in both arrival delays and departures delays.

We have split the arrival delay data into train and test sets in 80:20 ratio for training and testing. So that we train the model and predict the outcomes.

Same method has been applied on departure delay data to predict the outcome.







**Conclusion:**

In conclusion, our flight delay analysis and prediction project provided valuable insights into the causes and effects of flight delays and identified areas for improvement in airline operations and customer experience. By considering factors such as weather, air traffic control, and airline scheduling, airlines can minimize the impact of flight delays and improve their overall performance.