****

**Module 4 Final Project**

**ALY6110**

**Professor: Anis Sharafoddini**

* Farhan Khaliq
* Joumana Obeid
* Aashish Punjabi
* Siyan Shen
* Thu Tran

July 1st , 2023

**Introduction:**

Air travel is integral to modern transportation because it lets people and businesses move around quickly and efficiently. But flight delays and cancellations can cause a lot of trouble and cost customers and airlines a lot of money. One of the requirements to improve the efficiency and reliability of air travel services is to know what causes these problems and their effects. In this project, we want to investigate the 2015 flight delay and cancellation dataset from the Bureau of Transportation Statistics at the U.S. Department of Transportation. By thoroughly analyzing this flight dataset, we will uncover insights, analyze distributions, detect patterns, and develop accurate forecasts based on regression techniques. The flight details include departure times, airline information, delays, and distances in the dataset.

**Overview:**

This project answers the following questions:

* In the dataset, how are departure delays distributed?
* What are the primary reasons for flight cancellations for each airline?
* How do different airlines distribute cancellation reasons?
* Predict/Forecast "DEPARTURE\_DELAY”,based on the independent variables in the dataset.

**Methodology:**

A four-stage analysis will be performed:

Data Cleaning, Exploratory Data Analysis, Visualizations, and Forecasting.

**Dataset Description:**

Each entry of the flights.csv file corresponds to a flight, and we see that more than 5'800'000 flights have been recorded in 2015. These flights are described according to 31 variables. These are some of the meaningful variables (1):

* **YEAR, MONTH, DAY, DAY\_OF\_WEEK**: dates of the flight
* **AIRLINE**: An identification number assigned by US DOT to identify a unique airline
* **ORIGIN\_AIRPORT** and **DESTINATION\_AIRPORT**: code attributed by IATA to identify the airports
* **SCHEDULED\_DEPARTURE** and **SCHEDULED\_ARRIVAL** : scheduled times of take-off and landing
* **DEPARTURE\_TIME** and **ARRIVAL\_TIME**: real times at which take-off and landing took place
* **DEPARTURE\_DELAY** and **ARRIVAL\_DELAY**: difference (in minutes) between planned and real times
* **DISTANCE:** distance (in miles)

**Analysis:**

**Data Cleaning:**

Data Types Matching.

**Findings:**

* **Departure summary statistics**

A screenshot of a computer

Description automatically generated

*figure1.Departure Delay Summary*

 **Count:** The total number is 5,732,926 delays.

 **Mean:** The average departure delay is approximately 9.37. This indicates that, on average, flights tend to depart slightly later than their scheduled departure time.

 **Standard Deviation:** The standard deviation of the departure delay is approximately 37.08. This measures the dispersion and indicates it is quite large.

 **Minimum:** The minimum departure delay is -82, indicating that some flights depart earlier than their scheduled time.

 **25th Percentile:** 25% of departure delay values fall below -5. This means that a quarter of flights experience -5 minute difference between the scheduled time and actual arrival time.

 **Median (50th Percentile):** The median departure delay is -2, indicating that 50% of flights experience -2 minutes delay.

 **75th Percentile:** 75% of departure delay values is 7. This means that three-quarters of flights are delayed by 7 minutes or more.

 **Maximum departure delay:** The maximum departure delay is 1988, indicating that some flights experience 1988, minute departure delays.

**Distribution of Delays per Airline:**

A screenshot of a computer

Description automatically generated

*figure2.Distribution of Delay per Airline*

This figure displays the distribution of delays categorized into three groups: delays of less than 5 minutes, delays in the range of 5 to 45 minutes, and delays greater than 45 minutes. It is evident that regardless of the airline, delays exceeding 45 minutes are relatively rare, constituting only a small percentage. However, the proportion of delays within these groups varies depending on the airline. For instance, SkyWest(OO) Airlines has a 30% lower number of delays greater than 45 minutes than delays in the range of 5 to 45 minutes. On the other hand, Southwest (WN)Airlines delays exceeding 45 minutes are four times less frequent than delays of 5 to 45 minutes. This indicates a relatively better performance in minimizing long delays.

**Distribution of Delays Per Airline-The Ranking:**

A screenshot of a computer

Description automatically generated

*figure3.Distribution of The Ranking*

This figure displays the normalized distribution of delays, modeled using an exponential distribution function f(x) = a \* exp(-x/b) (1)(2). The parameters a and b, which describe each airline's distribution, are provided in the upper right corner of each panel. It's critical to note that the normalization of the distribution applies to the histograms but not the model function. This results in an approximate integral of f(x)dx ≈ 1. The coefficients a and b are correlated, with a ∝ 1/b, meaning that either value is sufficient to describe the distribution. Based on the values of a or b, we can rank the airlines accordingly: lower values of a indicate a higher proportion of significant delays, while higher values of a indicate airlines that prioritize punctuality.

**Box Plot of Departure Delay:**

A screenshot of a computer

Description automatically generated

*figure4.Distribution of Departure Delays*

The delays range from 0 to 15000 minutes with some outliers around 20000 minutes.

**Relation between Distance and Delay:**

A screenshot of a computer

Description automatically generated

*figure5.Relation Between Variables, Esp. Distance and Delay*

A screenshot of a computer

Description automatically generated with medium confidenceWe see the correlation between departure delays and some factors in the dataset. We see a weak negative correlation between distance, day\_of\_the\_week , month and departure delays.

*figure6.Distribution of Day of the week*

*A screenshot of a computer

Description automatically generated*

*figure7.Distribution of Day of Distance*

We can see that most flights are for short distances.

This can be the cause for the weak correlations.

*A screenshot of a computer

Description automatically generated with medium confidence*

*Figure8.Distribution of Day of Month*

**Delay Reason Count Per Airline:**

A screenshot of a computer

Description automatically generated

*Figure9.Delay Reason Counts of Every Airline*

A screenshot of a computer

Description automatically generated

*Figure10.Key Statistics About Delays Information*

Overall, Southwest Airlines (WN) emerges as a prominent player in flight delays across multiple categories, including departure delays, weather-related delays, airline-related delays, and late aircraft delays. This indicates the airline's significant impact on delay statistics.

While Frontier Flight (F9) and JetBlue Airways both recorded high numbers of delays, JetBlue Airways Corporation (B6) is among the top six in delay reasons counts.

It is worth noting that different airlines excel in specific delay categories. This suggests variations in their operational strategies, infrastructure, and response to different delay factors. However, the most common delay factors are “Air System Delay and Airline Delay”.

**Cancelation Reason Count Per Airline:**

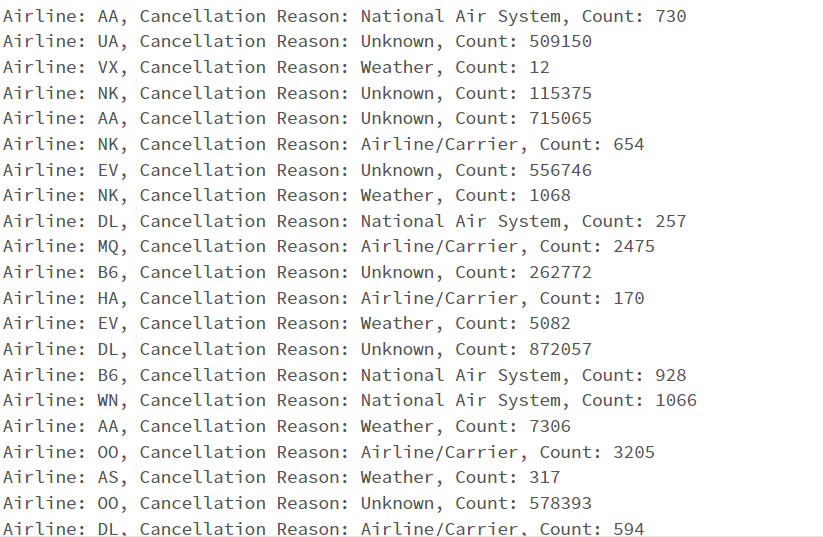
**A screenshot of a computer

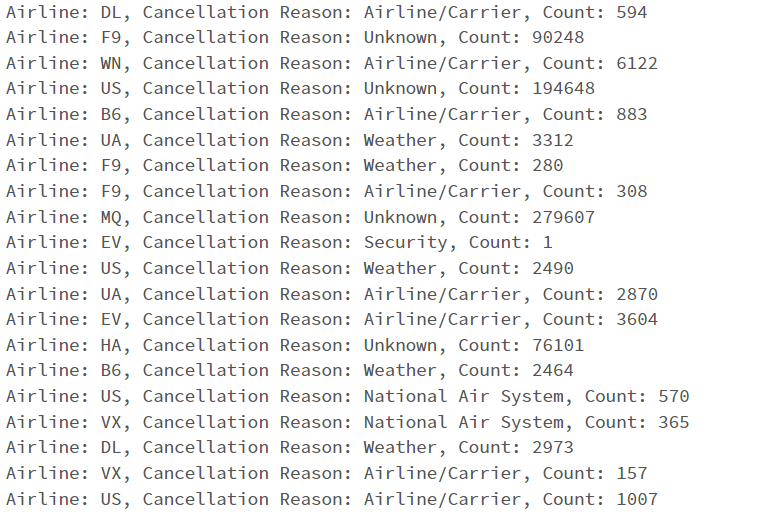
Description automatically generated**

*Figure11.Key Statistics About Delays Information*

The bar graph shows the cancellation reason per airline.

We see that the weather is at the top of cancellations.

**

**

*A screenshot of a computer

Description automatically generated with medium confidence*

**

*Figure12.Cancelation Counts of Every Airline*

A screenshot of a computer

Description automatically generated

*Figure13.Key Statistics About Cancelation Information*

In the provided dataset, the airline with the highest cancellation count is WN (Southwest Airlines) The most common cancellation reasons across all airlines are "Unknown", "Weather" and "Airline/Currier".

**Regression Model:**

A screenshot of a computer

Description automatically generated

*Figure14.Regression Model*

After comparing ARIMA and Multilinear Regression Model, we selected Multilinear regression model because the RMSE of Multilinear Regression Model was much lower than RMSE of ARIMA RMSE.

A regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

We selected these variables based on our intuition that these are important factors in the regression model.

The dependent variable:

* 'DEPARTURE\_DELAY'

The independent variables:

* 'DISTANCE'
* 'AIRLINE'
* 'ORIGIN\_AIRPORT'
* 'AIR\_SYSTEM\_DELAY'
* 'SECURITY\_DELAY'
* 'AIRLINE\_DELAY'
* 'LATE\_AIRCRAFT\_DELAY'
* 'WEATHER\_DELAY'

**Results:**

The regression model has an RMSE (Root Mean Squared Error) of 16.39, indicating the average prediction error of the model. The R-squared value is 0.941, suggesting that the model explains 94.1% of the dependent variable variance. The model coefficients represent the effect of each independent variable on the dependent variable. The constant term is -3.7245, and the other variables (x1 to x8) have coefficients ranging from -0.0013 to 1.0650. The p-values associated with each coefficient indicate the statistical significance of the variables. Model performance and statistical assumptions, such as normality and multicollinearity, should be carefully considered.

**Conclusion:**

In this project, we thoroughly analyzed the 2015 flight delay and cancellation dataset from the Bureau of Transportation Statistics at the U.S. Department of Transportation. We aimed to gain insights into the causes of flight delays and cancellations and develop accurate forecasts using regression techniques.

From our analysis, we observed that the average departure delay was approximately 9.37 minutes, indicating that flights tend to depart slightly later than their scheduled departure time. The distribution of delays varied among different airlines, with some airlines performing better in minimizing long delays. We ranked the airlines based on their distribution of delays, where lower values of a indicated a higher proportion of significant delays.

We also investigated the reasons for flight cancellations and identified weather as the most common cancellation reason across all airlines. Southwest Airlines (WN) emerged as a prominent player in flight delays and cancellations, highlighting its impact on delay statistics.

Finally, we developed a multilinear regression model to predict departure delays based on several independent variables, including distance, airline, origin airport, and delay factors. The model achieved an RMSE of 16.39, indicating the average prediction error. The high R-squared value of 0.941 showed that the model explained 94.1% of the variance in departure delays.

In conclusion, our analysis provided valuable insights into the distribution of delays, cancellation reasons, and the relationship between departure delays and various factors. The regression model demonstrated the potential to predict departure delays accurately. The findings from this project can be used to improve the efficiency and reliability of air travel services, ultimately benefiting both customers and airlines.

**Resource:**

1. Author(s): Fabien Daniel Title: Predicting Flight Delays Tutorial Site Name: Kaggle URL: <https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial/>
2. Exponential distribution. (2020, October 15). Wikipedia. https://en.wikipedia.org/wiki/Exponential\_distribution
3. 5.Mount\_S3\_Buckets\_Databricks. (n.d.). Www.youtube.com. Retrieved June 23, 2023, from https://www.youtube.com/watch?v=JRcgoyE\_Tsc&t=776s
4. 3-Ding1, Y. (2017, August 1). *IOPscience*. IOP Conference Series: Earth and Environmental Science. <https://iopscience.iop.org/article/10.1088/1755-1315/81/1/012198>

**Appendix**

Please see Python code attached.