UNIVERSITY
OF LONDON

Programming for Data Science

ST2195

Analyzing Flight Delay Patterns in 2006 and 2007

Name: Cynthia Khong Qing

Student Number: 210407112

Table of Content

1. Introduction

Nowadays, people increasingly choose to travel by air which is convenient and time-saving. However, the most common problem faced by the airline traveler is the flight delays which resulted in waste of time and money for the passengers. Flight delays indicates a sign of inefficiencies in the aviation system. It is high cost for both airline companies and passengers. A study estimates that air transportation delays cost the airline industry and air travelers a total of $32.9 billion in the US in 2007, causing a $4 billion reduction in GDP. Predicting delays can improve airline operations and passenger satisfaction, leading to positive economic impacts.

The objective of this coursework is to analyze a dataset that contains the details of flight arrival and departure between January 2006 and December 2007. The dataset was obtained from the Harvard Dataverse at https://doi.org/10.7910/DVN/HG7NV7. The study aims to identify patterns and trends in flight delays. This analysis will involve using descriptive statistics, data visualization techniques, and machine learning algorithms to uncover insights from the dataset. The expected outcomes of this study are to provide recommendations for airlines and airport managers to improve their operations and reduce the incidence of flight delays and cancelations.

2. Data Wrangling

Importing Libraries

The following libraries are being used:

1. **numpy** is a Python library used for scientific computing on large multi-dimensional arrays
2. **pandas** is a Python library used for analyzing, cleaning, exploring and manipulating data
3. **sklearn** is a Python library that provides provides efficient interface and machine learning models
4. **seaborn** and **matplotlib** are both Python visualization libraries used for creating visualization of data

Data Reading and Cleaning

The data is downloaded from Harvard Dataset. This coursework used the data from the years 2006 and 2007.

Firstly, import these data using the function **read.csv** from **pandas**.

```
# Read Data
data_2006 = pd.read_csv(r"C:\Users\cynth\OneDrive\Documents\STUDY\Y2\Programming\coursework\2006.csv")
data_2007 = pd.read_csv(r"C:\Users\cynth\OneDrive\Documents\STUDY\Y2\Programming\coursework\2007.csv")
```

```
airports = pd.read_csv(r"C:\Users\cynth\OneDrive\Documents\STUDY\Y2\Programming\coursework\airports.csv")
carriers = pd.read_csv(r"C:\Users\cynth\OneDrive\Documents\STUDY\Y2\Programming\coursework\carriers.csv")
plane = pd.read_csv(r"C:\Users\cynth\OneDrive\Documents\STUDY\Y2\Programming\coursework\plane-data.csv")
```

Figure 2: Importing Dataset

Then, two dataset is being combined using the function concat from pandas.

```
#Combine two years into one data frame
df = pd.concat([data_2006, data_2007])
df.head()
```

Figure 3: Combine two dataset

Several functions are used to get the basic information of the dataset. The function **shape** is used to observe the rows and columns of the dataset.

```
df.shape
```

```
(14595137, 29)
```

Figure 4: The shape of the Dataset

The columns name of the data:

```
df.columns
```

```
Index(['Year', 'Month', 'DayofMonth', 'DayOfWeek', 'DepTime', 'CRSDepTime',
       'ArrTime', 'CRSArrTime', 'UniqueCarrier', 'FlightNum', 'TailNum',
       'ActualElapsedTime', 'CRSElapsedTime', 'AirTime', 'ArrDelay',
       'DepDelay', 'Origin', 'Dest', 'Distance', 'TaxiIn', 'TaxiOut',
       'Cancelled', 'CancellationCode', 'Diverted', 'CarrierDelay',
       'WeatherDelay', 'NASDelay', 'SecurityDelay', 'LateAircraftDelay'],
      dtype='object')
```

Figure 5: Columns Name

This dataset contains 14595137 rows of observations and 29 variables.

```
#Create a date for convenient
df['Date'] = pd.to_datetime(df[['Year','Month','DayofMonth']].rename(columns={'DayofMonth': 'Day'}))
df['Date']
```

```
#Create a Total Delay Column
df['Delay'] = df['ArrDelay']
#If we add DepDelay and ArrDelay there will be double counting delay as the correlation is high
#so we use ArrDelay as the overall delay
```

Figure 6: Creating new columns

Date and Delay columns are created for convenient use.

The Arrival Delay used as the total delay which is a better measure of delay for our analysis as it directly reflects the impact of the delay on the arrival time of the flight, which is the key concern for passengers.

3. Exploratory Data Analysis (EDA)

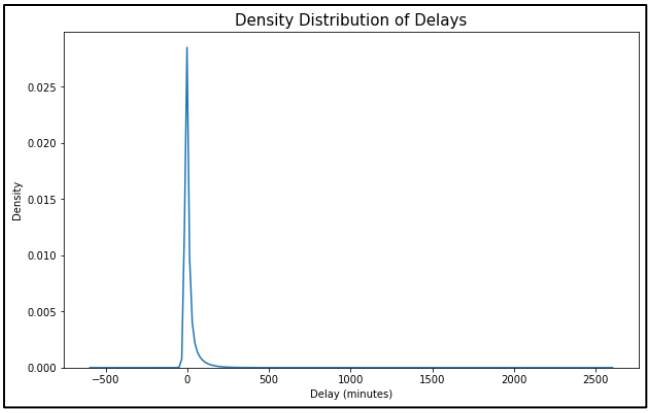


Figure 7: Density Distributions of Delays

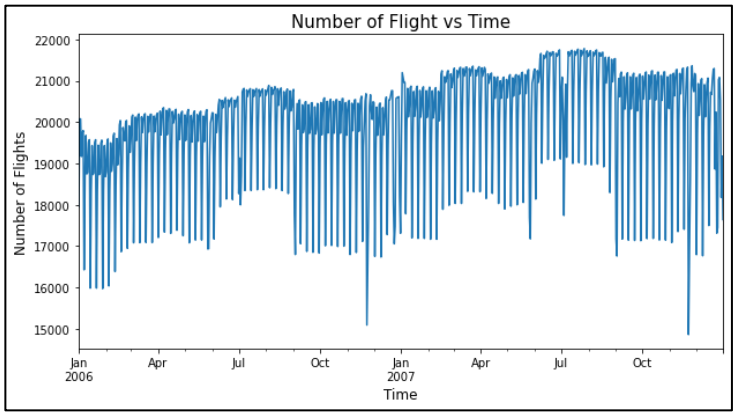
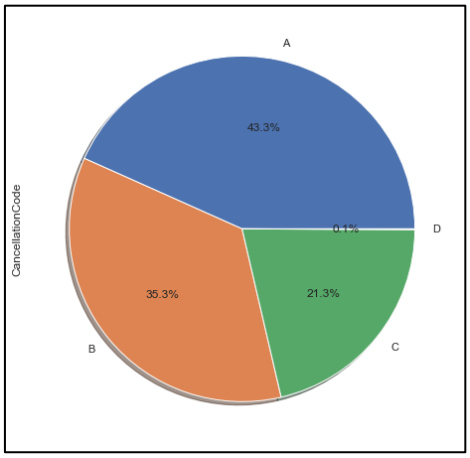


Figure 8: Number of Flight vs Time



A = carrier, B = weather, C = NAS, D = Security

Figure 9: Reason of Cancelled Flight

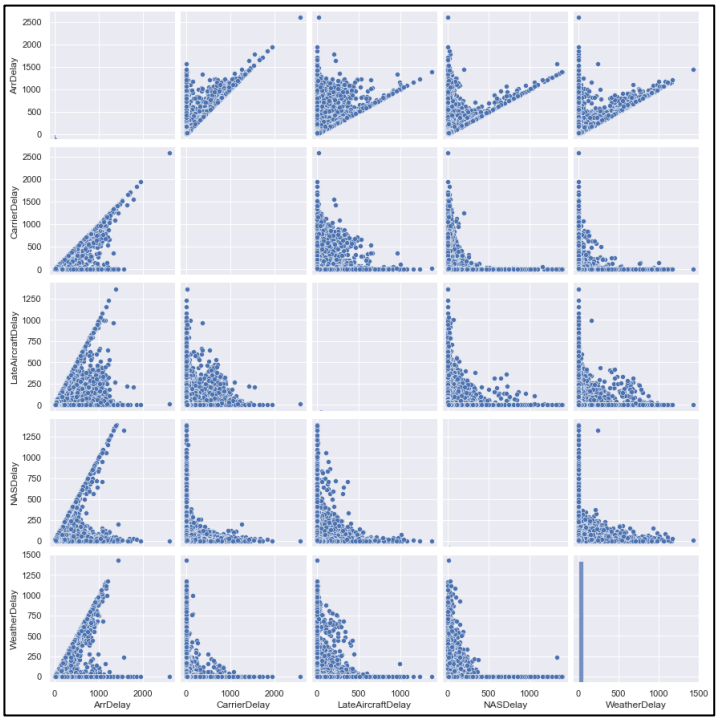


Figure 10: Correlation graph of Arrival Delay and Reason

Based on the four figures above, we can gain valuable insights into the dataset. Figure 7 shows most of the data is around 0, which means most of the flights had little or no delays. According to Figure 8, we can observe that there is an upward trend in the number of flights over 2006 and 2007. Figure 9 shows the reasons behind cancelled flights, which indicates that carrier and weather are the most common reasons for cancelling flights. Figure 10 reveals the correlation between arrival delay and the delay reason, and we can observe that carrier delay has a strong positive relationship with arrival delay, implying that carrier delay is the primary cause of delay in most cases.

4. Question and Analysis

3.1 Question 1

When is the best time of day, day of the week, and time of year to fly to minimise delays?

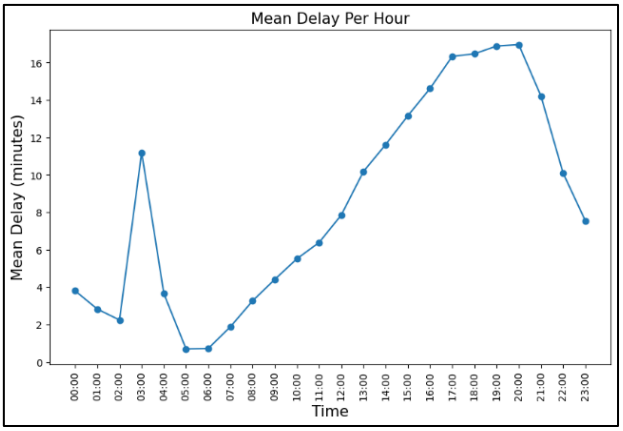


Figure 11:Mean Delay Per Hour

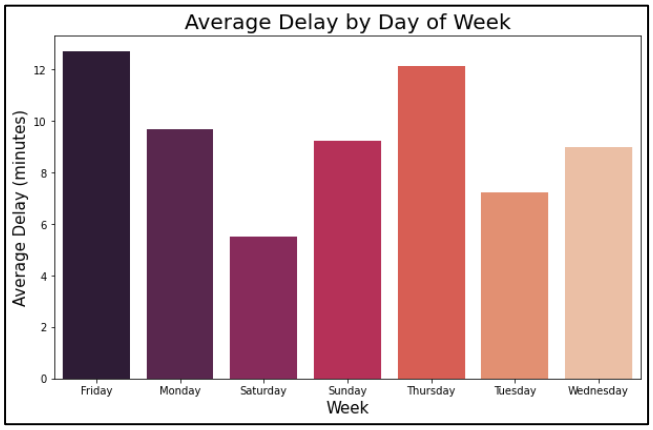


Figure 12: Average Delay by Day of Week

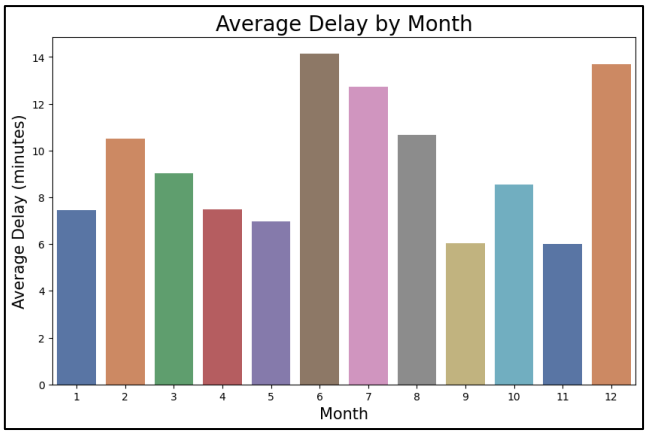


Figure 13: Average Delay by Month

Based on Figure 11, the time slot from 5.00am to 6.00am has the least mean delay in one day. According to Figure 1, 2Saturday is the day that having the least average delay in the week. Figure 13 shows that November has the least average delay compared to other months. Therefore, the best time of day, day of the week, and time of year to fly to minimize delays are 5.00am to 6.00am, Saturday and November respectively.

However, these time slots are having relatively fewer flights compared to other times. Overall, these findings could be useful for passengers to choose the optimal time to fly and avoid unnecessary delays and wasted time.

3.2 Question 2

Do older planes suffer more delays?

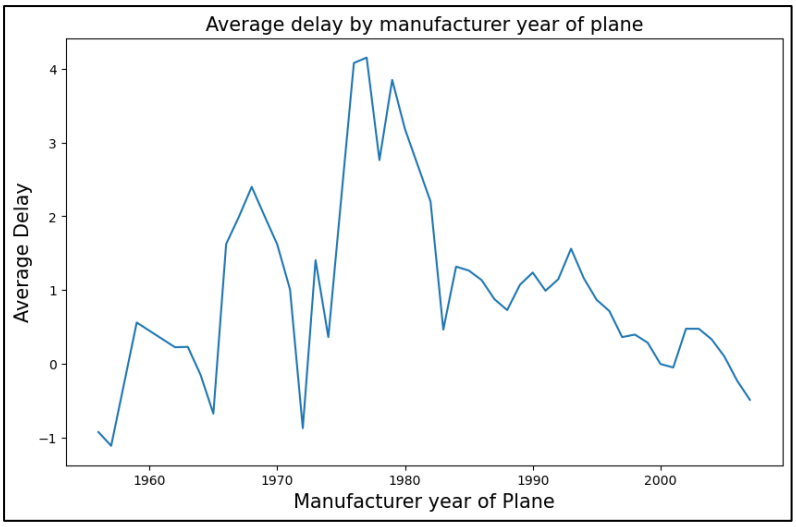


Figure 14: Average delay by manufacturer year of plane (line plot)

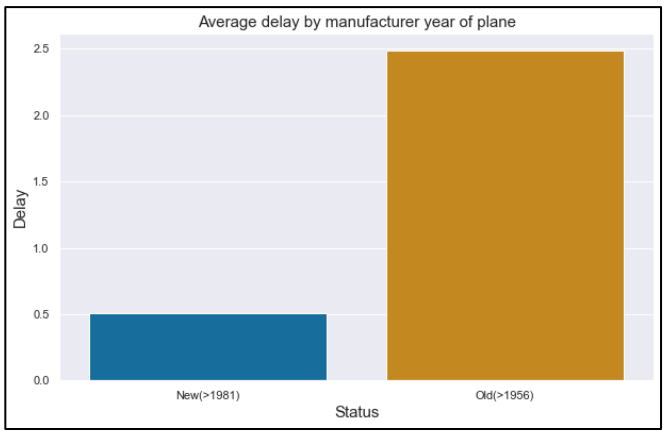


Figure 15: Average delay by manufacturer year of plane(bar plot)

According to Figure 14, we observe that planes manufactured before 1985 have a higher average delay compared to those manufactured after 1985. The delays for planes manufactured after 1985 are more stable over time. This result suggests that newer planes have a better performance in terms of delay.

From Figure 15, planes manufactured before 1981 are classified as 'older planes' while those manufactured after 1981 are classified as 'newer planes'. The result shows that older planes have a higher delay than newer planes. This further supports the finding that older planes suffer more delays.

Based on the above analysis, we can conclude that the manufacturer year of the plane is a significant factor in suffering delays. Older planes are more likely to experience delays compared to newer planes. This finding can be useful for airlines to make decision, such as retiring older planes and investing in newer ones to improve their operational efficiency and reduce delays.

3.3 Question 3

How does the number of people flying between different locations change over time?

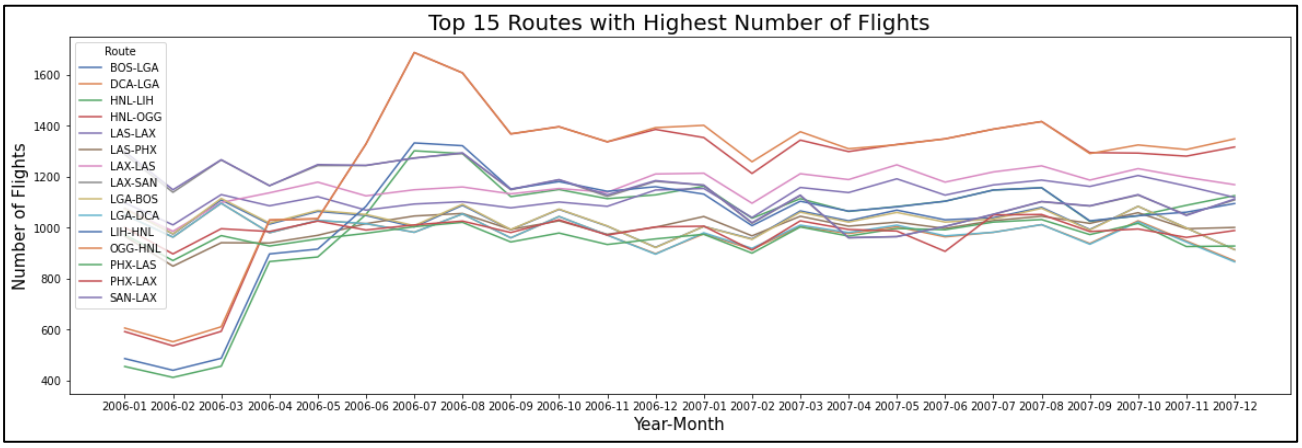


Figure 16: Top 15 Routes with Highest Number of Flights

The analysis of flight data reveals interesting patterns in the number of people flying between different locations over time. To gain insights into the most popular routes, we focused on the top 15 routes with the highest number of flights and studied their trends.

As shown in Figure 16, the number of flights for these routes remain stable between 800 and 1400 over time, with some exceptions. There are four routes, OGG-HNL, LIH-HNL, HNL-LIH, and HNL-OGG, experienced a sharp increase in the number of flights between 2006-03 and 2006-04, followed by another surge between 2006-05 and 2006-07. LIH-HNL and HNL-LIH emerged as the top routes among the others.

Overall, these findings suggest that LIH and HNL were the most popular origin and destination routes since 2006-03. The sharp increases in the number of flights for these routes during 2006-03 to 2006-04 and 2006-05 to 2006-07 could be attributed to several factors such as seasonal demand, competitive pricing, and the introduction of new flight routes. Based on these insights, airline companies can make more informed decisions to optimize their routes, pricing strategies, and marketing efforts to attract more customers.

3.4 Question 4

Can you detect cascading failures as delays in one airport create delays in others?

```
N308SW    8406
N478HA    8046
N479HA    7944
N485HA    7932
N480HA    7928
N484HA    7851
N481HA    7794
N475HA    7726
N487HA    7721
N477HA    7713
Name: TailNum, dtype: int64
```

Figure 17 : Top 10 Tail Number having the most flights

	TailNum	Cascading Failures	Total Flights	Cascading Failures(%)	Cascading Failures with delay in first Airport(%)	Cascading Failures with delay in first and second Airport(%)
0	N478HA	492	8046	5.828983	54.725788	66.714083
0	N479HA	469	7944	5.903827	60.516129	71.932515

Figure 18: Table of Tail Number having Cascading Failures

From figure 17, two tail numbers were randomly selected from the top 10 flights with the highest number of flights. We can observe that around 5% to 6% of flights experience cascading failures based on figure 18

Moreover, the table in figure 18 shows that the delays in the first airport have resulted in cascading failures for 50% to 60% of cases, while delays in both the first and second airport have caused cascading failures in 65% to 72% of cases. These results indicate that delays in one airport can have a significant impact on the entire flight network, causing delays in other flights as well.

These observations have shown the significant of controlling delay of flights which could cause serious cascading failures. It is important to have efficient management and coordination between airports to minimize the risk of cascading failures and delays.

Overall, the analysis have suggests that delays in one airport creates delays in another, which could result in serious disruption to air travel. It is important for airlines to ensure the smooth operations and take proactive measure to minimize the risk of cascading failures.

5. Machine Learning

5.1 Question 5

Use the available variables to construct a model that predicts delays.

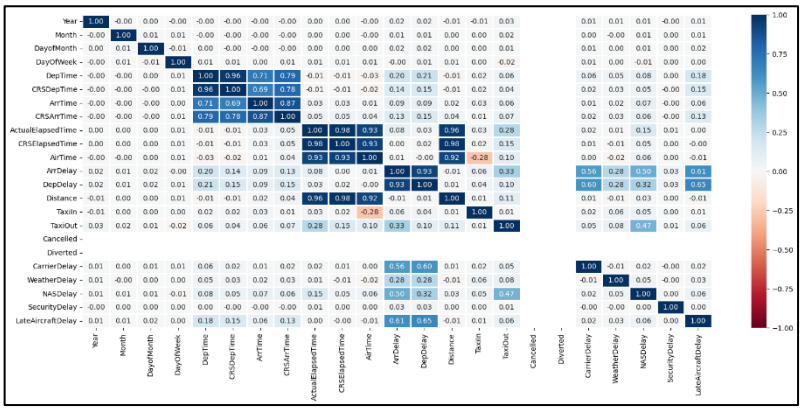
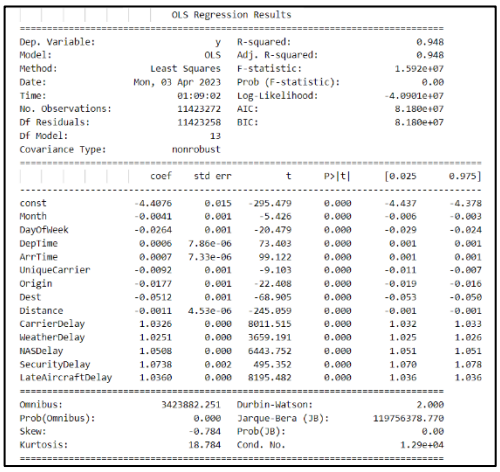


Figure 19: Correlation between each variable



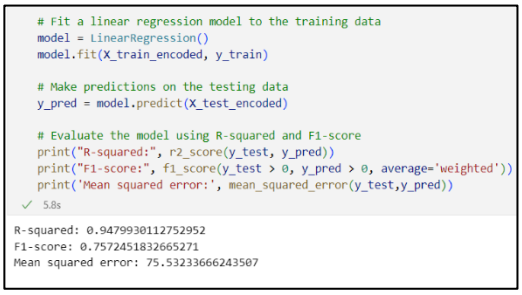
OLS Regression Results

Dep. Variable:	y	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	1.592e+07
Date:	Mon, 03 Apr 2023	Prob (F-statistic):	0.00
Time:	01:09:02	Log-Likelihood:	-4.0901e+07
No. Observations:	11423272	AIC:	8.180e+07
Df Residuals:	11423258	BIC:	8.180e+07
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.4076	0.015	-295.479	0.000	-4.437	-4.378
Month	-0.0041	0.001	-5.426	0.000	-0.006	-0.003
DayOfWeek	-0.0264	0.001	-20.479	0.000	-0.029	-0.024
DepTime	0.0006	7.86e-06	73.403	0.000	0.001	0.001
ArrTime	0.0007	7.33e-06	99.122	0.000	0.001	0.001
UniqueCarrier	-0.0092	0.001	-9.103	0.000	-0.011	-0.007
Origin	-0.0177	0.001	-22.408	0.000	-0.019	-0.016
Dest	-0.0512	0.001	-68.905	0.000	-0.053	-0.050
Distance	-0.0011	4.53e-06	-245.059	0.000	-0.001	-0.001
CarrierDelay	1.0326	0.000	8011.515	0.000	1.032	1.033
WeatherDelay	1.0251	0.000	3659.191	0.000	1.025	1.026
NASDelay	1.0508	0.000	6443.752	0.000	1.051	1.051
SecurityDelay	1.0738	0.002	495.352	0.000	1.070	1.078
LateAircraftDelay	1.0360	0.000	8195.482	0.000	1.036	1.036

Omnibus:	3423882.251	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	119756378.770
Skew:	-0.784	Prob(JB):	0.00
Kurtosis:	18.784	Cond. No.	1.29e+04

Figure 20: OLS Regression Result



```
# Fit a linear regression model to the training data
model = LinearRegression()
model.fit(X_train_encoded, y_train)

# Make predictions on the testing data
y_pred = model.predict(X_test_encoded)

# Evaluate the model using R-squared and F1-score
print("R-squared:", r2_score(y_test, y_pred))
print("F1-score:", f1_score(y_test > 0, y_pred > 0, average='weighted'))
print('Mean squared error:', mean_squared_error(y_test,y_pred))
```

✓ 5.8s

```
R-squared: 0.9479930112752952
F1-score: 0.7572451832665271
Mean squared error: 75.53233666243507
```

Figure 21: R-squared, F1-score and Mean squared error

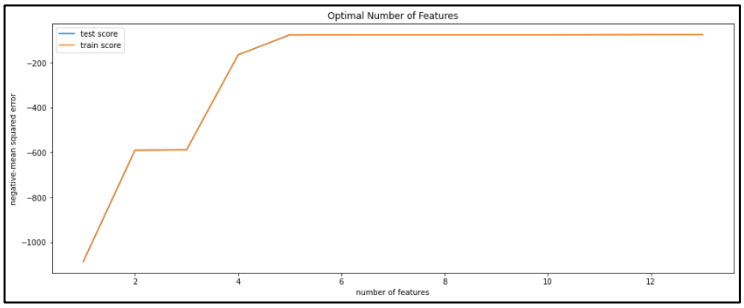


Figure 22: Optimal Number of Features

Machine learning is a powerful tool for predicting flight delays by analyzing data and identifying patterns. In this analysis, a Linear Regression model was used to model the relationship between a dependent variable (flight arrival delay) and multiple independent variables. The model was fit to the data after constructing a correlation matrix to avoid multicollinearity and dropping unnecessary high-correlation features. Firstly, a correlation matrix between each variable is constructed between each variable to avoid multicollinearity. High-correlation features were dropped to avoid redundancy and improve the accuracy of the model. After this pre-processing step, the Linear Regression model can be fit to the remaining data and used to make predictions. Figure 20 and 21 is the OLS Regression result. The model achieved an R-squared value of 0.95, indicating that 95% of the variance in the data is explained by the given predictors. F1 score of 0.76 indicates this model provides a good fit with the current predictors as compared to no predictor. All predictor had p-value below 0.05, stating their statistical significance. Cross-validation was used to fine-tune the feature selection. Based on Figure 22, cross validation suggests that current number of features is the best result. Hence, we conclude R-square, F1-score and Mean squared error are 0.95, 0.75 and 75 respectively with given predictors

'Month', 'DayOfWeek', 'DepTime', 'ArrTime', 'UniqueCarrier', 'Origin', 'Dest', 'Distance', 'CarrierDelay', 'WeatherDelay', 'NASDelay', 'SecurityDelay', 'LateAircraftDelay' in predicting 'ArrDelay'.

6. Conclusion

In conclusion, the given dataset is explored and cleaned before analysis and the best possible machine learning technique is used. We can conclude that most flights experienced little or no delay. However, flight delays can have negative impacts on passengers, resulting in wasted time and money. The primary cause of delay is the carrier delay, airline company should prioritize addressing this issue. The four questions are guidelines analysing the data. Based on the analysis, the best times to fly to avoid delays are between 5:00am and 6:00am, on Saturdays, and in November. Older planes tend to experience more delays than newer ones. The LIH and HNL airports are the busiest departure and destination points. It should be emphasized that delays can have cascading effects, causing further delays at other airports. Overall, this analysis provides valuable insights for airlines to improve their operations and minimize delays.

7. References

Yuemin Tang (October 2021) Airline Flight Delay Prediction Using Machine Learning Models https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725#:~:text=It%20can%20be%20seen%20that,the%20Decision%20Tree%20is%200.9778.

Stefanovič, P., Štrimaitis, R., & Kurasova, O. (2020). Prediction of flight TIME deviation for Lithuanian airports using supervised machine learning model. Computational Intelligence and Neuroscience, 2020, 1–10. https://doi.org/10.1155/2020/8878681

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/stable/about.html#citing-scikit-learn

Rob J Hyndman and George Athanasopoulos (April 2018) Forecasting: Principles and Practice (2nd ed) https://otexts.com/fpp2/