University of London

ST2195 Programming for Data Science

Coursework (Part 2)

Prepared by:

Elizabeth Leonny Efendi

(220657172)

Table of Contents

ASA Statistical Computing and Graphics Data Expo Analysis

The analysis utilizes the 2009 ASA Statistical Computing and Graphics Data Expo dataset spanning from 1995 to 2004. The data was sourced from CSV files and processed using both Python and R programming languages.

1. What are the Best Times and Days of the Week to Minimize Delays Each Year?

Note: All plots produced by both Python and R are the same, thus this section will only include any one of them.

1.1 Data Cleaning and Exploratory Data Analysis (EDA)

The data was collected from CSV files stored in a specified directory. The CSV files were imported and processed to create DataFrames containing the flight data. Thus, missing values were handled by dropping rows with NaN values in specific columns relevant to the analysis such as departure delay and arrival delay column. Additionally, a new column 'DepHour' was created to extract the hour from the scheduled departure time (CRSDepTime), and an 'AvgDelayTime' column was calculated as the average delay between arrival and departure delays.

1.2 Visualization and Analysis of Best Times to Minimize Delays Each Year

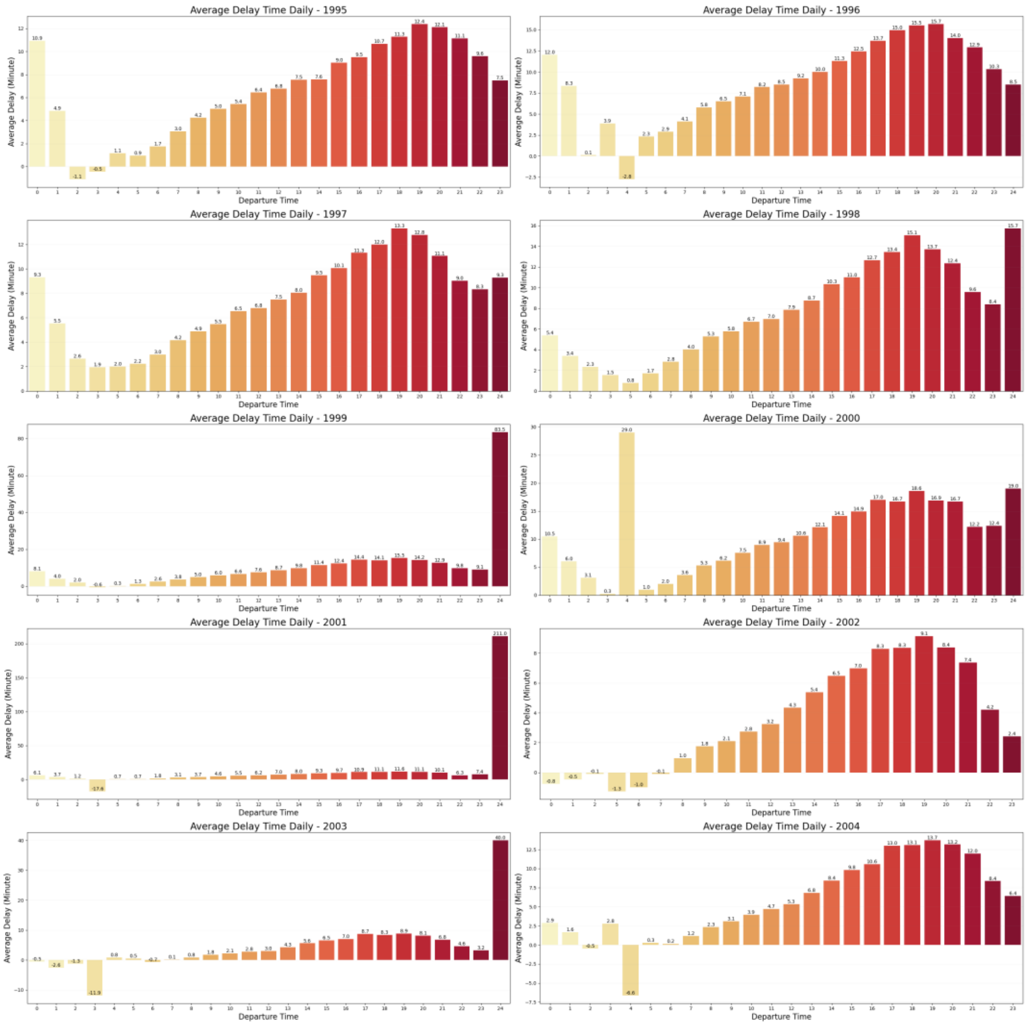
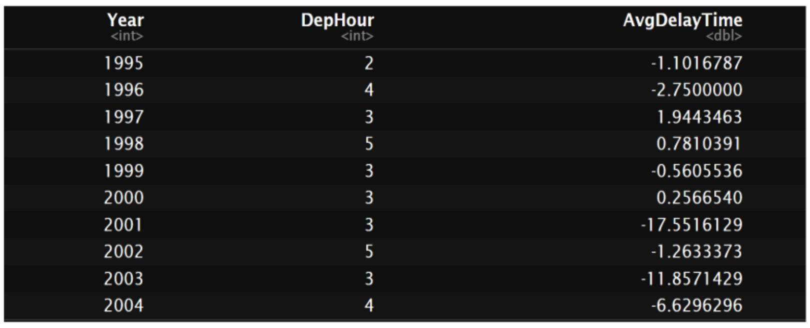


Figure 1: Best Times to Minimize Delays Each Year (Python)

The results for best times to minimize delays each year are visualized using bar plots as shown in *Figure 1*, with x-axis representing departure hours in 24-hours format and the y-axis representing the average delay time in minutes, which was calculated as the average delay between arrival and departure delays. Each bar on the plot represents the average delay for a specific hour of departure. The plots help to identify optimal departure times that align with lower delay probabilities, thus enhancing passenger experience and reducing operational disruptions. The graphs also show that in some years, between midnight and morning, the average delay time is negative. This indicated that, on average, flights were arriving or departing earlier than scheduled. This could be due to improved airline operations, better weather conditions, efficient scheduling, or other factors contributing to smoother flight experiences during those times. In 1995, 2 AM was identified as the time with the lowest average delay. Meanwhile, 4 AM in 1996 and 2004, and 5 AM in 1998 and 2002 was noted as the most optimal time of the year. Conversely, for the years 1997, 1999, 2000, 2001, and 2003, 3 AM showed the lowest average delay times, highlighting potential operational advantages during these early morning hours. Generally, morning hours stand out as optimal times with lower average delay times. However, certain years exhibit notable deviations from this pattern. In 2000, at 4 AM, flights were experiencing a substantial delay of 29 minutes, possibly due to increased air traffic or specific disruptions during that hour. Contrastingly, in 2001 at 3 AM, there was a significant negative delay of -17.6 minutes, meaning flights were departing or arriving early on average.



Year <int>	DepHour <int>	AvgDelayTime <dbl>
1995	2	-1.1016787
1996	4	-2.7500000
1997	3	1.9443463
1998	5	0.7810391
1999	3	-0.5605536
2000	3	0.2566540
2001	3	-17.5516129
2002	5	-1.2633373
2003	3	-11.8571429
2004	4	-6.6296296

Figure 2: Table Best Times to Minimize Delays Each Year (R)

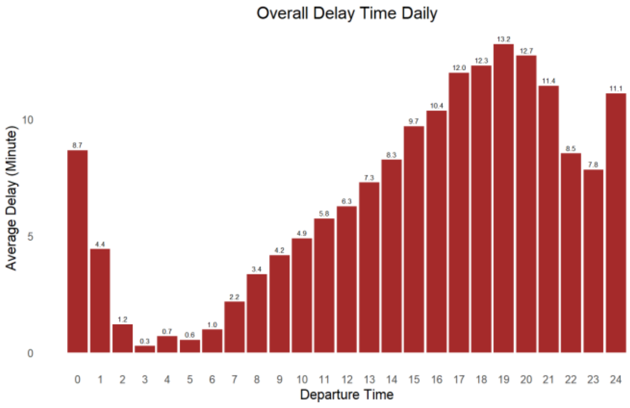


Figure 3: Overall Best Times to Minimize Delays (R)

Overall, the findings indicated a consistent pattern of lower average delay times during early **morning hours**, between **2 AM to 5 AM**. In most years, 3 AM is the best time to minimize delays, as shown in *Figure 2* and *Figure 3*. At this hour, there are also many instances of negative average delay times, indicating that there are few technical errors or disruptions in operations. Therefore, it is likely that these flights arrive or depart early.

1.3 Visualization and Analysis of Best Days to Minimize Delays Each Year

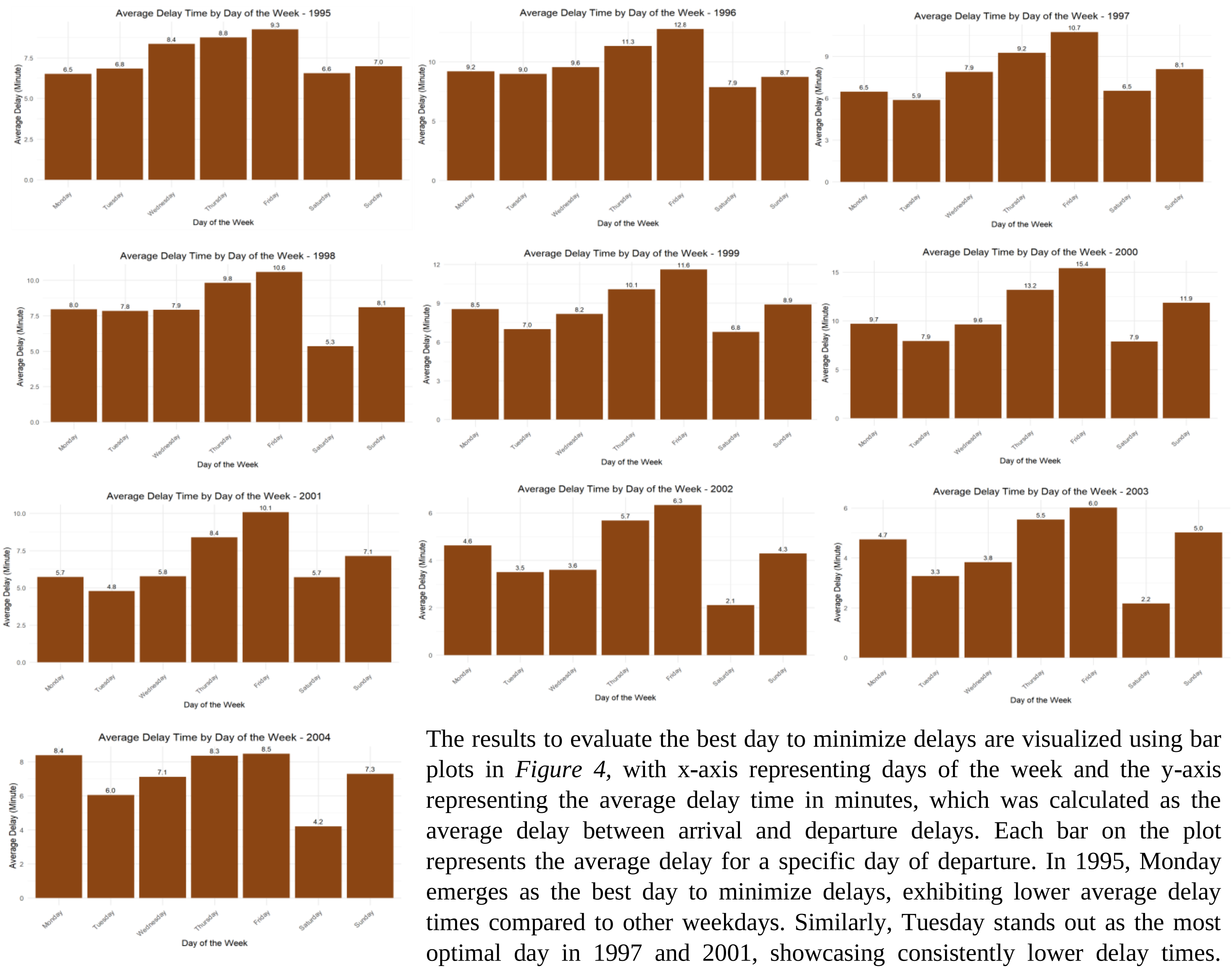
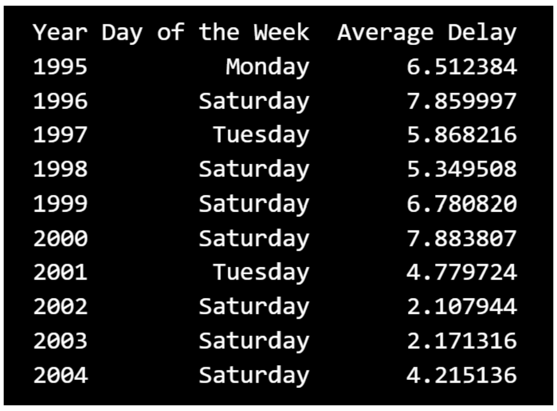


Figure 4: Best Day to Minimize Delays (R)

The results to evaluate the best day to minimize delays are visualized using bar plots in *Figure 4*, with x-axis representing days of the week and the y-axis representing the average delay time in minutes, which was calculated as the average delay between arrival and departure delays. Each bar on the plot represents the average delay for a specific day of departure. In 1995, Monday emerges as the best day to minimize delays, exhibiting lower average delay times compared to other weekdays. Similarly, Tuesday stands out as the most optimal day in 1997 and 2001, showcasing consistently lower delay times. However, the trend shifts in the remaining years, 1996, 1998, 1999, 2000, 2002, 2003, and 2004, with Saturday consistently demonstrating the lowest average delay times compared to other weekdays. disruptions.



```
Year Day of the Week  Average Delay
1995          Monday       6.512384
1996        Saturday       7.859997
1997         Tuesday       5.868216
1998        Saturday       5.349508
1999        Saturday       6.780820
2000        Saturday       7.883807
2001         Tuesday       4.779724
2002        Saturday       2.107944
2003        Saturday       2.171316
2004        Saturday       4.215136
```

Figure 5: Table Best Days to Minimize Delays Each Year (Python)

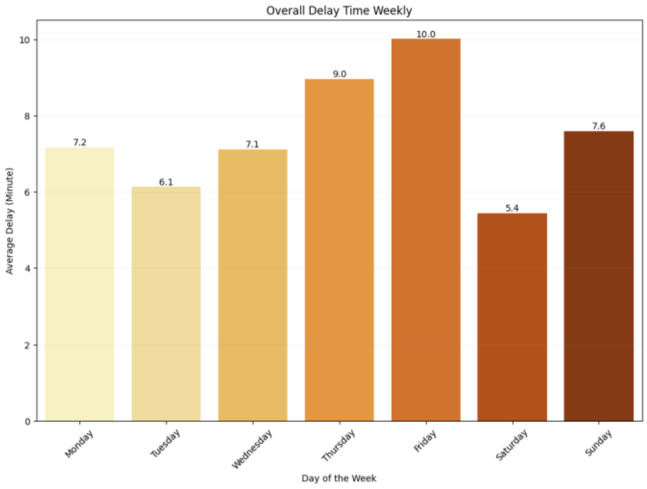


Figure 6: Overall Best Day to Minimize Delays (Python)

Overall, **Saturdays** consistently emerge as the best day to minimize delays across all years, as it consistently has the lowest average delay time. On Tuesdays, the decrease in business-related flights allows for smoother airport operations, while airlines resolve early-week issues. This mid-week stability contributes to fewer delays. Similarly, Saturdays see reduced business and commuter traffic, with a focus on leisure travel, leading to less congestion and more flexible schedules for travelers. Proactive maintenance scheduling and optimized resource allocation further contribute to these days experiencing lower delays, making them preferable options for scheduling flights with minimal disruptions. The findings highlight the significance of adopting a proactive approach to address potential challenges and capitalize on opportunities for smoother airport operations. Ultimately, prioritizing days like Tuesday and Saturday for flight scheduling can lead to a more seamless travel experience for passengers while mitigating operational disruptions.

2. Evaluate Whether Older Planes Suffer More Delays on a Year-to-year Basis

2.1 Data Cleaning and Exploratory Data Analysis (EDA)

The supplementary dataset used in this analysis is the "plane-data.csv" file, which contains information about planes, including their manufacturing years. Upon inspecting the 'year' column, it was observed that some entries were incorrect such as 'None' and '0000'. These were treated as missing values and handled accordingly. The cleaned plane-data file is merged with flight data based on a common identifier, the tail number on each plane. This merging process allows us to combine information about each plane's manufacturing year with flight-specific data, such as average delays. The average delay time was determined by averaging arrival and departure delays.

2.2 Visualization and Analysis

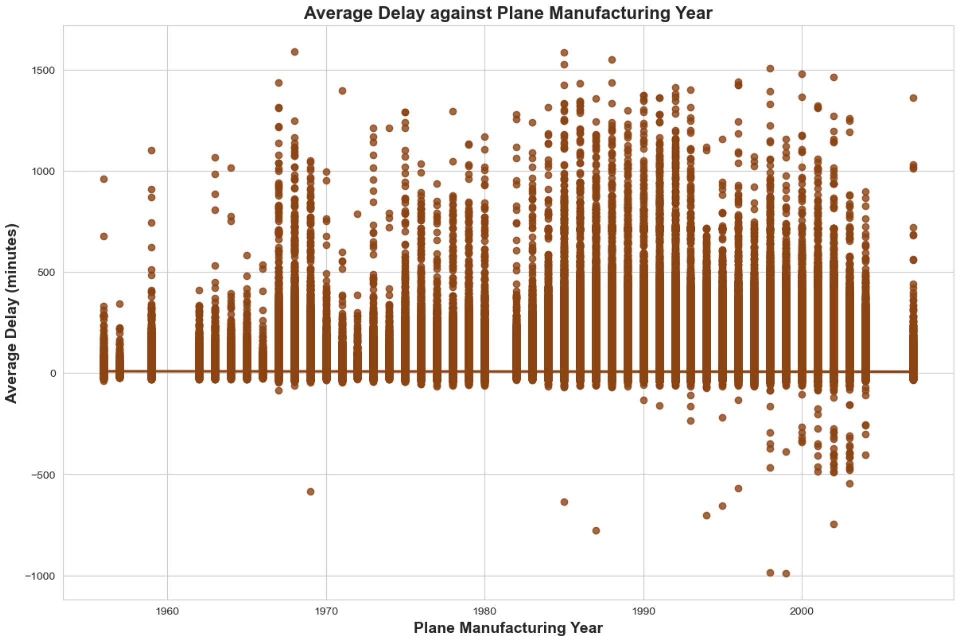


Figure 7: Plane Ages vs Average Delays (Python)

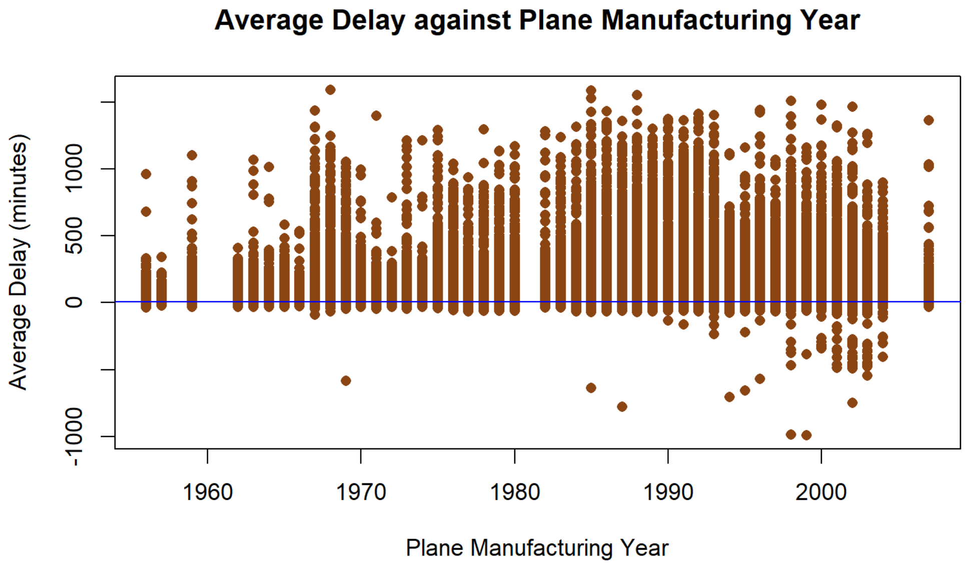
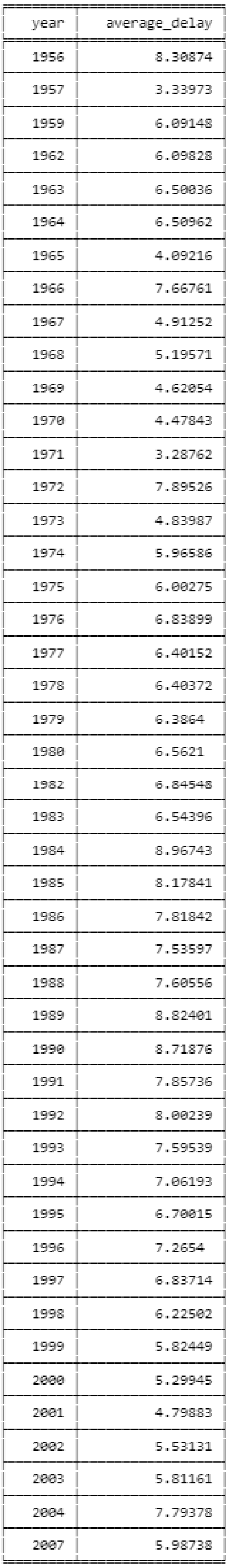


Figure 8: Plane Ages vs Average Delays (R)

A scatter plot with a regression line was generated to visualize the relationship between plane manufacturing year and average delay in *Figure 7* and *Figure 8*. The analysis outcome suggests that there is **no linear correlation** between the duration of delays and the year of a plane's manufacture. It can tentatively be concluded that there may not be a significant relationship between a plane's age (manufacturing year) and the likelihood of experiencing more delays.

This suggests that older planes do not necessarily suffer markedly more delays compared to newer ones. However, the weak correlation observed may be influenced by other variables, such as flight routes, weather conditions, airport congestion, air traffic control, etc.

Figure 9 provides a detailed breakdown of the average delay associated with each plane's manufacturing year.



year	average_delay
1956	8.30874
1957	3.33973
1959	6.09148
1962	6.09828
1963	6.50036
1964	6.50962
1965	4.09216
1966	7.66761
1967	4.91252
1968	5.19571
1969	4.62054
1970	4.47843
1971	3.28762
1972	7.89526
1973	4.83987
1974	5.96586
1975	6.00275
1976	6.83899
1977	6.40152
1978	6.40372
1979	6.3864
1980	6.5621
1982	6.84548
1983	6.54396
1984	8.96743
1985	8.17841
1986	7.81842
1987	7.53597
1988	7.60556
1989	8.82401
1990	8.71876
1991	7.85736
1992	8.00239
1993	7.59539
1994	7.06193
1995	6.70015
1996	7.2654
1997	6.83714
1998	6.22502
1999	5.82449
2000	5.29945
2001	4.79883
2002	5.53131
2003	5.81161
2004	7.79378
2007	5.98738

Figure 9: Table Plane Ages vs Average Delays (Python)

3. Logistic Regression Model for the Probability of Diverted US Flights

Note: Most plots produced by both Python and R are similar, thus this section will only include one of them.

3.1 Data Cleaning and Exploratory Data Analysis (EDA)

The data is collected from CSV files containing flight information such as departure time, arrival time, distance, carrier details, etc. Due to RAM limitation, the retrieved data is limited to 2 million rows. Supplementary dataset is also collected including airport information. Missing data is handled, which replaces missing numeric values with the median. Categorical variables (e.g., UniqueCarrier) are transformed into numerical values for modeling. The airport data is then merged into the main dataset based on the ‘Origin’ and ‘Dest’ airport codes, creating new features for origin and destination latitude and longitude. Also, relevant features for the predictive model are selected, including departure time, arrival time, distance, latitude, longitude, carrier information, and year. The Y-axis represents the coefficient values obtained from the logistic regression model.

3.2 Model Building and Training

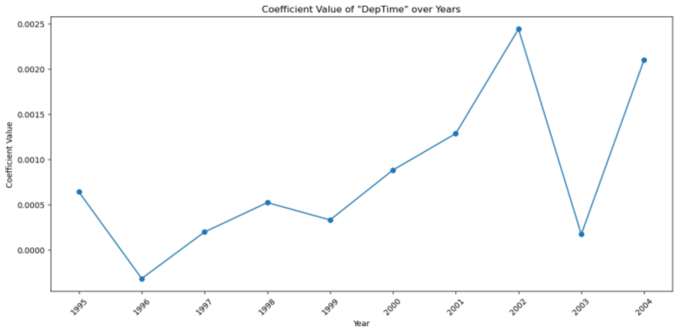
A step-by-step process was created to prepare our data using a tool called Scikit-Learn. This process includes filling in missing information and turning categories like airline names into numbers that can be understood better by computers. Here, logistic regression model is used because it is effective at predicting outcomes that have two possibilities, such as “yes” or “no”.

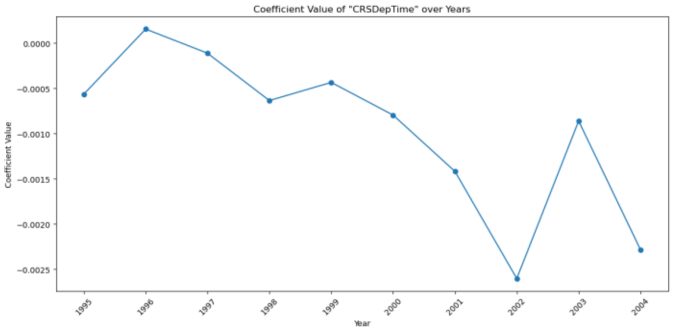
During the train-test split, the data is divided into two parts: one for teaching the model how to predict (training), and another to assess its performance (testing). Approximately 80% of the data is used for the training, and the remaining 20% is for testing. For model training, the model is trained for each important piece of information, such as departure time, separately for each year between 1995 to 2004. Subsequently, the model’s predictions are analyzed to understand the impact of each factor on the likelihood of a flight being diverted.

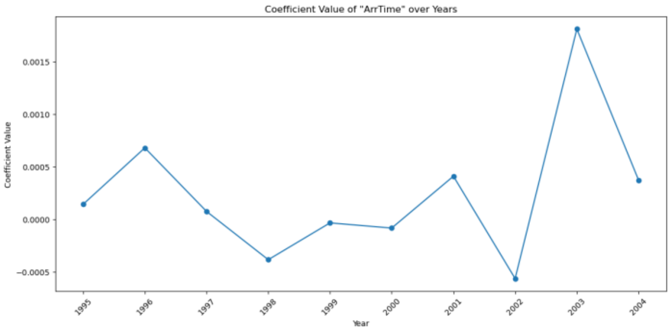
3.3 Differences in R and Python Code

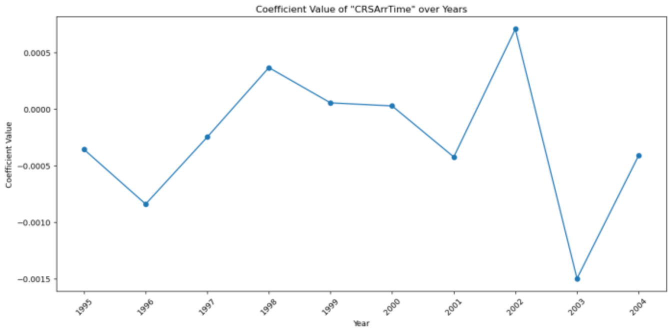
There are slight differences in coefficient values between the R and Python plots, which are due to how each modeling environments process information and makes predictions. Each environment uses slightly different methods to analyze the data and make predictions, which can result in different coefficient values. Each environment (R and Python) uses its own algorithms and approaches to build predictive models. While both R and Python utilize logistic regression for these plots, the underlying calculations and optimizations can differ. Also, during the model training process, each modelling environment may prioritize different aspects of the data, even when given the same dataset. This variation can result in different model outcomes.

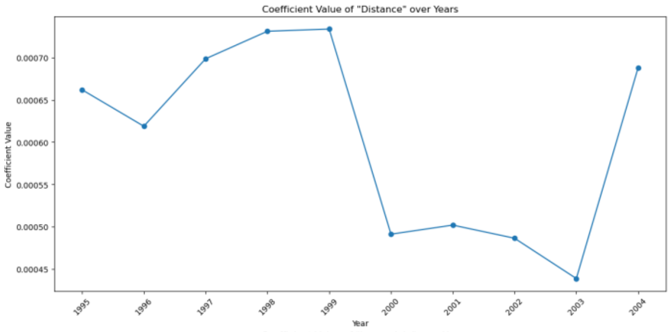
3.4 Visualization and Analysis

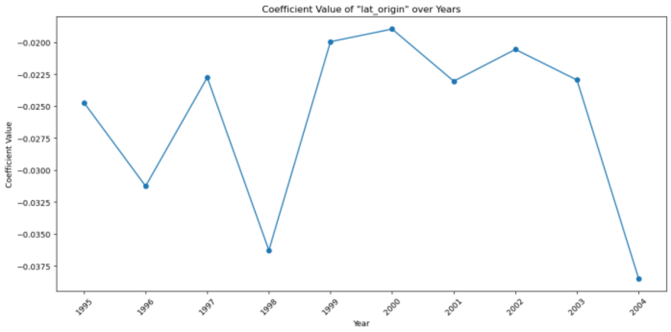


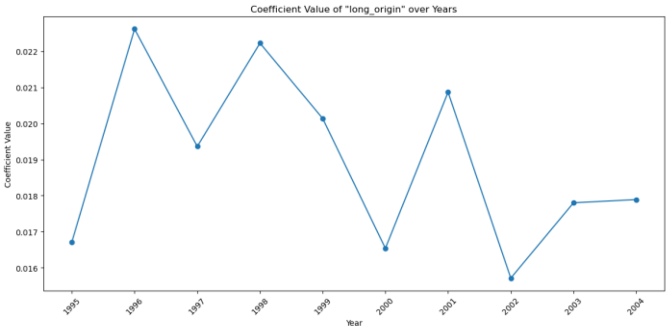


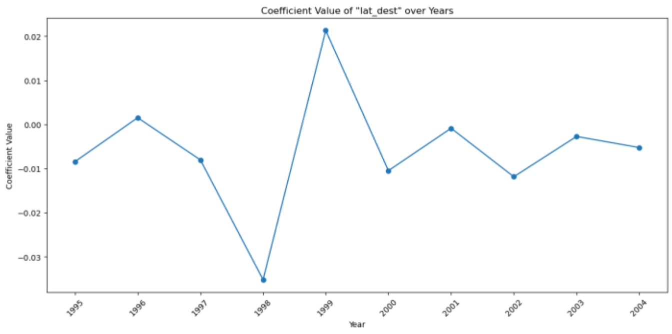












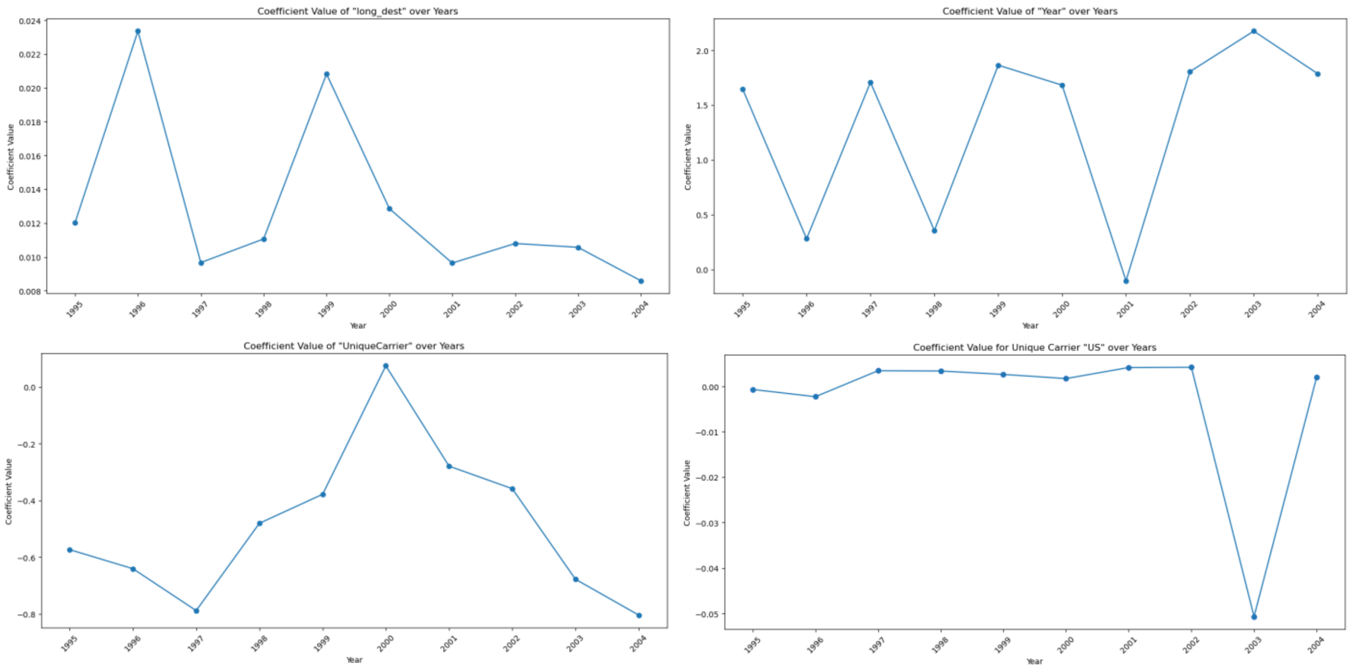


Figure 10: Logistic Regression Model (Python)

Each graph represents the trend of a specific feature's impact on flight diversion over the years. The graphs show the coefficient values of key features such as:

1. DepTime: Actual departure time
2. CRSDepTime: Scheduled departure time
3. ArrTime: Actual arrival time
4. CRSArrTime: Scheduled arrival time
5. Distance: Distance covered by a flight between its origin and destination airports
6. Lat_origin: Geographical latitude coordinate of the flight's origin airport
7. Long_origin: Geographical longitude coordinate of the flight's origin airport
8. Lat_dest: Geographical latitude coordinate of the flight's destination airport
9. Long_dest: Geographical longitude coordinate of the flight's destination airport
10. Year: The year in which the flight took place
11. UniqueCarrier: Unique identifier for the airline operating the flight

The coefficient value for **DepTime** exhibits a mostly positive trend, indicating that departures at certain times are likely to result in flight diversions. Notably, the coefficient dips slightly below zero in 1996 but then rises to its highest in 2002. This suggests that flights departing around the peak times in 2002 were more prone to diversions, while flights in 1996 are less likely to be diverted. In contrast, the coefficient values for **CRSDepTime** show a mostly negative trend. This implies that departures scheduled at specific times were less likely to experience diversions. However, there's a spike in 1996 where the coefficient rises slightly above zero, and the lowest coefficient is observed in 2002, indicating a period where scheduled departures were particularly less prone to diversions.

The coefficient values for **ArrTime** experienced fluctuations over the years. 2002 stands out as the year with the lowest coefficient, suggesting that flights arriving at certain times during that year were less likely to be diverted. Conversely, 2003 shows the highest coefficient value, indicating a higher likelihood of diversions for flights arriving at specific times during that year. **CRSArrTime** also shows variations in coefficient values. The highest coefficient is observed in 2002, indicating increased diversion likelihood for flights with scheduled arrivals during that year. On the other hand, 2003 displays the lowest coefficient value, implying a comparatively lower probability of diversions for flights with scheduled arrivals during that period.

The coefficient values for flight **Distance** exhibit a fluctuating pattern over the years. Notably, there's a sharp drop in coefficient value in 2000, followed by a decreasing trend until 2003, and then a notable increase in 2004. This suggests that flight distance had varying impacts on the likelihood of diversions across different years. The coefficient values for **lat_origin** and **long_origin** exhibit zigzag patterns but different ranges of coefficient value. Specifically, lat_origin shows negative coefficients which are more likely to be diverted, while long_origin shows positive coefficients which less likely to experience diversions.

Similar observations can be made for **lat_dest** and **long_dest**. These coefficients also show mixed trends, reflecting the complexity of geographical factors in flight diversion predictions. Lat_dest exhibit zigzag pattern in both positive and negative ranges. This indicates that flights destined for specific latitudes or longitudes had varying probabilities of diversions across different years. On the other hand, long_dest also shows a zigzag pattern but only in positive range. The coefficient values for the **Year** feature show a predominantly positive range, except for 2001. This implies that the year of operation had an impact on diversion likelihood, with certain years experiencing higher probabilities of diversions.

The overall trend for **UniqueCarriers** shows a mountain-like shape with mostly in the negative value range, indicating varying impacts on diversion likelihood. For instance, the lowest coefficients in 1997 and 2004 suggest that flights operated by specific carriers during those years were less likely to be diverted. Conversely, the highest coefficient in 2000 indicates a period where flights by certain carriers were more prone to diversions. The trend for **carrier 'US'** shows a stable probability of flights being diverted from 1995 to 2002 at around 0. However, in 2003, there was a notable increase in diversion likelihood, which returned to previous levels by 2004. This suggests temporary challenges for 'US' flights in 2003, followed by a recovery.

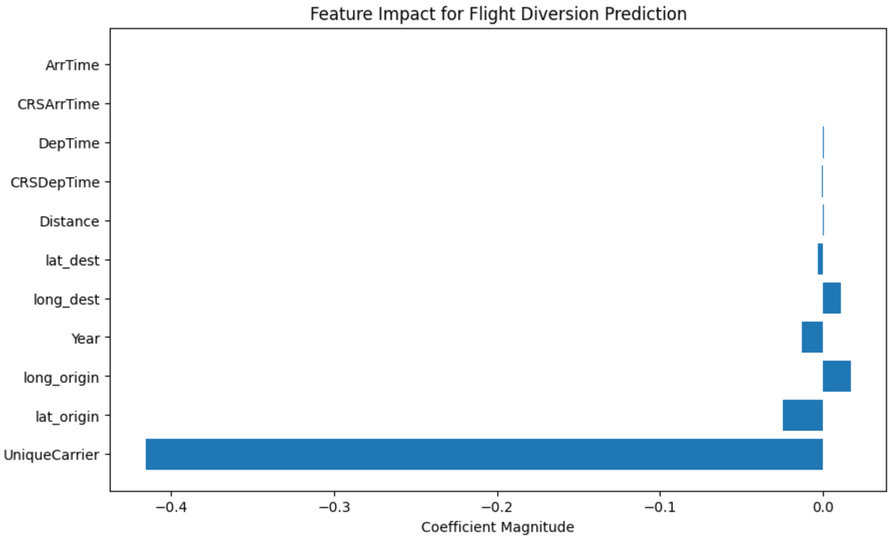
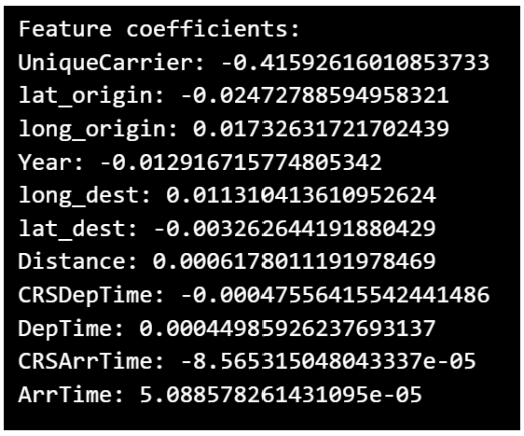


Figure 11: Feature Impact for Flight Diversion Prediction (Python)



```
Feature coefficients:
UniqueCarrier: -0.41592616010853733
lat_origin: -0.02472788594958321
long_origin: 0.01732631721702439
Year: -0.012916715774805342
long_dest: 0.011310413610952624
lat_dest: -0.003262644191880429
Distance: 0.0006178011191978469
CRSDepTime: -0.00047556415542441486
DepTime: 0.00044985926237693137
CRSArrTime: -8.565315048043337e-05
ArrTime: 5.088578261431095e-05
```

Figure 12: Feature Overall Coefficients (Python)

As shown in *Figure 11* and *Figure 12*, a UniqueCarrier in general was strongly correlated with a reduced likelihood of flight diversion, as indicated by its substantial negative coefficient. Additionally, the variables showing positive coefficients (indicating an increasing likelihood of diversion) are 'long_origin' and 'long_dest'. The positive coefficients on both these variables suggest that flight routes with some longitudes, both at the origin and destination airports, tend to have a higher likelihood of being diverted. On the other hand, certain variables such as the lat_origin, lat_dest, Distance, CRSDepTime, ArrTime, etc. were found to have relatively minor impacts on flight diversion prediction. These features exhibited coefficients close to zero or with negligible effects, suggesting a limited influence on the likelihood of flight diversion according to this logistic regression model.

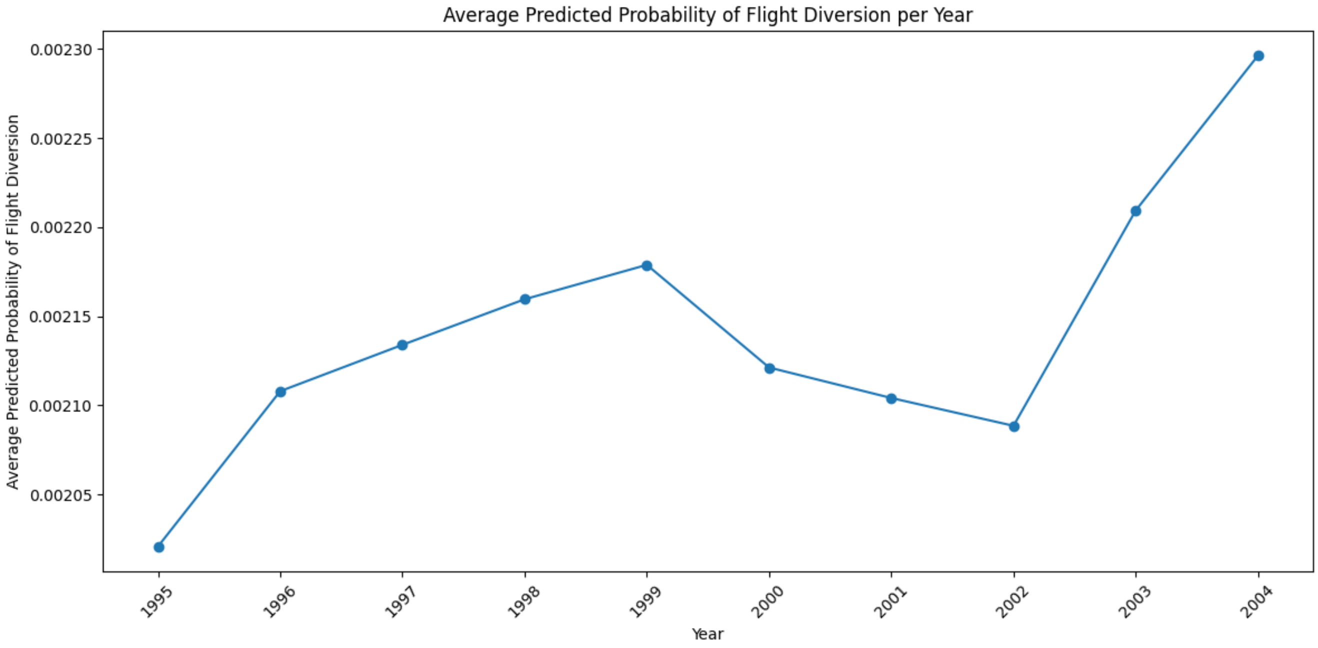


Figure 13: Average Predicted Probability of Flight Diversion per Year (Python)

As shown in *Figure 13*, our analysis using a predictive model revealed that flights in the year 2004 had the highest average predicted probability of experiencing diversion, with an average probability of approximately 0.0023 per flight, equivalent to a 0.23% chance. This finding suggests that conditions or factors specific to 2004 may have contributed to increased flight diversions.

4. Conclusion

4.1 Best Times and Days to Minimize Delays

- The best times to minimize delays are typically in the early morning (2 AM – 5 AM), especially around 3 AM.
- The best days to minimize delays are Saturdays, as it consistently shows the lowest average delays, followed by Tuesdays and Mondays in certain years.

4.2 Evaluation of Plane Manufacturing Year and Delay Correlation

- No linear correlation. Plane age does not necessarily cause more delays than newer planes.

4.3 Logistic Regression Model for Diverted Flights

- All factors, including departure time, arrival time, flight distance, and geographical coordinates (latitudes and longitudes), affect diversions, but the extent of their influence varies.
- The analysis of various features' impacts on flight diversion revealed distinct trends over the years. Feature longitude of both origin and destination airports exhibited a positive trend, and emerged as the most influential factor, indicating an increasing likelihood of diversions.
- The year 2004 stands out as a period with heightened average predicted probabilities of flight diversion.