ST2195 COURSEWORK REPORT

Programming for Data Science

LORENCIA LO

STUDENT NUMBER : 220458845

Table of Contents

Introduction... 3

(a). What are the best times and days of the week to minimise delays each year?...................... 4

(b). Evaluate whether older planes suffer more delays on a year-to-year basis......................... 6

(c) For each year, fit a logistic regression model for the probability of diverted US flights using as many features as possible from attributes of the departure date, the sched- uled departure and arrival times, the coordinates and distance between departure and planned arrival airports, and the carrier. Visualize the coefficients across years.................................... 7

Conclusion.. 9

Introduction

This report examines the dataset of flight arrivals and departures from the 2009 ASA Statistical Computing and Graphics Data Expo. with a specific emphasis on the consecutive years spanning 1995 to 2004. It comprehensively outlines the process of data analysis, encompassing tasks such as data extraction, manipulation, and the generation of both tables and visual representations.

(a). What are the best times and days of the week to minimise delays each year?

To assess the optimal times and weekdays for minimizing delays, our analysis centers on identifying the time frames with the lowest average delay in minutes. This serves as a key indicator of the most efficient window for air travel, facilitating timely and smooth journeys.

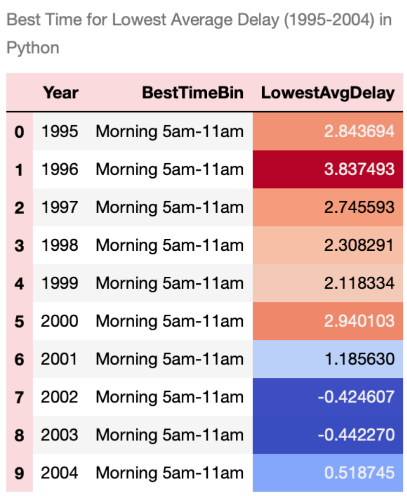
The process begins with sourcing data from CSV files named by year, stored within a directory named `dataverse_files`, containing flight data for the specified years, including departure and arrival delays. Both Python and R scripts are employed to read the data, with Python using `pandas.read_csv()` and R using `read_csv()` from the `tidyverse` package along with `lubridate` for handling dates and times.

Both scripts clean data by removing rows with missing values to maintain accuracy. Python uses the `dropna()` method to expunge rows with any missing values, which purges incomplete records from the dataset. In contrast, R leverages `drop_na()` after mutating and transforming the dataset to remove only the rows missing crucial data points required for the analysis. This ensures that every calculation of average delays operates on a complete case, thereby preserving the integrity of the dataset for an accurate identification of the best times. Additionally, R strategically excludes NA values from the average delay calculations with the parameter `na.rm = TRUE` during the summarization phase, which ensures that NAs do not skew the results, rather than prematurely eliminating data rows.

Departure times are categorized into Midnight (12am-5am), Morning (5am-11am), Noon (11am-5pm), and Night (5pm-12am) using functions in both languages. This categorization helps in calculating the average delay within each time slot, identifying the best travel times. In Python, this categorization is facilitated by applying the function to the departure hour extracted from `DepTime`, while R uses `case_when()` within a `mutate()` call for similar categorization, handling NA values explicitly.

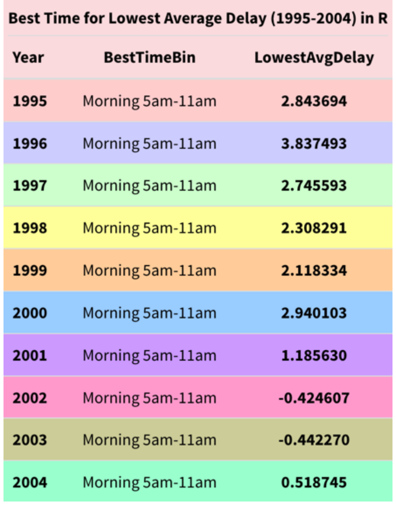
Both scripts calculate the average delay for flights within each time bin by averaging departure and arrival delays, aiming to find the time bin with the lowest delay. Python and R use ‘idxmin()’ and ‘slice_min()’, respectively, to select the time with the smallest average delay. Additionally, to find the best day for minimizing delays, both languages incorporate a new column, ‘AvgDelay’, which averages departure and arrival delay times per flight. This leads to the crucial identification of the best day for travel by pinpointing the day with the lowest historical average delay. Both Python and R further enhance interpretability by transforming numeric day codes into their respective weekday names.

Finally, visualization plays a pivotal role in presenting the findings. Python leverages ‘matplotlib’, while R utilizes ‘ggplot2’, both providing visually appealing and informative representations of the data to aid in decision-making and understanding.



Best Time for Lowest Average Delay (1995-2004) in Python

	Year	BestTimeBin	LowestAvgDelay
0	1995	Morning 5am-11am	2.843694
1	1996	Morning 5am-11am	3.837493
2	1997	Morning 5am-11am	2.745593
3	1998	Morning 5am-11am	2.308291
4	1999	Morning 5am-11am	2.118334
5	2000	Morning 5am-11am	2.940103
6	2001	Morning 5am-11am	1.185630
7	2002	Morning 5am-11am	-0.424607
8	2003	Morning 5am-11am	-0.442270
9	2004	Morning 5am-11am	0.518745

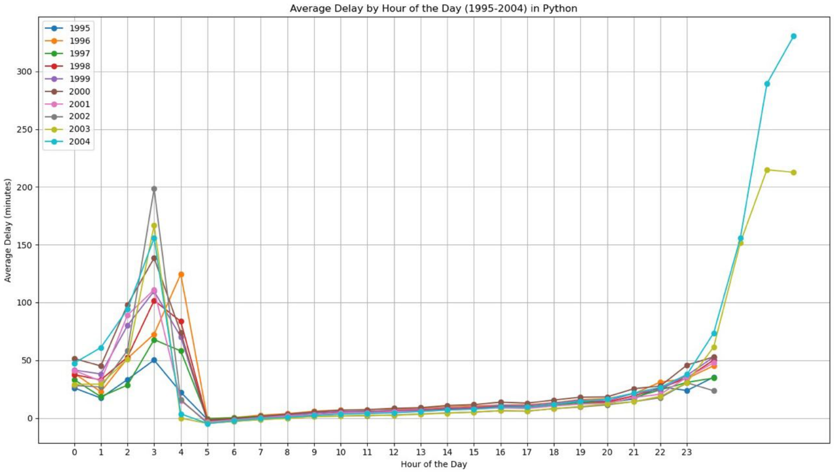


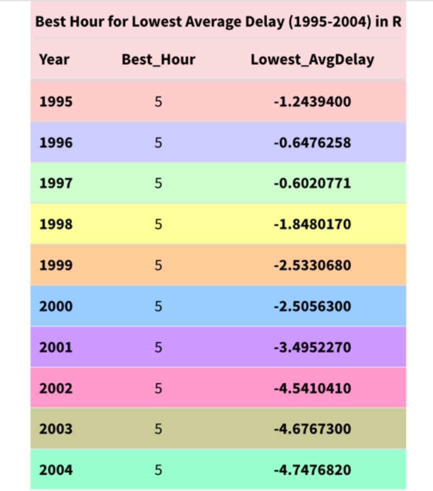
Best Time for Lowest Average Delay (1995-2004) in R

Year	BestTimeBin	LowestAvgDelay
1995	Morning 5am-11am	2.843694
1996	Morning 5am-11am	3.837493
1997	Morning 5am-11am	2.745593
1998	Morning 5am-11am	2.308291
1999	Morning 5am-11am	2.118334
2000	Morning 5am-11am	2.940103
2001	Morning 5am-11am	1.185630
2002	Morning 5am-11am	-0.424607
2003	Morning 5am-11am	-0.442270
2004	Morning 5am-11am	0.518745

The dataset covers the years 1995 to 2004 and identifies the most advantageous times and specific hours for achieving the least average delay. Throughout this period, **the early morning hours from 5 AM to 11 AM** regularly offer **the smallest average delays** compared to other times. Yearly, the minimum average delays within this timeframe start at 2.84 minutes in 1995 and gradually reduce to 2.12 minutes by 1999.

From 2002 onward, the delays turn negative, indicating arrivals that are earlier than scheduled, with delays recorded at -0.42 minutes in 2002, worsening slightly to -0.44 minutes in 2003, and then improving to 0.52 minutes in 2004.

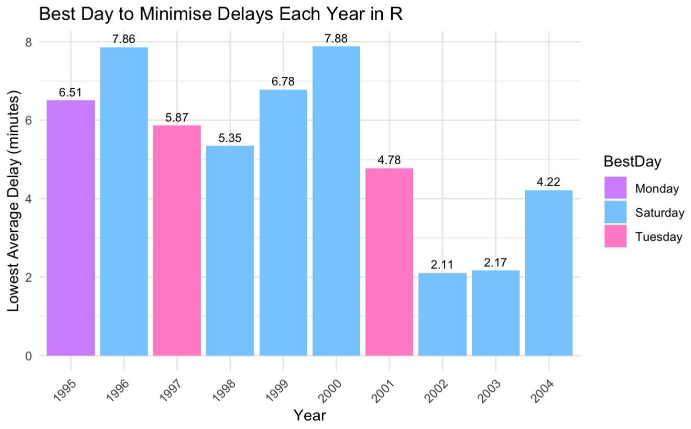




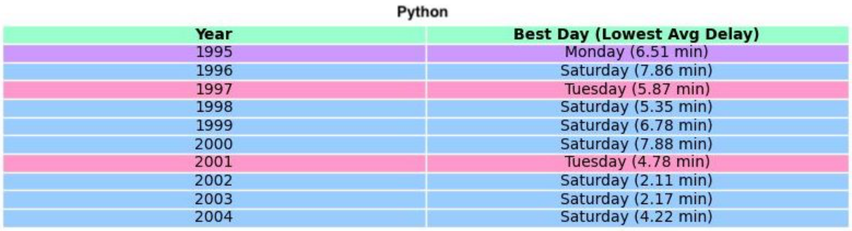
Best Hour for Lowest Average Delay (1995-2004) in R

Year	Best_Hour	Lowest_AvgDelay
1995	5	-1.2439400
1996	5	-0.6476258
1997	5	-0.6020771
1998	5	-1.8480170
1999	5	-2.5330680
2000	5	-2.5056300
2001	5	-3.4952270
2002	5	-4.5410410
2003	5	-4.6767300
2004	5	-4.7476820

Further specificity is provided for **the best hour** within this range, pinpointing **5 AM** as the time with the lowest average delay for each respective year. The 'Lowest_AvgDelay' column shows negative values, suggesting that flights during this hour typically depart earlier than scheduled. The values range from -0.6476258 in 1996 to -4.7676820 in 2003, illustrating an increasing trend towards earlier departures as years progress.



The analysis of flight delay data spanning from 1995 to 2004 reveals intriguing insights into the best days for minimizing average delays across each year. In 1995, **Monday** emerges as the optimal day, boasting the lowest average delay of 6.51 minutes. Subsequent years present varying outcomes, with **Saturday** dominating as the preferred day in 1996, 1998, 1999, 2000, 2002, 2003, and 2004, showcasing average delays ranging from 2.11 to 7.88 minutes. Conversely, 1997 and 2001 showcase **Tuesday** as the best day, with average delays of 5.87 and 4.78 minutes, respectively. Notably, these findings underscore the fluctuating nature of flight delays over different days of the week throughout the years under analysis. Such insights can inform strategic decision-making in travel planning and scheduling, aiding stakeholders in mitigating delays and optimizing operational efficiency.

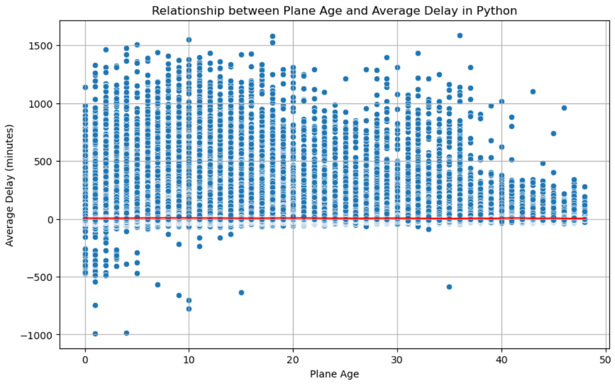


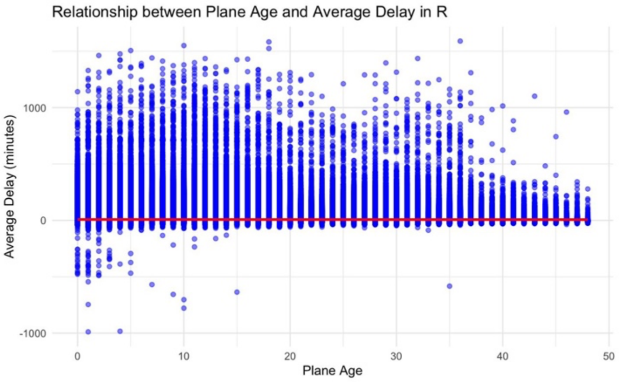
Python

Year	Best Day (Lowest Avg Delay)
1995	Monday (6.51 min)
1996	Saturday (7.86 min)
1997	Tuesday (5.87 min)
1998	Saturday (5.35 min)
1999	Saturday (6.78 min)
2000	Saturday (7.88 min)
2001	Tuesday (4.78 min)
2002	Saturday (2.11 min)
2003	Saturday (2.17 min)
2004	Saturday (4.22 min)

(b). Evaluate whether older planes suffer more delays on a year-to-year basis.

Two analytical approaches are utilized in the investigation: a scatter plot and Pearson correlation coefficient calculation. The scatter plot illustrates the potential connection between airplane age and the average delay they encounter, whereas the Pearson correlation coefficient quantifies the strength of this relationship in a quantitative manner.





The analysis commences with the loading of flight data covering the years 1995 to 2004, alongside plane data, retrieved from CSV files stored in the designated directory. Upon ensuring data integrity by removing missing values from the plane dataset, a merge operation is performed between the flight and plane data, facilitated by the 'TailNum' identifier. Standardization of the 'Year' column in the flight data to align with the 'year' column in the plane data ensures consistency across datasets. Subsequently, data filtering is applied to eliminate rows with 'None' values and exclude entries with plane years falling outside the operational range of 1950 to 2024. These meticulous data preparation steps guarantee the utilization of relevant and accurate data for subsequent analysis.

Following data preparation, the analysis moves towards investigating the relationship between plane age and average delay. This involves computing the plane age by subtracting the plane's manufacturing year from the year of operation, and calculating the average delay for each flight as the mean of departure and arrival delays. Visualization of the relationship is achieved through a scatter plot, illustrating individual flight data points with plane age on the x-axis and average delay on the y-axis. Additionally, a trend line is overlaid on the scatter plot to provide insights into the overall trend.

Based on the scatter plot, there is not a strong or significant linear relationship between the age of planes and the average delay of flights. This implies that older planes do not necessarily experience more delays. However, it's worth noting that the trend line appears flat, indicating that the average delay does not noticeably increase or decrease with plane age.

```
Pearson correlation coefficient: -0.00603591806031737
```

Furthermore, upon computing the Pearson correlation coefficient between plane age and delay within the flight data frame, the obtained value of approximately -0.006 indicates an exceedingly weak negative linear correlation. This suggests that a higher plane age does not necessarily result in increased delay occurrences.

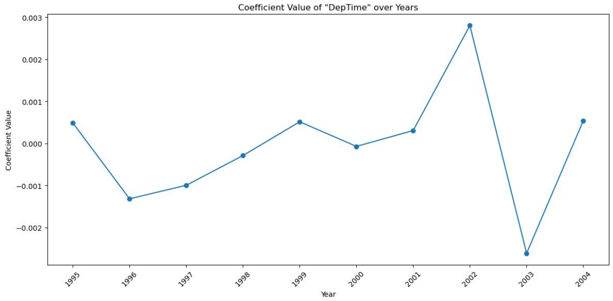
(c) For each year, fit a logistic regression model for the probability of diverted US flights using as many features as possible from attributes of the departure date, the sched- uled departure and arrival times, the coordinates and distance between departure and planned arrival airports, and the carrier. Visualize the coefficients across years.

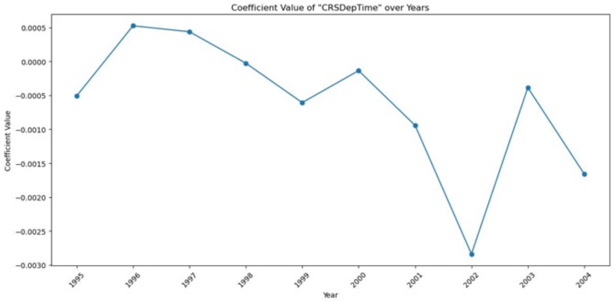
This part focuses on constructing and interpreting logistic regression models to predict flight diversions, utilizing a comprehensive set of features to capture the multifaceted nature of flight operations.

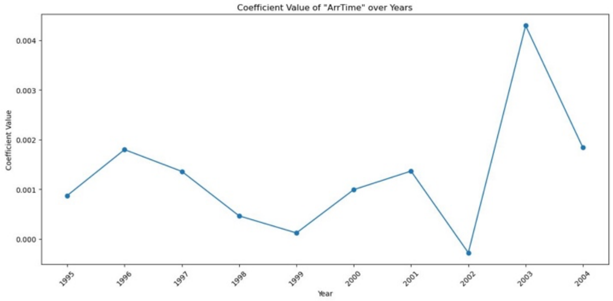
The report outlines the procedures for managing and analyzing a dataset on US flight operations from 1995 to 2004, focusing on flight diversions. The project begins with data acquisition and preparation, which sets the foundation for further analysis. This step involves loading the flight data and a separate airports dataset. Due to the dataset's size, the analysis is limited to the first 1,000,000 rows for each year. Selected features for modeling include departure and scheduled and arrival times, distance, geographical coordinates of the origin and destination airports, and the unique carrier identifier. The dependent variable indicates whether a flight is diverted and is the binary outcome for logistic regression models.

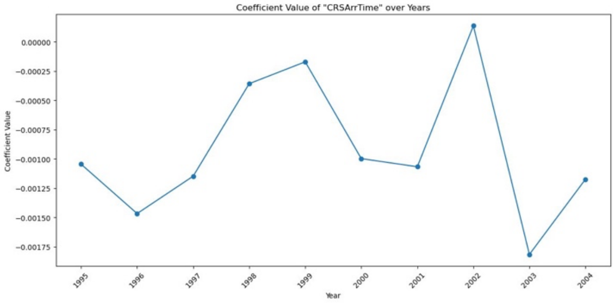
Data cleaning and merging are significant parts of data preparation. The airports dataset, containing latitude and longitude coordinates for airports identified by IATA codes, merges with the flight data. This merge happens twice for each flight entry, once for the origin airport coordinates and once for the destination. This requires careful renaming of latitude and longitude columns to differentiate between origin and destination coordinates, adding essential geographical information for the analysis. The dataset then splits into training and testing sets to evaluate model performance on unseen data. This split follows the standard practice in data science and ensures both subsets represent the full dataset.

The analysis fits logistic regression models to the data, chosen for the binary nature of the target variable (flight diversion). Models fit separately for each year to account for temporal variations in the predictors' influence on the likelihood of flight diversion. This method examines the stability and significance of each feature over time and addresses challenges in model fitting, like data sparsity and convergence, with strategic model adjustments and careful data subset handling.

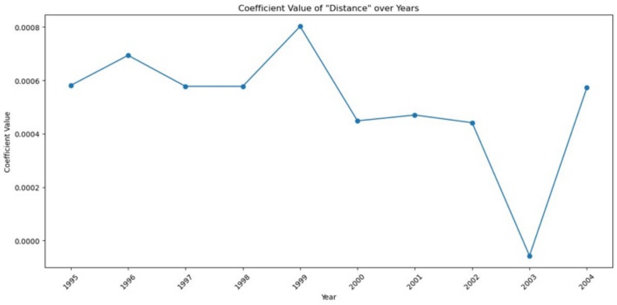

Coefficient Value of "DepTime" over Years


Coefficient Value of "CRSDepTime" over Years

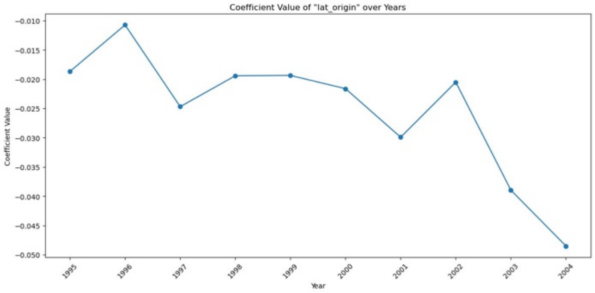

Coefficient Value of "ArrTime" over Years

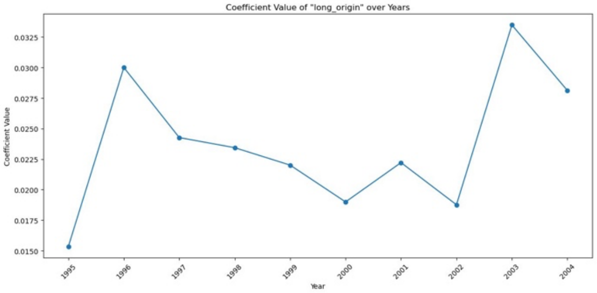

Coefficient Value of "CRSArrTime" over Years

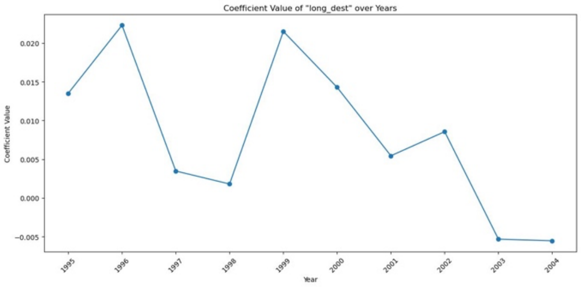
The collection of plotted coefficients represents a decade-long analysis of factors influencing the probability of flight diversions in the US. **Temporal variables such as departure and arrival times, both scheduled ('CRSDepTime', 'CRSArrTime') and actual ('DepTime', 'ArrTime'), exhibit fluctuations in their coefficients year over year,** indicating that the timing of flights bears a variable influence on diversion likelihood. These fluctuations might be reflective of peak airport congestion times, changes in air traffic patterns, or alterations in airline scheduling practices.

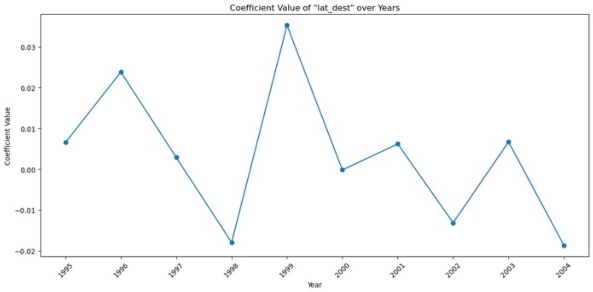


The graph of the distance coefficient illustrates changes that may indicate how flight duration and the associated exposure to variable conditions can influence diversion risk**. A longer flight has more opportunities for issues that could lead to diversion**, such as changes in weather or operational challenges.

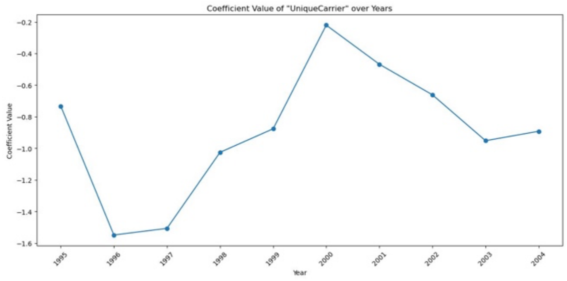


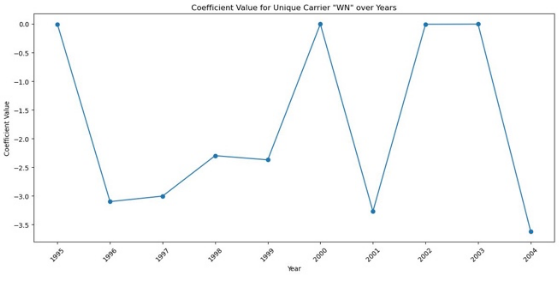


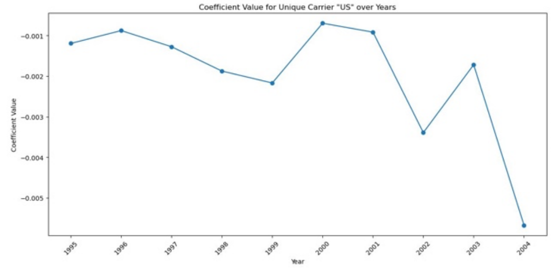


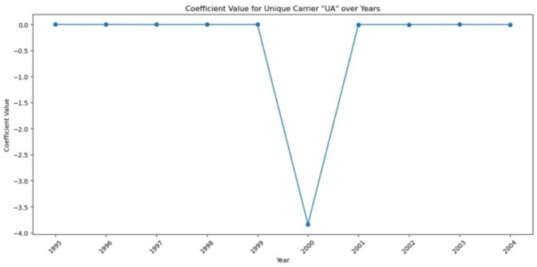


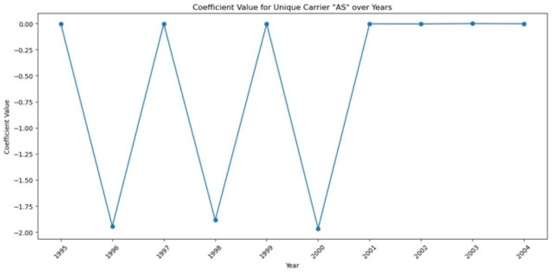
For geographical coordinates, represented by **'long_dest', 'lat_dest', 'long_origin', and 'lat_origin', the coefficients fluctuate from year to year and shift in magnitude and direction**. These changes point to the evolving impact of geographic factors on the propensity for a flight to be diverted, which may be influenced by regional weather, topography, or the strategic routing decisions of airlines.

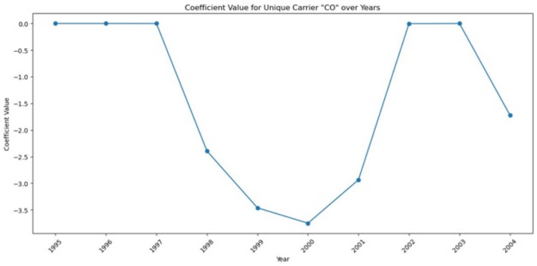





Coefficient Value for Unique Carrier "US" over Years


Coefficient Value for Unique Carrier "UA" over Years


Coefficient Value for Unique Carrier "AS" over Years


Coefficient Value for Unique Carrier "CO" over Years

The series of graphs provided illustrate the logistic regression coefficients for the 'Unique Carrier' variable across different years. These coefficients measure the impact of the carrier on the likelihood of a flight being diverted, with each plot corresponding to a unique carrier in the dataset. **The coefficients vary significantly from one year to the next**, demonstrating that **the effect of a particular airline on flight diversions can change over time**. For instance, some carriers show a steep negative coefficient in certain years, which could suggest a higher than average diversion rate during that period, potentially due to operational issues, fleet changes, or other internal factors specific to that airline.

The dramatic spikes or dips in coefficients, as seen in some of the graphs, may indicate years where the respective carrier faced unique challenges or changes. This could include external factors such as regulatory changes, strikes, economic downturns, or significant weather events that disproportionately affected that carrier.

In contrast, relatively stable coefficients for certain periods could suggest consistent operational performance with respect to flight diversions. It is also worth noting that the scale and direction of these coefficients are crucial, **a positive value indicates a higher probability of diversion**, **whereas a negative value indicates a lower probability relative to the baseline probability of diversion for all flights** in the model.

Conclusion

In conclusion, the analysis delves into the intricacies of flight delays and diversions over a span of ten years. It reveals that early morning flights at 5 AM are most likely to depart on schedule and weekly trends show variability with certain days like Saturdays generally seeing fewer delays, except in specific years where other weekdays performed better. Furthermore, there is no significant correlation exists between aircraft age and delay duration. This implies that factors other than aircraft age are more significant in determining delay times. Moreover, logistic regression models are utilized to examine the impact of various factors on the likelihood of flight diversions. Coefficients for departure times, geographical coordinates, and other attributes fluctuate year over year, reflecting the dynamic nature of airline operations and external influences on diversion risks. Overall, the multifaceted analysis offers a valuable resource for decision-making in the aviation industry, enabling stakeholders to optimize for punctuality, reduce delays, and improve the reliability of flight operations.