ST2195 Programming for Data Science

Flight Data Analysis - Python

Student Number: 200615248

Table of Contents

1. Executive Summary

This report is designed to deliver formal explanations and insights to the Key Questions for readers unfamiliar with programming.

1.1. Introduction

The 2009 ASA Statistical Computing and Graphics Data Expo dataset for flight arrival and departure is a large dataset of about 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed that is available for download from Harvard Dataverse.

1.2. Methods & Data Source

The dataset for flight arrival and departure contains around 120 million records containing details for all commercial flights on major carriers within the USA, from October 1987 to April 2008. The data includes the flight date, departure time, arrival time, carrier name, flight origin, flight destination, distance, et cetera.

Spyder programming software for Python language was used to explore and conduct statistical and diagnostic analysis to discover information and trends in the data. An HTML file, converted from Jupyter Notebook, was created to provide reproducible codes, visualization, and explanations into the steps for uncovering insights from the dataset to Answer the Key Questions. Only records from January 1995 to December 2000 were used in the analysis.

1.3. Key Questions

- When is the best time of day, day of the week, and time of year to fly to minimise delays?
- Do older planes suffer more delays?
- How does the number of people flying between different locations change over time?
- Can you detect cascading failures as delays in one airport create delays in others?
- Use the available variables to construct a model that predicts delays.

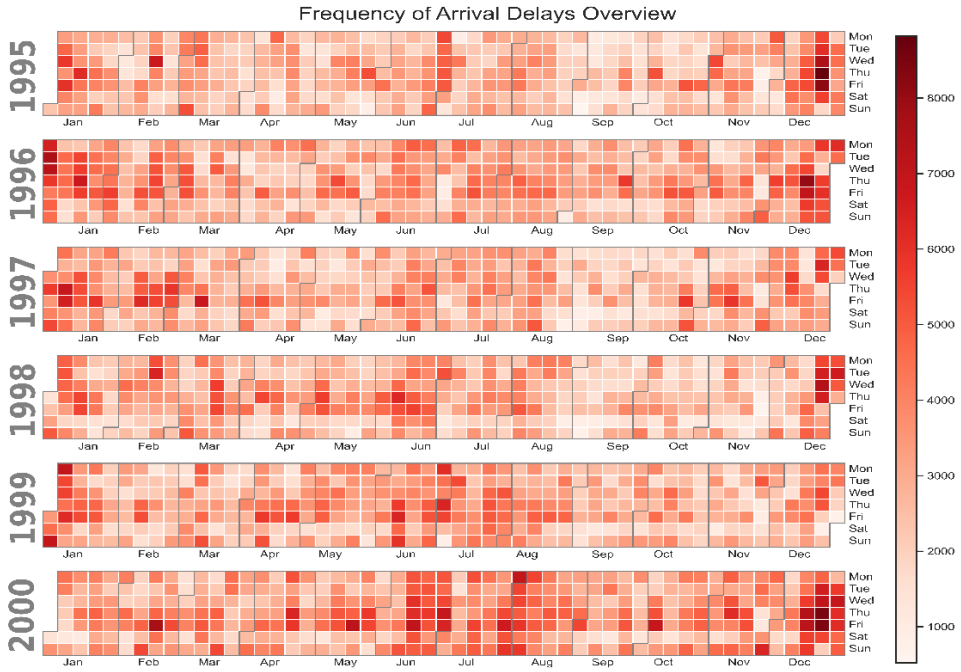
1.4. Variables Description

Name	Description
Year	1995 - 2000
Month	Jan
DayofMonth	31
DayOfWeek	1 (Monday) – 7 (Sunday)
DepTime	Actual departure time (local, hhmm)
CRSDepTime	Scheduled departure time (local, hhmm)
ArrTime	Actual arrival time (local, hhmm)
CRSArrTime	Scheduled arrival time (local, hhmm)
UniqueCarrier	Unique carrier code
FlightNum	Flight number
TailNum	Plane tail number
ActualElapsedTime	In minutes
CRSElapsedTime	In minutes
AirTime	In minutes
ArrDelay	Arrival delay, in minutes
DepDelay	Departure delay, in minutes
Origin	Origin IATA airport code
Dest	Destination IATA airport code
Distance	In miles
TaxiIn	Taxi in time, in minutes
TaxiOut	Taxi out time, in minutes
Cancelled	1 (Cancelled) – 0 (Not cancelled)
CancellationCode	Cancel reason (A = carrier, B = weather, C = NAS, D = security)
Diverted	1 (Diverted) – 0 (Not diverted)
CarrierDelay	in minutes
WeatherDelay	in minutes
NASDelay	in minutes
SecurityDelay	in minutes
LateAircraftDelay	in minutes

2. When is it Best to Fly to Minimize Delays?

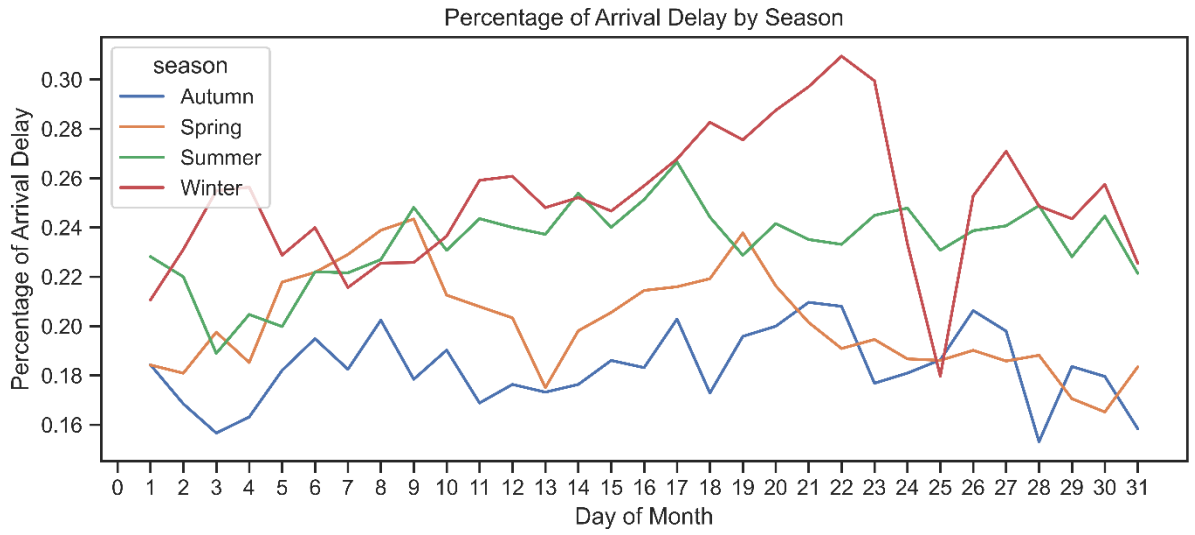
In this analysis, "Delay" is classified where time delayed in minutes exceeds a grace period of 15 minutes. Intuitively, flights can arrive on-time despite having a late departure. Therefore, arrival delay will be used to analyze when is best to fly to minimize the delay.

2.1. General Overview of Arrival Delays



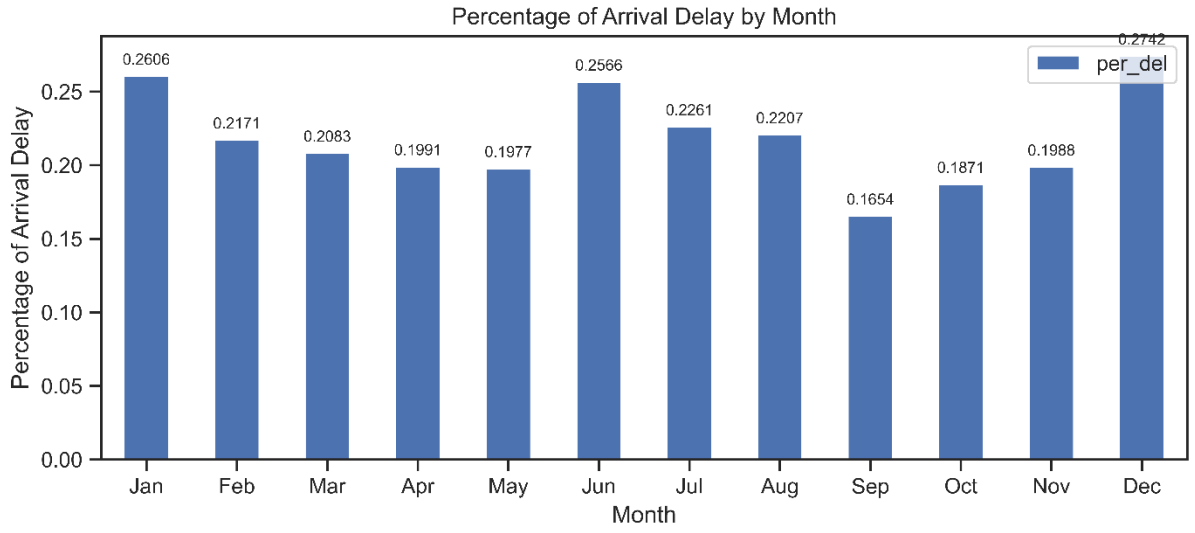
The color gradient denotes the number of arrival delays. By observation, September to November has the least number of arrival delays.

2.2. Best Season to Fly



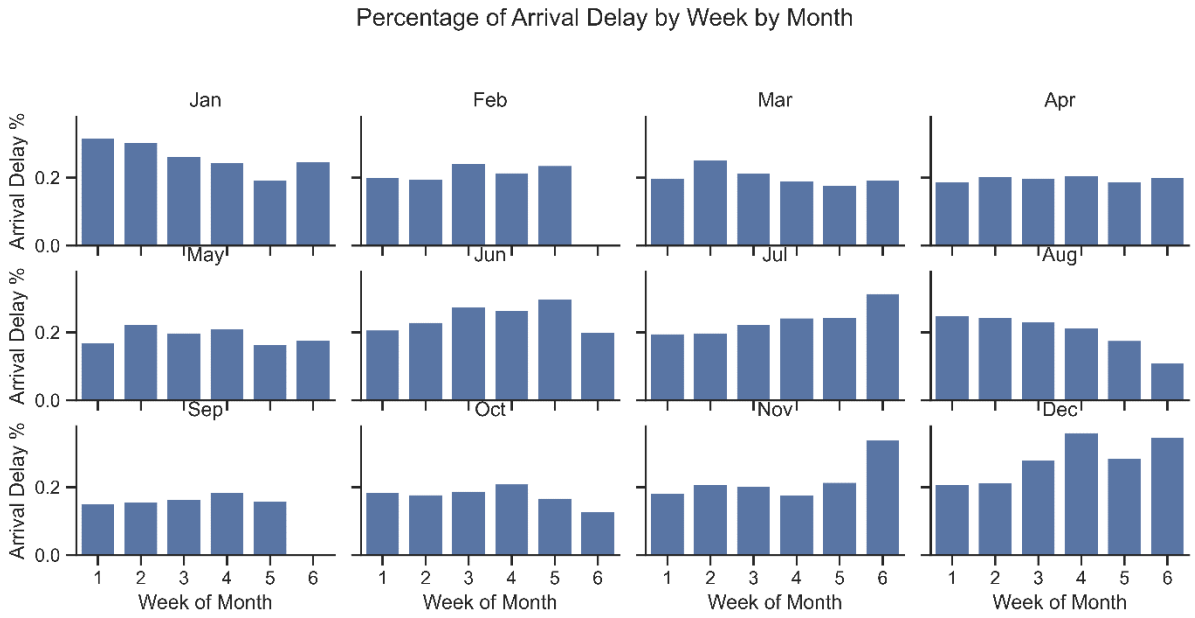
Autumn (the most bottom line) has the lowest percentage of flights delayed on arrival and hence is the Best Season to fly.

2.3. Best Month to Fly



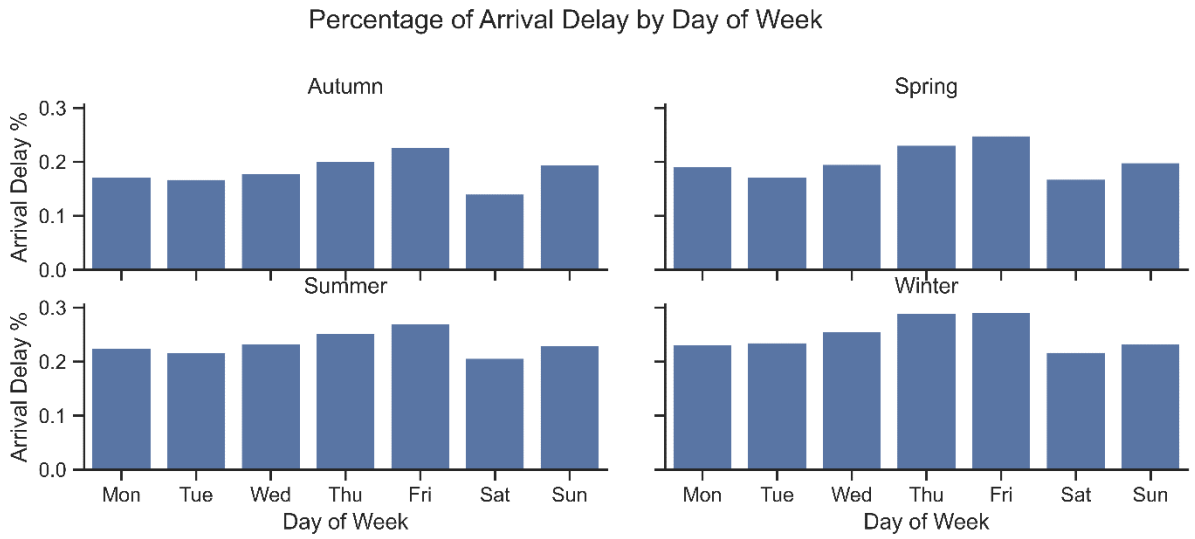
The best Month(s) to fly is in May (Spring), September, and October (Autumn), with the chance of delay for flights on arrival being 19.77%, 16.54%, and 18.71%, respectively.

2.4. Best Week to Fly



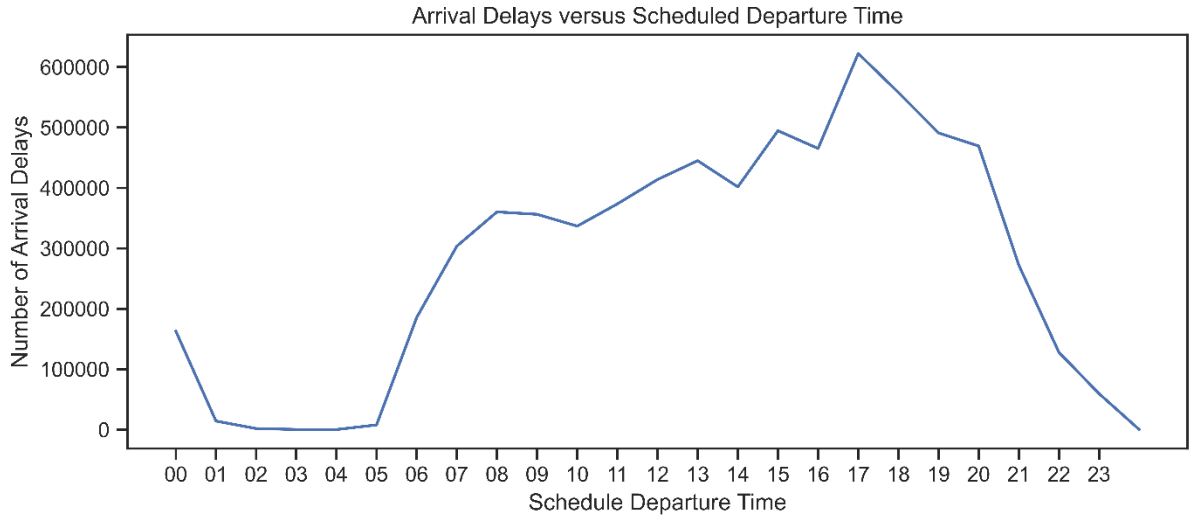
Statistically, the best week to travel during the Autumn Season is in the first week. The average percentage of arrival delays in the first week of the Autumn Season is 17.27%.

2.5. Best Day of Week to Fly



It is best to fly on the weekends. Most notably, the best day of the week to travel during the Autumn Season is on Saturday, with a low of 14.10% of flights delayed on average. In fact, the best day of the week to fly in each Season is on a Saturday.

2.6. Best Time to Fly



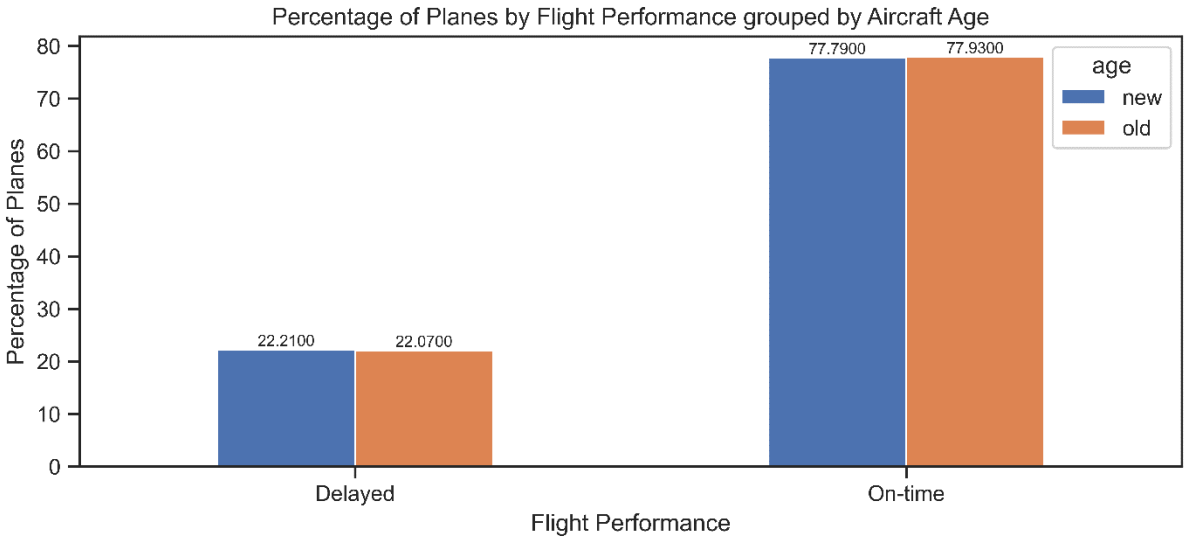
According to the chart, flight delays were low in the early morning and picked up after 10:00, with the highest number of delays occurring between 17:00 - 20:00. The delays decline towards late night.

Minimize the chance of flight delays by flying in the morning and avoid flights that depart between 16:00 - 19:00 hours.

3. Do Older Planes Suffer More Delays Than Newer Planes?

The age of a plane depends on multiple factors besides its chronological age. This analysis refers to 11 years - the average age of U.S. commercial aircraft - as the standard for old planes.

3.1. Percentage of Planes Percentage of Planes based upon Flight Performance



There seems to be no association between the planes' age and flight performance. The Chi-square test of association was conducted to check for statistical significance in the difference between the planes' age and flight performance.

3.2. Chi-square Test for Association

H0: There is no association between the planes' age and flight performance

H1: There is such an association

```
p value is 0.002813509411296891
Dependent (reject H0)
```

As the p-value = 0.28%, is less than the 1% significance level, the test result is highly significant. The results indicate enough evidence to reject the null hypothesis and conclude an association between the planes' age and flight performance.

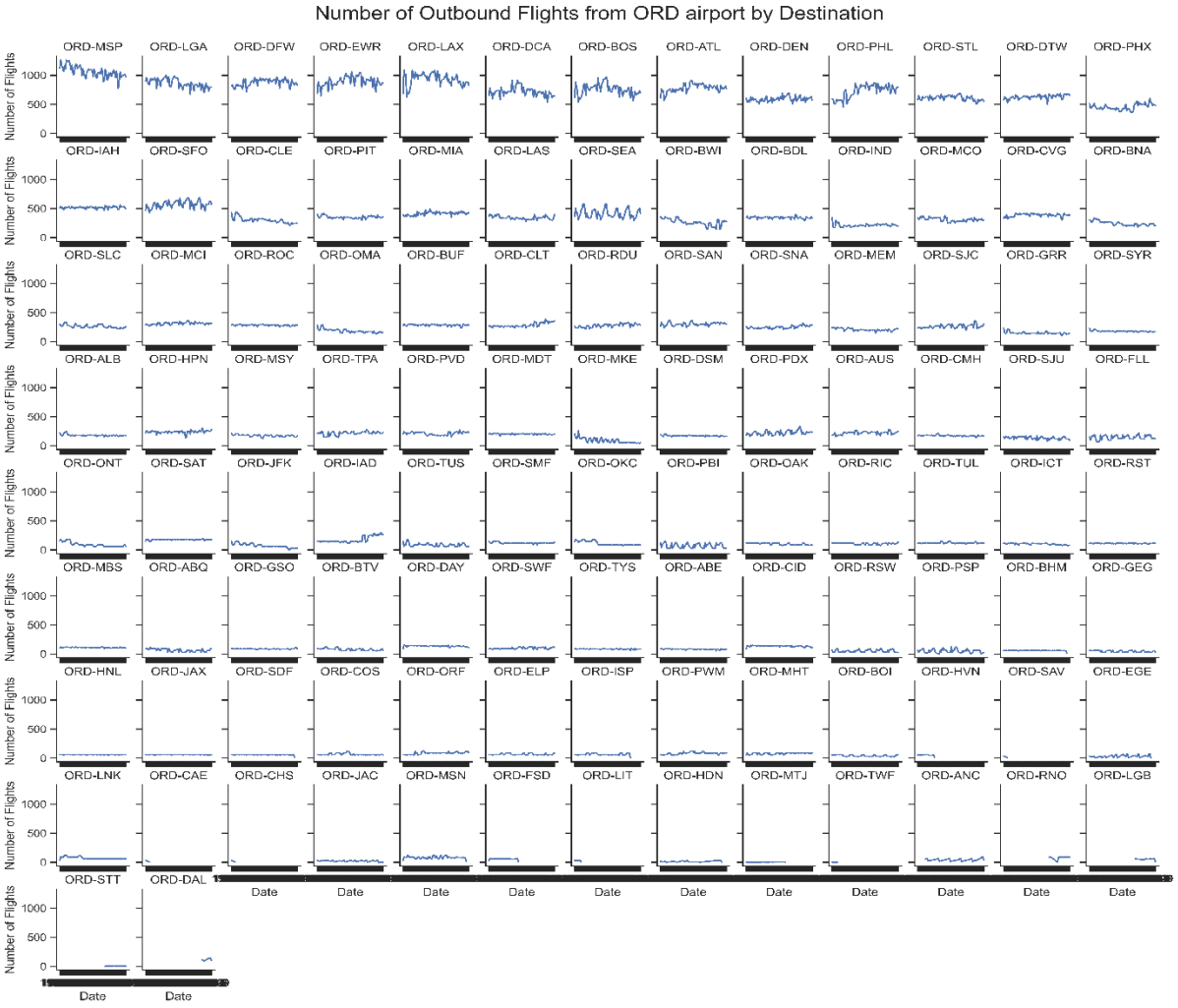
Although the difference in on-time performance and planes' age is statistically significant, the graphical representation of performance for both the age groups looks similar with a 0.14% difference in on-time arrivals.

Hence, there is a small, consistent association between the planes' age and their performance due to the statistical significance, but biologically insignificant. Therefore, it is unlikely that older planes do suffer from more delays.

4. How Does the Number of People Flying Between Different Location Change Over Time?

There is no information on the number of passengers abroad on each plane. Hence, the number of flights is used as a proxy for the indication of popularity. ORD - Chicago O'Hare International airport has the highest number of outbound flights and hence is used as the sampling frame for this analysis.

4.1. Multiple Time Series Chart



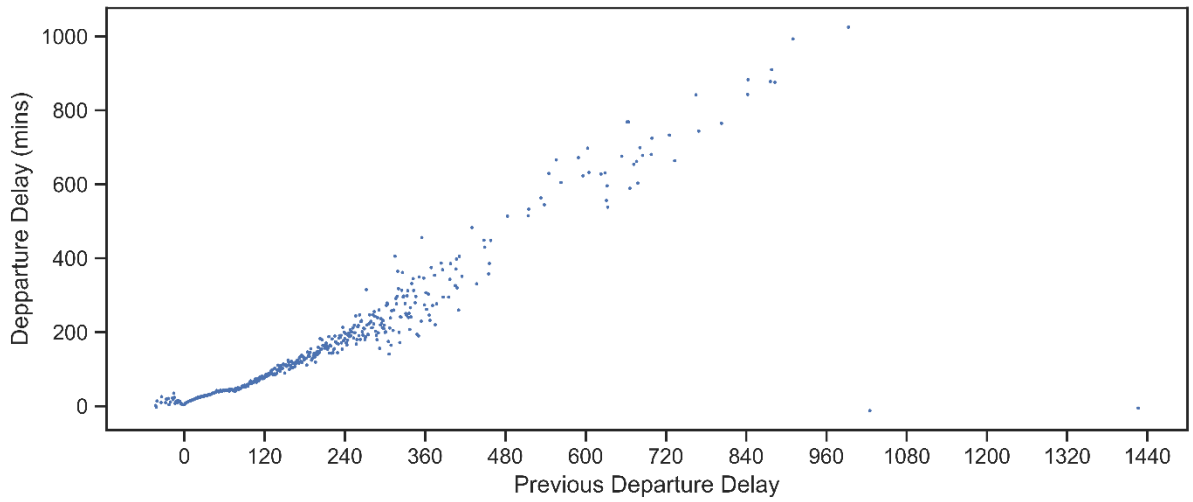
The visualization illustrates the trend of outgoing flights from ORD (Chicago) over the years, from 1995 to 2000 (left to right). The majority of the routes showed consistent trends.

Flights from Chicago to Minneapolis (ORD-MSP) have been consistently decreasing over the years, and flights from Chicago to Philadelphia (ORD-PHL) have been increasing. Meanwhile, flights to Seattle (ORD-SEA) are low at the beginning and end of each year and highest mid-year.

This result is exclusively for flights outbound from ORD airport, which does not make a good representation of all the airports. However, the same analysis using different origin airports of interest can be re-performed to uncover the flight patterns for the airport.

5. Can you Detect Cascading Failures as Delays in One Airport Create Delays in Others?

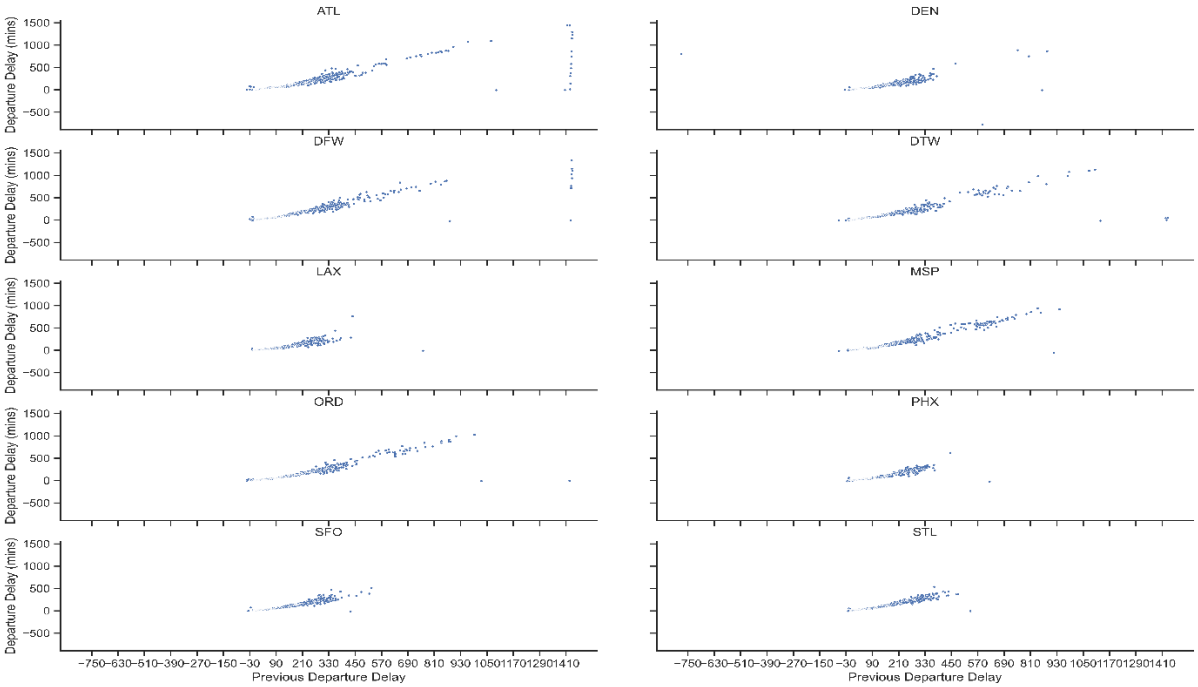
5.1. Relationship Between the Avg. Delay Against Time Delayed for Previous Flights



The scatter diagram shows a positive relationship between the previous delay and the subsequent flights' departure delay time.

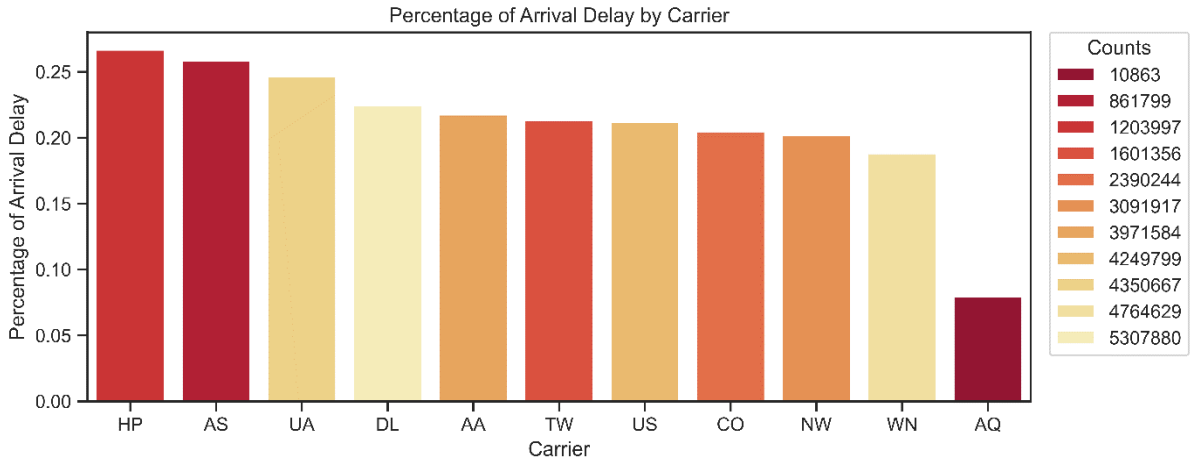
The increase in variability after the 480 (mins) mark indicates the strength of the relationship cooling off. Suggests flights with shorter delays have stronger effects on the subsequent flights' disability to depart on-time while flights with longer delays have a poorer effect. This interpretation is valid since long-delayed flights can be interspersed, with flights leaving on time.

5.2. Overview of Cascading Delays for the Top 10 Busiest Airports



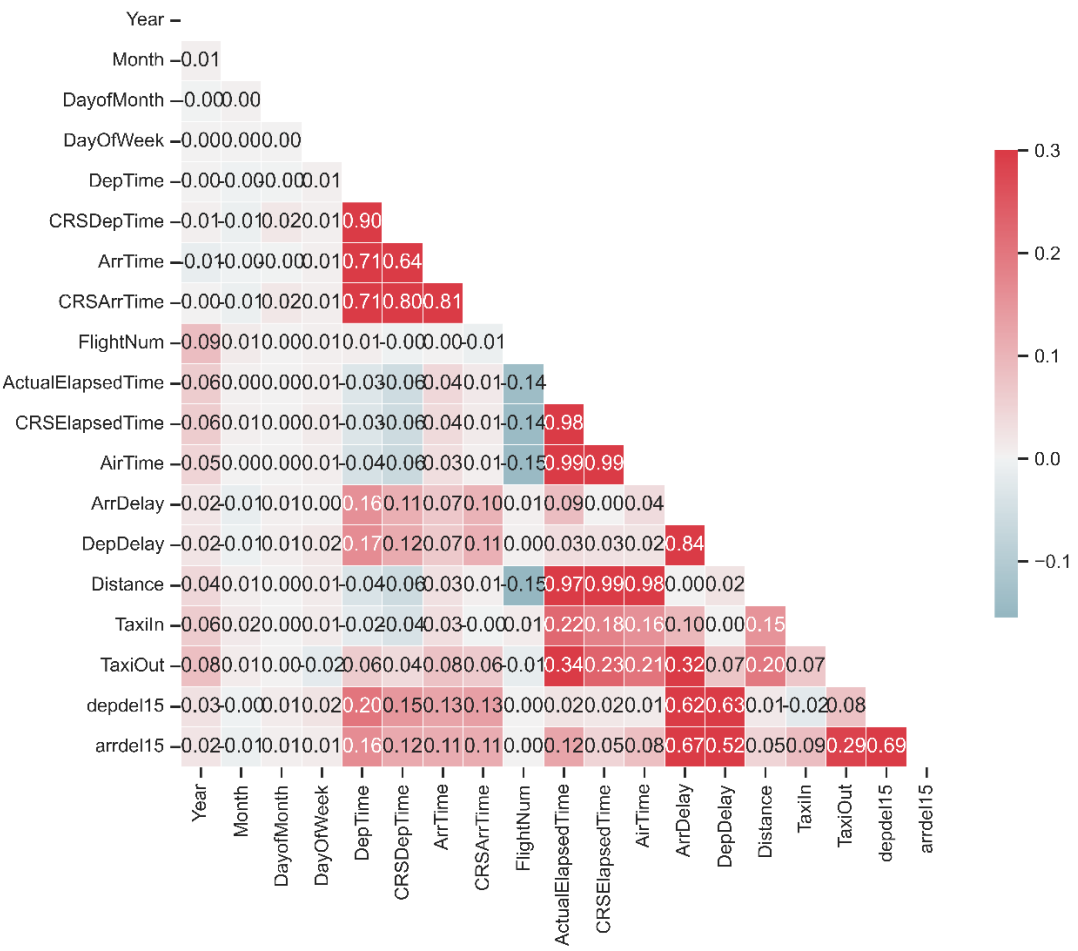
6. Use the Available Variables to Construct a Model that Predicts Delays

6.1. Arrival Delay based upon Unique Carrier



There is no obvious trend between the number of flights and arrival delays. However, by observation, some carriers (darker colored) have a higher percentage of arrival delays even though they are lower in terms of the number of flights. Thus, carrier is a factor in determining arrival delays.

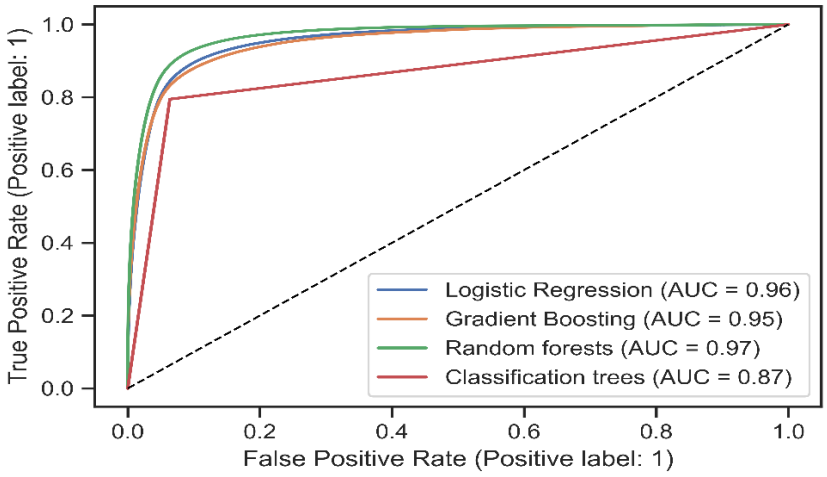
6.2. Correlation Analysis



6.3. Data Modelling and Performance of Classification Models

```
 #   Column             Dtype
---  ------             -----
 0   season             object
 1   UniqueCarrier      object
 2   DepTime            float64
 3   depdel15           int64
 4   ActualElapsedTime  float64
 5   Distance           float64
 6   TaxiOut            int64
```

The variable in the column denote the factors selected for use by the modeling process in predicting arrival delays.



The ROC curve illustrates the trade-off between the "True Positive Rate" and "1 - False Positive Rate". Performance of the classifiers are indicated by the distance between the curve and the upper-left corner, the shorter the distance, the higher the accuracy and the better the performance.

Overall, "Random Forest" has the highest accuracy in predicting delays. Hence, "Random Forest" is the better model out of the three.

7. Conclusions

- To minimize delay travel on the first week of September (Autumn) from 05:00 – 09:00 hours, Saturday.
- There is a small, consistent association between the planes' age and performance due to the statistical significance, but biologically insignificant. Therefore, it is unlikely that older planes do suffer from more delays.
- The number of people flying to different locations overtime is generally consistent with some minor exceptions in a few locations.
- Since flight schedules are aligned between the origin and destination of a flight, the impact of cascading delays in one airport on another airport can be interpreted implicitly through the relationship between previous delays and the subsequent flights' departure delay time.
- Random Forest is the better model of the other three used in this analysis in predicting delays.