ST2195 Programming for Data Science

Flight Data Analysis - R

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 12 gigabytes, uncompressed, 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.

RStudio programming software for R language was used to explore and conduct statistical and diagnostic analysis to discover information and trends in the data. An HTML file, converted from RMarkdown, 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?

The term "delay" in this analysis is classified where time delayed exceeds a grace period of 15 minutes. Despite having late departures, flights may arrive on time. Hence arrival delay will be used to analyze when is best to fly to minimize delays.
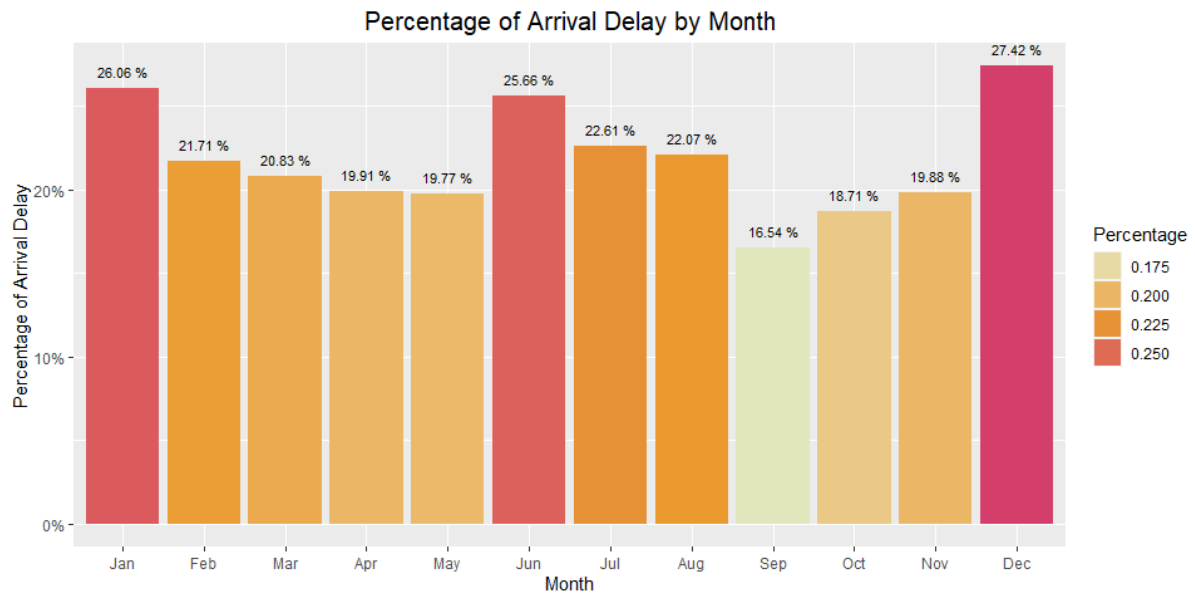
2.1. General Overview of Arrival Delays
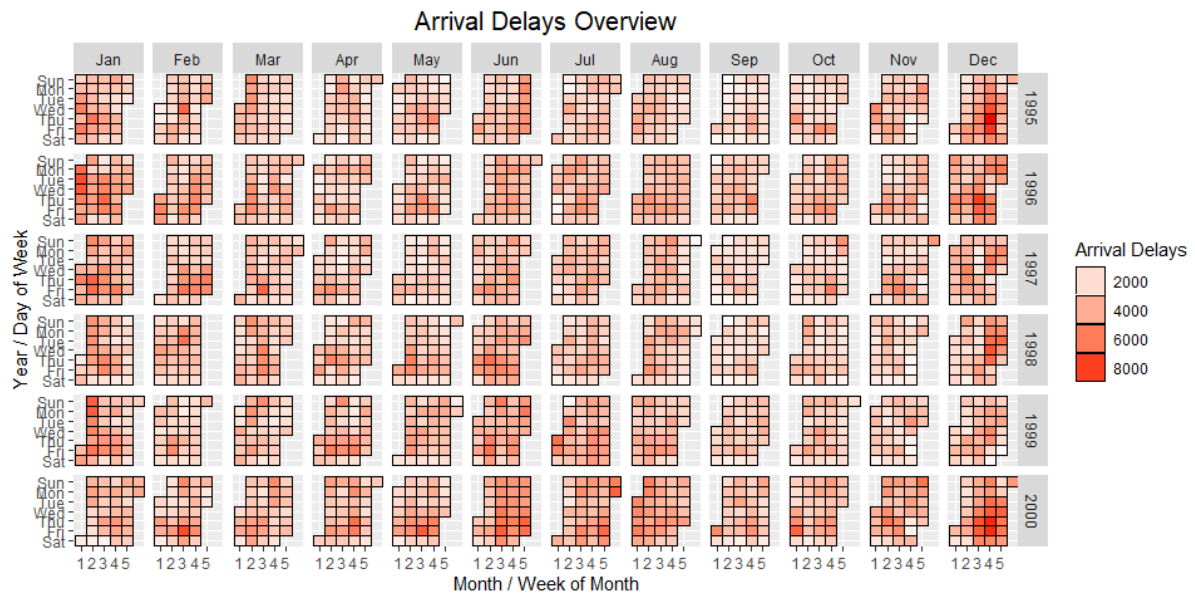
The color gradient denotes the number of arrival delays. By observation, September to November has the least number of arrival delays.

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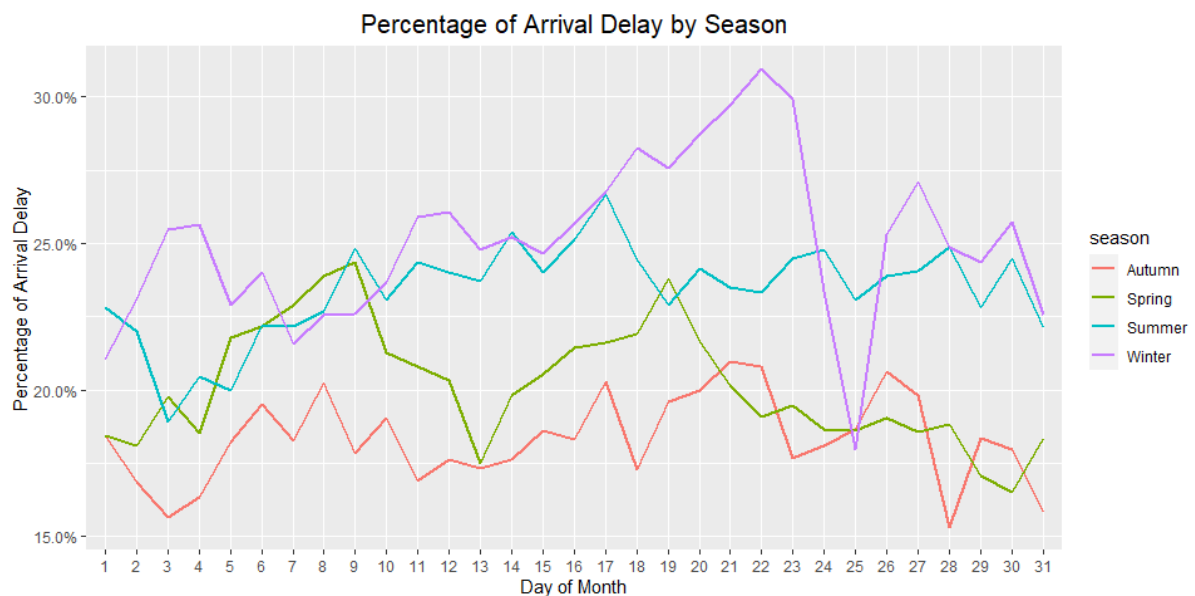
Autumn (the most bottom line) has the lowest percentage of flights delayed on arrival and hence is the Best Season to fly.

2.3. Best Month to Fly

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

2.4. Best Week to Fly
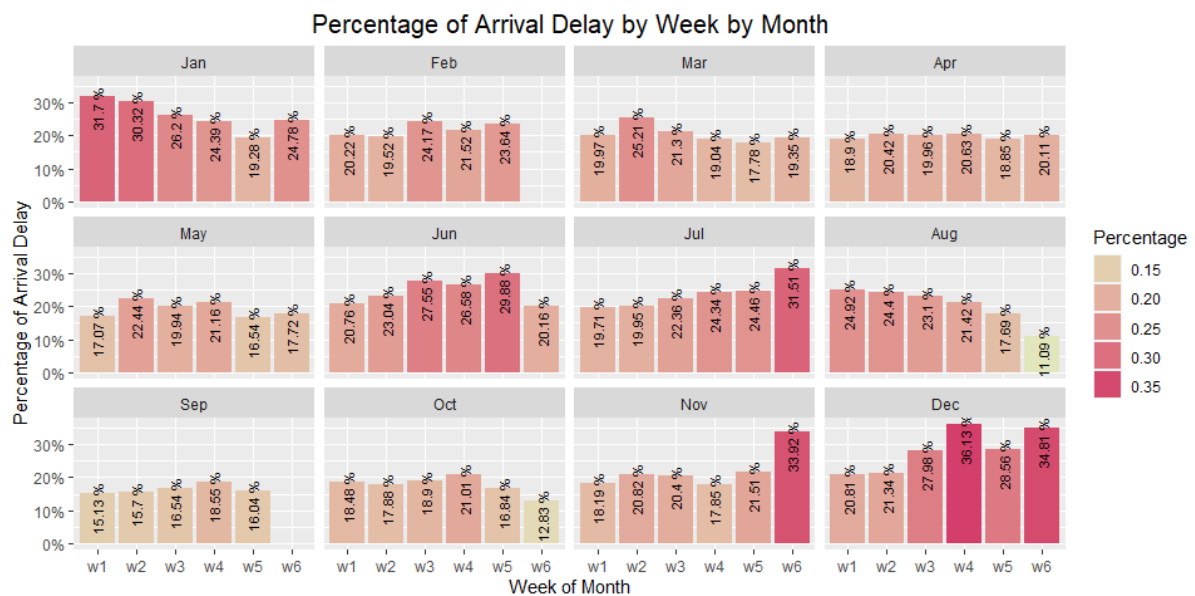
During the Autumn season, the first week of the Month is when it is best to fly. Statistically, the average percentage of arrival delays in the first week of the Autumn Season is 17.27%.

2.5. Best Day of Week to Fly

Percentage of Arrival Delay by Day of Week

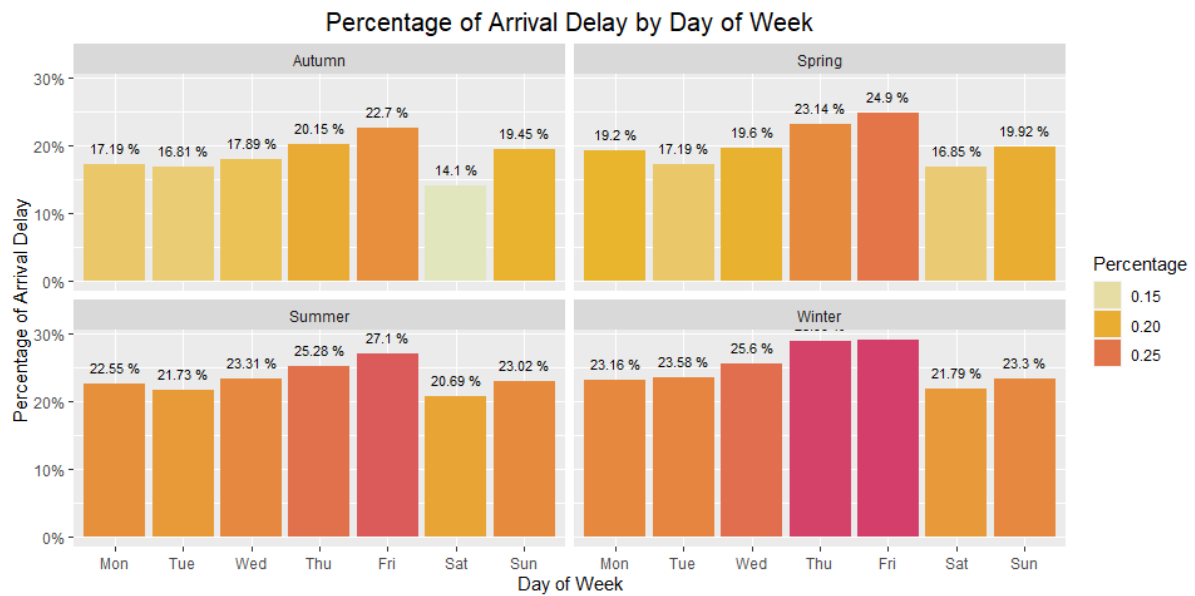Diving deeper into the best day of the week to fly, the Saturdays of the Autumn season has percentage flights delayed on average 14.10%. Generally speaking, Saturdays are the best day of the week to fly regardless of the season.

2.6. Best Time to Fly

Arrival Delays versus Scheduled Departure Time
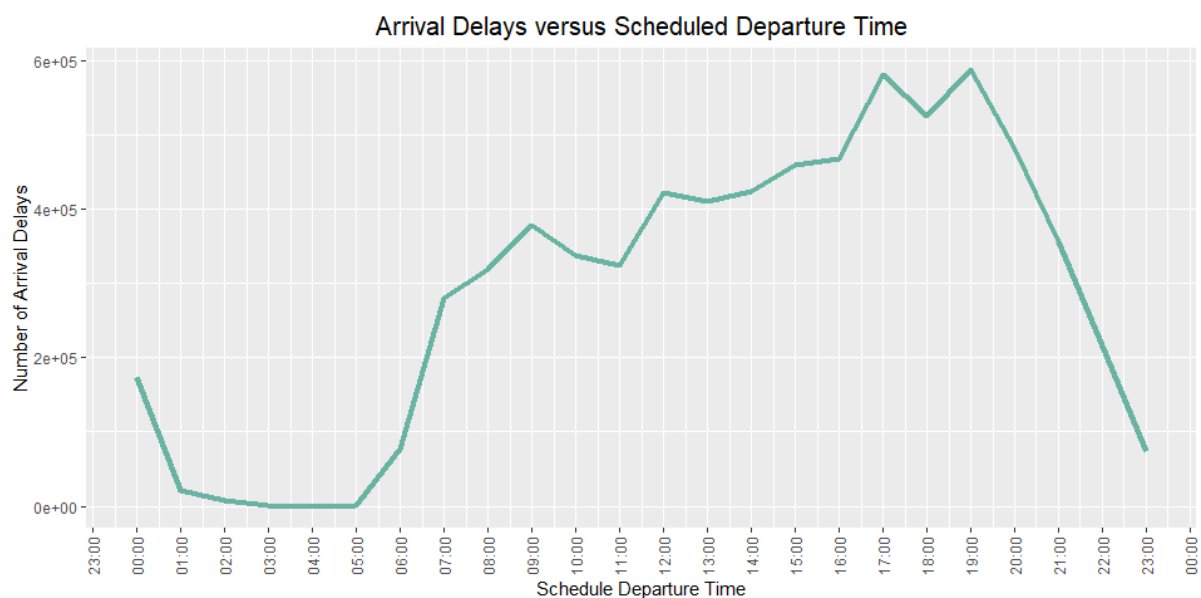
By observation, flight delays were low in the early morning and increased after 10:00 hours. Between 17:00 - 20:00 has the highest number of flights delayed, with the numbers declining towards midnight.

Reduce the odds of flight delays by flying in the morning and avoiding flights that depart during 16:00 - 19:00 hours.

3. Do Older Planes Suffer More Delays Than Newer Planes?
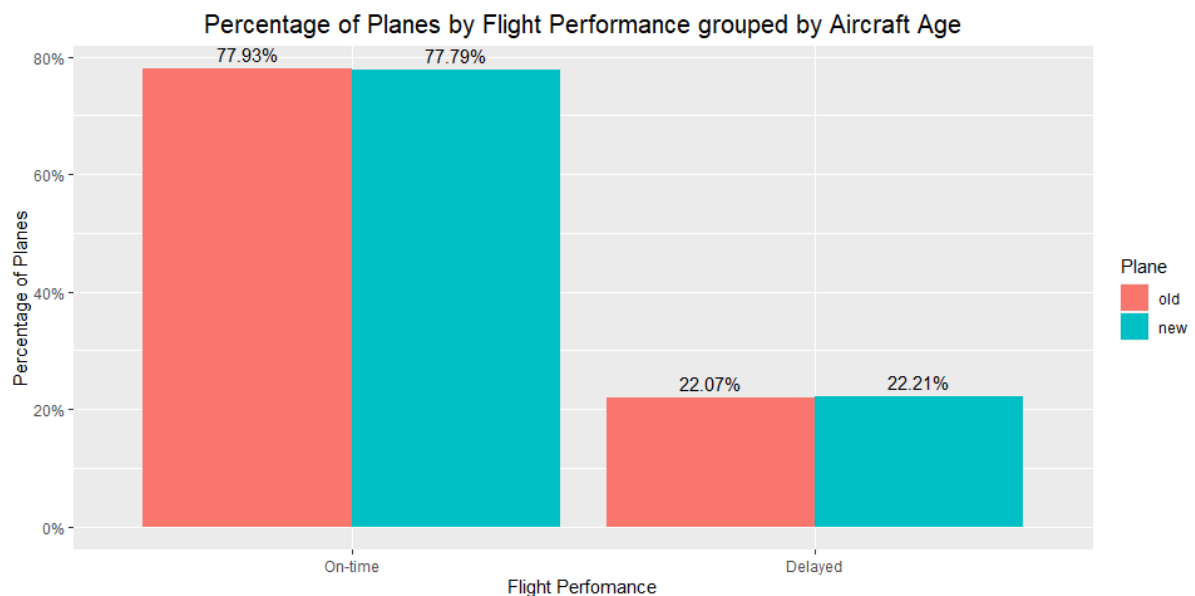
There are multiple factors besides its chronological age to consider when evaluating the age of a plane. This analysis will take 11 years, the average age of a U.S. commercial aircraft as a guide for "old" planes.

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

Based on the visualization, there appears to be no association between the plane age and flight performance. To further confirm this result, we check for any statistical significance between the plane age and flight performance using the Chi-square test for association.

3.2. Chi-square Test for Association

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

H1: There is such an association

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  xtbl_count
## X-squared = 8.9246, df = 1, p-value = 0.002814
```

As the p-value = 0.28%, the null hypothesis is rejected at the 1% significance level. The results are highly significant and provide strong evidence for rejecting the null hypothesis to conclude an association between plane age and flight performance.

Although the plane age and flight performance are statistically significant, its graphical visualization for the association is similar for both the age group, with only a 0.14% difference.

In conclusion, the association between plane age and flight performance is small, consistent, and biologically insignificant. Hence, 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

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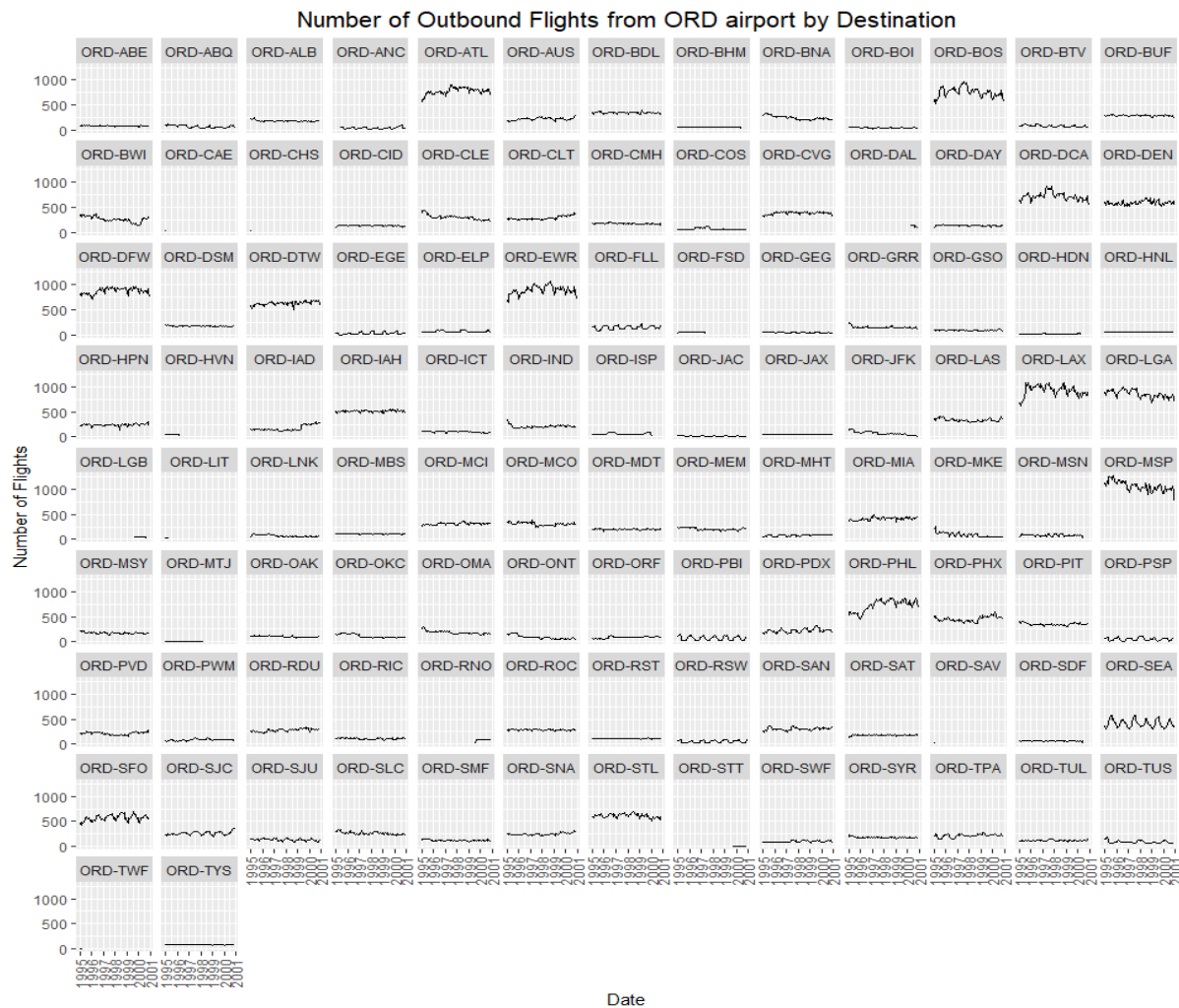
The visualization illustrates the trend of outgoing flights from ORD (Chicago) over the years, from 1995 to 2000 (left to right). The majority of the routes showed consistent trends.

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
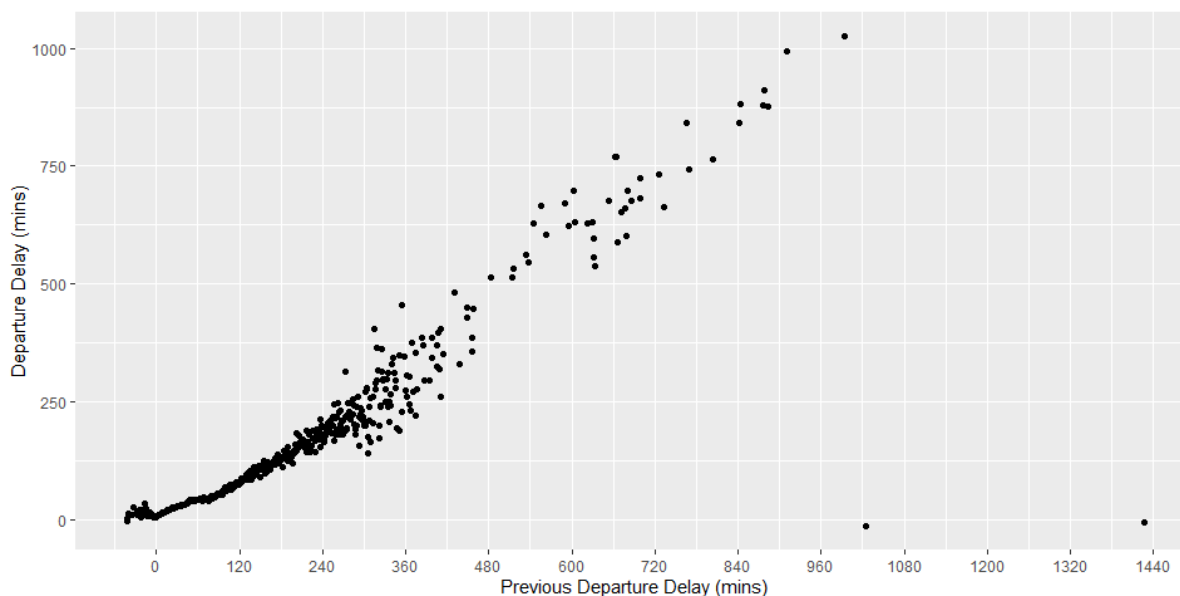
The above scatter diagram illustrates the relationship (positive) between the previous delay and the subsequent flights' departure delay time (avg.) for flights outbound from Chicago O'Hare International airport between 19:00 – 22:00 hours.

Indicated by the increase in variability, the strength of the relationship cools off around the 480 (mins) mark. Suggesting that flights' ability to depart on time increases with the duration delayed for the previous flight since flights with longer delays can intersperse 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.
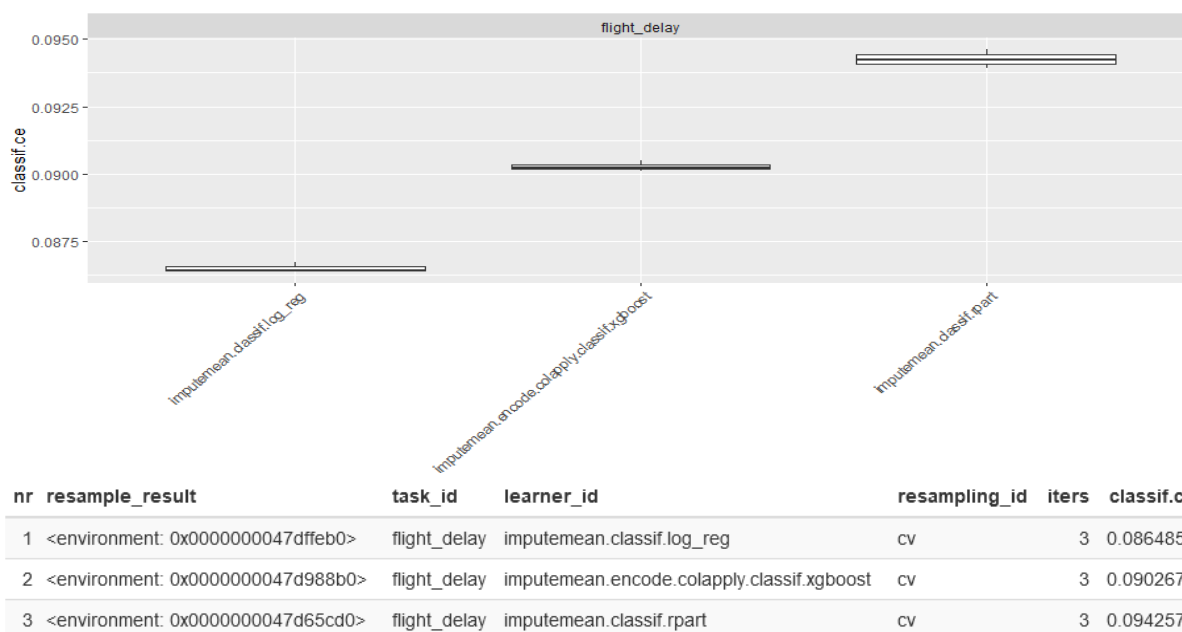
6.2. Correlation Analysis

6.3. Data Modelling and Performance of Classification Models

```
## * Target: ArrDel15
## * Properties: twoclass
## * Features (7):
##   - int (4): ActualElapsedTime, DepTime, Distance, TaxiOut
##   - fct (2): DepDel15, UniqueCarrier
##   - ord (1): season
```

"Features" denote the factors selected for use by the modeling process in predicting arrival delays.

nr	resample_result	task_id	learner_id	resampling_id	iters	classif.ce
1	<environment: 0x0000000047dffeb0>	flight_delay	imputemean.classif.log_reg	cv	3	0.0864856
2	<environment: 0x0000000047d988b0>	flight_delay	imputemean.encode.colapply.classif.xgboost	cv	3	0.0902674
3	<environment: 0x0000000047d65cd0>	flight_delay	imputemean.classif.rpart	cv	3	0.0942570

The baseline algorithm selected is Logistic Regression due to its simplicity and speed. The other two algorithms were selected for comparison.

Overall, the three models have a similar distribution in errors. But "log_reg" has the lowest error in predicting delays. Hence "log_reg" 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.

❖ The association between plane age and flight performance is small, consistent, and biologically insignificant. Hence, 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.

❖ Logistic regression is the better model of the other two used in this analysis in predicting delays.