Data Analytics Final Project

AUTHORS
Trifanov Matvei
Pasindu Perera

ABSTRACT
In this project we analyzed flight data from New York City airports to uncover key factors influencing flight delays and build predictive models for both regression and classification tasks. Using five datasets containing flight details, weather conditions, airline information, and more, we conducted thorough data cleaning, preprocessing, and exploratory analysis to ensure data quality and uncover meaningful insights. For the regression task, we focused on predicting flight arrival delays (arr_delay) based on features such as flight distance, weather variables, and airline carriers. Our refined regression model, evaluated using metrics like RMSE and MAE, provided insights into the key contributors to delays, though inherent variability in the data posed challenges. A classification task was also undertaken to categorize flights, offering additional perspectives on the data. Overall, this project demonstrates the application of supervised learning techniques to real-world data, highlighting both the opportunities and limitations of predictive modeling in complex domains.

▶ Code

1 Introduction

This project uses dataframes about flights, specifically a sample of domestic flights departing from the three major New York City airports in 2013.

There are 5 tabular dataframes. Some frequencies statistics:
14 **airlines**,
1251 **airports**,
435352 **flights**,
4840 **planes**,
26204 **weather** observations.

The main dataframe `flights` has the following variables:

`year`, `month`, `day`: Date of departure.
`dep_time`, `arr_time`: Actual departure and arrival times (format HHMM or HMM), local time zone.
`sched_dep_time`, `sched_arr_time`: Scheduled departure and arrival times (format HHMM or HMM), local time zone.
`dep_delay`, `arr_delay`: Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.
`carrier`: Two letter carrier abbreviation. See `airlines` dataframe to get the full name.
`flight`: Flight number.
`tailnum`: Plane tail number. See `planes` for additional metadata.
`origin`, `dest`: Origin and destination. See `airports` for additional metadata.
`air_time`: Amount of time spent in the air, in minutes.
`distance`: Distance between airports, in miles.
`hour`, `minute`: Time of scheduled departure broken into hour and minutes.

`time_hour`: Scheduled date and hour of the flight as a *POSIXct* date. Along with *origin*, can be used to join flights data to `weather` data.

2 Data Cleaning

In this project, we used a structured strategy to handle missing values across datasets, carefully balancing data preservation and noise. For instance, datasets with many variables with missing values, such as `flights` and `weather`, we developed a threshold.

- Remove rows in which all critical variables are missing. These rows contain no useful information and add little to the analysis.
- Impute rows with partial missing values: In rows where only a few variables are missing (e.g., less than 50%), we used imputation to preserve the remaining information.

▶ Code

Missing values in the flights

Column	Missing_Values
dep_time	10738
dep_delay	10738
arr_time	11453
arr_delay	12534
tailnum	1913
air_time	12534

▶ Code

Missing values in weather

Column	Missing_Values
temp	25536
dewp	25536
humid	25536
wind_dir	1220
wind_speed	1033
wind_gust	1033
precip	24611
pressure	25632
visib	24

For example, in the `weather` dataset, which contains missing values in 9 out of 16 variables:

- Removed rows where 5 or more (≥50%) of the 9 variables were missing.

- Retained and imputed rows where fewer than 5 variables were missing. Numerical variables were imputed using the mean, as it maintains the central tendency of the data without introducing bias.

This approach provides a good balance between data preservation and quality maintenance, guaranteeing that incomplete but valuable rows are not deleted unnecessarily while avoiding excessive imputation of very incomplete rows.

▶ Code

Missing values in flights_cleaned

	x
X	0
year	0
month	0
day	0
dep_time	0
sched_dep_time	0
dep_delay	0
arr_time	0
sched_arr_time	0
arr_delay	0
carrier	0
flight	0
tailnum	0
origin	0
dest	0
air_time	0
distance	0
hour	0
minute	0
time_hour	0

▶ Code

Missing values in airports_clean

	x
X	0
faa	0
name	0
lat	0

	x
lon	0
alt	0
tz	0
dst	0
tzone	0

▶ Code

Missing values in planes_clean

	x
X	0
tailnum	0
year	0
type	0
manufacturer	0
model	0
engines	0
seats	0
speed	0
engine	0

▶ Code

Missing values after cleaning weather dataset

	x
X	0
origin	0
year	0
month	0
day	0
hour	0
temp	0
dewp	0
humid	0
wind_dir	0
wind_speed	0
wind_gust	0
precip	0

	x
pressure	920
visib	0
time_hour	0

▶ Code

```
 chr [1:1492] NA NA NA NA NA NA NA NA "1008,8" "1008,5" "1008,1" "1008,1" ...
```

▶ Code

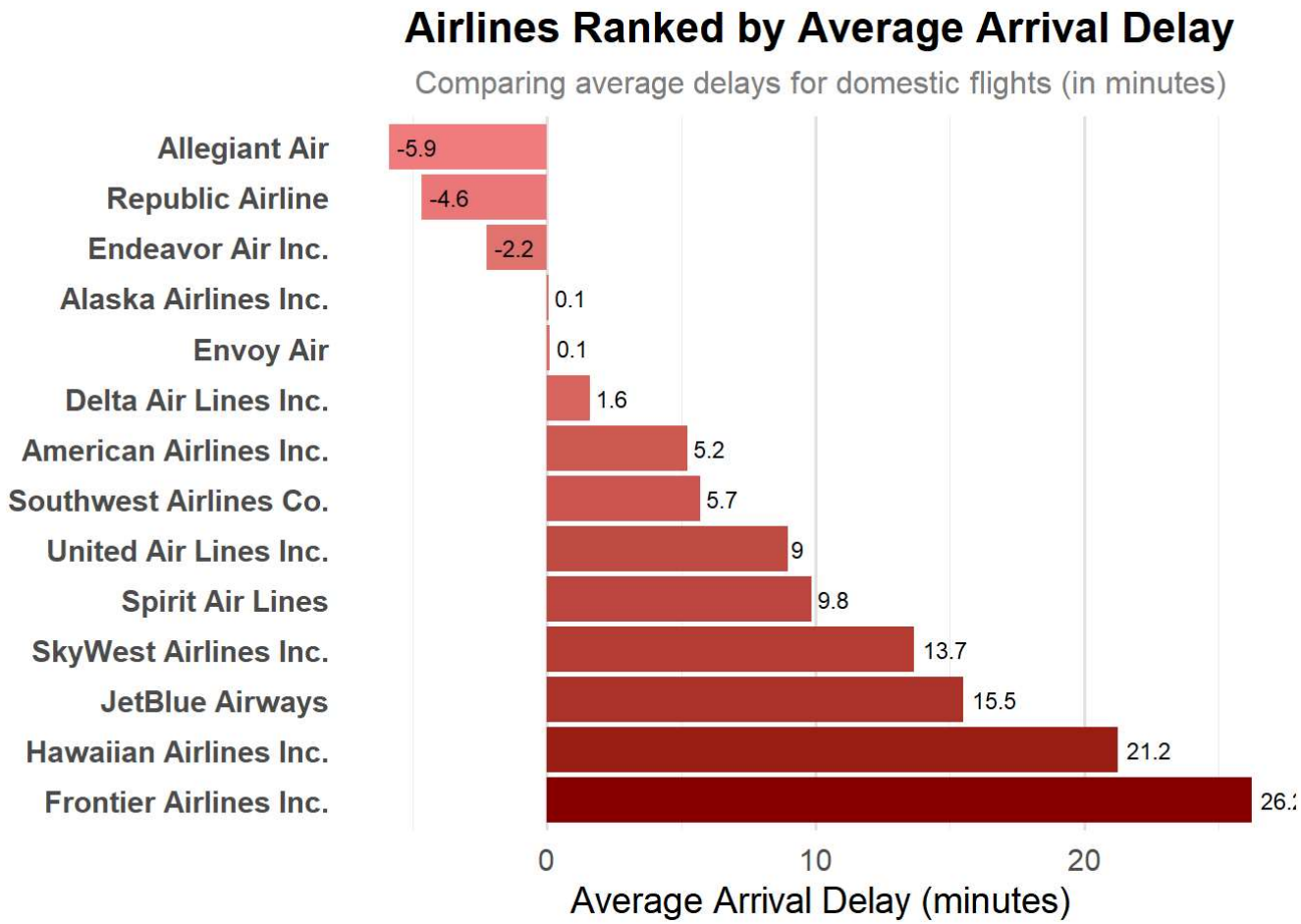
Misssing values in weather_clean

	x
X	0
origin	0
year	0
month	0
day	0
hour	0
temp	0
dewp	0
humid	0
wind_dir	0
wind_speed	0
wind_gust	0
precip	0
pressure	0
visib	0
time_hour	0

3 Exploratory Data Analysis

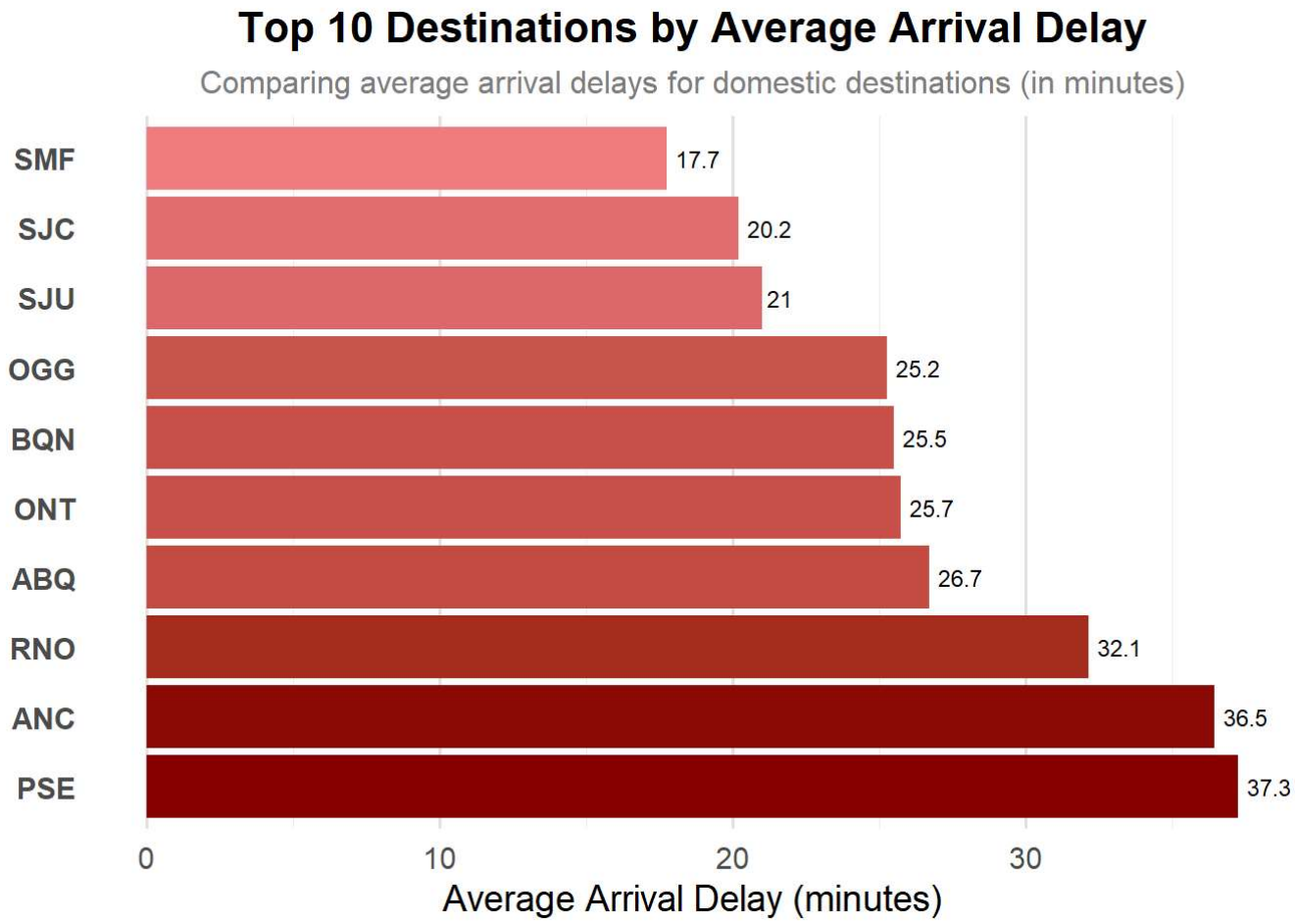
In the Exploratory Data Analysis (EDA) section, the focus is on summarizing and visualizing the dataset to uncover key patterns, relationships, and trends. This process helps identify significant insights and supports the later modeling tasks. Using visual tools such as bar charts, line plots, and correlation matrices, the EDA reveals variability in airline punctuality, the impact of weather conditions on flight delays, and delay trends across different destinations, timeframes, and carriers. These insights lay a foundation for understanding the data and addressing the research questions effectively.

▶ Code



The chart ranks airlines by average arrival delays, with Allegiant Air showing the best performance (-5.9 minutes) and Frontier Airlines the worst (26.2 minutes). It highlights variability in airline punctuality, with some airlines arriving early while others face significant delays.

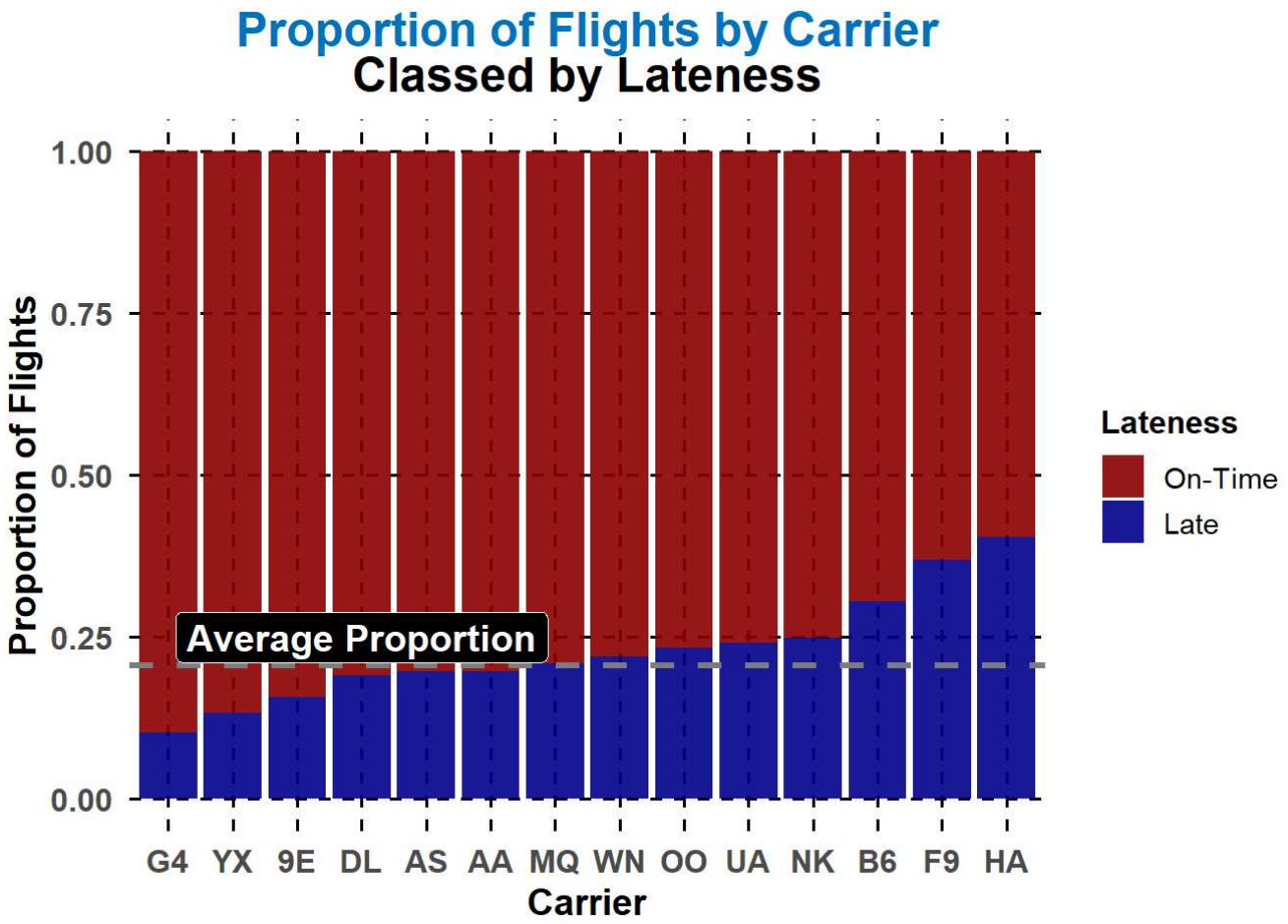
▶ Code



The chart shows the top 10 domestic destinations with the highest average arrival delays. Sacramento (SMF) has the shortest delay among these destinations (17.7 minutes), while Ponce (PSE) faces the highest average delay (37.3 minutes). This analysis highlights airports with significant delays, offering insights for better resource allocation and scheduling.

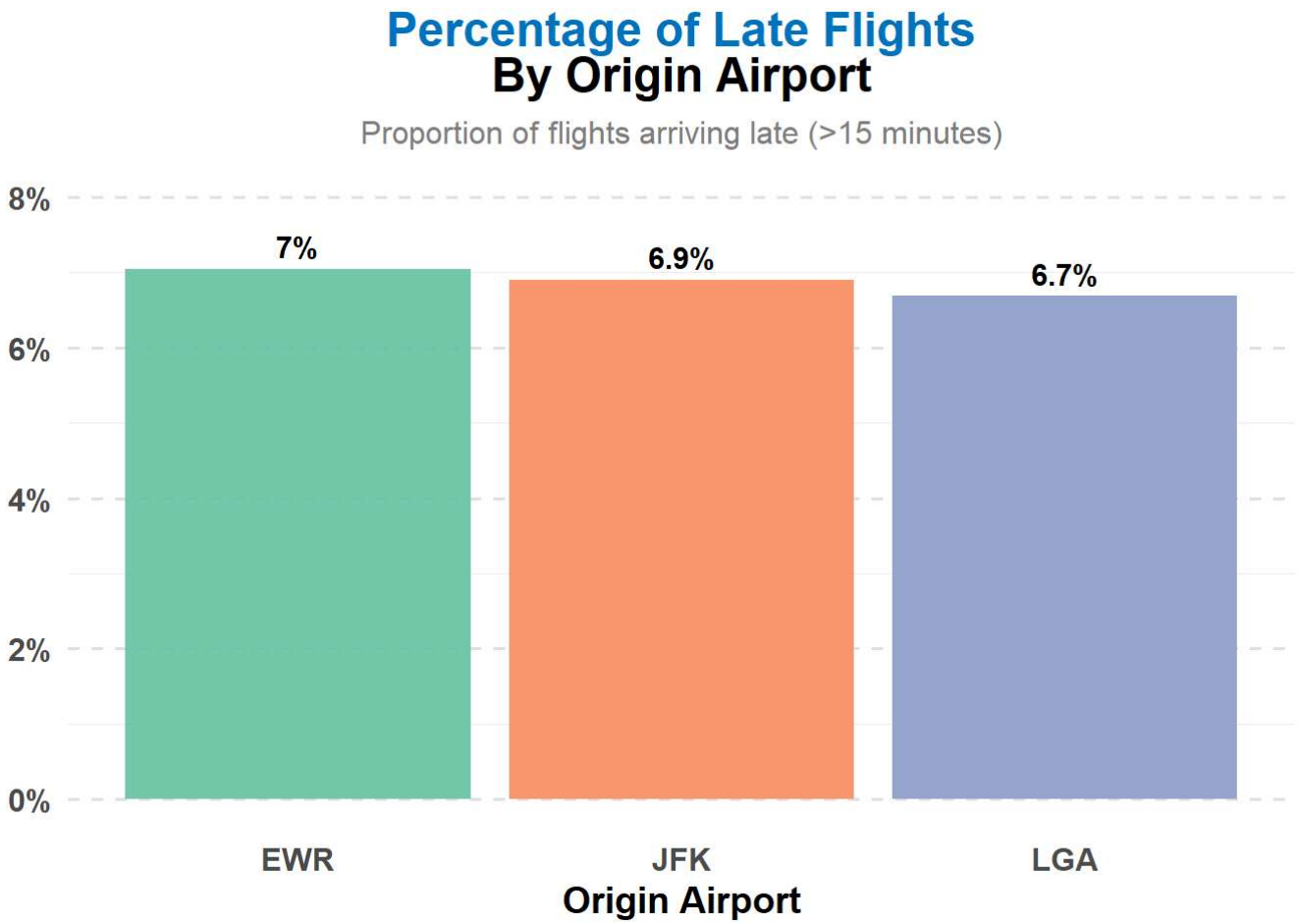
▶ Code

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
```



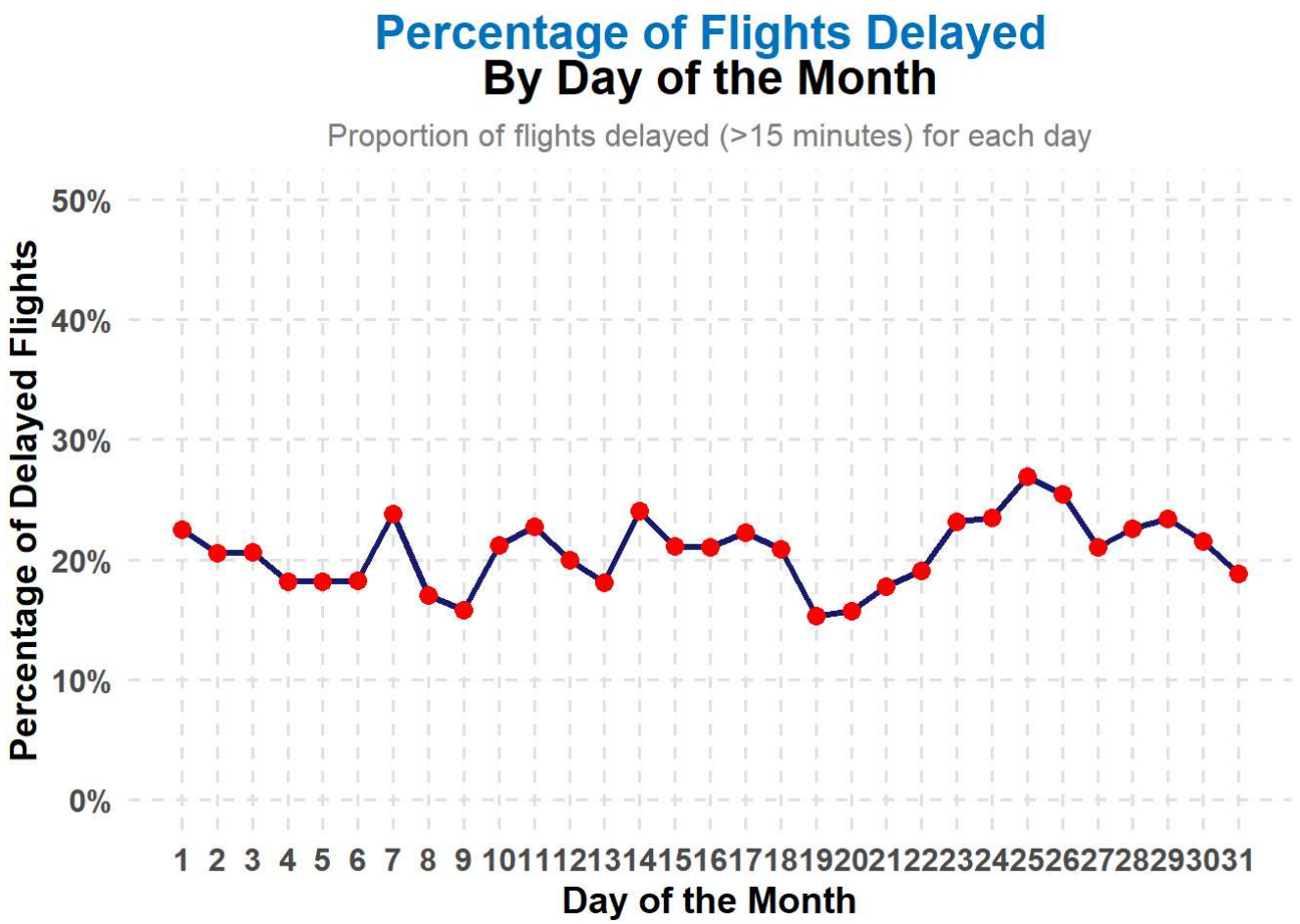
The chart illustrates the proportion of flights classified as on-time or late for each airline carrier. Green bars represent on-time flights, while red bars show late flights. The dashed line indicates the average proportion of late flights across all carriers. Airlines like G4 and YX have a higher proportion of on-time flights, whereas carriers like F9 and HA have a larger share of delayed flights, highlighting variability in punctuality across airlines.

▶ Code



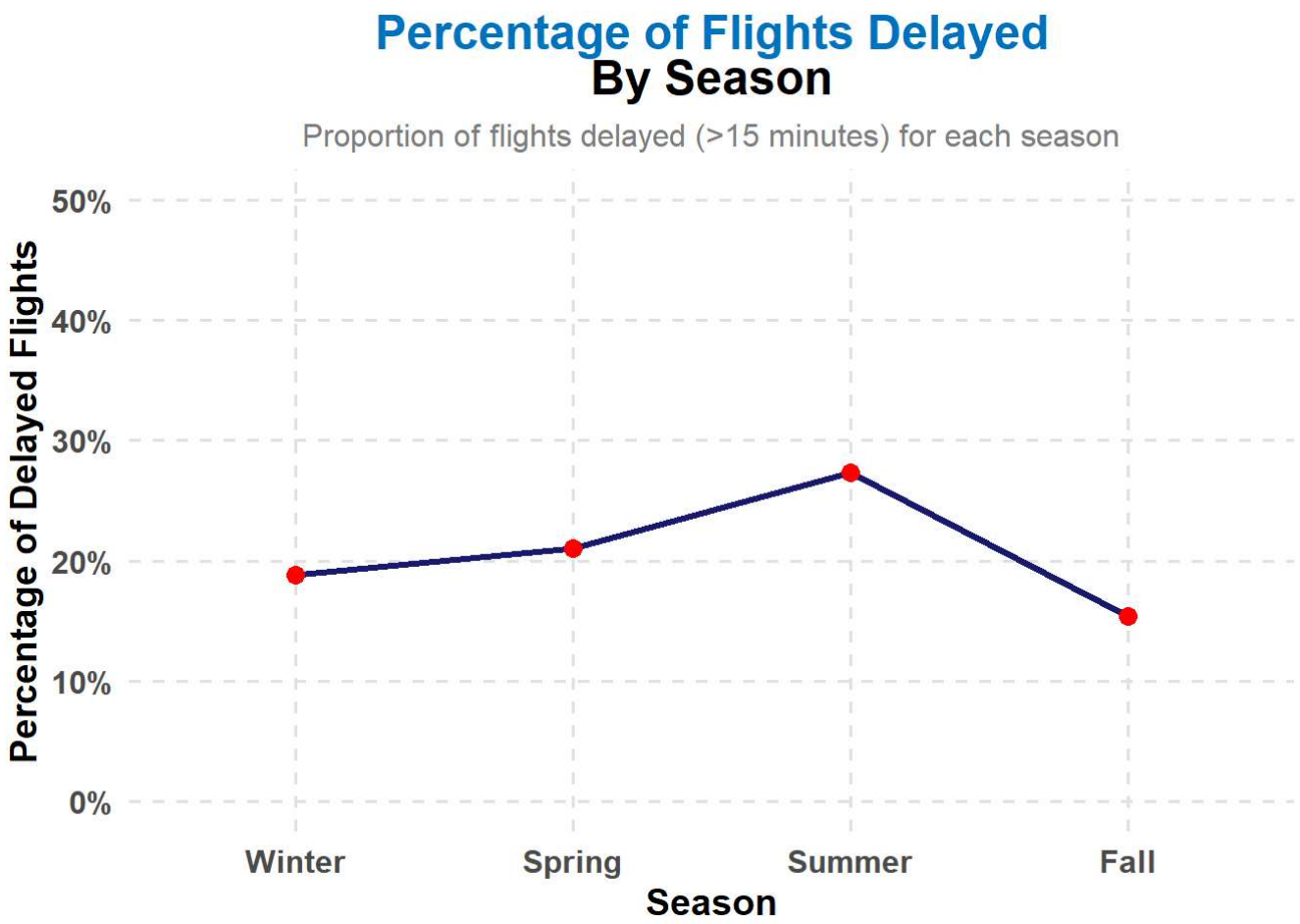
The chart displays the percentage of late flights by origin airport, highlighting the proportion of flights arriving more than 15 minutes late. Each bar represents an origin airport, with the height indicating the percentage of delayed flights. The highest percentage of late flights is evident at specific airports, emphasizing variations in punctuality influenced by airport-specific factors. The percentages are clearly labeled, aiding quick interpretation.

▶ Code



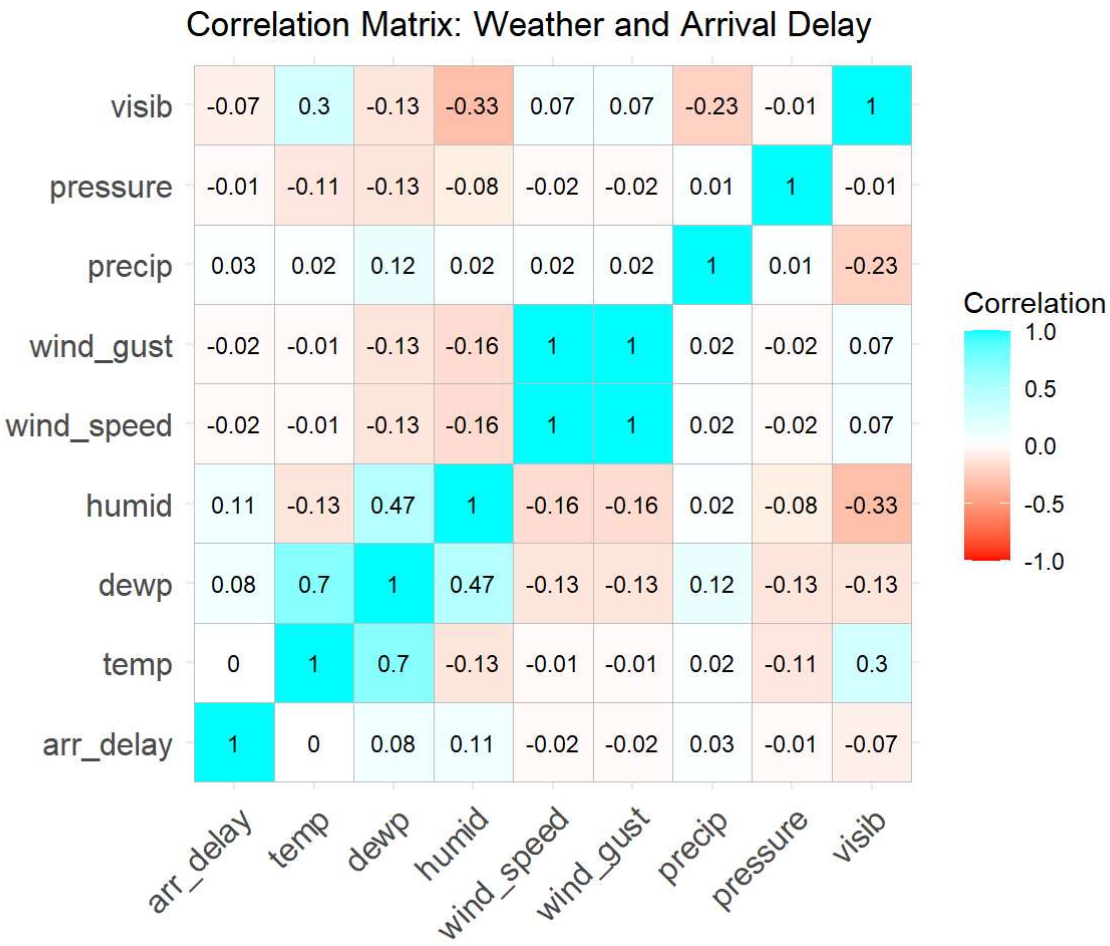
The line chart illustrates the percentage of flights delayed by day of the month. The trend fluctuates, with the proportion of delays generally ranging between 20% and 30%. Peaks on specific days suggest variations in operational or external factors affecting punctuality. The visualization highlights potential patterns in delays over the month.

▶ Code



The line chart displays the percentage of flights delayed by season. Delays peak during the summer, with approximately 25% of flights delayed, while the fall has the lowest delay percentage, around 15%. This seasonal trend reflects variations in factors like weather conditions and travel demand.

▶ Code



The correlation matrix illustrates the relationships between weather variables and arrival delay. Most correlations with arr_delay are weak, with humid showing the strongest positive correlation (0.11), indicating higher delays with increased humidity. Other variables like visib and pressure exhibit minimal or negligible correlation with arrival delays, suggesting limited direct weather impact. The results highlight that while weather factors contribute, they alone do not heavily dictate delays.`

4 Data Preprocessing

In the preprocessing step, we prepared the dataset for modeling by transforming variables and normalizing numerical features. We left-joined flights data with weather data using `time_hour` and `origin` as keys. Rows with missing values in critical weather variables (`wind_speed`, `visib`, `precip`, `temp`) were removed to ensure data quality. To handle categorical variables, we grouped less significant carriers into an "Other" category for simplicity. Finally, numerical features such as `distance`, `wind_speed`, `visib`, `precip`, and `temp` were normalized using min-max scaling to improve model performance and comparability across features.

▶ Code

5 Supervised Learning

5.1 Regression task

For the regression task, we aimed to predict flight arrival delays `arr_delay` based on various features, including flight `distance`, weather conditions, and `carrier` information. Our goal was to build and

evaluate a regression model that could identify the key factors contributing to delays and provide reasonable predictions. We explored feature engineering, refined our model by addressing multicollinearity and insignificant predictors, and assessed its performance using metrics like *RMSE* and *MAE* to ensure practical insights.

▶ Code

```
# A tibble: 22 × 5
   term            estimate std.error statistic  p.value
   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)        14.5       2.29     6.30  3.06e-10
 2 distance_norm     -29.3       3.72    -7.89  3.10e-15
 3 wind_speed_norm    -2.23      2.75    -0.813 4.17e- 1
 4 visib_norm        -17.8       1.41   -12.6   2.15e-36
 5 precip_norm         8.77      7.92     1.11  2.68e- 1
 6 temp_norm          25.4       3.69     6.87  6.48e-12
 7 carrierAA           4.67      2.02     2.31  2.10e- 2
 8 carrierAS           5.94      3.67     1.62  1.06e- 1
 9 carrierB6          20.4       1.75    11.6   4.94e-31
10 carrierDL           4.99      1.84     2.71  6.70e- 3
# ℹ 12 more rows
```

▶ Code

```
# A tibble: 20 × 5
   term               estimate std.error statistic  p.value
   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)           13.9       2.06     6.76  1.43e-11
 2 distance_norm        -29.3       3.72    -7.89  3.14e-15
 3 visib_norm           -18.2       1.36   -13.4   1.16e-40
 4 temp_norm             25.9       3.67     7.05  1.86e-12
 5 carrierAA              4.72      2.02     2.33  1.98e- 2
 6 carrierAS              6.07      3.67     1.65  9.84e- 2
 7 carrierB6             20.3       1.75    11.6   5.06e-31
 8 carrierDL              5.00      1.84     2.72  6.61e- 3
 9 carrierF9             29.7       9.61     3.09  2.00e- 3
10 carrierG4             -5.17     11.2     -0.461 6.45e- 1
11 carrierHA             48.6      14.8      3.28  1.05e- 3
12 carrierMQ            -17.8      17.0     -1.05  2.92e- 1
13 carrierNK             11.2       2.93     3.82  1.34e- 4
14 carrierOO              4.75      3.67     1.29  1.96e- 1
15 carrierUA              6.50      1.79     3.63  2.82e- 4
16 carrierWN              4.73      3.20     1.48  1.40e- 1
17 carrierYX             -8.72      1.62    -5.39  7.19e- 8
18 part_of_dayEvening    14.2       1.13    12.6   4.13e-36
19 part_of_dayMorning   -11.2       1.06   -10.6   2.76e-26
20 part_of_dayNight     -16.4       5.18    -3.17  1.52e- 3
```

▶ Code

```
# A tibble: 11 × 5
   term                           estimate std.error statistic  p.value
   <chr>                             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                        46.0       5.72     8.05  8.88e-16
```

```
 2 distance_norm                      -8.17      5.23    -1.56  1.18e- 1
 3 part_of_dayEvening                 11.6       1.81     6.39  1.73e-10
 4 part_of_dayMorning                 -9.45      1.71    -5.52  3.48e- 8
 5 part_of_dayNight                  -13.5      12.0     -1.13  2.58e- 1
 6 visib_norm                        -52.6       5.96    -8.82  1.25e-18
 7 temp_norm                         -56.1      14.3     -3.92  8.91e- 5
 8 distance_norm:part_of_dayEvening   16.6       7.66     2.16  3.07e- 2
 9 distance_norm:part_of_dayMorning   -8.30      7.18    -1.16  2.47e- 1
10 distance_norm:part_of_dayNight     12.8      44.3      0.289 7.73e- 1
11 visib_norm:temp_norm               89.8      14.9      6.03  1.67e- 9
```

▶ Code

```
RMSE = 57.06862
```

▶ Code

```
MAE = 33.76829
```

▶ Code

```
R-squared = 0.04544999
```

5.2 Classification task

#Question 1: Predicting Late Arrivals Based on Departure Conditions

▶ Code

```
'data.frame':   339120 obs. of  46 variables:
 $ X                    : int  1 3 4 5 6 8 9 10 11 12 ...
 $ year                 : int  2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...
 $ month                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ day                  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ dep_time             : int  1 31 33 36 503 524 537 547 549 551 ...
 $ sched_dep_time       : int  2038 2344 2140 2048 500 530 520 545 559 600 ...
 $ dep_delay            : num [1:339120, 1] 3.498 0.615 2.943 3.96 -0.198 ...
 $ arr_time             : int  328 500 238 223 808 645 926 845 905 846 ...
 $ sched_arr_time       : int  3 426 2352 2252 815 710 818 852 901 859 ...
 $ arr_delay            : num  205 34 166 211 -7 -25 68 -7 4 -13 ...
 $ carrier              : chr  "UA" "B6" "B6" "UA" ...
 $ flight               : int  628 371 1053 219 499 981 206 225 800 93 ...
 $ tailnum              : chr  "N25201" "N807JB" "N265JB" "N17730" ...
 $ origin               : chr  "EWR" "JFK" "JFK" "EWR" ...
 $ dest                 : chr  "SMF" "BQN" "CHS" "DTW" ...
 $ air_time             : num [1:339120, 1] 2.529 0.542 -0.379 -0.694 0.137 ...
 $ distance             : num [1:339120, 1] 2.154 0.843 -0.491 -0.701 0.146 ...
 $ hour                 : int  20 23 21 20 5 5 5 5 5 6 ...
 $ minute               : int  38 44 40 48 0 30 20 45 59 0 ...
 $ time_hour            : chr  "2023-01-01 20:00:00" "2023-01-01 23:00:00" "2023-01-01
21:00:00" "2023-01-01 20:00:00" ...
 $ is_late              : Factor w/ 2 levels "0","1": 2 2 2 2 1 1 2 1 1 1 ...
 $ season               : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ dep_hour             : num [1:339120, 1] 1.41 2.04 1.62 1.41 -1.75 ...
 $ dep_minute           : num [1:339120, 1] 0.484 0.79 0.586 0.994 -1.453 ...
```

```
 $ dep_time_of_day        : chr  "Evening" "Night" "Evening" "Evening" ...
 $ origin_JFK             : int  0 1 1 0 0 0 0 0 1 0 ...
 $ origin_LGA             : int  0 0 0 0 0 0 0 0 0 1 ...
 $ carrier_AA             : int  0 0 0 0 1 1 0 0 0 0 ...
 $ carrier_AS             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_B6             : int  0 1 1 0 0 0 0 0 1 1 ...
 $ carrier_DL             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_F9             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_G4             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_HA             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_MQ             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_NK             : int  0 0 0 0 0 0 0 1 0 0 ...
 $ carrier_OO             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_UA             : int  1 0 0 1 0 0 1 0 0 0 ...
 $ carrier_WN             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_YX             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dep_time_of_day_Evening: int  1 0 1 1 0 0 0 0 0 0 ...
 $ dep_time_of_day_Morning: int  0 0 0 0 1 1 1 1 1 1 ...
 $ dep_time_of_day_Night  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ distance_airtime_ratio : num [1:339120, 1] 0.852 1.557 1.294 1.01 1.066 ...
 $ dep_delay_squared      : num [1:339120, 1] 12.2328 0.3779 8.6621 15.6779 0.0394 ...
 $ distance_squared       : num [1:339120, 1] 4.6406 0.7108 0.2408 0.4911 0.0214 ...
```

▶ Code

```
'data.frame':   84779 obs. of  46 variables:
 $ X                      : int  2 7 23 24 36 38 45 50 52 55 ...
 $ year                   : int  2023 2023 2023 2023 2023 2023 2023 2023 2023 2023 ...
 $ month                  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ day                    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ dep_time               : int  18 520 611 611 644 652 701 704 704 706 ...
 $ sched_dep_time         : int  2300 510 530 620 645 700 705 705 700 711 ...
 $ dep_delay              : num [1:84779, 1] 1.188 -0.069 0.504 -0.42 -0.272 ...
 $ arr_time               : int  228 948 923 913 936 947 953 845 1005 953 ...
 $ sched_arr_time         : int  135 949 839 930 959 1033 1016 853 1018 957 ...
 $ arr_delay              : num  53 -1 44 -17 -23 -46 -23 -8 -13 -4 ...
 $ carrier                : chr  "DL" "B6" "B6" "UA" ...
 $ flight                 : int  393 996 646 311 972 4 585 729 535 871 ...
 $ tailnum                : chr  "N830DN" "N2043J" "N948JB" "N27273" ...
 $ origin                 : chr  "JFK" "JFK" "EWR" "EWR" ...
 $ dest                   : chr  "ATL" "BQN" "FLL" "FLL" ...
 $ air_time               : num [1:84779, 1] -0.379 0.564 0.238 0.16 0.137 ...
 $ distance               : num [1:84779, 1] -0.315 0.843 0.118 0.118 0.146 ...
 $ hour                   : int  23 5 5 6 6 7 7 7 7 7 ...
 $ minute                 : int  0 10 30 20 45 0 5 5 0 11 ...
 $ time_hour              : chr  "2023-01-01 23:00:00" "2023-01-01 05:00:00" "2023-01-01
05:00:00" "2023-01-01 06:00:00" ...
 $ is_late                : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 1 1 1 1 ...
 $ season                 : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ dep_hour               : num [1:84779, 1] 2.04 -1.75 -1.75 -1.54 -1.54 ...
 $ dep_minute             : num [1:84779, 1] -1.4528 -0.943 0.0764 -0.4333 0.841 ...
 $ dep_time_of_day        : chr  "Night" "Morning" "Morning" "Morning" ...
 $ origin_JFK             : int  1 1 0 0 0 1 0 1 0 0 ...
 $ origin_LGA             : int  0 0 0 0 0 0 1 0 0 1 ...
 $ carrier_AA             : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
 $ carrier_AS             : int  0 0 0 0 0 1 0 0 0 0 ...
 $ carrier_B6             : int  0 1 1 0 0 0 0 1 0 1 ...
 $ carrier_DL             : int  1 0 0 0 0 0 1 0 0 0 ...
 $ carrier_F9             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_G4             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_HA             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_MQ             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_NK             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_OO             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_UA             : int  0 0 0 1 1 0 0 0 1 0 ...
 $ carrier_WN             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ carrier_YX             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dep_time_of_day_Evening: int  0 0 0 0 0 0 0 0 0 0 ...
 $ dep_time_of_day_Morning: int  0 1 1 1 1 1 1 1 1 1 ...
 $ dep_time_of_day_Night  : int  1 0 0 0 0 0 0 0 0 0 ...
 $ distance_airtime_ratio : num [1:84779, 1] 0.83 1.495 0.495 0.738 1.066 ...
 $ dep_delay_squared      : num [1:84779, 1] 1.41036 0.00476 0.25386 0.17652 0.07415 ...
 $ distance_squared       : num [1:84779, 1] 0.0991 0.7108 0.0139 0.0139 0.0214 ...
```

▶ Code

```
Warning: glm.fit: algorithm did not converge

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

▶ Code

```

Call:
glm(formula = is_late ~ dep_delay + distance + air_time + origin_JFK +
    origin_LGA + carrier_DL + carrier_AA + dep_time_of_day_Evening +
    dep_time_of_day_Night + distance_airtime_ratio + dep_delay_squared +
    distance_squared, family = "binomial", data = train_data)

Coefficients:
                          Estimate Std. Error    z value Pr(>|z|)
(Intercept)             -1.420e+14  2.327e+05 -6.100e+08   <2e-16 ***
dep_delay                1.617e+15  1.724e+05  9.378e+09   <2e-16 ***
distance                -1.961e+15  7.876e+05 -2.490e+09   <2e-16 ***
air_time                 1.883e+15  7.690e+05  2.449e+09   <2e-16 ***
origin_JFK              -3.374e+13  3.023e+05 -1.116e+08   <2e-16 ***
origin_LGA              -8.815e+13  2.914e+05 -3.025e+08   <2e-16 ***
carrier_DL              -7.511e+13  3.508e+05 -2.141e+08   <2e-16 ***
carrier_AA              -5.164e+12  4.077e+05 -1.267e+07   <2e-16 ***
dep_time_of_day_Evening  7.011e+12  2.730e+05  2.568e+07   <2e-16 ***
dep_time_of_day_Night    9.414e+13  4.660e+06  2.020e+07   <2e-16 ***
distance_airtime_ratio  -9.427e+11  2.865e+04 -3.291e+07   <2e-16 ***
dep_delay_squared       -5.130e+13  1.531e+04 -3.350e+09   <2e-16 ***
distance_squared         2.246e+13  7.743e+04  2.901e+08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  345331  on 339119  degrees of freedom
```

```
Residual deviance: 1864322  on 339107  degrees of freedom
AIC: 1864348

Number of Fisher Scoring iterations: 25
```

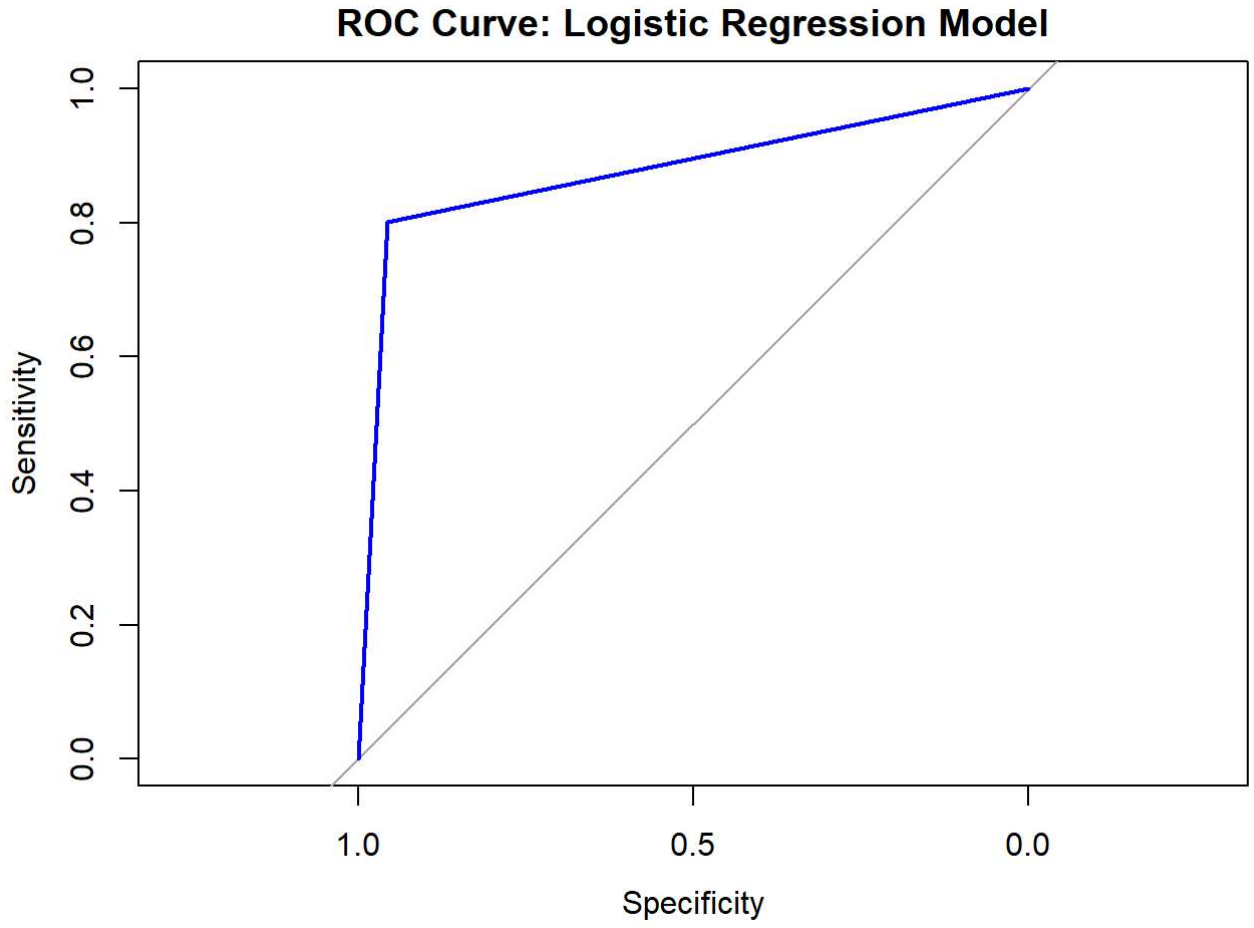
▶ Code

```
Confusion Matrix and Statistics

          Reference
Prediction OnTime  Late
    OnTime  64298  3487
    Late     2982 14012

               Accuracy : 0.9237
                 95% CI : (0.9219, 0.9255)
    No Information Rate : 0.7936
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7646

 Mcnemar's Test P-Value : 3.697e-10

            Sensitivity : 0.8007
            Specificity : 0.9557
         Pos Pred Value : 0.8245
         Neg Pred Value : 0.9486
             Prevalence : 0.2064
         Detection Rate : 0.1653
   Detection Prevalence : 0.2005
      Balanced Accuracy : 0.8782

       'Positive' Class : Late
```

▶ Code

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

▶ Code



▶ Code

```
AUC: 0.8782046
```

▶ Code

```
Confusion Matrix and Statistics

          Reference
Prediction OnTime  Late
    OnTime  64298  3487
    Late     2982 14012

               Accuracy : 0.9237
                 95% CI : (0.9219, 0.9255)
    No Information Rate : 0.7936
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7646

 Mcnemar's Test P-Value : 3.697e-10

            Sensitivity : 0.8007
            Specificity : 0.9557
         Pos Pred Value : 0.8245
         Neg Pred Value : 0.9486
             Prevalence : 0.2064
         Detection Rate : 0.1653
```

```
    Detection Prevalence : 0.2005
       Balanced Accuracy : 0.8782

        'Positive' Class : Late
```

▶ Code

```
Setting levels: control = OnTime, case = Late
Setting direction: controls < cases
```

▶ Code

```
AUC: 0.8782046
```

The ROC curve evaluates the performance of the logistic regression model in predicting flight delays. The Area Under the Curve (AUC) measures the model's ability to distinguish between delayed and on-time flights. A high AUC indicates good predictive accuracy, with the curve significantly above the diagonal line (random guess). This visualization highlights the model's effectiveness in classification tasks.

6 Insights and Discussion

Model Performance Comparison: The regression and classification models demonstrated varying levels of effectiveness in addressing the research questions. The regression model predicted flight arrival delays with moderate accuracy, as evaluated using metrics such as RMSE and MAE. It identified key contributors to delays, including departure delay, distance, and specific time-of-day variables. However, the inherent variability and noise in the dataset limited the model's precision.

The logistic regression model for classification performed well, achieving an AUC of 0.878, indicating strong discrimination between on-time and delayed flights. The confusion matrix revealed a balanced accuracy of 87.8%, with high sensitivity and specificity, highlighting the model's effectiveness in correctly identifying delayed flights without compromising on-time predictions.

Key Findings and Insights:

Impact of Departure Delays: Both models consistently identified departure delay as the most significant predictor of arrival delay, emphasizing its cascading effect on subsequent flights.

Weather Influence: Correlation analysis showed a modest relationship between weather variables (e.g., humidity) and delays, but their overall contribution to the models was minimal.

Seasonal Trends: Delays peaked during the summer, likely due to increased air traffic and weather disruptions, while fall had the lowest delay rates.

Airline Variability: Airlines differed significantly in their punctuality, with Allegiant Air performing best and Frontier Airlines facing the highest delays, as reflected in EDA visualizations.

Destination-Specific Delays: Certain destinations, such as Ponce (PSE), experienced higher average delays, indicating potential operational or environmental challenges.

Limitations: *Data Quality:* Despite data cleaning efforts, missing and imputed values for critical variables may have introduced biases, particularly for weather-related predictors.

Exclusion of Categorical Features: Some categorical variables, like specific flight routes or events, were simplified, which may have overlooked complex relationships.

Temporal Scope: The analysis is limited to a specific year (2013), potentially reducing generalizability to other time-frames with different traffic or weather patterns.

Complex Interactions: The models primarily captured linear relationships, which might not fully represent the nonlinear or interactive effects present in the data.

In conclusion, while the models provided valuable insights into factors influencing flight delays and showed reasonable predictive power, addressing these limitations and incorporating more granular data could further enhance their performance and utility for real-world applications.