Modeling and Forecasting US Airline Flight Delays
2020 TAMIDS Data Science Competition

Big Data Energy
Johnathan Lo & Isaac Ke
Advisor: Dr. Huiyan Sang

April 8, 2020

Contents

Chapter 1

Introduction

Reliable transportation supports a strong economy by facilitating the rapid and timely exchange of goods and services and bolstering tourism revenue. In the United States in 2018, the transportation industry accounted for $648 billion per year, which was 3.16% of the GDP [cite]. Worldwide, the aviation industry contributes $2.7 trillion (3.6%) of the world's GDP. In fact, it is projected that global air transporation will support $5.7 trillion of the global economy [cite]. A key metric for evaluating the efficiency of airline industry production is flight delay time. In 2018, flight delays led to an economic loss of 31.2 billion dollars[cite]. For individual companies, delays can influence consumer choice, and for the industry itself, unmitigated delays can impel consumers to switch to substitute goods, such as automotive or rail-based transport.

Therefore, a major goal of this project is to analyze flight delays and diagnose areas for improvement. We intend to create models using the provided dataset as well as publicly available data that can accurately predict future delays. In doing so, we can hopefully uncover significant and controllable covariates that can help guide airline companies to reduce flight delays.

Chapter 2

Executive Summary

2.1 Problem and Approach

2.2 Data Preprocessing

2.3 Exploratory Analysis

2.4 Model Formulation

2.5 Model Selection

2.6 Applications and Conclusions

Chapter 3

Motivation, Data Description, and Software

3.1 Motivation

As stated in the introduction, flight delays can have a wide-ranging effect on the economy. Most airline companies have already done everything in their power to mitigate and reduce delays. We are interested in finding whether delays can be further alleviated, and whether those variables can be controlled by airline companies. To the extent that some delays are unavoidable or difficult to predict, we are also interested in devising methods to minimize the impact of those delays, whether by reducing the number of passengers affected, offering alternate routes to affected passengers, or discounting tickets. Overall, for the benefit of airline companies, consumers, and society-at-large, we should minimize flight delays, or the impact thereof.

3.2 Data Collection

Our data was provided as .csv files by the competition organizers. The primary dataset was composed of roughly 11 million observations of 50 variables. Each observation was a distinct flight that occurred between 1/1/2018 and 6/30/2019, and the 50 covariates included origin, destination, quarter, arrival delay, departure delay, distance, and many more variables pertaining to each flight. Auxiliary datasets included information on flight routes, airports, and market share.

In addition to these data, we also sought out additional data to enhance our analysis. We obtained geographic coordinates for each airport from *openflights.org* and historical

weather data from the NOAA databases through the NCDC API [cite]. The geographic coordinates are given in decimal format, and our weather data describes meteorological events near the origin and destination of each flight. Importantly, data *along* the flight path was not obtained, due to time constraints and complexity. A full list of covariates along with brief descriptions can be found in Supplementary Table 1.

3.3 Software

All analyses were performed in R v3.6.3 using the RStudio IDE [cite]. Packages used include, but are not limited to, *ggplot2*, *ggmap*, *dplyr*, *caret*, *rnoaa*, and *tseries*. Individual datasets were loaded as *data.frame* objects and combined using *merge* along with various *dplyr* commands . In addition, Microsoft Power BI was utilized in order to tidy the data and perform computationally-intensive data rearranging. The final dataset can be found as a .csv file in Supplementary Data 1.

Chapter 4

Exploratory Data Analysis

4.1 Data Wrangling

Our dataset was drawn from four different main sources - flight delays and airfare data, geographic coordinates from *openflights.org*, and weather data from NOAA. Flight delays and geographic coordinates were combined by merging on both common origin and destination names. The resulting data frame was then combined with fare/market data by common routes, year, and quarter. Adding weather data was more challenging in that the observations related information collected by weather stations, and not the airports themselves. Thus, weather station coordinates were cross referenced with airport coordinates to find the closest active weather station to each airport. Due to this constraint, 10 airports, corresponding to 99,980 observations (a negligable amount given the overall size of our data), were dropped due to the lack of NOAA weather stations nearby. Weather data was then merged with the rest of the data on common dates and airports, with separate variables for weather at origin and weather at destination. The final dataset containing all four sets of information is what will be referenced in this paper hereafter. Small tweaks to these data were made on a model-to-model basis, depending on the type of unique problem selected-whether it be classification or regression, for example.

The data contained a number of variables with numerical values but that could be interpreted either as quantitative or categorical variables. Using substantive knowledge, a number of numeric variables were converted to factors, including day of week, month, quarter, and route number. Additionally, our data contained many missing values. Where applicable, for categorical variables, these values were replaced by adding an additional factor level *Unk*. For quantitative variables, missing values were replaced with

either 0, or observations were removed, based on substantive knowledge of the variable characteristics. Finally, we discovered that the flight delay dataset somewhat bewilderingly assigned canceled flights a delay time of 0. Canceled flights were removed from the dataset and analyzed separately. Thoughout this report, we further to flight delays as specifically arrival delays as opposed to departure delays. In all, the final dataset consisted of 10,614,150 observations of 100 variables.

4.2 Distribution of Flight Delays

A histogram of all arrival delays is shown in Fig 1. Clearly, the data is strongly right-skewed. To correct the skewness, a cube-root transformation was performed, but subsequent Shapiro-Wilk test provided strong evidence against normality for this transformation, so it was abandoned. To heuristically assess dependence between covariates and response, we examined various conditional distributions.

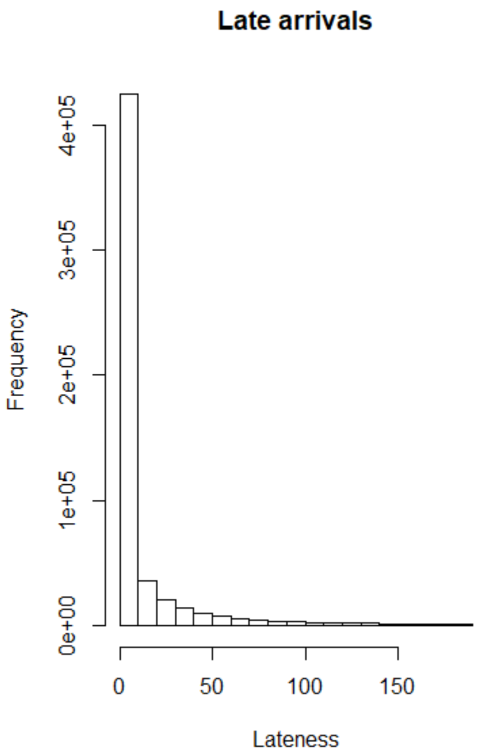

Figure 4.1: Fig 1 - Histogram of arrival time for all observations in the dataset

4.2.1 Geographic Distribution of Flight Delays

The distribution of flight delays across the geographic United States is shown in Fig 2. As we can see, XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX, XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX.

4.2.2 Temporal Distribution of Flight Delays

In Fig 3, we demonstrate the distribution of flight delays by time of day, day of week, month, and quarter. As we can see, XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX, XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX.

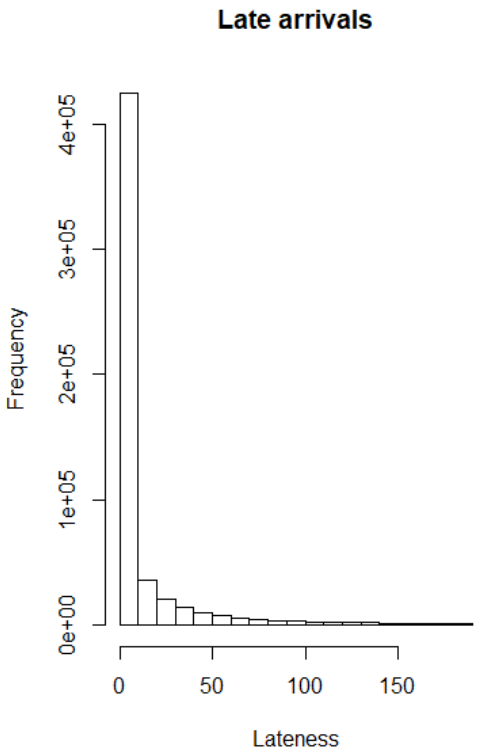

Figure 4.2: Fig 2 - Figures showing the geographic distribution of flight delays by delay length and airport usage

4.2.3 Weather-based Distribution of Flight Delays

Our weather data provided us primarily with data on precipitation and temperature. Other events, such as heavy fog and ice, were observed, but sparse. We show scatterplots of flight delays by precipitation and temperature in Fig 4. In Fig 5, histograms of flight delays for some unusual weather events are given. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX, XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX. XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX.

4.2.4 Carrier-based Distribution of Flight Delays

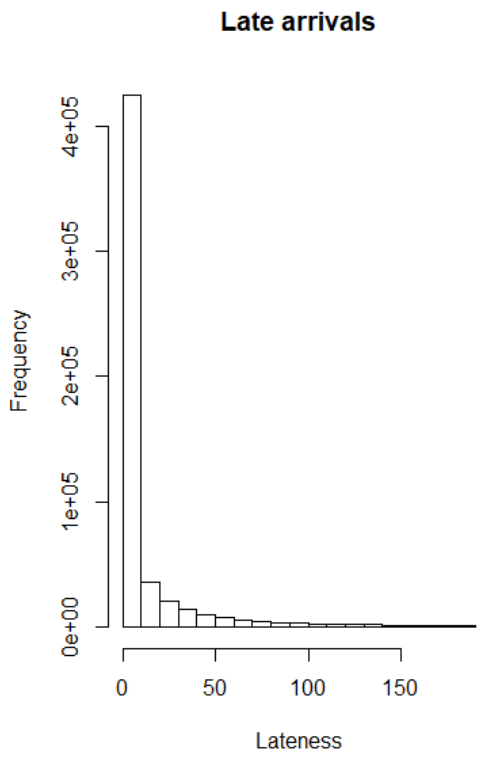

Figure 4.3: Fig 3 - Histograms of ARR DELAY by time of day, day of week, month, and quarter

For each of the 19 carriers, flight delays was plotted. Histograms for each are given in Fig 6. We also hypothesized that there might also be differences in flight delays between different carriers. A one-way ANOVA was conducted, and p-values corrected using Tukey's HSD, shown in Fig 7.

4.2.5 Airport-based Distribution of Flight Delays

We were also interesting in detecting if individual airports showed differences in flight delays, shown in Fig 8.

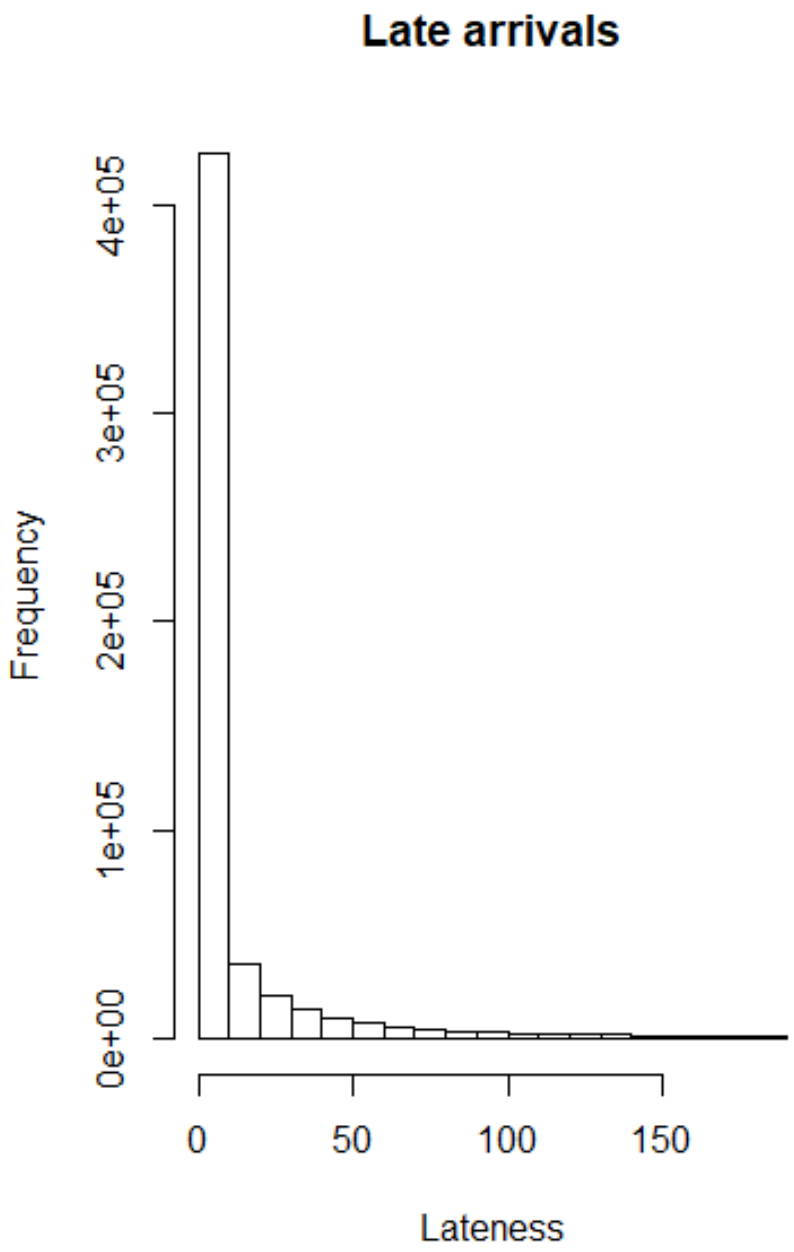

Figure 4.4: Fig 4 - scatterplots of ARR DELAY by precipitation and temperature

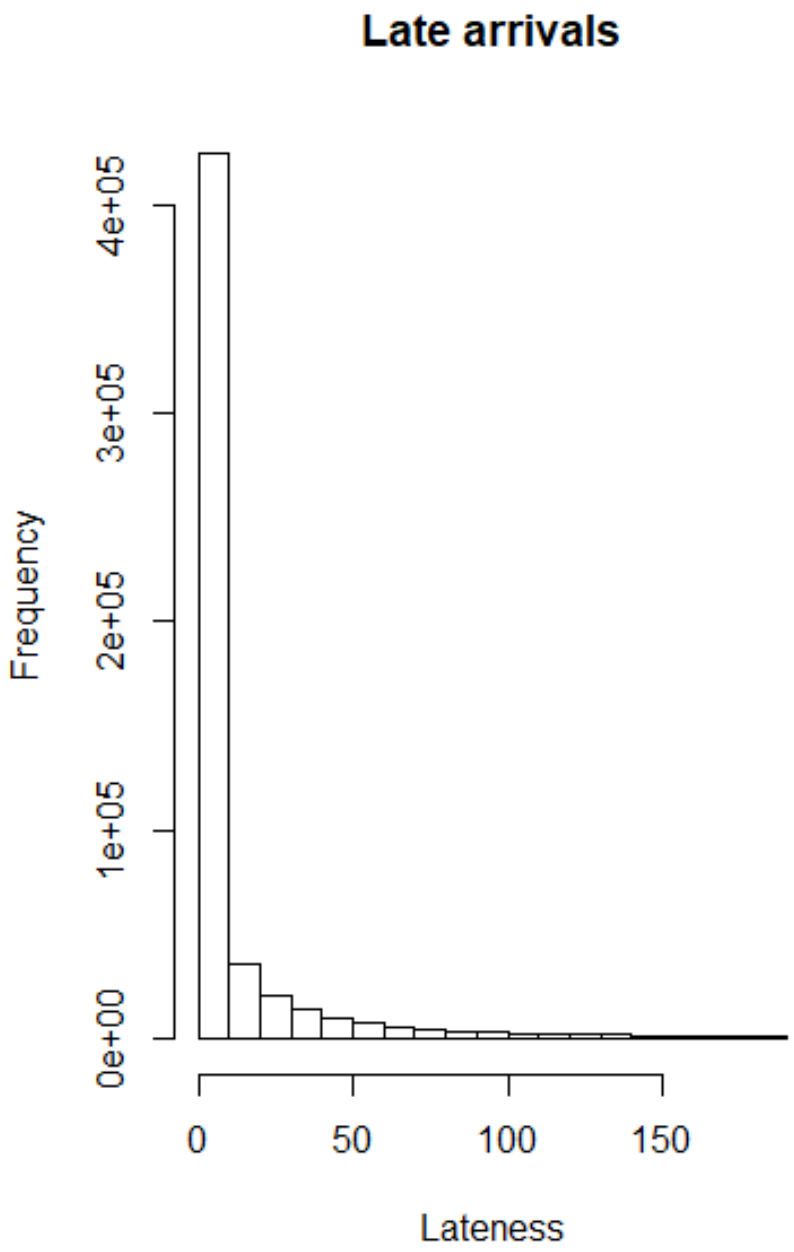

Figure 4.5: Fig 5 - histograms of ARR DELAY for each unusual weather event

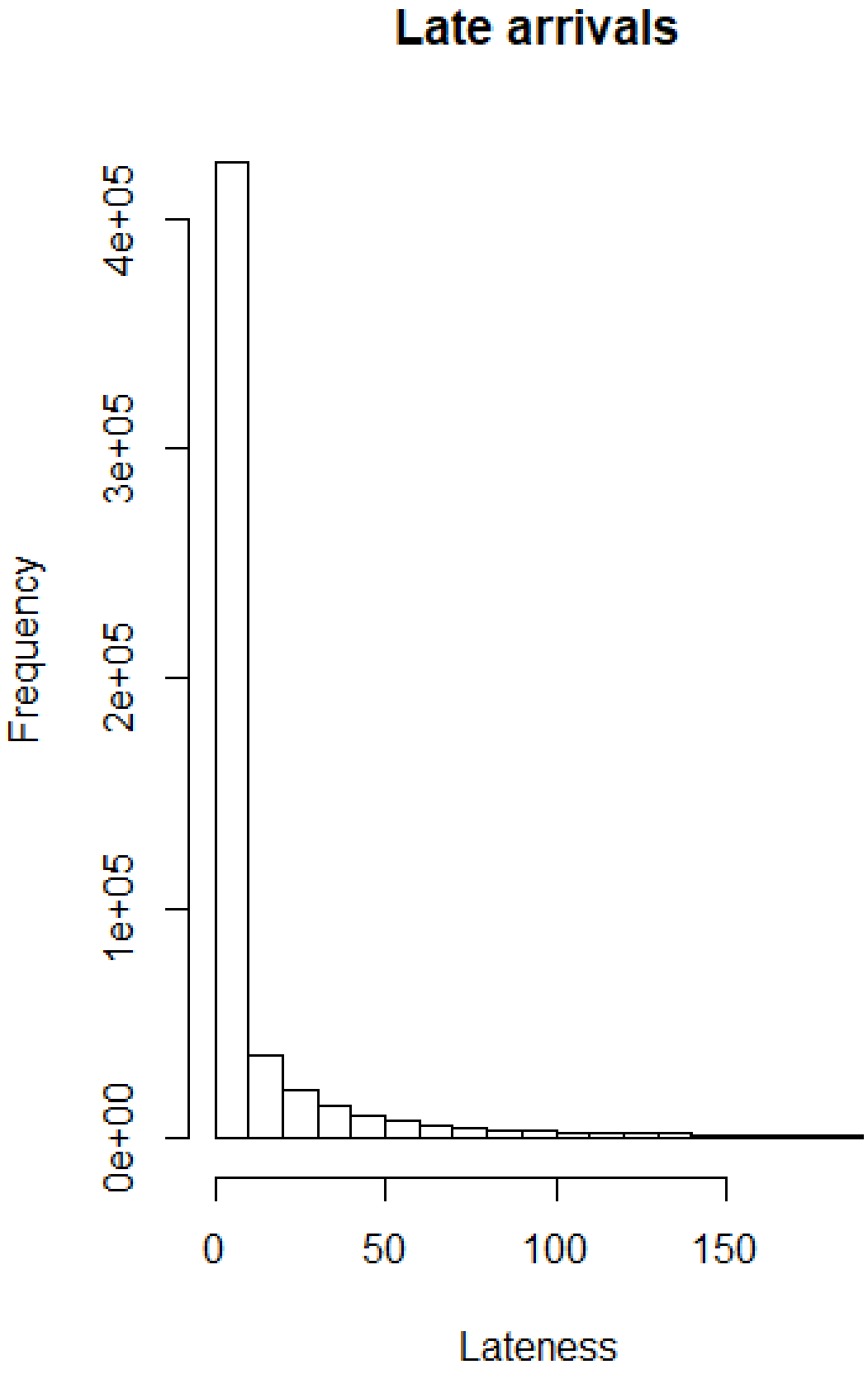

Figure 4.6: Fig 6 - Histograms for the airlines

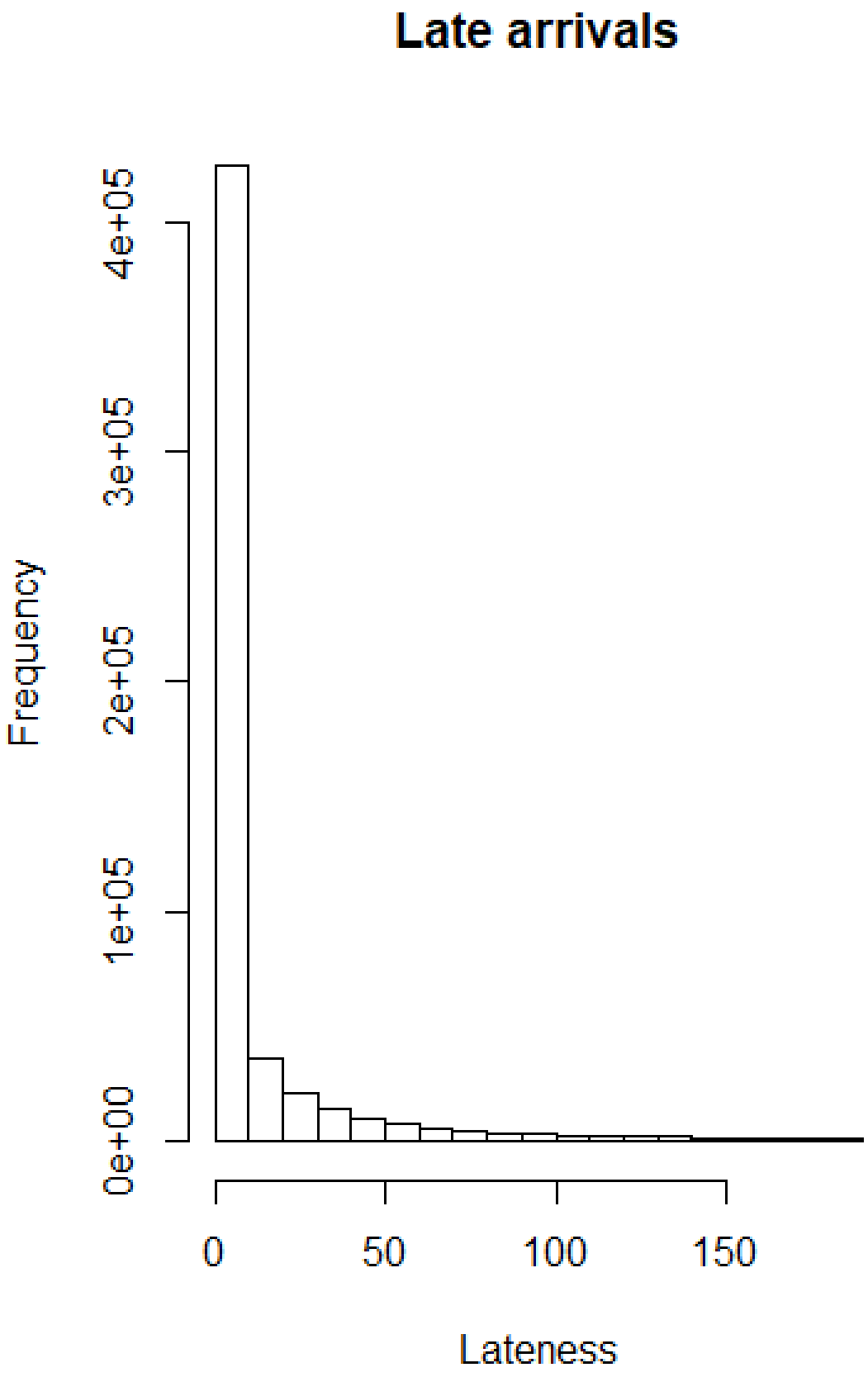

Figure 4.7: Fig 7 - airline ANOVA

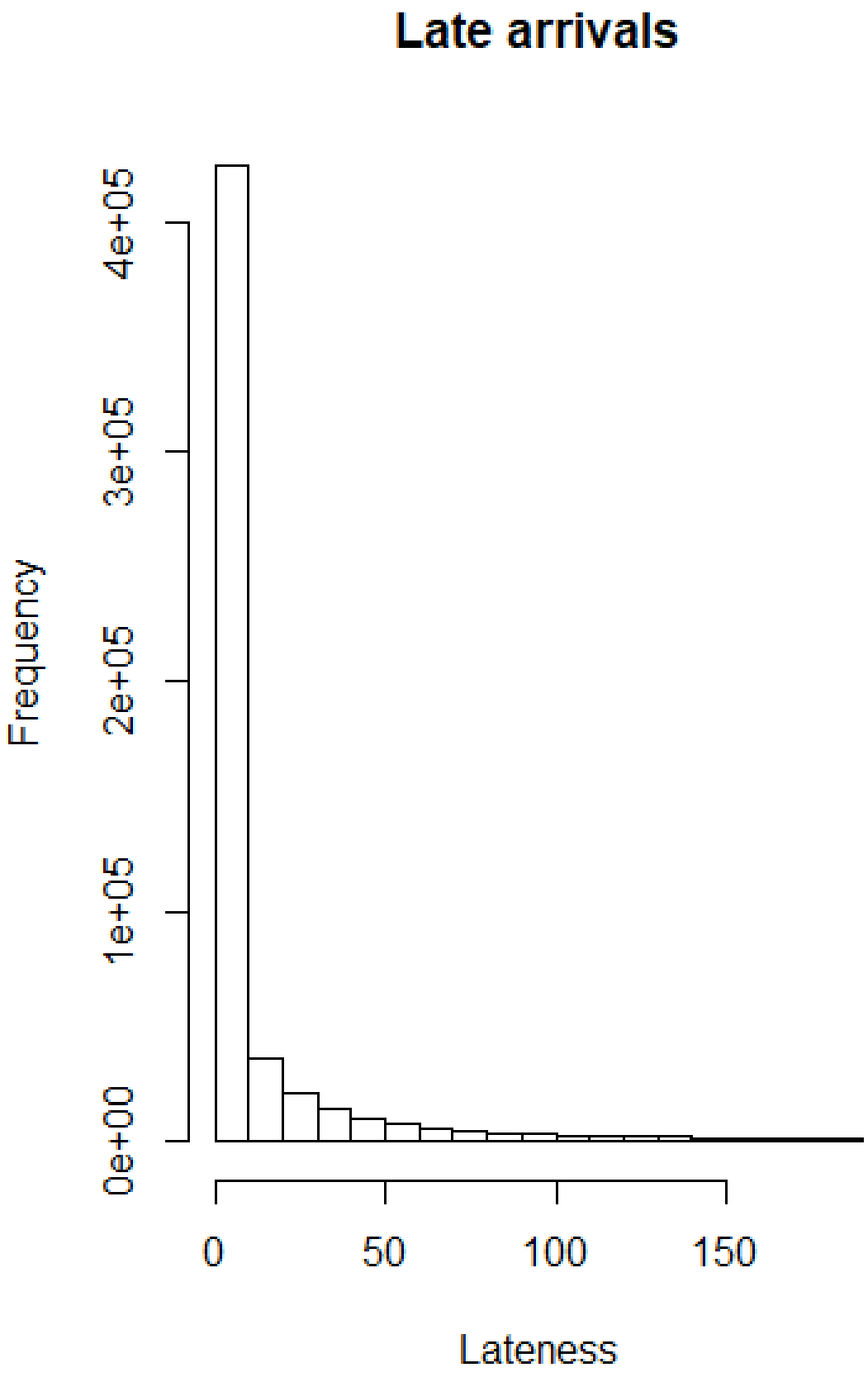

Figure 4.8: Fig 8 - Airport histograms

Chapter 5

Model Formulation

Chapter 6

Model Selection

Chapter 7

Forecasting Flight Delays for 2019 Q3

Chapter 8

Business Recommendations

Chapter 9

Closing Thoughts

Chapter 10

Appendix

10.1 References

10.2 Additional Figures, Tables, Code, and Data

Bibliography

[Mangel u. Clark 1988] Mangel, Marc ; Clark, Colin W.: *Dynamic Modeling in Behavioral Ecology*. Princeton, New Jersey : Princeton University Press, 1988

[Sandholm 2010] Sandholm, William H.: *Population Games and Evolutionary Dynamics*. Cambridge, Massachussetts : The MIT Press, 2010

[Sarah P. Otto 2007] Sarah P. Otto, Troy D.: *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press, 2007. – ISBN 0691123446,9780691123448

[Sigmund 1993] Sigmund, Karl: *Games of Life: Explorations in Ecology, Evolution and Behavior*. Dover Edition. Mineola, New York : Dover Publications, 1993, 2017

[Smith 1982] Smith, John M.: *Evolution and the Theory of Games*. Oxford, Great Britain : Cambridge University Press, 1982