Big Data Energy 2020 TAMIDS Competition

Johnathan Lo & Isaac Ke

3/28/20

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 US, the transportation industry accounts for XXX billion dollars per year, which is XXX% of GDP [cite]. Of that economic product, XXX% is accounted for by the airline industry [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 XXX 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 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 mitigated, 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 to us as csv files by the competition organizers. The primary dataset was composed of over 10,000,000 observations of 50 variables. Each observation was a distinct flight that occurred between 1/1/2018 and XX/XX/2019, and the 70 covariates included origin, destination, quarter, arrival delay, departure delay, distance, and many more variables pertaining to each flight. An auxiliary dataset included pricing data given for each route, by quarter.

In addition to the data provided to us by the competition organizers, we also sought out additional data to enhance our dataset. We obtained geographic coordinates for each

airport from XXXXX [cite], and weather data from 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 [cite]. Packages used include, but are not limited to, *ggplot2*, *dplyr*, *caret*, *rnoaa*, and *Isaac put stuff here*. Individual datasets were loaded as *data.frame* objects and combined using *merge*. 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 4 different sources - flight delays and fare data, provided to us by the competition organizers, geographic coordinates from XXXXX, and weather data from NOAA. Flight delays and geographic coordinates were combined by merging on common origin and destination names. The resulting dataframe was then combined with fare data by common routes, year, and quarter. Adding weather data was more challenging in that the observations relate 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 99980 observations, 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 4 sets of information is what will be referenced in this paper hereafter.

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, the flight delay data somewhat bewilderingly assigned canceled

flights a delay time of 0. Canceled flights were removed from the dataset and analyzed separately. In all, the final dataset consisted of 10614150 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.

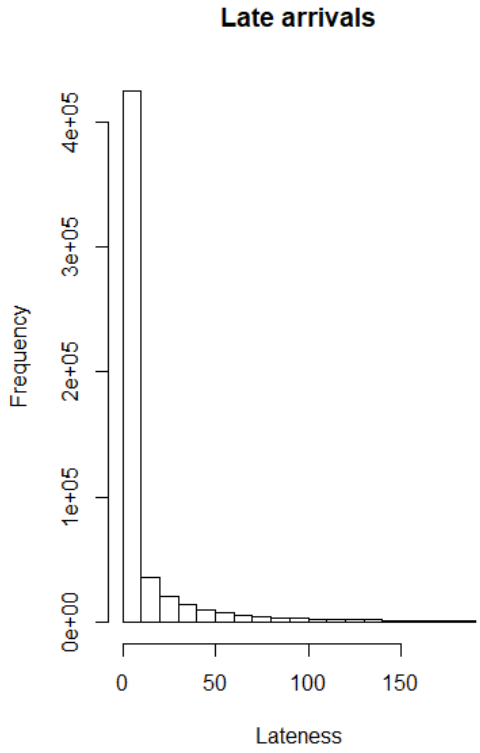


Figure 4.1: Fig 1 - Histogram of arrival time for all observations in the dataset

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 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 XXXXX XXXXX.

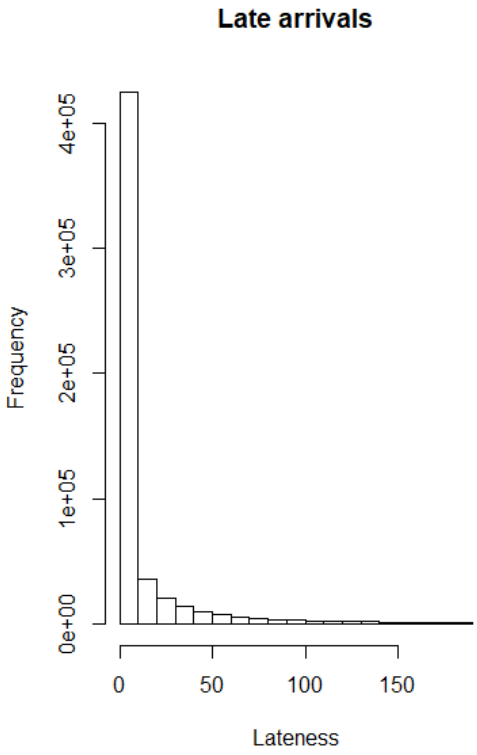


Figure 4.2: Fig 2 - Figures showing the geographic distribution of flight delays by delay length and airport usage

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

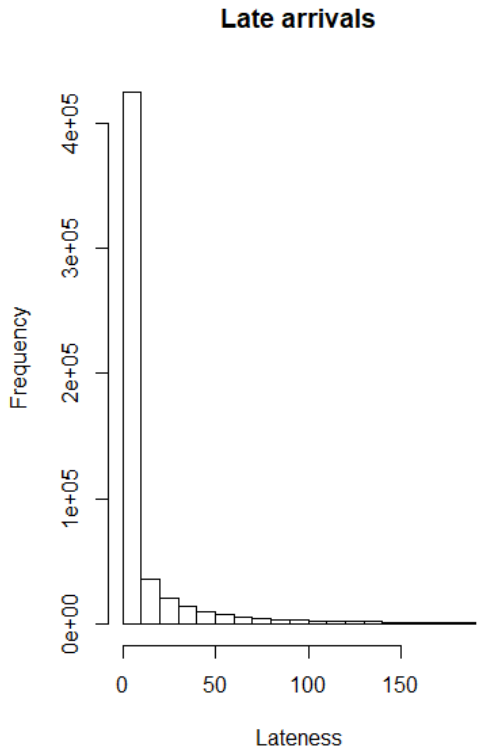


Figure 4.3: Fig 3 - Histograms of ARR DELAY by time of day, day of week, month, and quarter

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.

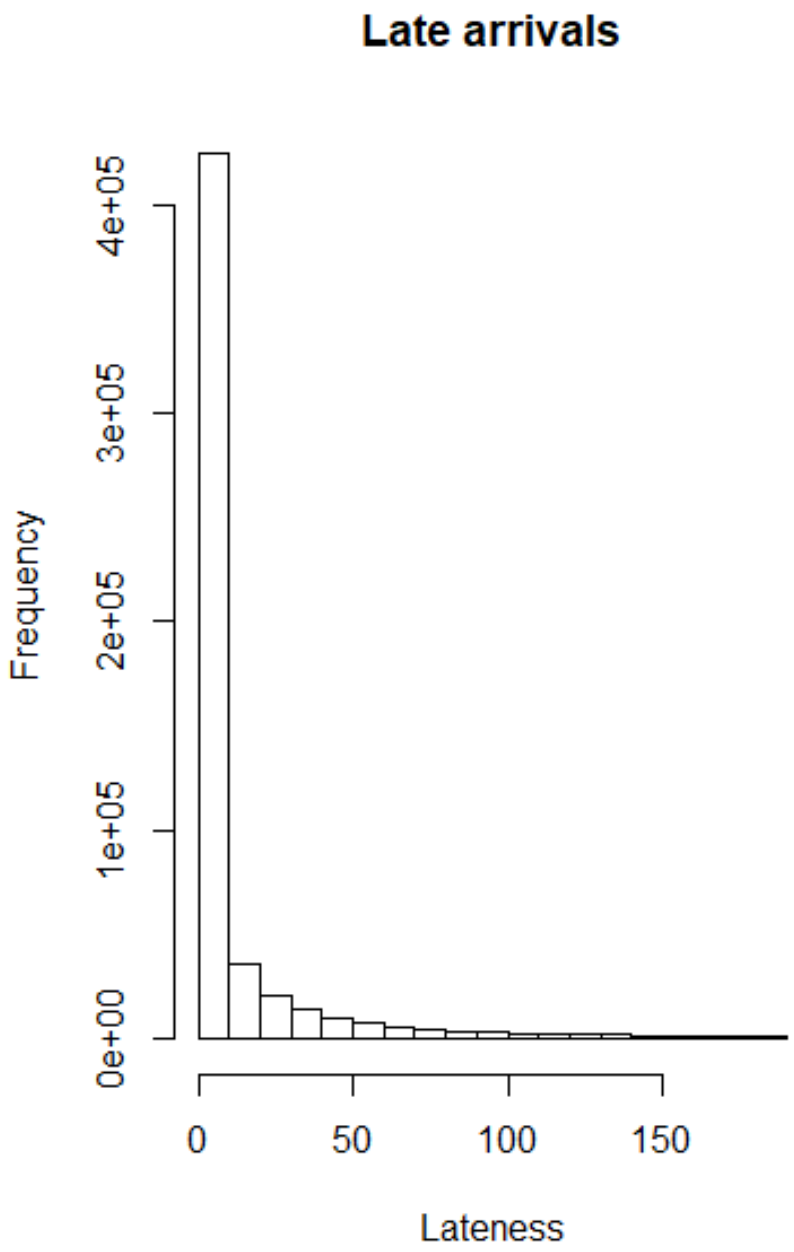


Figure 4.4: Fig 4 - scatterplots of ARR DELAY by precipitation and temperature

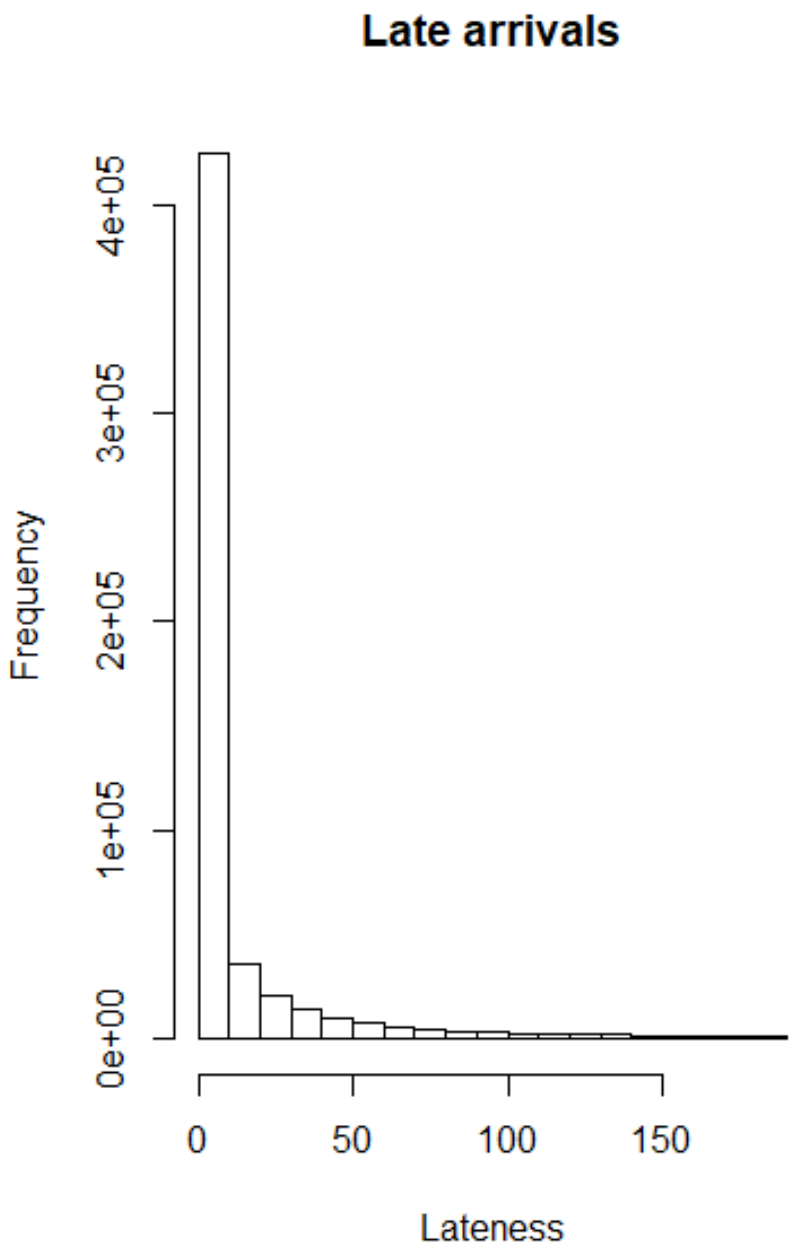


Figure 4.5: Fig 5 - histograms of ARR DELAY for each unusual weather event

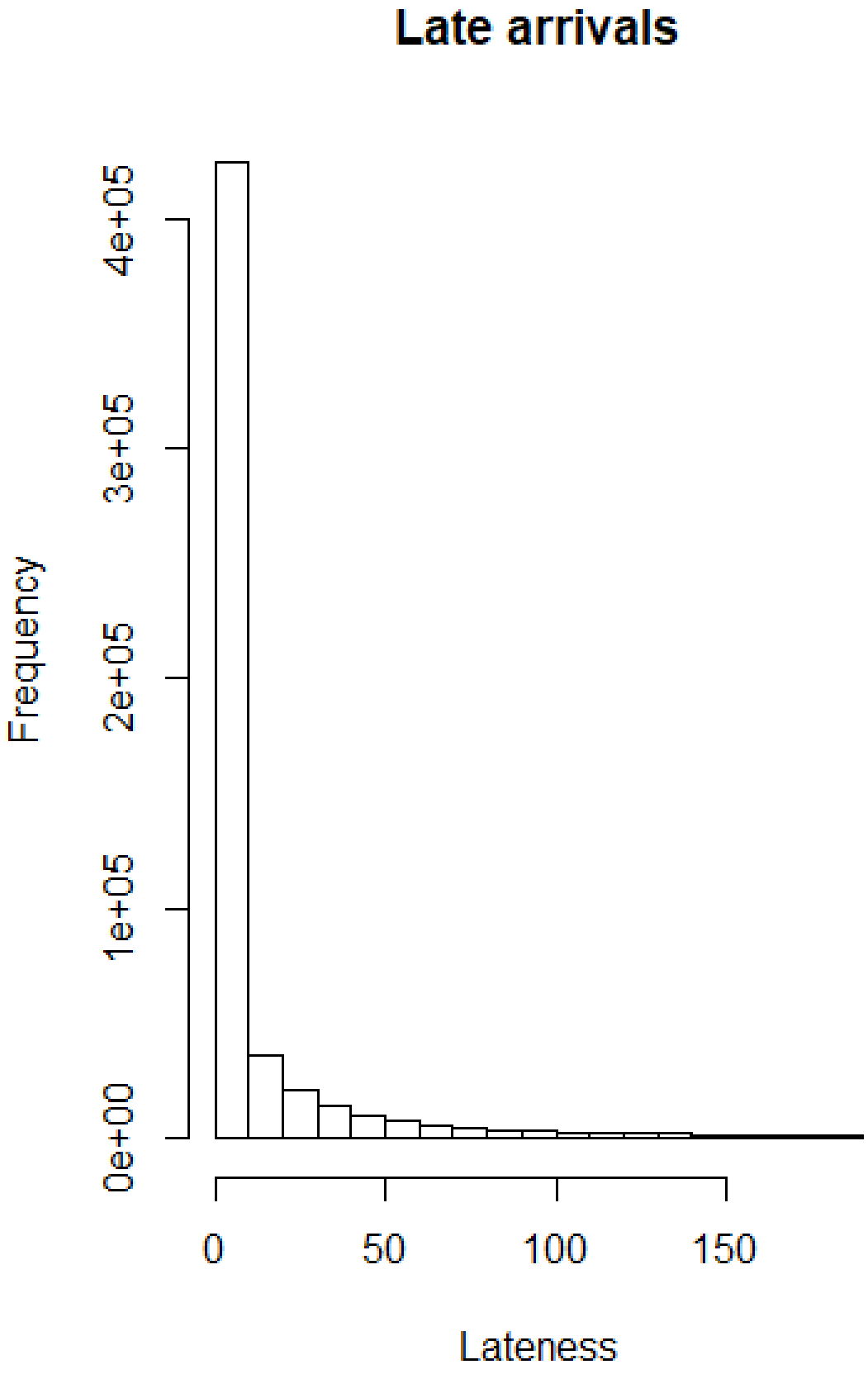


Figure 4.6: Fig 6 - Histograms for the airlines

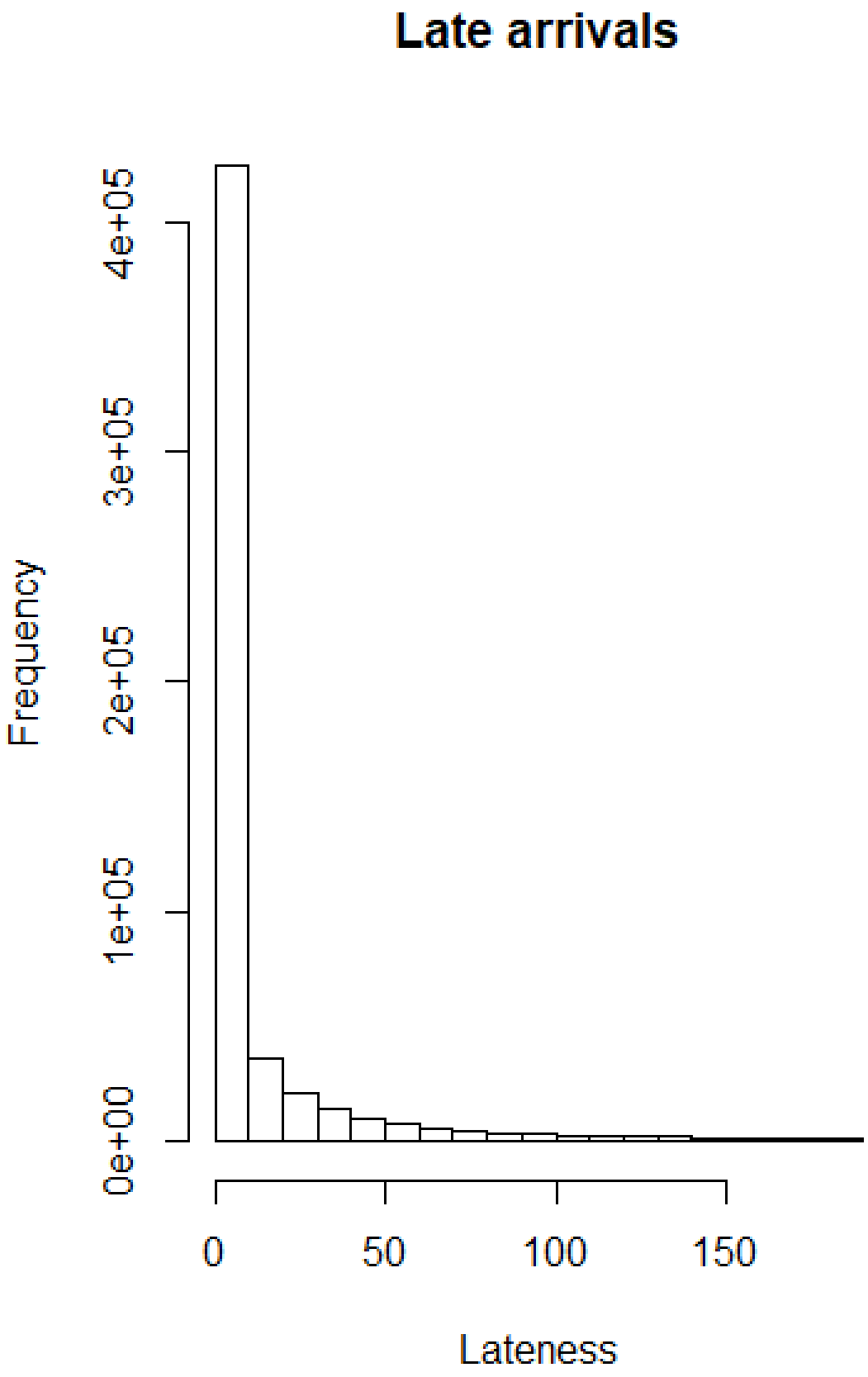


Figure 4.7: Fig 7 - airline ANOVA

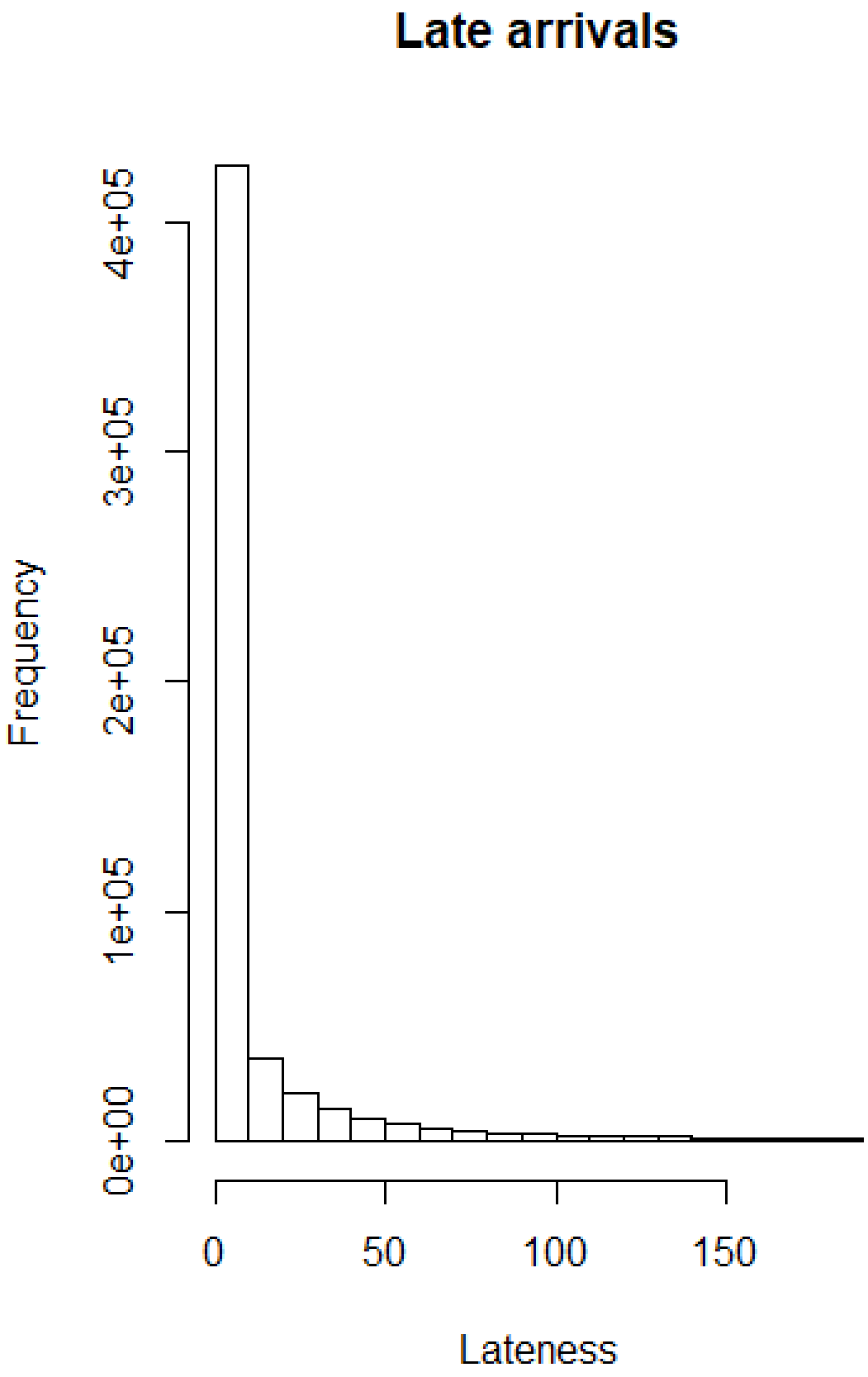


Figure 4.8: Fig 8 - Airport histograms

Chapter 5

Model formulation

5.1 Constructing a parametric distribution for delays

We were interested in deducing the marginal distribution of arrival times from the data. Although it would have been relatively facile to estimate a valid empirical distribution, we decided that a parametric distribution would be more useful and intuitive. By establishing a set of parameters, further work could be directed towards estimating parameters under certain combinations of covariate values. With mostly categorical data, it then becomes feasible to estimate parameters for certain combinations of interesting variables. Parameterization also allows the density functions of the distribution to be readily expressed analytically.

As seen in Fig 1, the marginal distribution is strongly right skewed, and thus our first plan of action was to attempt a transformation to correct the skewness. A number of transformations were considered. With negative values in the data, log or square root transformations could only be applied by first shifting the data to be strictly positive. One way to do this would be to simply add the smallest (most negative) number to each of our delay times; however, it was decided that this approach rendered our results somewhat uninterpretable. Consider, for example, a new observation where the flight arrived earlier (with a flight delay time more negative) than any other flight in the dataset. Such an observation would not be supported in a distribution of log-transformed values. We also briefly considered cube root transformation, but as seen in Fig 9, it does not result in normality or resemblance to any familiar parametric distribution. In lieu of transforming the data, we considered several well-known skewed distributions, but none of them fit well or appeared sensible.

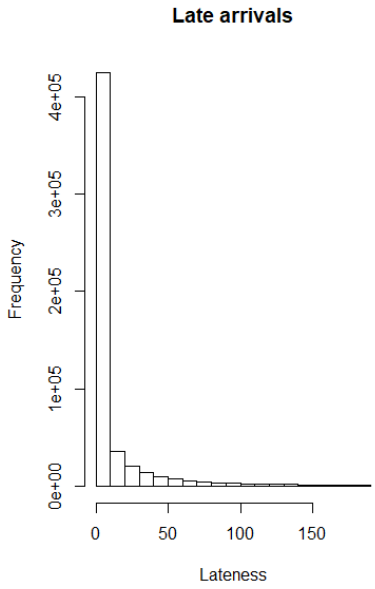
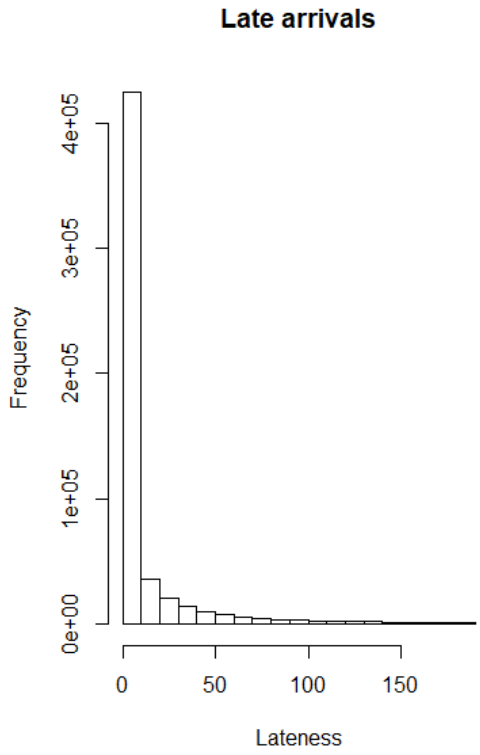


Figure 5.1: Fig 9 - density plot of cube root transformation vs normal distribution with same mean and variance

Thus, we instead looked to construct a mixture distribution. The primary issue that we had been confronted with this far was finding a distribution that appeared to have sensible parameterizations. Our search was rooted in the premise that not all delays are created equal; we suspected that the majority of delays are "run-of-the-mill" events that do not result from any extraordinary circumstances in particular, while a minority of delays have true, substantial causes. This is similar to the rationale for a zero-inflated Poisson in manufacturing processes, in that most machines are in good working order and do not produce any products with defects, but some machines with defects will produce defective products by a Poisson distribution. As such, we decided upon a mixture of the form

$$Y = UV + (1 - U)T$$

with $U \sim Ber(p)$, $V \sim N(\mu, \sigma^2)$, and $T \sim exp(\lambda)$. In this model, U describes whether or not a delay with "true, substantial causes" occurs, V describes the distribution of arrival times when no extenuating circumstances occur, and T describes arrival times under defined circumstances that result in lateness. Using MLE, we were able to generate estimates of the parameters for the marginal data, as shown in Fig 10.



To validate this distribution, a QQ plot was made of the observed arrival times against the mixture distribution under the parameter estimates from the MLE (Fig. 11). From this plot, it can be observed that our mixture distribution is in fact able to describe the observed values of arrival time quite accurately. Note that the arrival times are given as discrete values (as minutes), hence the "jumps" from percentile to percentile. It is important to note here, however, that our distribution does fail to describe data points in the extreme upper range of observations.

An additional approach to validating this distribution would be through conditional density estimation.

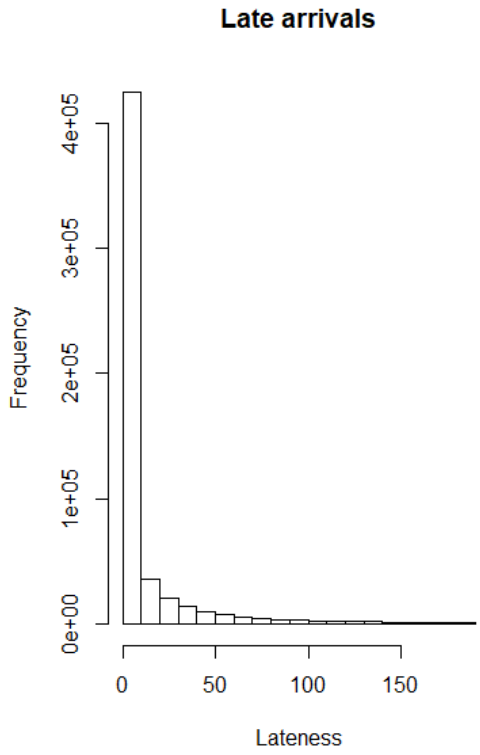


Figure 5.3: Fig 11 - QQ plot of of data vs theoretical distribution

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, 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