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.

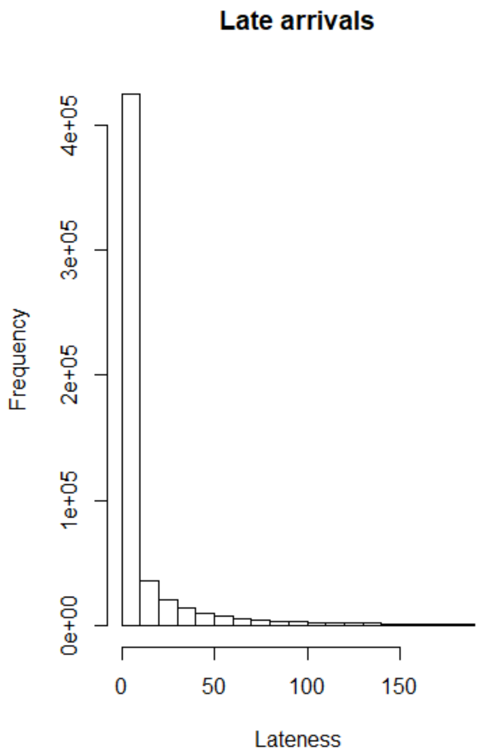


Figure 4.1: Fig 1 - Histogram of arrival time for all observations in the dataset

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 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.

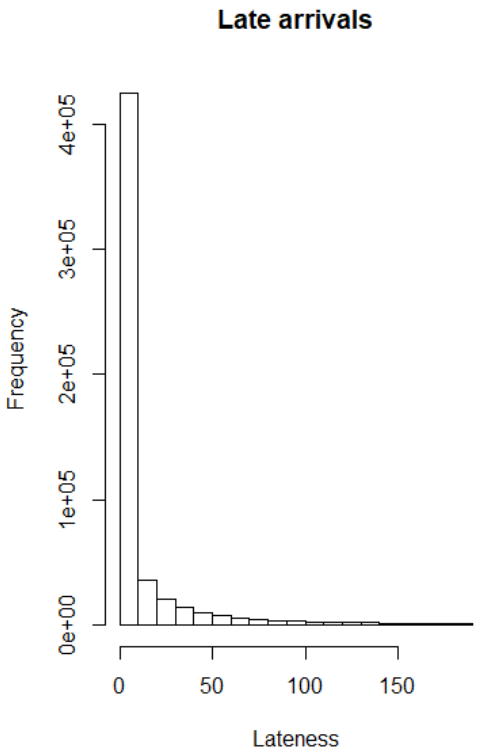


Figure 4.2: Fig 2 - Figures showing the geographic distribution of flight delays by delay length and airport usage

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

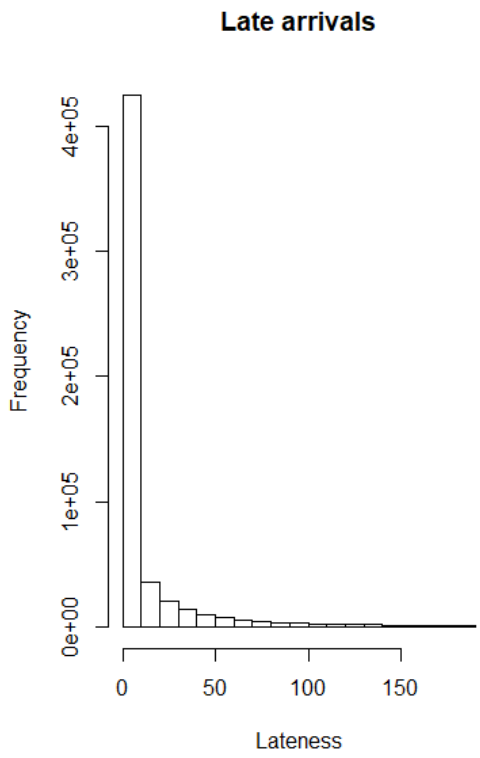


Figure 4.3: Fig 3 - Histograms of ARR DELAY by time of day, day of week, month, and quarter

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.

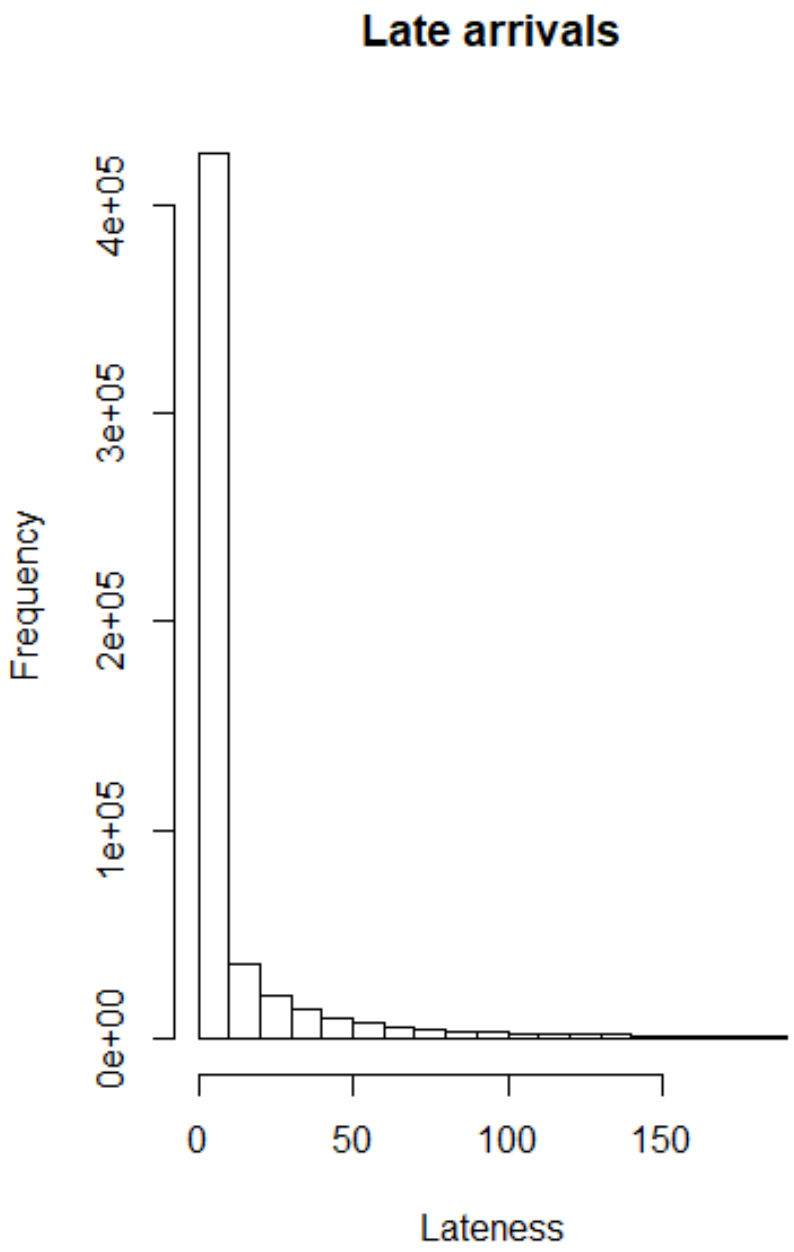


Figure 4.4: Fig 4 - scatterplots of ARR DELAY by precipitation and temperature

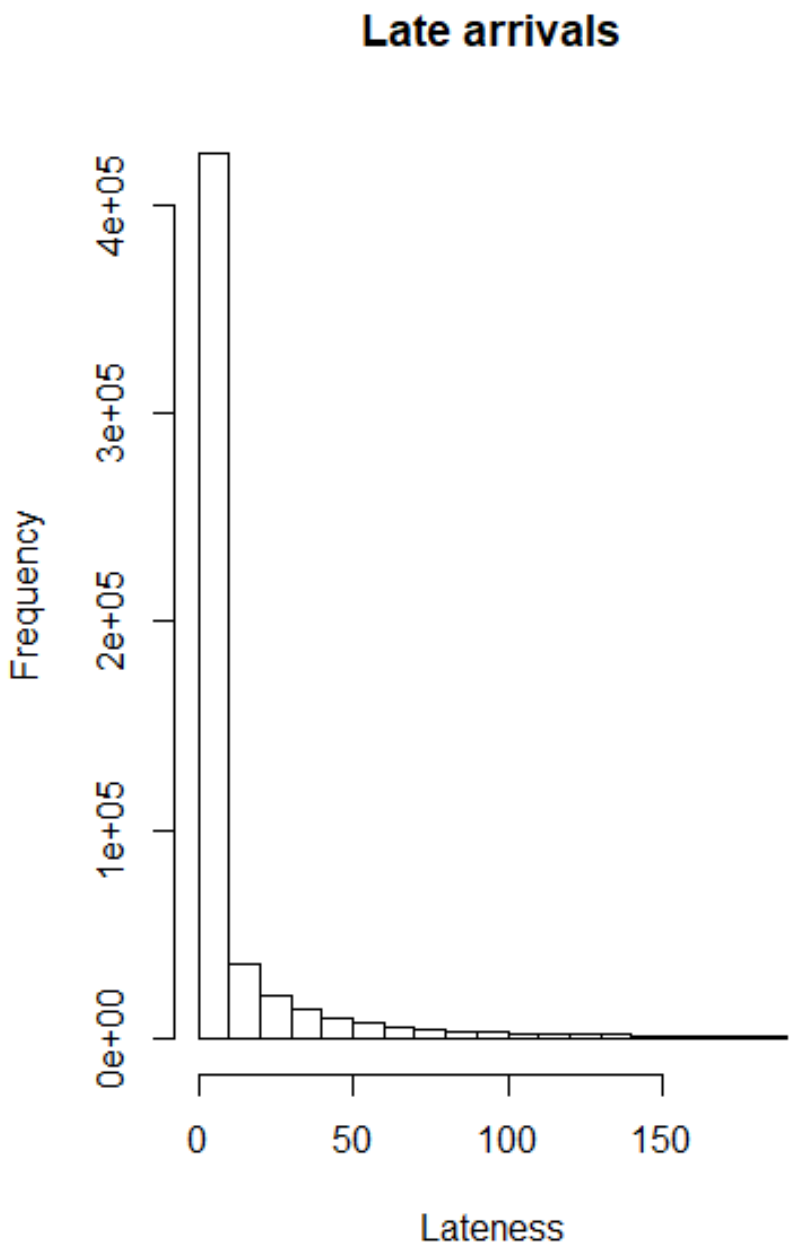


Figure 4.5: Fig 5 - histograms of ARR DELAY for each unusual weather event

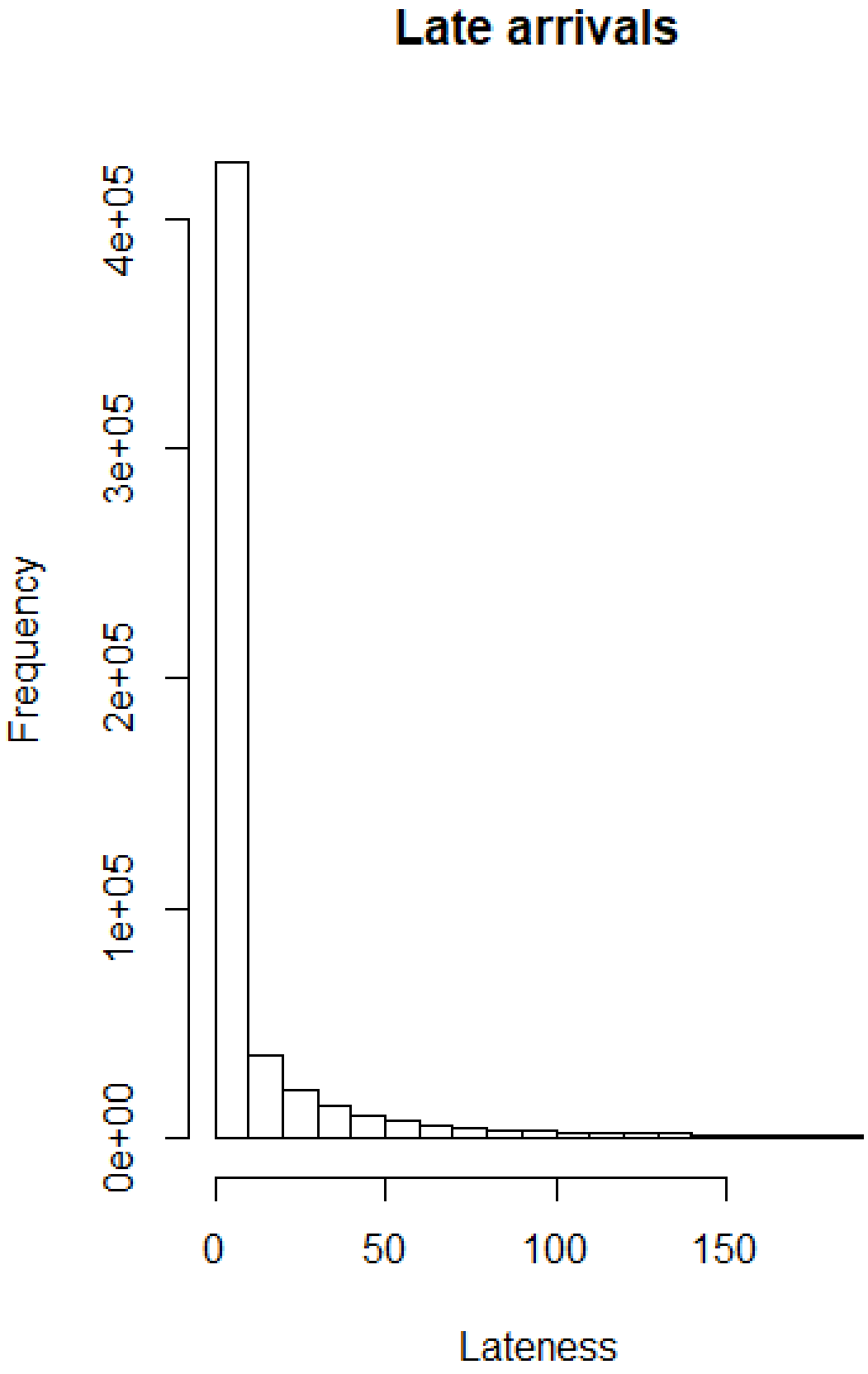


Figure 4.6: Fig 6 - Histograms for the airlines

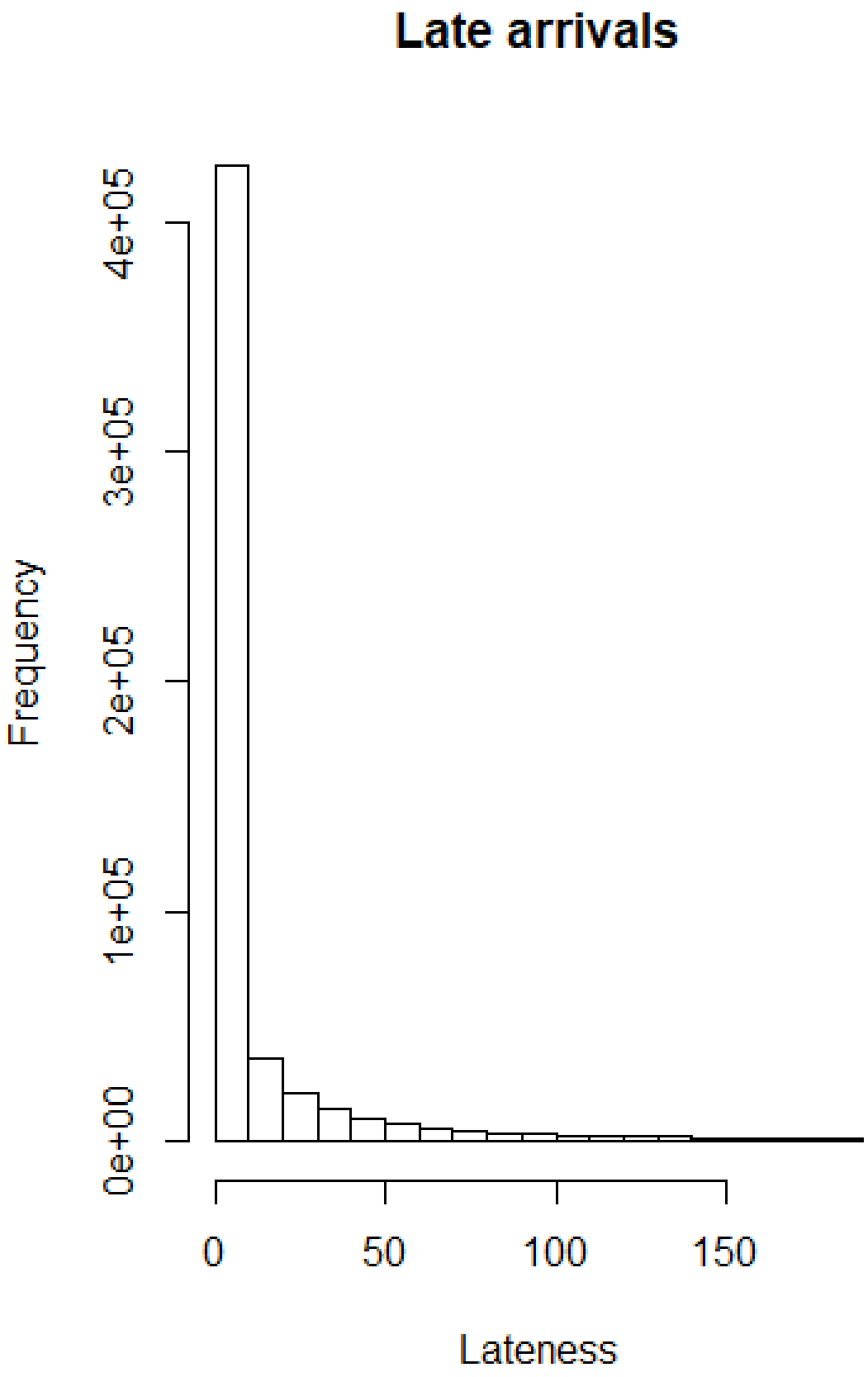


Figure 4.7: Fig 7 - airline ANOVA

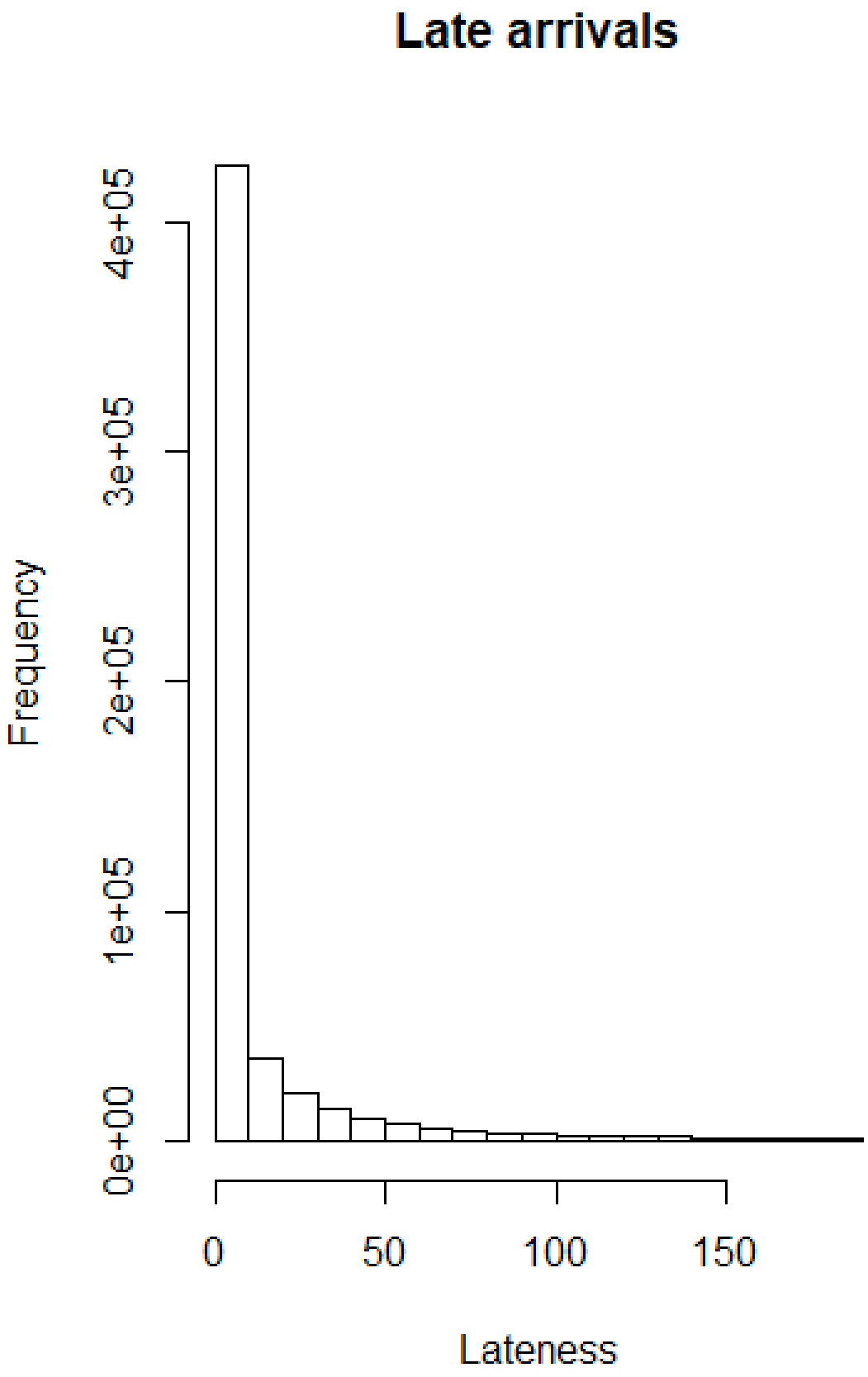


Figure 4.8: Fig 8 - Airport histograms

Chapter 5

Model formulation and assessment

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.

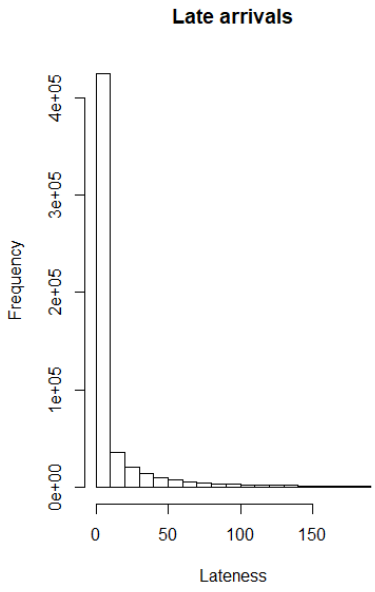


Figure 5.1: Fig 9 - density plot of cube root transformation vs normal distribution with same mean and variance

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 $U \sim Ber(p)$, $V \sim exp(\lambda)$, and $T \sim N(\mu, \sigma^2)$. In this model, U describes whether or not a delay with "true, substantial causes" occurs, T describes the distribution of arrival times when no extenuating circumstances occur, and V describes arrival times under defined circumstances that result in lateness. The CDF of this distribution is given by

$$F_X(x) = \sum \alpha_i F_i(x) = p(1 - e^{-\lambda x}) + (1 - p)\left(\frac{1}{2}(1 + erf(\frac{x - \mu}{\sigma\sqrt{2}}))\right)$$

e.g. a simple mixture of two distributions. The pdf is given by

$$f_X(x) = F'_X(x) = p(\lambda e^{-\lambda x}) + (1 - p)\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\right)$$

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 fails to describe data points in the extreme upper range of observations, p¡.001. Some delays were quite extreme, with values in excess of 1000 minutes. To prevent such outliers from having outsize effect on our downstream analyses, we further pared the dataset here to include only data below the 99th percentile.

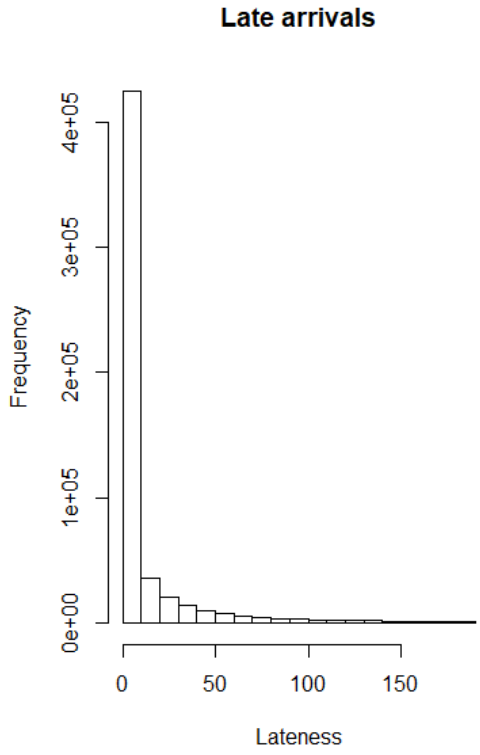


Figure 5.2: Fig 10 - Histogram of samples from distribution given by MLE, and histogram of values from data (fig 9)

Further, we see that conditional density estimation agrees with our theoretical parametric distribution. In Fig 12, we show comparisons of the empirical conditional density vs densities given by our theoretical distribution with parameter estimates from

MLE. Therefore, we suggest that our theoretical distribution can be used to describe arrival times across all combinations of factor levels.

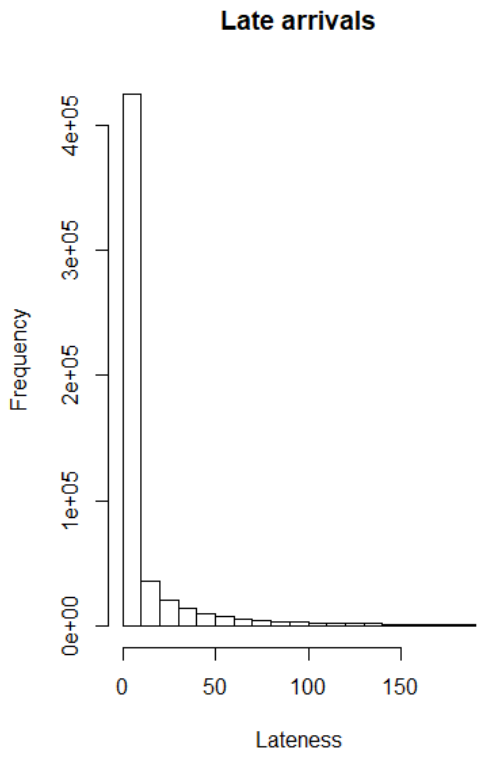
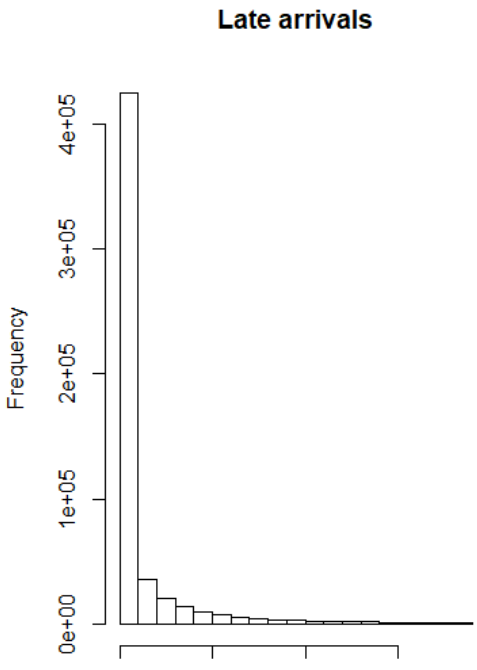


Figure 5.3: Fig 11 - QQ plot of of data
vs theoretical distribution

5.2 Linear model using OLS

Ideally, when constructing a linear model using OLS estimates, we want to satisfy four principal assumptions; namely, a linear relationship between predictors and response, independence of errors, homoscedasticity, and normality of the errors. However, having seen that the conditional distributions of the response are strongly *non-normal*, and that there exists no simple transformation to restore normality, the fourth assumption appears to be violated. It is



important now, to remember that, in fact, the Gauss-Markov theorem does *not* require normality of the errors for the OLS estimates to be the best linear unbiased estimator, and that normally distributed errors are required only for conducting inference on the model via the t-distribution. For merely constructing a model equation and generating predictions that minimize mean squared error, normality of the errors is not strictly required. Further, inference on the model parameters can still be conducted through a variety of alternate methods. Most commonly, M-estimation is used to fit heavy-tailed data. With a model given by

$$Y = X\beta + e$$

and residuals given by

$$\hat{e_i} = y_i - \hat{y}_i$$

we can define a minimizable objective function

$$f(\hat{\beta}) = \sum \rho(y_i - x_i^T \hat{\beta})$$

where ρ is a function of the error such that $\rho(e) \geq 0$, $\rho(0) = 0$, $\rho(e) = \rho(-e)$, and monotone on its absolute value, e.g. $\rho(e_i) \geq \rho(e_{i-1}$ Having determined a general form for the conditional distribution of the response, we have several ways to bypass this we proceeded to fit a generalized linear model to our data, where

5.3 Logistic regression

Another approach we wanted to try was to predict whether or not a flight would be delayed, regardless of time. To do this we did a logistic regression. Here are plots. Here are the problems. Here is how we tried to fix it. Here is how we failed.

5.4 Dynamic linear model

From our exploratory data analysis, we noted that several different categorical variables related to time showed significant effects on the mean arrival time. We thought the data might have an underlying time series, so our approach here was to construct a time series and then perform regression on the residuals.

Last but not least, we From our data exploration, we see that for major covariates, we do not see a change in the general shape of the distribution itself, only small changes in the parameters. Based on this, we make the simplifying assumption that the conditional distribution is identical and independent across all combinations of factor levels, keeping in mind the limitations of this approach. Furthermore, simplified model, use p instead of location parameter, show moments, describe strategy for producing conditional distributions of response based on 1st, 2nd, 3rd moments from data. Give equations. If the fixed effect model worked, then we would have fit that model and then fit a linear regression on the residuals to try to capture all effects.

Chapter 6

Forecasting Flight Delays for 2019 Q3

6.1 Challenges

Predicting flight delays can be challenging particularly because it is a subject that has already been investigated no doubt thoroughly by data scientists under the employ of the airline companies themselves. Many influential observations are the result of extreme values that happen rarely, while most observations are not so extreme. Black swan stuff, disparity in sample sizes, difficulty of incorporating models that can accurately predict rare, but highly impactful events (e.g. blizzards, tornadoes)

6.2 Using our forecast model

Using our regression model we predict such and such requiring the following list of covariates. Based on time series data we have the following.

Chapter 7

Business Recommendations

7.1 Differences between carriers

Look at our ANOVA. Some of you guys suck.

7.2 Important variables to keep an eye on

The following variables are important to note. Some of these variables are not under control. Some variables are under control. What to do in each case. Some variables that may seem to be important but are not.

Chapter 8

Closing Thoughts

8.1 Retrospect

we should've done this, should've used google cloud, should've clearly segmented the workflow and not gotten ahead. Alternative distributions, e.g. starting at flight takeoff and adjusted for distance

8.2 Improving future analyses

Our cool distribution, with calculated moments, use it for good.

Chapter 9

Appendix

9.1 References

9.2 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