**4.2**

The dataset provided information on flights from 17 different carriers along 6,684 unique routes being flown between a combination of 362 airports in the United States (including Alaska and Hawaii and a couple airports in US territories such as Guam). 3,814,366 out of the 10,915,495 flights (~35\%) had some form of arrival delay. A histogram of all arrival delays is shown in \underline{Fig 4.1}. We observe that most “early” flights are not incredibly early compared to the frequency at which late flights can become “significantly” delayed, say more than 30 minutes. Clearly, the data is strongly right-skewed. To correct the skewness (to meet assumptions for our analysis), a cube-root transformation was performed, but subsequent Shapiro-Wilk tests provided strong evidence against normality for this transformation, so we turned to other methods, as will be discussed in chapter 5. To heuristically assess dependence between covariates and late arrivals, we examined various conditional distributions, as follows.

4.2.1 Geographic

In \underline{Fig 4.21}, the routes that have various average intervals of delay time are shown. Note straight lines are drawn for simplicity; the actual flight most likely flew a non-linear path toward the destination. Also, early arrivals are given a delay time of 0 in the computation of the mean. We observe many things. Almost all routes have an average delay that is positive. This spotlights the nature of this project: to focus on remedying delays to improve customer satisfaction and increase airline revenue. For the routes with a mean delay between 30 and 45 minutes, most of the delays are clustered on the eastern US with most of the routes either beginning or ending in the San Francisco, New York, or southern Florida regions. For more severe delays of 45+ minutes, these routes encompass more cross-country flights. Furthermore, it can be seen that an airport in the northeast, most likely JFK in New York, is involved in a lot of severe flight delays. Next, in \underline{Fig 4.22}, the average delays at certain airports is depicted. We see further evidence that the more problematic delays are centered in bigger cities, particularly those on the east and west coast. \underline{Fig 4.23} illustrates the relationship between the popularity of an airport and the amount of delays. As expected, airports that crank out more flights have a higher average delay time. This raises the question: Does more flights simply give airports a higher probability of having delays (by random chance), or does an increase in flights also bring in other factors that \textit{cause} an increase in flight delays. In more broad terms, what factors correlate with delays, and which ones can be controlled by the airlines?

4.2.2 Temporal

\underline{Fig 4.31}, \underline{Fig 4.32}, \underline{Fig 4.33}, and \underline{Fig 4.34} display the distribution of flight delays for quarter, month, day of the week, and time of day, respectively. Note that because our data contained a year and a half’s worth of observations, the frequency of delays for, say quarter 1, are higher than other time periods that were not recorded twice. The histograms appear roughly symmetric with a slight right-skew. Specifically, early arrivals appear to follow a normal distribution while it transitions to an exponential distribution once delays become positive. We will explore this specific observation in the first section of our next chapter. Regardless, the distributions across these time factors do not change significantly. This produces an initial assumption that time of the flight is independent of delay time. To further investigate this hypothesis, we conduct \textit{John talk about coplots and stuff here}.

4.2.3 Weather

This most obvious factor that most likely affects flight delays is the weather. Therefore, we analyzed the conditional distributions of flight delays given average temperature, precipitation, and the presence of various rare weather events. These are described in \underline{Fig 4.41}, \underline{Fig 4.42}, and \underline{Fig 4.43}, respectively. In the scatterplot depicting rain, we see a lack of evidence that more extreme temperatures correlate with increased delays. This would be evident if the scatterplot showed a parabolic pattern, which is not seen. Surprisingly, we see the same idea in the scatterplot of precipitation. Contrary to intuition, as the amount of precipitation increase, there is no visual subsequent increase in the delay time. Lastly, \underline{Fig 4.43} demonstrates that some dangerous weather events have a slightly bigger effect on delay times than others. For example, the distributions of “ice, sleet and hail” and “blowing or drifting snow” have more area in their right-tails. This is indicative of more occurrences where these events caused more significant delays. Besides these difference in tail-density, the overall shape remains very similar.

4.2.4 Carrier

Next, looking at delays by carrier would provide us with insight as to whether some carriers are better at mitigating delays. \underline{Fig 4.5} shows us that the distribution across carriers stays roughly the same. Apart from the differences in frequencies (with some airlines being more popular or providing more flight routes), the shape of the distribution is basically homogenous, especially when compared to the other conditional distributions of other factors. This means that delays across carriers \textit{behave} in the same manner. To analyze if the average delay time (as opposed to distribution) differs across airlines, we conducted ANOVA (analysis of variance). Specifically, we used Tukey’s HSD to make multiple comparisons across each combination of carriers to see what carriers differed from one another. A plot of the resultant 95% confidence intervals for the mean difference between airport delays is show in \underline{Fig 4.6}. We see that the majority of the intervals constructed do not fall within 0. Thus for those that didn’t, we conclude that they do indeed have a difference in mean arrival delays. All in all, we observe that the \textit{distribution} across carriers is similar, but the \textit{quantitative amount of delays} across carriers is different.

4.2.5 Airport

\underline{Fig 4.7} displays the histograms of delays for some of the most popular airports. As has been the trend thus far, the distribution across the airports does not change much save for the changes in frequency. Thus, we can conclude the behavior and process of flight delays is pretty universal and can be modeled with an explicitly defined distribution, as we will dive into in the next chapter.

4.2.6 | Making Up Lost Time En Route

One last thing we wanted to explore was the ability for pilots to make up lost time as a result of a departure delay. We presumed that a longer flight, distance-wise, would allow for more opportunity for a flight delay to be alleviated. In the air, harnessing favorable air currents or taking shortcuts can remedy the lost time they left the ground with. Surely, in \underline{Fig 4.8}, we see that flights over a longer distance have, on average, less arrival delays. Similarly, the shorter the flight, the more severe the arrival delay is. These observations support our hypothesis stated above. Of course, correlation does not imply causation, so more substantive knowledge on how pilots navigate the flight route would provide more clarification on this.

**LOGISTIC REGRESSION**

The majority of the models in data science can be put into three bins: regression, classification, and clustering. As we came to realize that our linear regression model was not useful enough, we turned to a classification method. In particular, we chose binary logistic regression to predict whether or not a flight would be delayed or not given certain pre-departure covariates. In retrospect, our motivation for pursuing logistic regression was two-fold. For one, we wanted to improve on the accuracy of our predictive model, so by moving away from regressing on a continuous variable, we shifted our focus to predicting a binary success or failure. Secondly, our conditional density estimates illustrate that given certain covariate levels, the shape of our distribution changes more than the mean (location) of our distribution. Thus, regressing on the parameter p from our Bernoulli random variable from our mixed distribution would help us explain these observed changes in our conditional density estimates. The marginal distribution of U in our model changes how thick-tailed our distributions become on either side of 0.

Using the \textit{glm()} function in R, we fit a generalized linear model using the logit function as the link function and the binomial distribution family as the probability distribution. A 5\% subset of our data frame was taken to shorten computation time. Even more, 80% of our data was allotted for training and 20% for testing. The summary of our fit model is show in \underline{Fig X.X}. Our covariates included various factors on the time as well as weather phenomena in the departure and arrival airport. We interpret our logistic regression output in two separate ways for numerical and categorical data. For numerical data, the specific factor level’s coefficient can be interpreted as the change in the log-odds of the “success” occurring – in our case a delay. Exponentiating these values gives you a more interpretable estimate. For example, from our model output, we can say that for every additional millimeter of snow, the odds of a delay occurring (versus not occurring) increases by a factor of e^0.1283 = 1.013. For categorial data, the interpretation is a little different. We illustrate by example. In our case, flying in April as opposed to January lowers one’s chances of having a flight delayed by a factor of e^-0.2646 = 0.767. Furthermore, we see that the majority of our coefficients are statistically significant as indicated by the asterisks next to each line. These significant predictors are thus being utilized in our model, as desired.

Since our logistic regression model outputted a \textit{probability} that a delay will occur, we needed to determine the optimal probability threshold to determine whether or not a delay will take place. We thus improved our model by plotting an ROC curve, or a “Receiver Operating Characteristics” curve (\underline{Fig X.X}. By maximizing the area under the curve, we are able to increase our accuracy. The y-axis gives the true positive rate while the x-axis gives 1 minus the true negative rate. Thus, by altering the probability threshold, we can then predict with better accuracy. Using this ROC curve and various functions, we found that the optimal probability threshold was 0.39. This means that any predictions that output the probability of a delay as being over 0.39, we say that a delay will occur.

Finally, after validating our model on the testing data, we generated a confusion matrix and other measures of performance as shown in \underline{Fig X.X}. Our accuracy was 67.62% where out of our testing set of 92,667 observations, it correctly predicted 62,187 of them. Observe from the confusion matrix that we have significantly more false positives than false negatives. Thus, we interpret this as our model is more on the pessimistic side and is more likely to forecast a delay when in reality there is not one.