ECO395M Final Project: Impact of Covid-19 on the flight delays and cancellation in California and Texas

Steven Kim and Shreekara Shastry

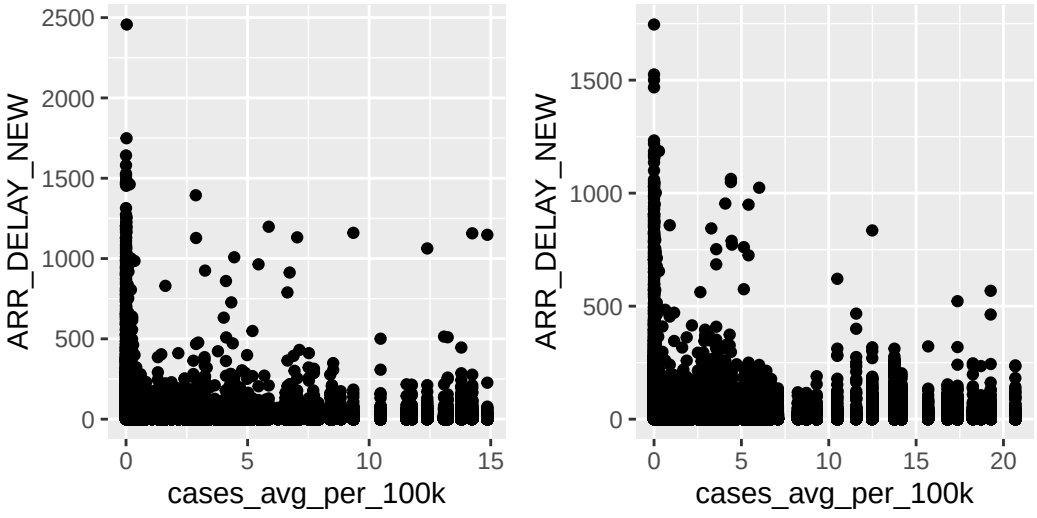
Abstract

This is our abstract.

Introduction

COVID-19 has affected our society in various ways. Among all of them, we decided to look at its effect on the flight delay and cancellations. Not only COVID-19 would have reduced the demand for the air travel, but also the counter-measures such as lockdowns are expected to negatively impacted the flights' on-time performance. Plus, if crew members had been exposed to the virus, they would have been required to quarantine which would likely cause the flight to be delayed or cancelled. Therefore, we would like to analyze the relationship between the COVID-19 cases and the delay and cancellation of the flights. We do this by asking three separate but related questions: "Is a flight more likely to be delayed with a higher number of cases of COVID-19?", "Is a flight more likely to be cancelled with a higher number of cases of COVID-19?", and "If a flight is delayed, is it expected to be delayed for longer with a higher number of cases of COVID-19?". The rest of our report goes in detail to explain how we estimated the effect.

Figure 4.



Methods

Data Description

We have used 2 separate datasets and combined them to form a data frame that we used in the further analysis. The first dataset is from The United States Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. The data collected is from January to June 2020 and contains relevant flight information (on-time, delayed, canceled, diverted flights) from the Top 10 United States flight carriers for 11 million flights. The second dataset is from the New York Times[2] which contains the state-wise data on the daily number of new cases and deaths, the seven-day rolling average, and the seven-day rolling average per 100,000 residents. We merged these two datasets based on the date and state to create a new dataset that we used in all the models. This combined dataset has in total of 2745847 observations with data from 375 different airports.

Data Wrangling

Data cleaning and preprocessing for the dataset was a four-step process.

1. Formatting the date field in both the individual datasets to match before performing a left join.
2. Merging the covid case dataset into the flight data based on date and state.
3. Factorizing the categorical variables from this combined dataset. `MONTH`, `DAY_OF_MONTH`, `DAY_OF_WEEK`, `MKT_UNIQUE_CARRIER`, `TAIL_NUM`, `ORIGIN`, `ORIGIN_STATE_NM`, `DEST`, `DEST_STATE_NM`, `ARR_DEL15`, `CANCELLED`, `CANCELLATION_CODE`, these are the categorical variables in the dataset.
4. Removing the variables that are not used in the analysis to have a cleaner dataset.

The cleaned data set is split into training and testing sets for California and Texas separately. There were 284941 observations with California as the origin state. So California data set has 227952 observations in the training set and 56989 observations in the testing set. Whereas, there were 287693 observations with Texas as the origin state and there are 230154 observations in the training set and 57539 observations in the testing set for Texas. In terms of the air traffic both California and Texas are comparable. The list of variables used in our models are below:

Variable Name	Description
`MONTH`	Month of Year
`DAY_OF_MONTH`	Text
`DAY_OF_WEEK`	Date of Month
`FL_DATE`	Day of Week (1: Monday, 7: Sunday)
`MKT_UNIQUE_CARRIER`	Airline Carrier Code (Look at appendix for detail)
`ORIGIN`	Flight Departure 3-Letter Airport Abbreviation
`ORIGIN_STATE_NM.`	Flight Departure State Name
`DEST`	Flight Arrival 3-Letter Airport Abbreviation
`DEST_STATE_NM`	Flight Arrival State Name
`ARR_DELAY_NEW`	Departure Delay Ignoring Early Departures (Listed as 0)
`ARR_DEL15`	Departure Delay Greater Than 15 Minutes (0: Not Greater Than 15, 1: Greater Than 15)
`CANCELLED`	0: Flight Not Cancelled, 1: Flight Cancelled

Variable Name	Description
DISTANCE	Distance Between Departure and Arrival Airports (in Miles)
cases_avg_per_100k	The average number of new cases per 100,000 people, reported over the most recentseven days of data. In other words, the seven-day trailing average.

Models of choice

With the dataset, we used various models to answer the questions. The dependent variables of interest are ARR_DEL15, ARR_DELAY_NEW, CANCELLED. First, we look at ARR_DEL15 and CANCELLED. They are binary variables, which call for classification models. We used a logit model to predict ARR_DEL15 and logit and tree models to predict CANCELLED. We used the tree model with a complexity parameter of 0.005 and minsplit of 30 since we have a lot of observations in both the cases of California and Texas. We adjusted the parameters by looking at the results and picking the adequate looking trees. They all have MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, and cases_avg_per_100k as features. We wanted to control for the origin and destination airports, but that would involve too many levels for categorical variables. Therefore, we included the destination state instead of the destination airport. We added month, day of week, carrier and distance to control for other causes than COVID cases.

Next, we look at ARR_DELAY_NEW. ARR_DELAY_NEW is 0 if the flight is not delayed and has the minutes of delay if the flight is delayed. Since the majority of the flights are not delayed, this variable has a lot of zeros compared to the other numbers. For this reason, we used zero-inflated poisson models, hoping that this would be more accurate. Although it is typically used for count data with many zeros and our variable of interest is time, I thought it is worth trying this model as we have discrete time values in minutes. We compare the results in this model with the basic linear model. For the zero-inflated poisson model, the first process generates zeros and the second process is governed by a Poisson distribution that generates counts, some of which may be zero. In this model building, the assumption is that the COVID cases would generate the non-zero counts. They both have MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, DISTANCE, and cases_avg_per_100k for same reasons as before, with ORIGIN and DEST_STATE_NM also excluded for having too many categories. We also use tree models to predict ‘ARR_DELAY_NEW. This has MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, and cases_avg_per_100kas features. We used the tree model with a complexity parameter of 0.0009 and minsplit of 30. We adjusted the parameters by looking at the results and picking the adequate looking trees.

Another way to look at the ARR_DELAY_NEW variable is to focus on the flights that are delayed. That is, we can see if the COVID cases would increase the delay time, given the flight is already delayed. We used random forest models to predict this, as it usually performed the best for the homework questions. These also have MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, DISTANCE, and cases_avg_per_100kas the features, without ORIGIN and DEST_STATE_NM, for having too many levels. We used default values for ntrees and mtry.

Results

Delay Rates - Logit Model

We looked if the Covid cases caused more delays by running a logit regression of ARR_DEL15 on MONTH, DAY_OF_WEEK, MKT_UNIQUE_CARRIER, ORIGIN, DEST_STATE_NM, DISTANCE, cases_avg_per_100k variables.

In the case of California, the model has an out-of-sample accuracy of 0.9108055. Whereas the logit model built on the Texas data has an out-of-sample accuracy of 0.8865125. These accuracies were calculated for a k value of 0.5.

For California, the partial effect of `cases_avg_per_100k` is 0.02087 that means if the `cases_avg_per_100k` increases by a unit, the log odds of flights delaying by 15 mins will incrase by 0.02087. This effect is statistically significant only at even 10% significance level. Whereas for Texas, the partial effect of `cases_avg_per_100k` is 0.03658 that means if the `cases_avg_per_100k` increases by a unit, the log odds of flights delaying by 15 mins will incrase by 0.03658. This effect is statistically significant at even 0.1% significance level and is slightly higher compared to Texas and this is clearly more significant compared to California.

Cancellation Rates - Logit Model

We looked at the effect of Covid cases on flight cancellations by running a logit regression of `CANCELLED` on `MONTH`, `DAY_OF_WEEK`, `MKT_UNIQUE_CARRIER`, `ORIGIN`, `DEST_STATE_NM`, `DISTANCE`, `cases_avg_per_100k` variables. In the case of California, the model has an out-of-sample accuracy of 0.8987173. Whereas the logit model built on the Texas data has an out-of-sample accuracy of 0.8880412. These accuracy rates were calculated for a k value of 0.5.

For California, the partial effect of `cases_avg_per_100k` is 0.9401 that means if the `cases_avg_per_100k` increases by a unit, the log odds of cancellation will increase by 0.9401. This effect is statistically significant at even 0.1% significance level. Whereas for Texas, the partial effect of `cases_avg_per_100k` is 0.9928 that means if the `cases_avg_per_100k` increases by a unit, the log odds of cancellation will increase by 0.9928. This effect is statistically significant at even 0.1% significance level and is slightly higher compared to California.

Cancellation Rates - Tree Model

For California, the initial decision in the tree is based on if the month was April. It is interesting to see the tree model capture the lock downs in April. Further decisions are made based on the type of carrier and distance of the flights. The importance plot shows that `avg_cases_per_100k` had an effect on the flight cancellations. In the case of Texas, the initial decision in the tree is based on if the month was April. It is interesting to see the tree model capture the lock downs in April even in the case of Texas. But the since the lock downs in Texas were not enforced as strictly as California, the left side of the decision tree has further classifications based on the flight carrier, day of the week and cases_avg_per_100k. When we look at the importance plot, Texas clearly deviates from the California. The `avg_cases_per_100k` is the second most important variable, and has played more effect on the flight cancellations.

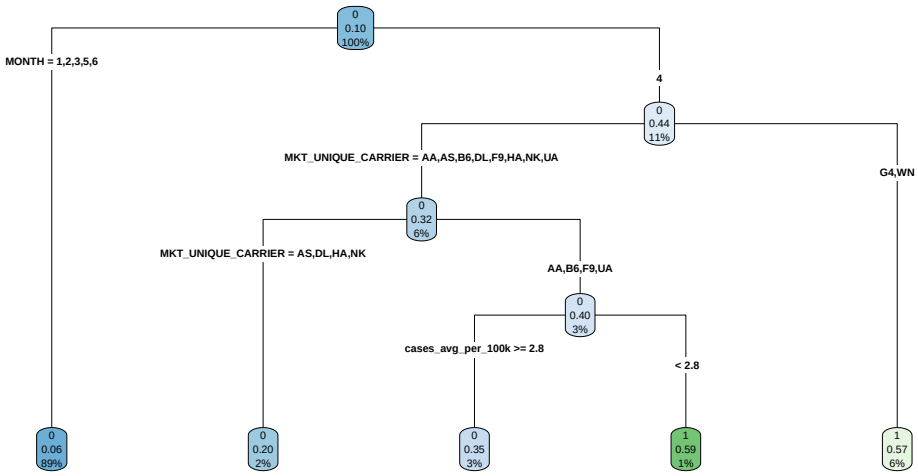
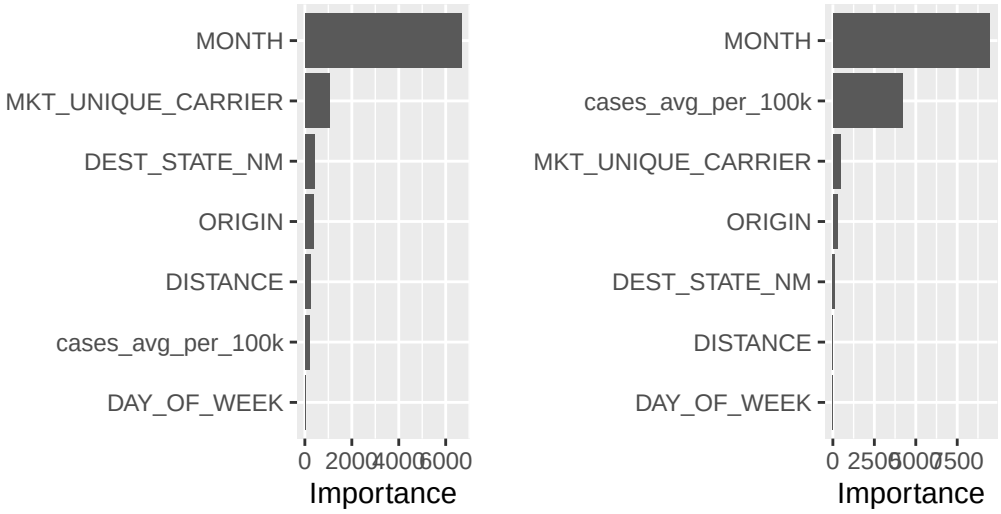


Figure 4.



Delay Minutes - Linear Regression

We looked at the effect of Covid cases on flight cancellations by running a linear regression of `ARR_DELAY_NEW` on `MONTH`, `DAY_OF_WEEK`, `MKT_UNIQUE_CARRIER`, `DISTANCE`, `cases_avg_per_100k` variables. In the case of California, the model has an out-of-sample accuracy of 33.80909. Whereas the linear regression model built on the Texas data has an out-of-sample accuracy of 31.96574. For California, the partial effect of `cases_avg_per_100k` is 0.0925041 that means if the `cases_avg_per_100k` increases by a unit, the arrival delay will increase by 0.0925041. This effect is not statistically significant at even 10% significance level. Whereas for Texas, the partial effect of `cases_avg_per_100k` is 0.2228197 that means if the `cases_avg_per_100k` increases by a unit, the arrival delay will increase by 0.2228197. This effect is statistically significant at even 0.1% significance level and is significantly higher compared to California.

Delay - Zero-Inflated Poisson Regression

Next, we used zero inflated poisson regression to see the effect of covid cases on the arrival delay. In both California and Texas the effect of `cases_avg_per_100k` is small and positive but statistically significant. The RMSE in the case of California is 33.72181 and Texas is 31.96069.

Delay - Tree Model

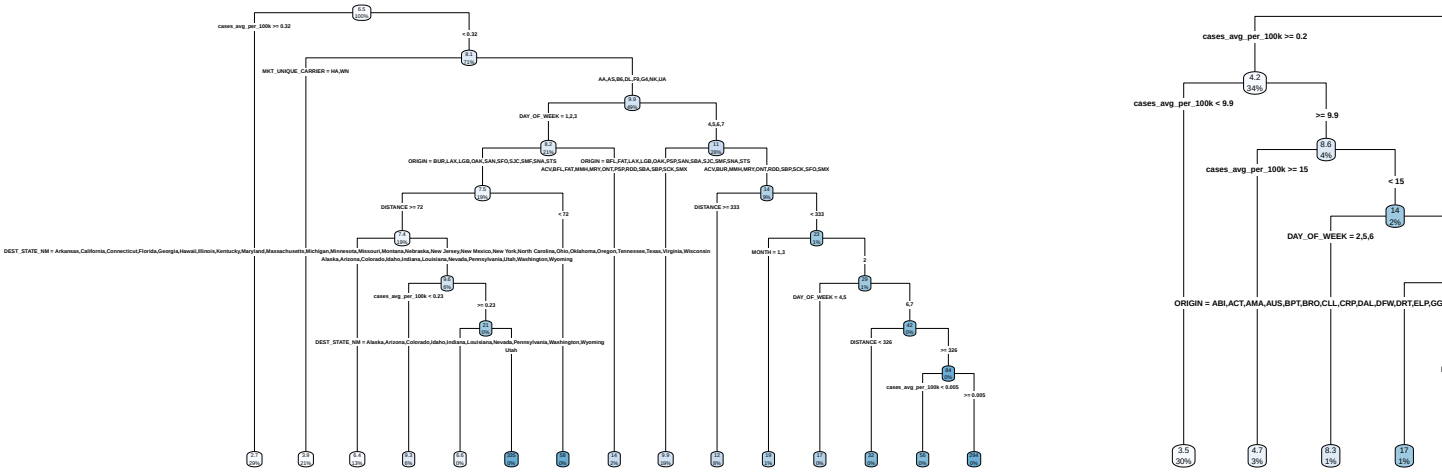
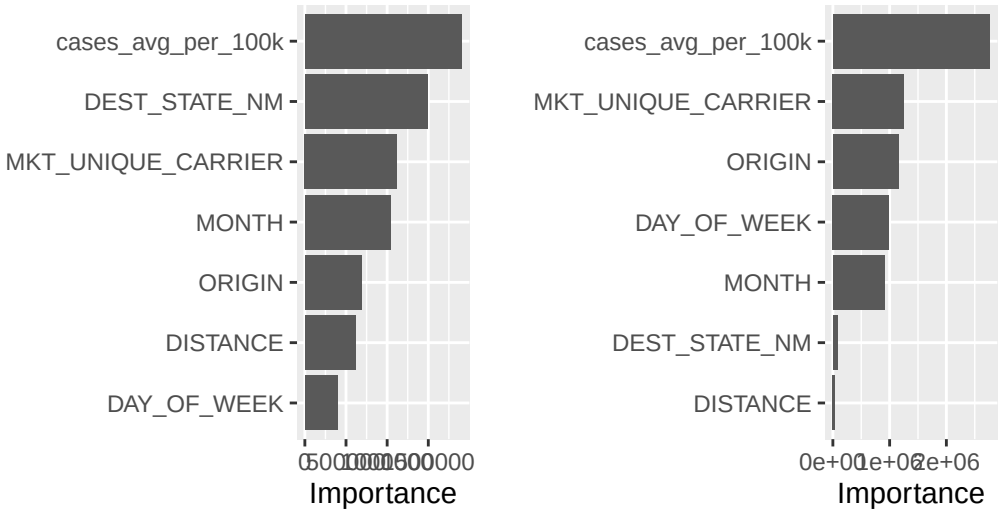


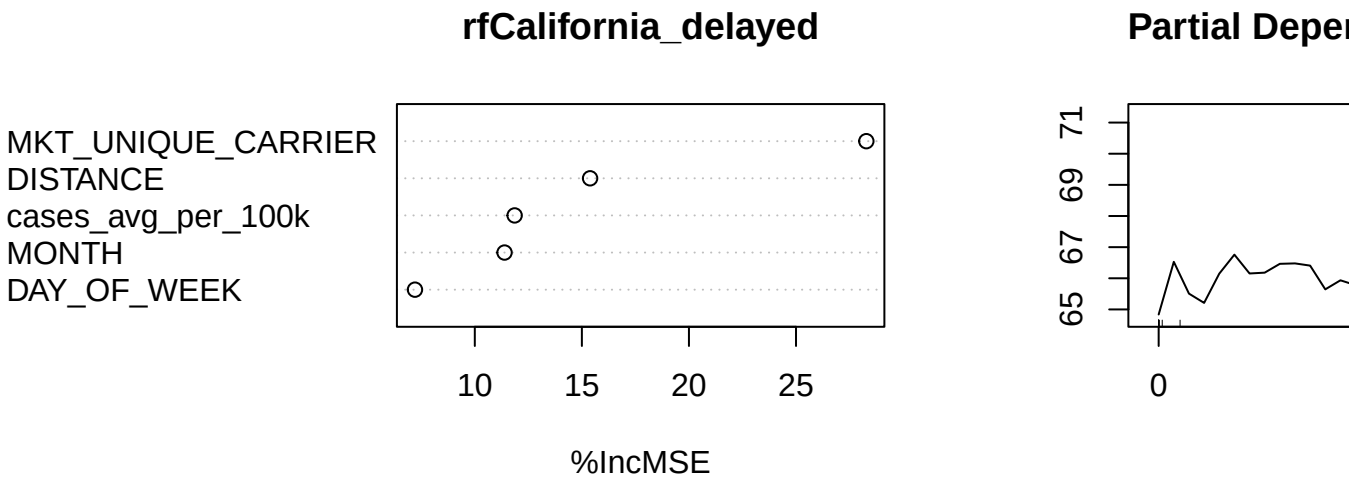
Figure 4.

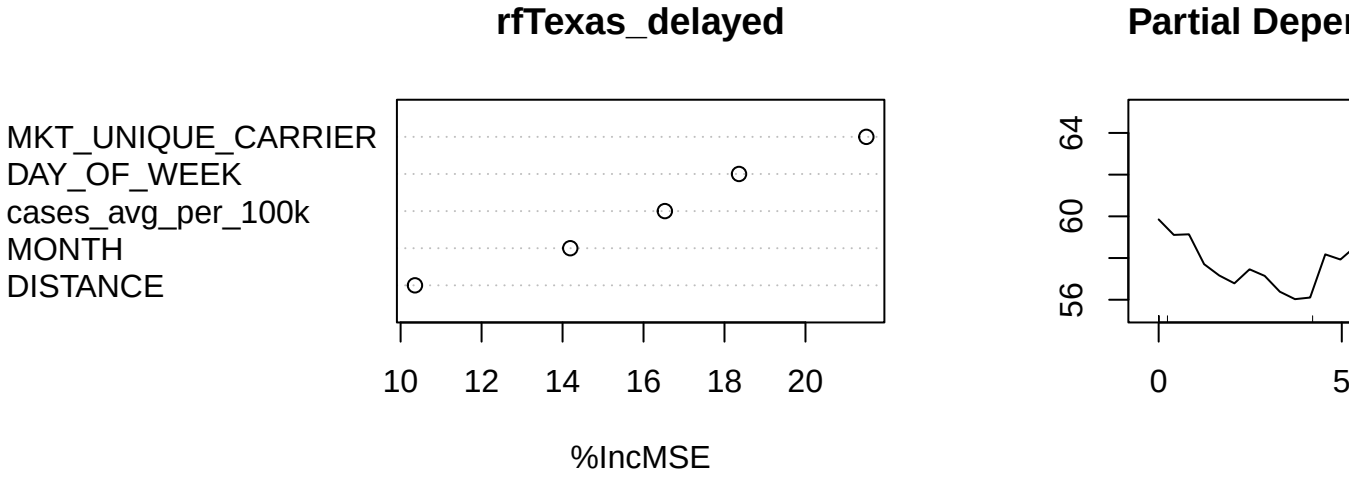


For California, the initial decision in the tree is based on if the average covid cases per 100k was less than 0.32 or not and rest of the decisions are made based on destination states. The importance plot shows that `avg_cases_per_100k` has the biggest effect on the flight cancellations. In the case of Texas, the initial decision in the tree is based on if the average covid cases per 100k was less than 0.2 or not which is less than in the case of California and rest of the decisions are made based on destination states. When we look at the importance plot, Texas follows California with `avg_cases_per_100k` being the most important variable. The out of sample RMSE value for California is 33.38648 and for Texas is 31.45617. The tree looks

to preform slightly better for Texas state data.

Delay Time Effect for Those Delayed - Random Forest Model






```
## $x
##  [1]  0.0000  0.2972  0.5944  0.8916  1.1888  1.4860  1.7832  2.0804  2.3776
## [10]  2.6748  2.9720  3.2692  3.5664  3.8636  4.1608  4.4580  4.7552  5.0524
## [19]  5.3496  5.6468  5.9440  6.2412  6.5384  6.8356  7.1328  7.4300  7.7272
## [28]  8.0244  8.3216  8.6188  8.9160  9.2132  9.5104  9.8076 10.1048 10.4020
## [37] 10.6992 10.9964 11.2936 11.5908 11.8880 12.1852 12.4824 12.7796 13.0768
## [46] 13.3740 13.6712 13.9684 14.2656 14.5628 14.8600
##
## $y
##  [1] 64.83761 66.52857 65.50805 65.21002 66.14835 66.75902 66.15472 66.18077
##  [9] 66.46424 66.47836 66.40737 65.64479 65.93530 65.78060 68.21837 65.36738
```

```
## [17] 65.88012 66.31344 67.72629 66.60254 67.46094 66.94728 69.53960 68.16159
## [25] 70.13286 70.03093 66.51145 65.43601 65.79776 65.22089 65.22258 66.51873
## [33] 66.44101 66.44373 65.76138 65.84585 66.43355 66.47593 66.34896 66.52443
## [41] 66.80128 71.22976 71.32942 69.55826 69.29064 68.64641 68.60552 68.56164
## [49] 65.66129 64.70771 64.70771

## $x
##  [1]  0.0000  0.4136  0.8272  1.2408  1.6544  2.0680  2.4816  2.8952  3.3088
## [10]  3.7224  4.1360  4.5496  4.9632  5.3768  5.7904  6.2040  6.6176  7.0312
## [19]  7.4448  7.8584  8.2720  8.6856  9.0992  9.5128  9.9264 10.3400 10.7536
## [28] 11.1672 11.5808 11.9944 12.4080 12.8216 13.2352 13.6488 14.0624 14.4760
## [37] 14.8896 15.3032 15.7168 16.1304 16.5440 16.9576 17.3712 17.7848 18.1984
## [46] 18.6120 19.0256 19.4392 19.8528 20.2664 20.6800
##
## $y
##  [1] 59.85179 59.10983 59.13921 57.70592 57.16318 56.78687 57.45965 57.14341
##  [9] 56.38461 56.02764 56.10484 58.17540 57.92884 58.55677 57.58092 56.89721
## [17] 56.77315 56.67342 56.22485 56.20235 56.20967 56.29556 56.68473 56.74513
## [25] 57.66368 57.68743 58.09061 59.02651 59.09998 59.10467 59.59342 59.96367
## [33] 65.19740 64.86176 57.32146 57.24227 57.19690 56.38110 55.63433 55.54647
## [41] 55.30670 55.30601 55.44057 55.54294 55.49754 55.33823 56.89528 56.89528
## [49] 57.32812 57.42215 57.42215

##                       %IncMSE
## MONTH              11.389241
## DAY_OF_WEEK         7.206308
## MKT_UNIQUE_CARRIER 28.283253
## DISTANCE           15.383438
## cases_avg_per_100k 11.860387

##                       %IncMSE
## MONTH              14.19052
## DAY_OF_WEEK        18.35906
## MKT_UNIQUE_CARRIER 21.50340
## DISTANCE           10.35469
## cases_avg_per_100k 16.52709
```

Random Forest model for the data with the delayed flights shows that for both California and Texas the `cases_avg_per_100k` is the third most important variable ahead of the `month`. Partial Dependence graph on `cases_avg_per_100k` for California seems to suggest that the delays exponentially increase from 65 mins to 85 mins when the `cases_avg_per_100k` reached 15. But in the case of Texas the range is smaller and the delays fluctuating between 54 and 64 mins. The out of sample RMSE value for California is 93.58049 and for Texas is 81.30241. This suggests that the randomForest model worked better with the Texas state data.

Conclusion

Appendix

Code	Airline Name
AA	American Airlines
AS	Alaska Airlines
B6	JetBlue
DL	Delta Air Lines
F9	Frontier Airlines
G4	Allegiant Air.
HA	Hawaiian Airlines
NK	Spirit Airlines
UA	United Airlines
WN	Southwest Airlines