DSC 630 Predictive Analytics
Milestone 5
Corinne Medeiros
Amy Nestingen
11/17/2020

Prediction of Travel in California and Covid-19 Impacts

Executive Summary

This study examines travel trends in the United States, specifically trips taken in the state of California, and aims to predict future trends. The scope of the data covers California trips from January 2019 through October 2020, with the prediction objective focused on the next 6 months for short term planning as well as the next 2 years for longer term planning. California Covid-19 data representing reported cases are also taken into account, offering further insight into travel during the current pandemic. Trends in positive cases parallel the dips observed in trips taken.

In order to predict the number of trips taken in upcoming months, we explored a variety of methods. Specifically, the technical forecasting approach in this study consists of several time series models, with the ARIMA model producing the best results. However, we have found that it is very difficult to predict the travel trends in California due to the uncertainty surrounding Covid-19 cases and deaths. There are many different inputs that can affect the number of Covid-19 cases and deaths, such as the distribution of a vaccine, mask mandates, attending school in-person, and the willingness of people to adhere to health guidelines. Because of these unknowns, it is a challenge to model future travel trends.

Intro and Background of the Problem

Many small businesses depend heavily on tourist sales. When people travel, they spend money on hotels, restaurants, shops, and experiences. In the past, we have seen travel decrease during recessions caused by wars, financial crises, epidemics, and energy crises. We are seeing that same trend, possibly even more severely, during the Covid-19 pandemic. Although the tourism industry has historically proven to bounce back quickly from recessions, people are still eager to predict when travel will return to normal and tourist shops will regain their sales.

Data.gov hosts frequently updated trip data from The Bureau of Transportation Statistics (BTS). These records originate from an “anonymized national panel of mobile device data,” and are estimated by the Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland (BTS.gov., 2020). Specifically, trips are measured by distance (in miles) and are defined as “movements that include a stay of longer than 10 minutes at an anonymized location away from home” (BTS.gov., 2020). When more than one stop occurs before a return to home, this is recorded as multiple trips. Trips involve all varieties of transportation including cars, trains, public transportation, and flights.

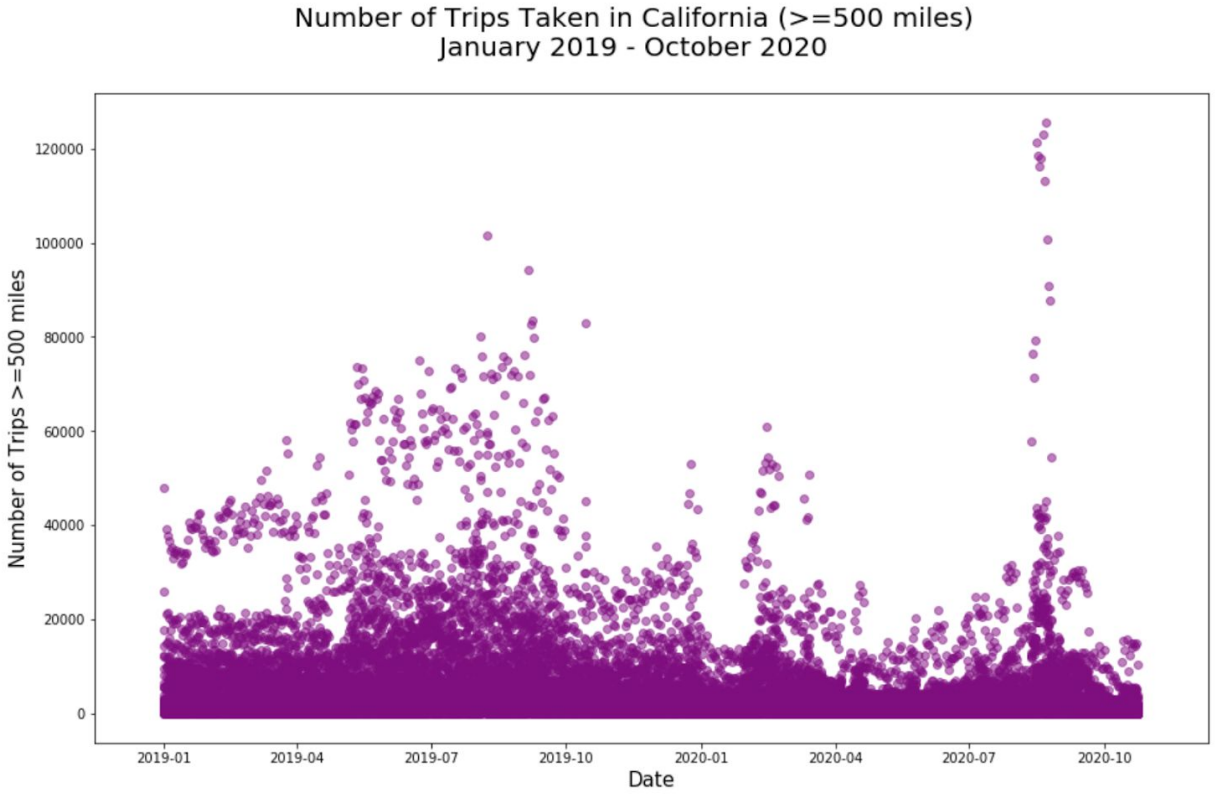
In terms of patterns, the percentage of people who stayed at home in 2019 was consistent. It hovered around 20% plus or minus 1%. In contrast, there has been a sharp increase in 2020 for the percentage of people staying home, with the highest percentage being 28% in April. Covid-19 cases have been on the rise since April of 2020, with only a slight leveling out in September before another climb.

The number of trips people have taken in 2019 compared to the trend in 2020, along with Covid-19 observations, will give us insight on when travel will return to normal. It will also provide insight on what travel will look like in the coming months. In turn, this information could help small businesses put plans in place to hire back the people they had to lay off and also stock up on goods for the return of their customers. It can assist in decision-making, inventory needs, and preparation for potentially new and improved travel experiences as we continue to navigate these unprecedented circumstances.

Methods

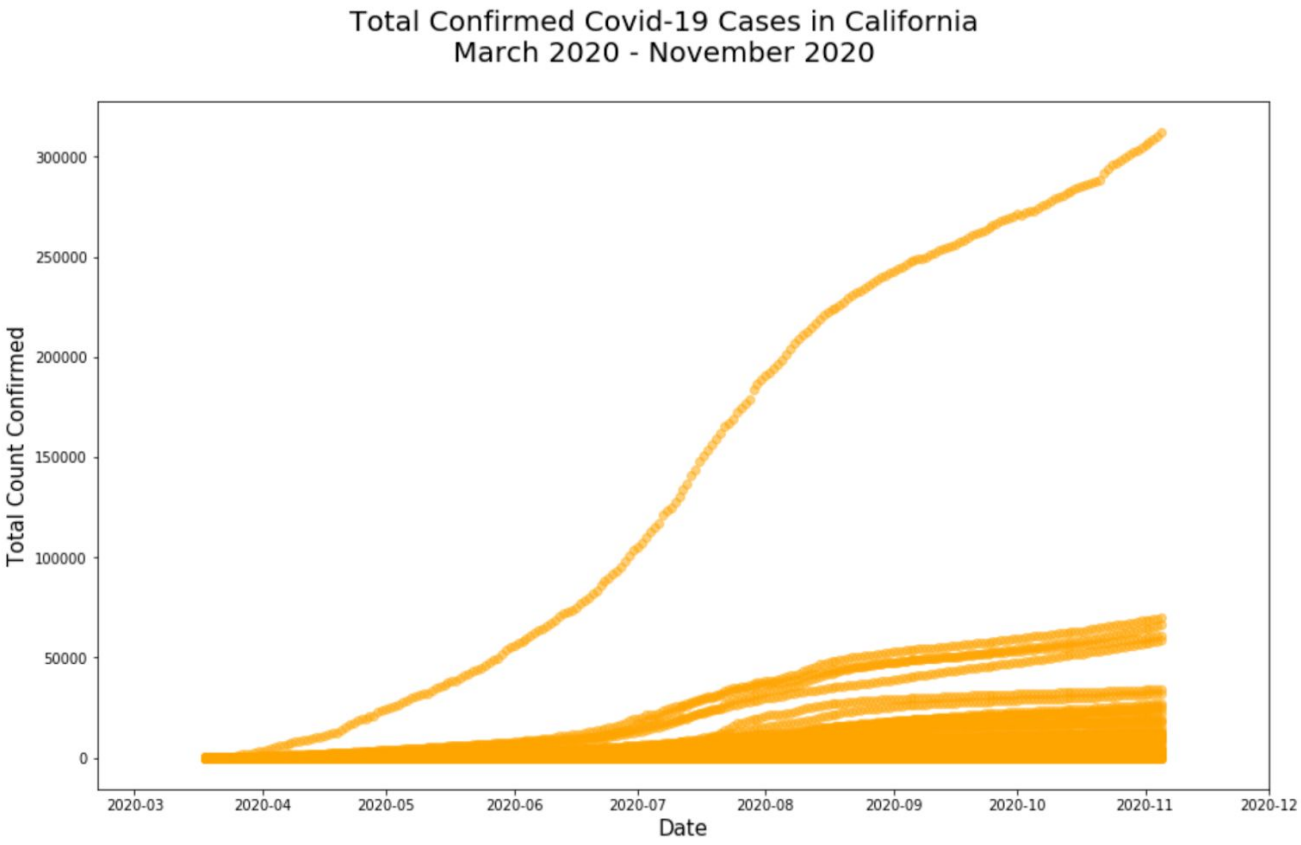
Our data cleanup methods with Python consisted of removing rows with missing data as well as the outliers. Printed summaries from the raw data attributes displayed negative values within the population not staying at home and also within the number of trips taken, so these outliers were removed. We also cleaned up spelling discrepancies in the names of the counties to ensure there was no duplicate information. After acquiring updated data further along in our analysis, the same spelling cleanup wasn't required.

Visually, we got a sense of the patterns in our data through boxplots and scatterplots. Using all of the observations in one graph proved to be convoluted and ineffective, so in order to make our data more manageable and focused, we decided to narrow our analysis to only California counties, and only trip distances covering over 500 miles.



With a more manageable subset of data, we were able to see the general ups and downs of travel displayed. The overall shape of the trend is what we originally expected. There are more trips in the first half of the graph during 2019, with a drop across most of 2020, and a dramatic spike during the summer in August 2020. Finally, there is a drop at the end of the data during the Fall as flu season starts up and Covid-19 cases begin to rise again.

Keeping in mind the current circumstances surrounding the pandemic, we supplemented our analysis with additional data reporting Covid-19 statistics in California in order to compare the trends. California Open Data provides confirmed positive cases and deaths reported by local health departments in California and is updated daily. These observations include a cumulative number of laboratory-confirmed positive cases, as well as Covid-related deaths from the California Department of Public Health (CDPH), starting from March 19, 2020 (California Open Data, 2020).



With our cleaned data sets from Python, we moved into R to merge them and began the modeling phase. Combining the Covid-19 data with our California trips data allowed us to visualize and integrate the number of Covid-19 cases and deaths along the same timeline as the number of trips taken.

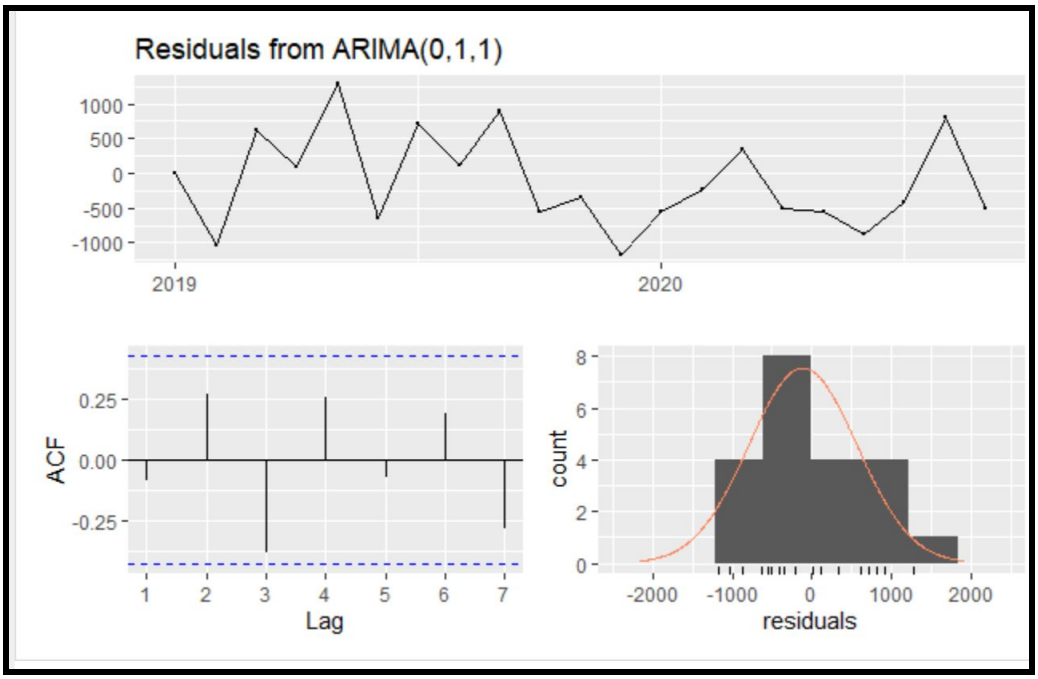
We wanted to review multiple methods when determining the best model. First, we broke the data out into test data and train data so we would not overfit the algorithm. We ran many time series methods on our train data to predict the number of trips to be taken in California in the coming months.

The models we ran included Naive, Simple Exponential Smoothing, ARIMA, Holt's Trend and TBATS. The Naive model is the simplest forecasting tool. It is used as a benchmark to evaluate other algorithms. The Naive method predicts the next value equal to the last observed value. Simple Exponential Smoothing is slightly more

complex than the Naive model. This method uses weighted averages so more weight is given to more recent data points and less weight is given to older data points. The next model we ran was ARIMA. ARIMA stands for Autoregressive Integrated Moving Average. This tool is a type of smoothing model for stationary time series. This method contains three parts. The first is Autoregressive which is the weighted sum of the past values in the series. The Moving Average part is the weighted sum of the past forecasted errors. Finally the Integrated part is the difference of the model. Holt's Trend is an extension of the simple exponential smoothing method. It calculates a forecast based on a forecast equation and 2 smoothing equations. The TBATS was derived from exponential smoothing models. It takes into account seasonality in the data.

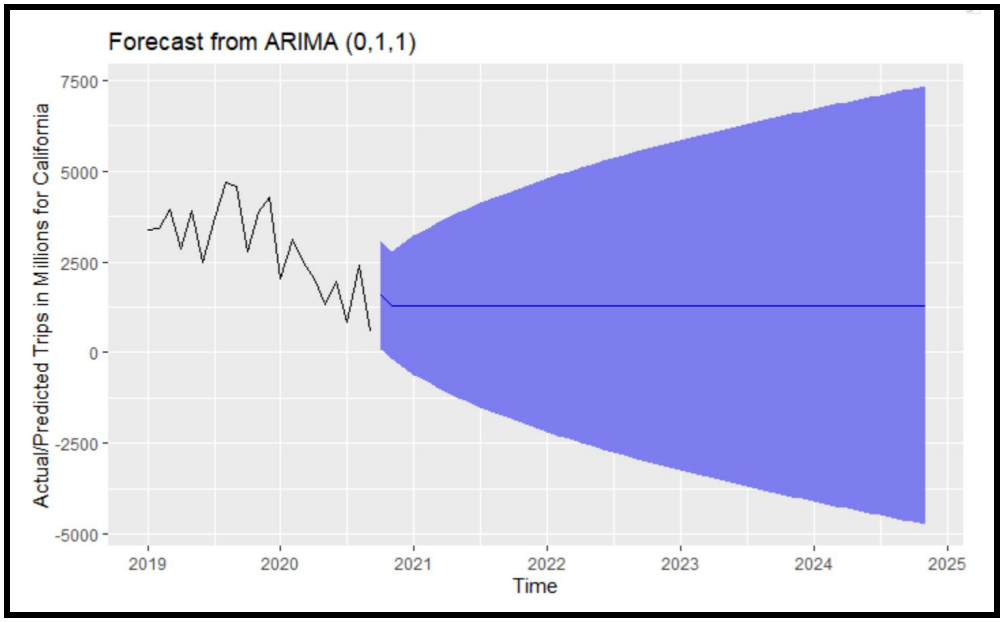
Results

The main criteria that we based the models on are mean absolute percentage error (MAPE). The lower the Mape, the better the model. The Naive process was by far the worst. Simple Exponential Smoothing, ARIMA, and TBATS were all similar but ARIMA was consistently the winner. The Mape for the Naive Model was 5,629. This is a very poor Mape but this was to be expected for the Naive method. The Simple Exponential Smoothing Mape was 161. This is a dramatic drop from the Naive process. The ARIMA Mape was 158. This turned out to be the lowest Mape but was very close to the other models excluding the Naive. Holt's trend had a Mape of 255. This was the second highest Mape from the methods that were tested. The TBATS model had a Mape of 162. This was narrowly beaten out by the ARIMA.

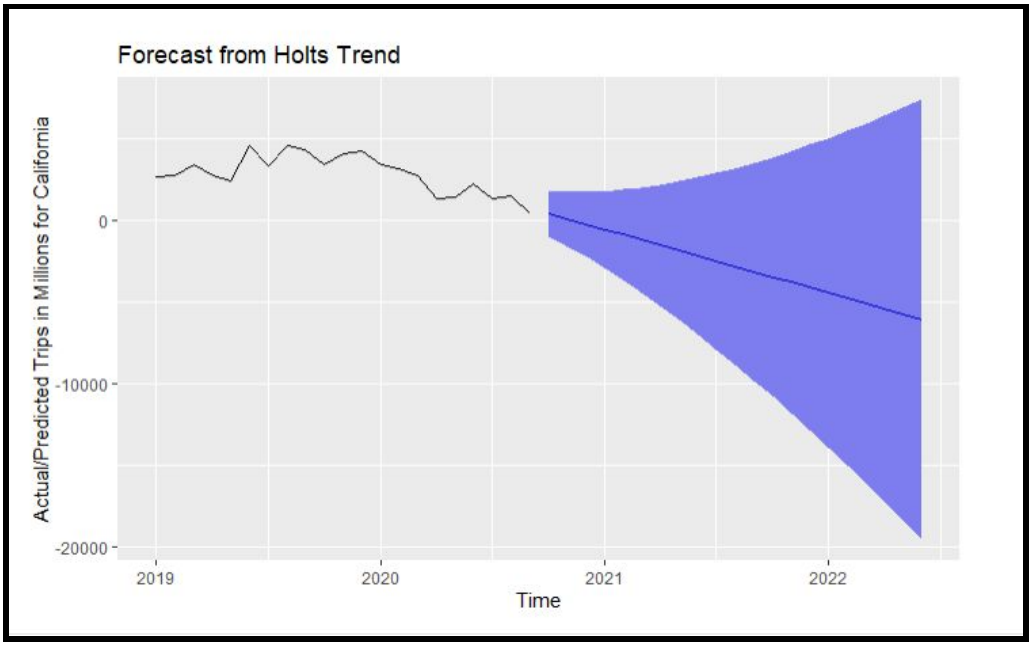


Since the ARIMA model was the best choice, we wanted to explore that further. Using an auto fit ARIMA method, we found ARIMA (0,1,1) was the best choice for our data. The ARIMA (0,1,1) is a simple exponential smoothing tool with growth. ARIMA models use p, d, and q terms which are the numbers in parentheses next to ARIMA. The p is the Autoregressive term. This is the number of lags that will be used as predictors. The q is the Moving Average term, which is the number of lagged forecast errors. The d is the difference factor which is the number of times the data was differenced so it would be stationary. For our model, ARIMA (0, 1, 1), the autoregressive coefficient is 0, the differencing coefficient is 1, and the moving average coefficient is 1. The figure below shows the graphs that support the decisions of these coefficients. The residuals are stationary meaning the variance is consistent through the timeline. For the lag graph, there are no points that fall outside of the blue variance line. This shows the correct choice was made for the autoregressive and moving average terms.

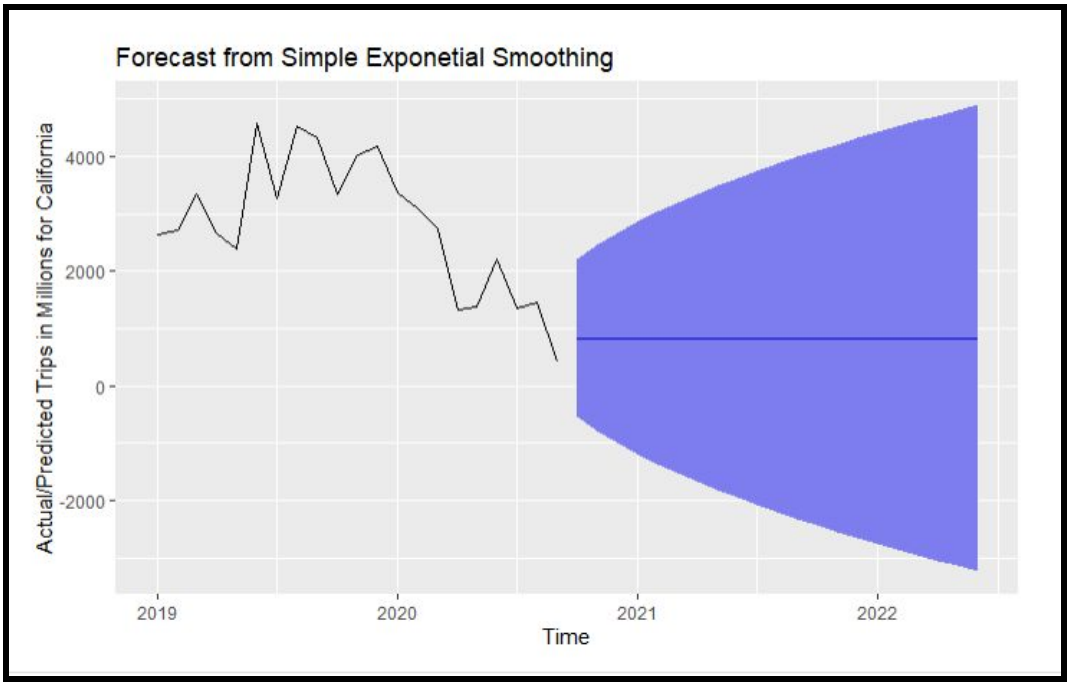
The figure below shows us the predicted number of trips traveled into the future in millions. Based on the graph below, it is hard to predict the number of trips to be traveled. The purple shade provides a 90% confidence interval. The confidence interval does go below zero but this is not realistic. It is estimated the number of trips per month will be around 1400 million but this will be hard to predict.



We compared the ARIMA forecast with a few of the other models that we tested. The Holt's Trend Forecast decreased over time because that is the current trend of travel in California. This forecast is worse than the ARIMA forecast because it is too aggressive in it's decrease.



The forecast for the SImple Exponential Smoothing model looks similar to the ARIMA forecast. The difference is that the ARIMA forecasts 1400 million trips per month and the Simple Exponential Smoothing model predicts about 900 million trips per month. The Simple Exponential Smoothing prediction is too low.

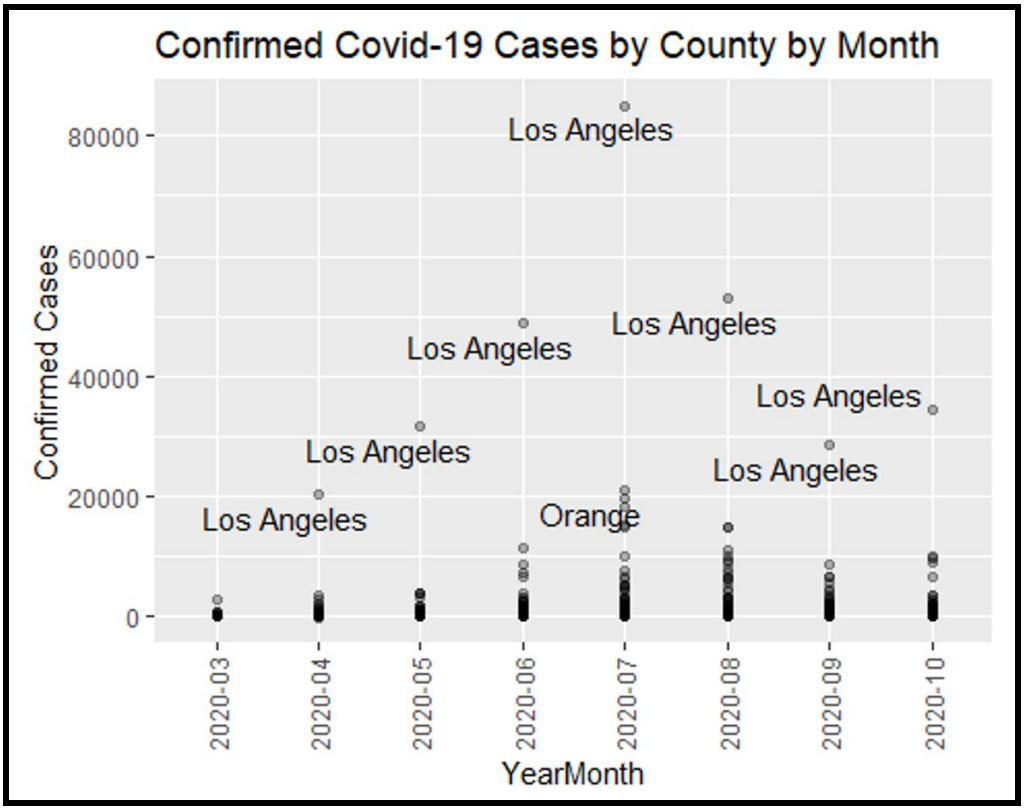


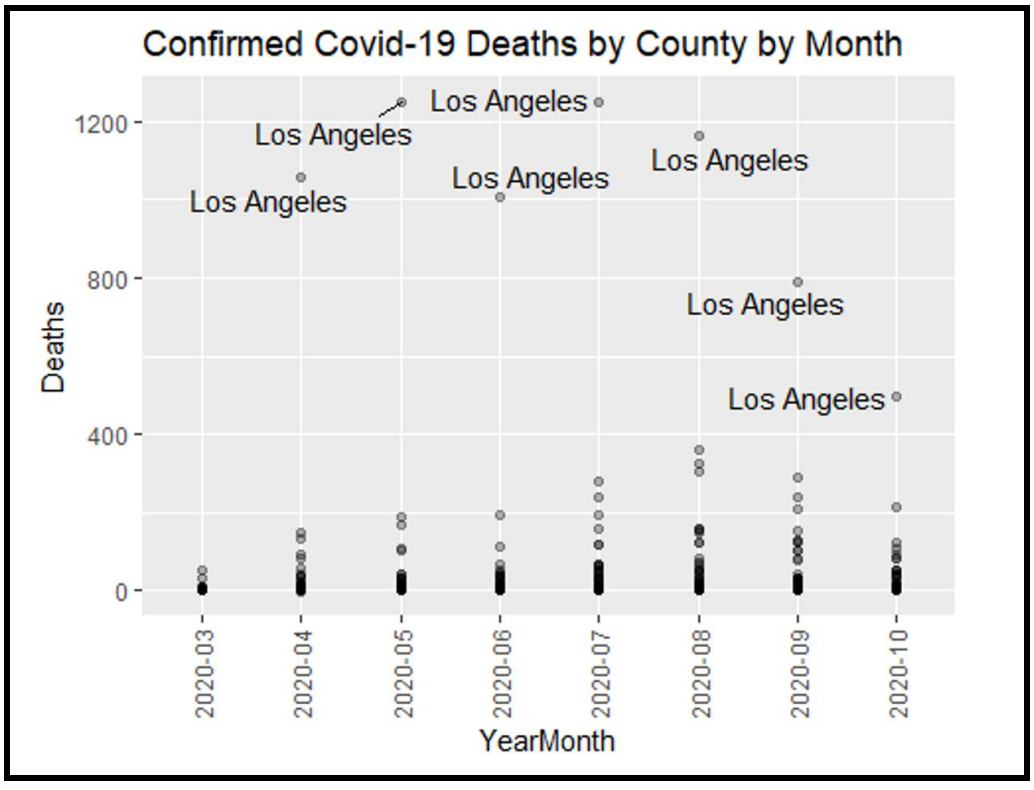
Discussion and Conclusion

We do know the current trend of Covid-19 cases and deaths which will drive a lot of the travel in the coming months. We saw a travel dip when Covid-19 first infiltrated California in March and April. If cases and deaths keep decreasing, we would expect travel to increase, but recently the Covid-19 numbers have continued rising. September and October have very similar numbers of cases with October having slightly more. We are not seeing the downward trend that we saw from July through September.

Towards the end of our analysis, we downloaded and ran updated data from both Data.gov and California Open Data through our preliminary analysis steps and our models to check for any new trends in trips being taken and Covid-19 statistics. The additional 3,230 observations illustrated a further rise in positive cases and another dramatic dip in trips taken after the summer months. Thankfully, even with positive cases rising, the number of deaths has continued to decrease in November.

It will be interesting to see whether or not the number of cases stays the same, and if the number of deaths stays the same or keeps decreasing in the coming months. Los Angeles continues to have the highest number of deaths and cases for California. This makes sense because it is the most populous county. Different safety measures should be taken in Los Angeles county versus more rural counties. With a more dense population, Los Angeles must have stricter protocol to keep the Covid-19 cases and deaths down.


Confirmed Covid-19 Cases by County by Month
Confirmed Cases
80000
60000
40000
20000
0
Los Angeles
Los Angeles
Los Angeles
Los Angeles
Los Angeles
Los Angeles
Orange
Los Angeles
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
YearMonth


Confirmed Covid-19 Deaths by County by Month
Deaths
1200
800
400
0
Los Angeles
Los Angeles
Los Angeles
Los Angeles
Los Angeles
Los Angeles
Los Angeles
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
YearMonth

Hopefully the number of cases and deaths will continue to decrease. Some risks to this thinking would be if flu season increases the amount of people who are sick, resulting in another decrease in travel. A working and distributed vaccine could decrease cases and deaths and increase the number of trips being taken. There are a lot of aspects that will affect travel in the future.

Overall, it is challenging to make predictions based solely on past trip numbers. Using Covid-19 data helped with comparisons between trends observed and confirming expectations, but having more variables to take into consideration would be beneficial. For example, weather, school breaks, income, proximity to family, and profession could all come into play when people are deciding to take trips or not. Furthermore, having more insight into what type of trips are being taken within each record, such as road trips, air travel, or train travel could offer valuable information for the transportation industry and travel agencies.

Acknowledgments

We would like to acknowledge the support provided by our families during the preparation of this course project.

References

BTS.gov. (2020, November 9). Trips By Distance. Retrieved from https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

California ALL. (2020, November 9). Tracking COVID-19 in California. CA.gov. Retrieved from https://covid19.ca.gov/state-dashboard/

California Open Data. (2020). COVID-19 Cases. Retrieved from https://data.ca.gov/dataset/covid-19-cases/resource/926fd08f-cc91-4828-af38-bd45de97f8c3

Data.gov. (2020, September 29). Trips By Distance. Bureau of Transportation Statistics. Retrieved from https://catalog.data.gov/dataset/trips-by-distance

Forecasting: Principles and Practice. (n.d.). Retrieved November 08, 2020, from https://otexts.com/fpp2/arima-r.html

Prabhakaran, S. (2020, September 17). ARIMA Model - Complete Guide to Time Series Forecasting in Python. Retrieved November 08, 2020, from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Singh, D. (2019, July 12). Deepika Singh. Retrieved September 29, 2020, from https://www.pluralsight.com/guides/time-series-forecasting-using-r

Tourwriter. (2020). The travel industry is resilient: Historical events and how tourism has bounced back. Retrieved from https://www.tourwriter.com/travel-software-blog/covid-19-pt1/