Project Report

Effects of COVID-19

Prepared by: Team 5

Yuanying Li,

Chen Liang

Ketaki Joshi

Chenang Zhang

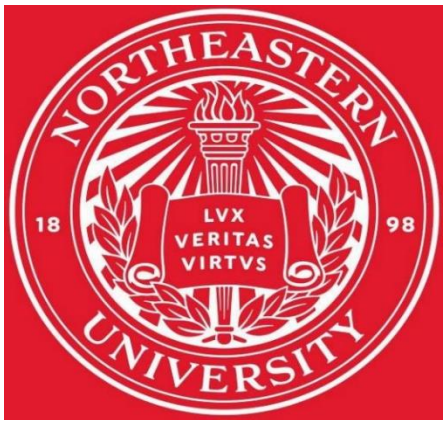
Course: ALY 6015

Under the Guidance Of:

Prof. Fidel Rodriguez

Northeastern University

May 15, 2020



Contents

1. Introduction

From the World Health Organization - On 31 December 2019.

WHO alerted several cases of pneumonia in Wuhan City, Hubei Province of China. The virus did not match any other known virus. This raised concern because when a virus is new, we do not know how it will affect people. The outbreak of COVID-19 was developed into a major international crisis and started influencing important aspects of daily life. For example:

Transportation: Companies around the globe rolled out mandatory remote work. Bans have been placed on hotspot countries, corporate travel has been reduced, and so on.

Grocery stores: In highly affected areas, people are starting to stock up on essential goods.

The pandemic has not only severely impacted humans 'daily life but also changed the lifestyle. After the outbreak of COVID-19, the government has started to take necessary precautions to slow down the spread of the virus. The crucial step was taken by declaring a shelter-in-place order/lockdown.

When the government rolled out a mandate to stay at home and maintain social distancing, it had a direct impact on day-to-day transportation. In the USA, the primary mode of transportation is owned vehicles (cars). Since the shelter-in-place order implemented, companies also rolled out an official mandate to work remotely. Schools, universities also closed the campuses and moved classes online. Vehicle traffic began to reduce. We were curious to find out the effect of COVID-19 on-road transportation, i.e. on vehicle miles traveled (VMT) and whether reduced VMT has any effect on fatal car accidents.

We divided our project into three parts. In the first part, we explored the current situations about the COVID-19 globally and presented our findings.

In the second part, as we introduced earlier, we were curious to find out the effect of COVID-19 on-road transportation. We found that on March 17, six San Francisco bay area counties firstly implemented shelter in place order. On the evening of March 19, California state became the first state in the nation to release a statewide order. This order required people to stay home and limit social interaction, except for essential activities. So, we narrowed our analysis for the state of California. We found that after shelter in place order was implemented, the average vehicle miles traveled was decreased.

We believed there is a proportional relationship between VMT and a fatal car accident. That is, the less miles drove, less chance of fatal accidents. So, keeping other things the same, in the third part, we use the prediction of a fatal car accident without COVID-19 to multiply with the decreased percentage of vehicle miles of travel to calculate how COVID-19 might saves people from the fatal car accident in California state.

In the fourth part, we incorporated the necessary improvements.

2. Research questions

1. What is the number of confirmed cases, deaths, recoveries, and active cases globally?
2. Prominently affected countries by COVID-19
3. Current situation in the USA
4. Which county in California has the largest vehicle miles traveled? Which county has the shortest?
5. Whether shelter in place order reduced the vehicle miles traveled during the COVID-19 period in California?
6. By how many percent the average vehicle miles traveled in California was dropped after the shelter in place order was implemented?
7. How will the decline in VMT affect the number of fatal accidents in California?
8. How many numbers of people saved?
9. Is there a directly proportional relationship between VMT and fatal car accidents in the past years?

3. Dataset(s) chosen

Part I: There are a lot of official and unofficial data sources on the web providing COVID-19 current situation. One of the most widely used dataset today is the one provided by the John Hopkins University's Center for Systems Science and Engineering (JHU CSSE). Here is the Github link of Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE.

Link: https://github.com/CSSEGISandData/COVID-19

Part II: We considered using the data to check if shelter in place order really reduced vehicle miles traveled and how much percent the miles dropped.

We got the vehicle miles traveled (VMT) data from StreetLight, a transportation analytics firm. The data shows the county-by-county (almost all states in the USA) VMT Metrics for more than 3,100 U.S. counties from March 1, 2020, to May 5, 2020, and includes the average VMT in January 2020.

https://learn.streetlightdata.com/vmt-data-counties

According to StreetLight's website, the company got data by ingesting, indexing, and processing over 100 billion anonymized location records from smartphones and navigation devices in connected cars and trucks. We have also seen some governments and media use StreetLight data for traffic analysis, so we assume that the data is reliable.

Part III: NHTSA is an agency of U.S. federal government, part of the Department of Transportation. We merged several datasets of California into one and thus the dataset is from 1999 Jan to 2018 Dec.

The dataset is from: https://www-fars.nhtsa.dot.gov/Crashes/CrashesTime.aspx.

Part IV: The data is from national VMT data for each month.

https://fred.stlouisfed.org/series/M12MTVUSM227NFWA.

Note: We are using national VMT's behavior to mimic the VMT behavior in California. We tried to find historical data for California VMT but failed because of limited time.

4. Method(s) chosen

Part I: Exploratory data analysis using plotly package

Part II: EDA, Hypothesis test, level of significance (alpha = 0.05)

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. We divided our dataset into two groups. One is before the shelter in place order is implemented, the other one is after the order is implemented. Two groups are mutually exclusive.

The null and alternate hypothesis is as follows:

H0: The average number of vehicle miles traveled (VMT) after shelter in place order was implemented is similar to the average VMT before the order was implemented.

H1: The average number of vehicle miles traveled after the shelter in place order was implemented is less than the average VMT before the order was implemented.

In this research, we used two types of hypothesis test: T-test and F-test. The T-test is used to determine if there is a significant difference between the means of two groups. F-test also tells whether two groups are similar or not based on their mean similarity and f-score.

Part III: A time-series regression to predict the fatal car accident after April 2020

Part IV: A time-series regression

5. **Part I: Current situation about Covid-19**

5.1: Importing & Understanding the Dataset

```
confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse
deaths_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_co
recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse
cases_country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-dat
print(confirmed_df.shape)
print(deaths_df.shape)
print(cases_country_df.shape)
print(cases_country_df.shape)
```

```
(266, 114)
(266, 114)
(187, 14)
(187, 14)
```

In this part, we used the above-mentioned data-frames to subset/filter the data for our use.

Here is a view of the columns in the confirmed_cases data frame.

5.2: Analyzing Coronavirus: Exploratory Analysis

Let us have a look at the global situation so far. That is the number of confirmed cases, deaths, recoveries, and active cases.

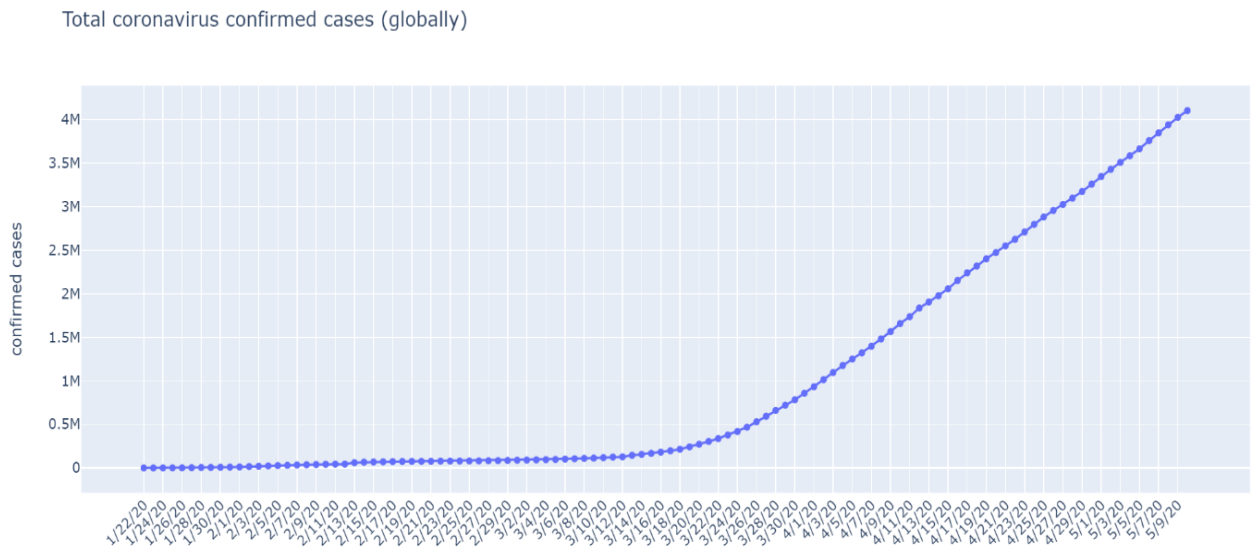
```
#Count the sum of confirmed, deaths, recovered and active cases of Covid-19 globally
global_data = cases_country_df.copy().drop(['Lat','Long_','Country_Region','Last_Update','Incident_Rate','People_Tested','People_Hospitalized','Morta
global_summary = pd.DataFrame(global_data.sum()).transpose()
global_summary.style.format( "{:,.0f}")
```

	Confirmed	Deaths	Recovered	Active
0	4,516,360	306,051	1,622,354	2,585,673

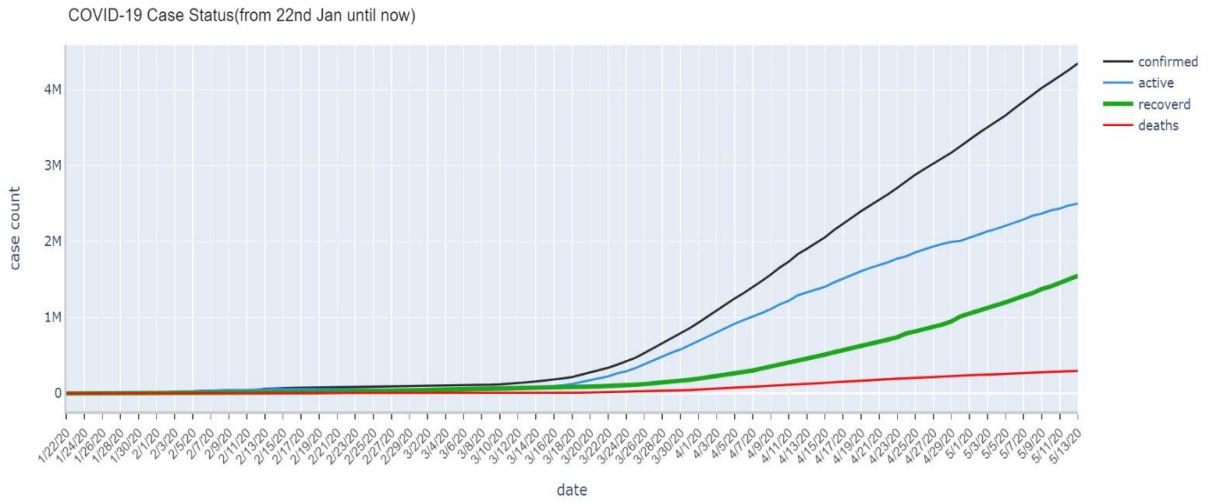
Here we can see the number of confirmed cases has crossed the figure of 4.5 million. If we look at the number of deaths because of COVID, it is 6.77%. The good news is, nearly 36% are recovered. The active cases are still above 2.5 million. This is scary. By these numbers, we can understand how deadly the impact of this virus is.

We plotted the confirmed cases using plotly.graph objects to find out how this spread has progressed over a period of time.

```
fig_1 = go.Figure(data = go.Scatter(x = confirmed_ts_summary.index, y = confirmed_ts_summary.values, mode = 'lines+markers'))
fig_1 = go.Figure(data = go.Scatter(x = confirmed_ts_summary.index, y = confirmed_ts_summary.values, mode = 'lines+markers'))
fig_1.update_layout(title = 'Total coronavirus confirmed cases (globally)',yaxis_title='confirmed cases',xaxis_tickangle = 315)
fig_1.show()
```



The sharp exponential curve can be seen on the right side of the graph. It shows the devastating rate at which the pandemic is spreading worldwide. Before further drill-down, we looked at the progression of recovered, death, and active cases as well.

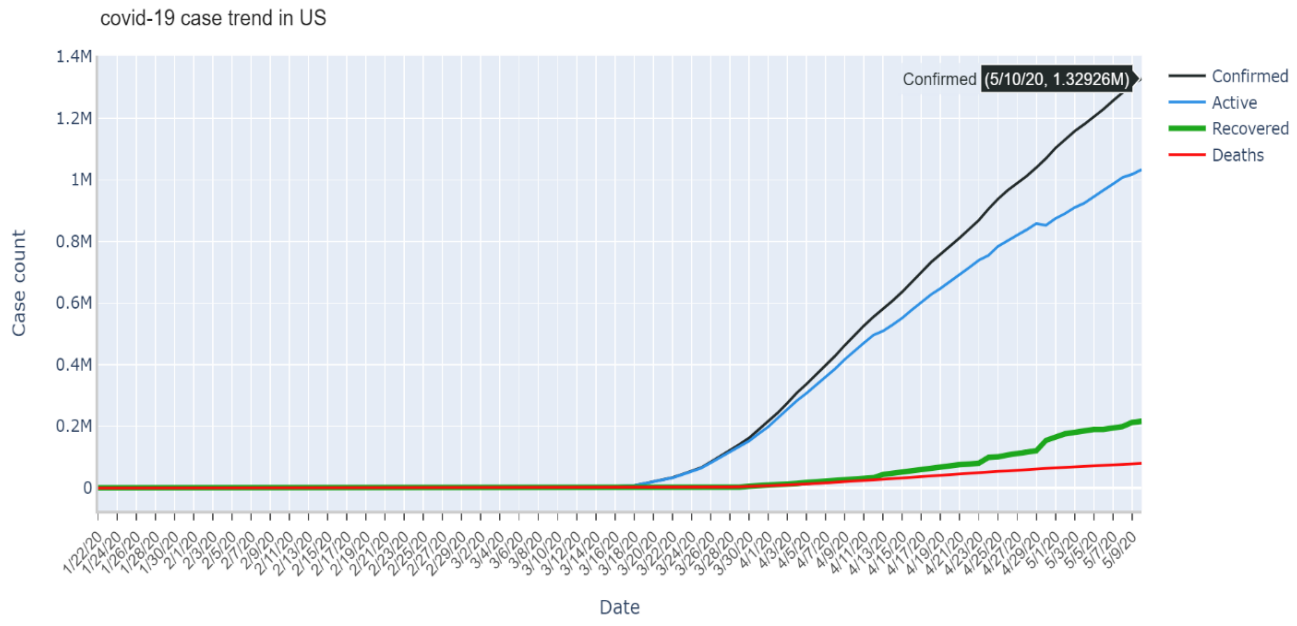


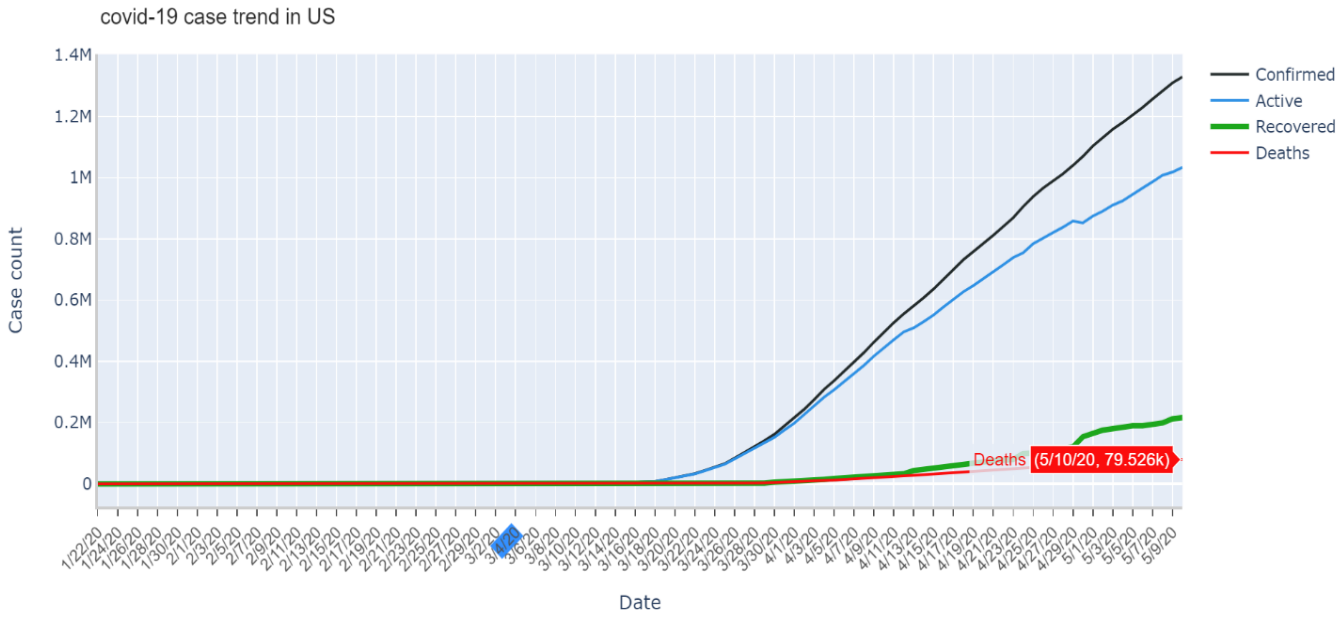
Drilling down at the country level, we found out which countries are prominently affected (the top 20 countries in terms of confirmed cases). The current situation is shown below. Also, notice the neat “bars” in each cell.

	Country_Region	Confirmed	Deaths	Recovered	Active	Incident_Rate	Mortality_Rate
0	US	1334951	79699	216169	1045589	405.185923	5.970182
1	Spain	224350	26621	136166	61563	479.843955	11.865835
2	United Kingdom	224327	32140	1007	191180	330.446612	14.327299
3	Russia	221344	2009	39801	179534	151.673566	0.907637
4	Italy	219814	30739	106587	82488	363.558310	13.984096
5	France	177094	26383	56328	94383	271.310666	14.897738
6	Germany	171999	7569	145600	18830	205.288734	4.400607
7	Brazil	163510	11207	64957	87346	76.924376	6.854015
8	Turkey	138657	3786	92691	42180	164.404237	2.730479
9	Iran	109286	6685	87422	15179	130.113297	6.116977
10	China	84010	4637	79171	202	5.980737	5.519581
11	Canada	70138	4992	32486	32660	185.277240	7.117397
12	India	69400	2254	21664	45482	5.028970	3.247839
13	Peru	67307	1889	21349	44069	204.134764	2.806543
14	Belgium	53449	8707	13697	31045	461.180077	16.290295
15	Netherlands	42987	5475	149	37363	250.874343	12.736409
16	Saudi Arabia	41014	255	12737	28022	117.809378	0.621739
17	Mexico	35022	3465	23100	8457	27.162997	9.893781
18	Pakistan	30941	667	8212	22062	14.007277	2.155716
19	Switzerland	30344	1845	26600	1899	350.610506	6.080279
20	Chile	30063	323	13605	16135	157.264445	1.074410

5.3: Analyzing Coronavirus: The USA Focus

From the above graph, we can see that, in terms of confirmed cases, the USA is on the top. So, we focused on the data points with respect to the USA. Let us have a look at how this virus has spread across so far by plotting the USA specific time-series and annotating those with the events manually.





One can easily see that there was hardly any spread in the USA until 03/18/2020. As of today, there are 1.32 million COVID-19 confirmed cases and 79, 526 deaths in the USA. On 17th March, some states began to issue a Shelter-In-Place order. And we have found some interesting changes during this quarantine.

Let us now look at the impact of COVID-19 on the road transportation in the USA.

6. Part II: Effect of COVID-19 on the vehicle miles traveled (VMT)

6.1: Analyzing VMT: Exploratory Analysis

During the period from March 1, 2020, to May 5, 2020, the average number of miles traveled was 2,396,748 with a standard deviation of 7,014,876. The maximum and a minimum number of miles traveled was 323,000,000 and 2,440, respectively.

count	192911.0
mean	2396748.0
std	7014876.0
min	2440.0
25%	311000.0
50%	757000.0
75%	1990000.0
max	323000000.0

Here we can see that the standard deviation is high. The standard deviation is so high because we considered almost all the states and counties in the USA. Moreover, all the states and counties do not have a similar population.

Also, shelter in place order has not been issued and implemented all over the USA on the same day. Some states issued and implemented shelter in place earlier, some states later and some states have not issued a shelter in place order.

These could be the reasons for the high standard deviation and difference between the minimum and the maximum number of miles traveled.

After exploring the dataset, we found that the maximum number of miles traveled was in Texas state and Harris county on March 7, 2020. The minimum number of miles was traveled in Massachusetts state and Nantucket county on April 26, 2020.

During our research, we found that, on March 19, 2020, California became the first state in the nation to declare a statewide shelter in place. So, we narrowed down our analysis for California state. Following is the five-point summary along with the average and standard deviation.

count	3696.0
mean	8532870.0
std	20362342.0
min	17000.0
25%	650000.0
50%	2135000.0
75%	6742500.0
max	275000000.0

We found that the minimum number of miles traveled was in Trinity county after the shelter in place was implemented and the maximum number of miles traveled was in Los Angeles county before the shelter in place was implemented.

	statefp10	countyfp10	state_name	county_name	ref_dt	county_vmt	jan_avg_vmt
13697	6	105	California	Trinity	2020-04-05	17000	90035

	statefp10	countyfp10	state_name	county_name	ref_dt	county_vmt	jan_avg_vmt
11489	6	37	California	Los Angeles	2020-03-06	275000000	237689973

We also found that Santa Clara County (including the other five counties) in the San Francisco Bay Area became the first in California state to implement the shelter in place starting March 17 to slow the spread of the virus.

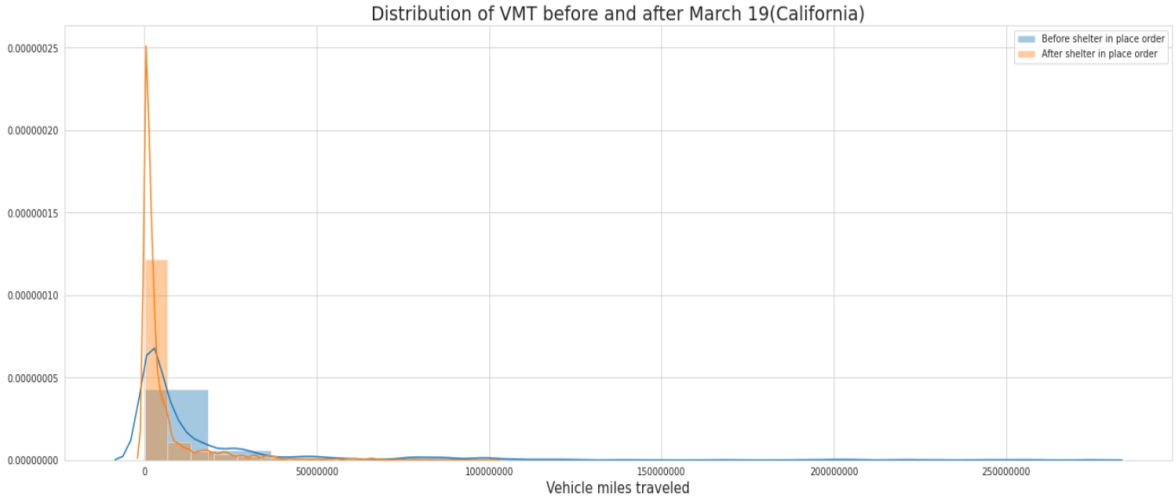
We decided to use Santa Clara, Los Angeles, and Trinity counties for further analysis.

The selection of Trinity and Los Angeles counties is based on the minimum and the maximum number of miles traveled, respectively. Here, we ignored the reference date.

Let us first look at for the state of California.

6.2: Hypothesis test for VMT in California State

On the evening of March 19, 2020, California became the first state in the nation to declare statewide order. The order immediately went into effect.



We found that the average VMT in California before and after shelter in place order implemented had a lot of overlap. The mode of VMT after the order implemented stands out. Two groups are independent and normally distributed.

```
# Compare samples by using F-test
stat, p = f_oneway(caBefore_df['county_vmt'], caAfter_df['county_vmt'])
print('Statistics=%.3f, p=%.3f' % (stat, p))

# Compare p-value with significance level alpha
alpha = 0.05
if p > alpha:
    print('Same distributions (fail to reject H0)')
else:
    print('Different distributions (reject H0)')
```

```
Statistics=247.791, p=0.000
Different distributions (reject H0)
```

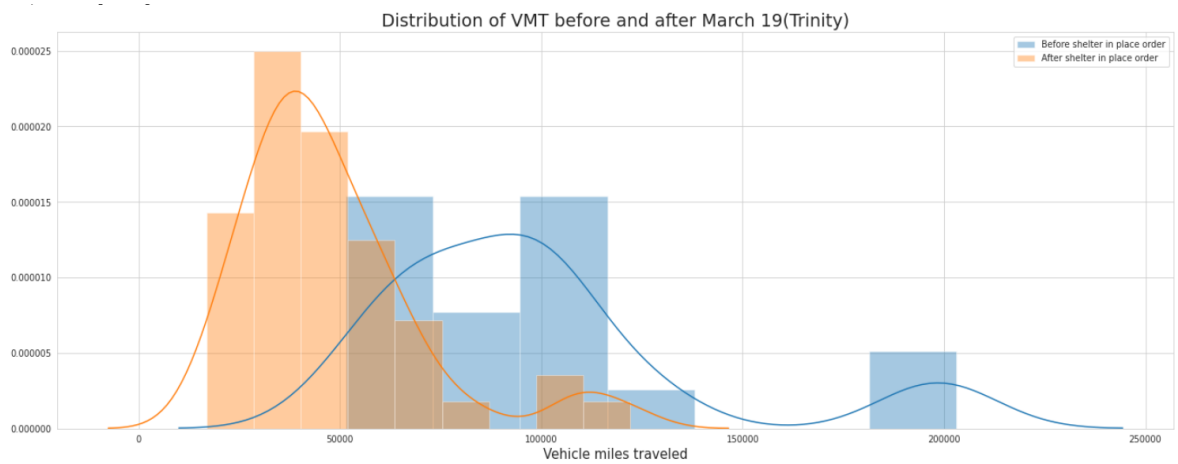
```
# Compare samples by using T-test
t_test, p =ttest_ind(caBefore_df['county_vmt'], caAfter_df['county_vmt'])
print("The t_test value =", t_test,"p-value=", p)

# Compare p-value with significance level alpha
alpha = 0.05
if p <= alpha:
    print("The average vehicle miles traveled are decreased (Reject Ho)")
else:
    print('The average vehicle miles traveled are same (Do not reject Ho)')
```

```
The t_test value = 15.7413789287316 p-value= 4.344304230335753e-54
The average vehicle miles traveled are decreased (Reject Ho)
```

From the F-test and T-test, we conclude that after the shelter in place order was implemented, the average vehicle miles traveled have been decreased.

6.3: Hypothesis test for VMT in Trinity County

The VMT in Trinity county before shelter in place order implemented ranged from 50,000 miles to 200,000 miles which is higher than the miles after the order implemented (ranged from 20,000 miles to 120,000 miles). Two groups are independent and normally distributed.

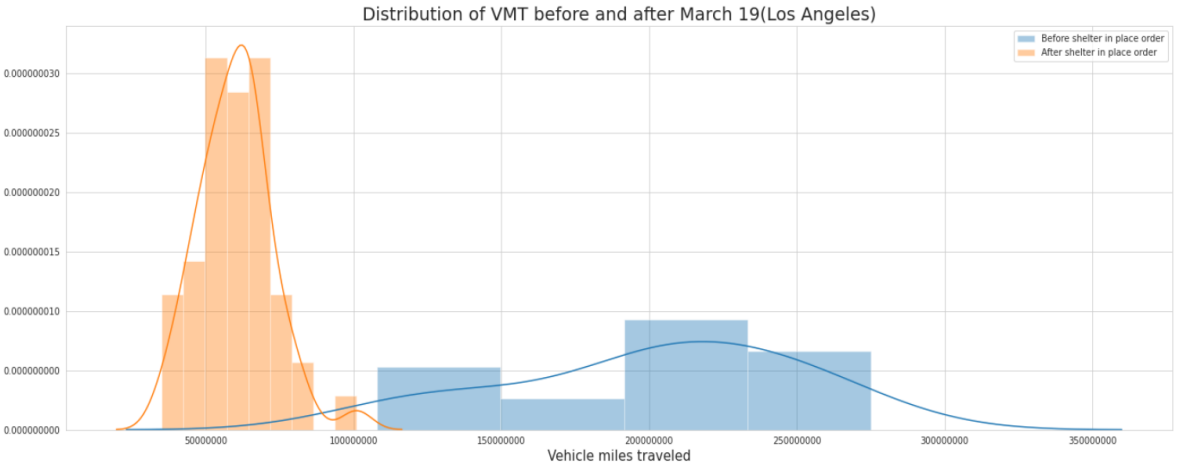



```
#Compare samples using t-test

t_test, p =ttest_ind(tBefore_df['county_vmt'], tAfter_df['county_vmt'])
print("The t_test value =", t_test,"p-value=", p)

#Interpret results
alpha = 0.05
if p <= alpha:
    print("The average vehicle miles traveled are decreased (Reject Ho)")
else:
    print('The average vehicle miles traveled are same (Do not reject Ho)')
```

```
The t_test value = 6.506581537034709 p-value= 1.3543537455941889e-08
The average vehicle miles traveled are decreased (Reject Ho)
```

Since the p-value is less than alpha, we rejected the null hypothesis. The vehicle miles traveled have reduced after the shelter in place order was implemented.

6.4: Hypothesis test for VMT in Los Angeles County



The VMT in Los Angeles county before shelter in place order implemented ranged from 110,000,000 miles to 255,000,000 miles which is higher than the miles after the order was implemented.

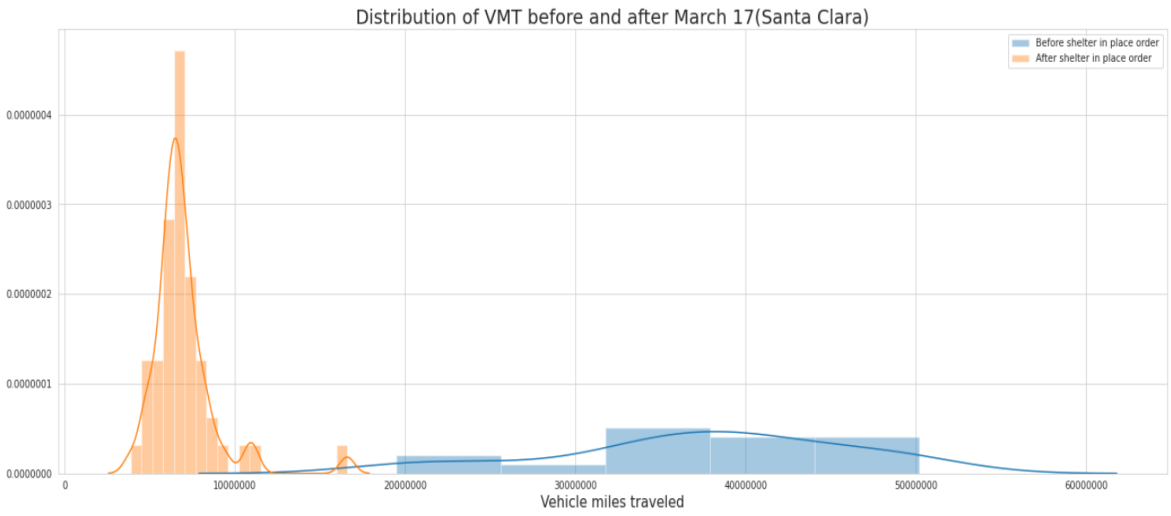
The number of miles that people drove every day in LA county after the order implemented was between 50,000,000 miles to 80,000,000 miles.

```
# Compare samples by using F-test
stat, p = f_oneway(laBefore_df['county_vmt'], laAfter_df['county_vmt'])
print('Statistics=%.3f, p=%.3f' % (stat, p))

# Interpret results
alpha = 0.05
if p > alpha:
    print('Same distributions (fail to reject H0)')
else:
    print('Different distributions (reject H0)')
```

```
Statistics=347.915, p=0.000
Different distributions (reject H0)
```

Since p < alpha, we rejected the null hypothesis. The shelter in place order was successful in reducing the vehicle miles traveled in Los Angeles county. F-statistic = 347.915 tells us that there is a significant difference in the group mean.

6.5: Hypothesis test for VMT in Santa Clara County



We found that the VMT in Santa Clara county before shelter in place order implemented ranged from 20,000,000 miles to 50,000,000 miles which is higher than the miles after the order was implemented. The number of miles that people drove every day in Santa Clara County after the order implemented was less than 10,000,000 miles.

```
# Compare samples by using F-test
stat, p = f_oneway(scBefore_df['county_vmt'], scAfter_df['county_vmt'])
print('Statistics=%.3f, p=%.3f' % (stat, p))

# Interpret results
alpha = 0.05
if p > alpha:
    print('Same distributions (fail to reject H0)')
else:
    print('Different distributions (reject H0)')
```

```
Statistics=547.216, p=0.000
Different distributions (reject H0)
```

Since p < alpha, reject the null hypothesis. Shelter in place order was successful in reducing the vehicle miles traveled in Santa Clara county. F-statistic = 547.216 tells us that there is a significant difference in the group mean.

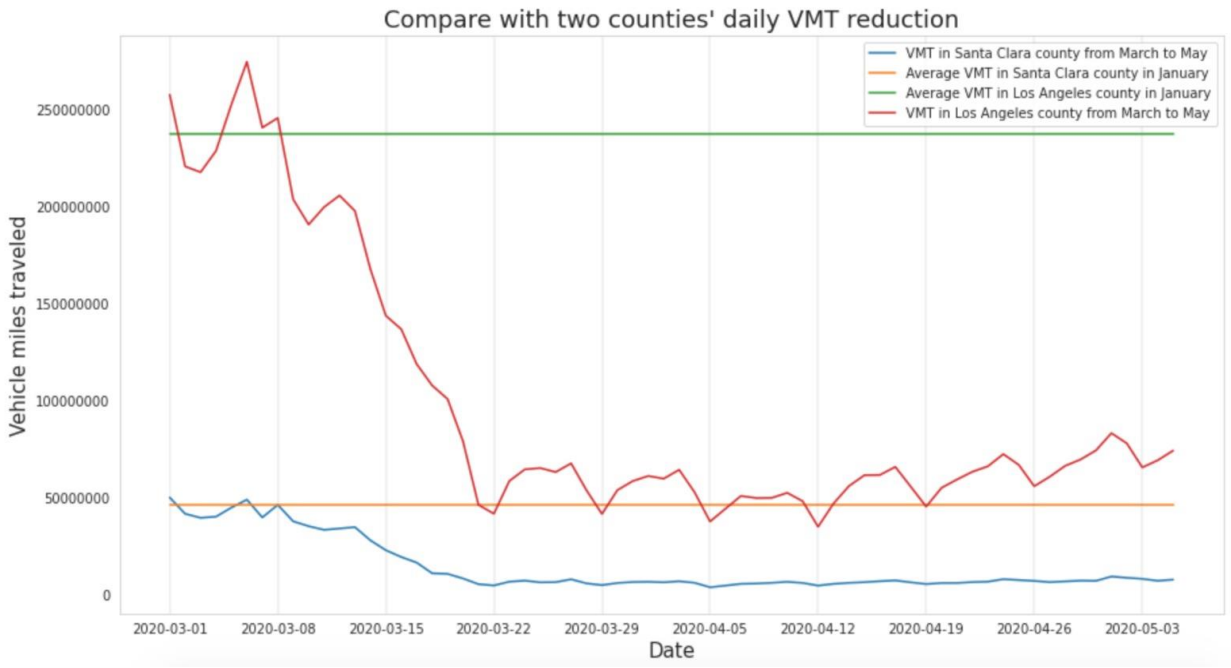
6.6: The average VMT percentage dropped

```
The percentage VMT decreased in California State:68.76%
The percentage VMT decreased in Trinity County:45.12%
The percentage VMT decreased in Los Angeles County:70.33%
The percentage VMT decreased in Santa Clara County:79.38%
```

We observed that in an initial span of 18 days after the shelter in place was implemented, there was a decrease of 68.76% in the average VMT in California State.

We did observe a significant percent decrease in the average VMT for the other three counties as well.

6.7: Graphical representation of daily VMT dropped in Santa Clara and Los Angeles County



The number of vehicle miles traveled in Los Angeles county is obviously larger than Santa Clara county, but they both had a significant decrease in March. Although California implemented the shelter in place order starting in mid-March, vehicle miles traveled has been declining since early March.

The reason is, as the number of COVID-19 confirmed cases increased, many companies in California required employees to work from home, and some school districts and universities announced closures in early March.

We can see the obvious waves in the plot of Los Angeles' VMT. The lowest point of the plot is usually on Sunday. This also means that some residents still drove for essential businesses on weekdays and followed the orders on weekends.

We are also seeing a rise in VMT as the shelter in place order continues, perhaps because people cannot stand staying indoors anymore.

We did not draw Trinity county on this plot because the VMT in Trinity county is so small that it tends to be a straight line when compared to other counties. We could not get valuable information.

7. Part III: Effect of COVID-19 on fatal car accidents

7.1: Clean the data

	Period	FatalCount
0	January	255
1	February	180
2	March	239
3	April	243
4	May	248

As you can see the index are not date-time, so we need to convert it to time series. We also made FatalCount float to make manipulation more convenient.

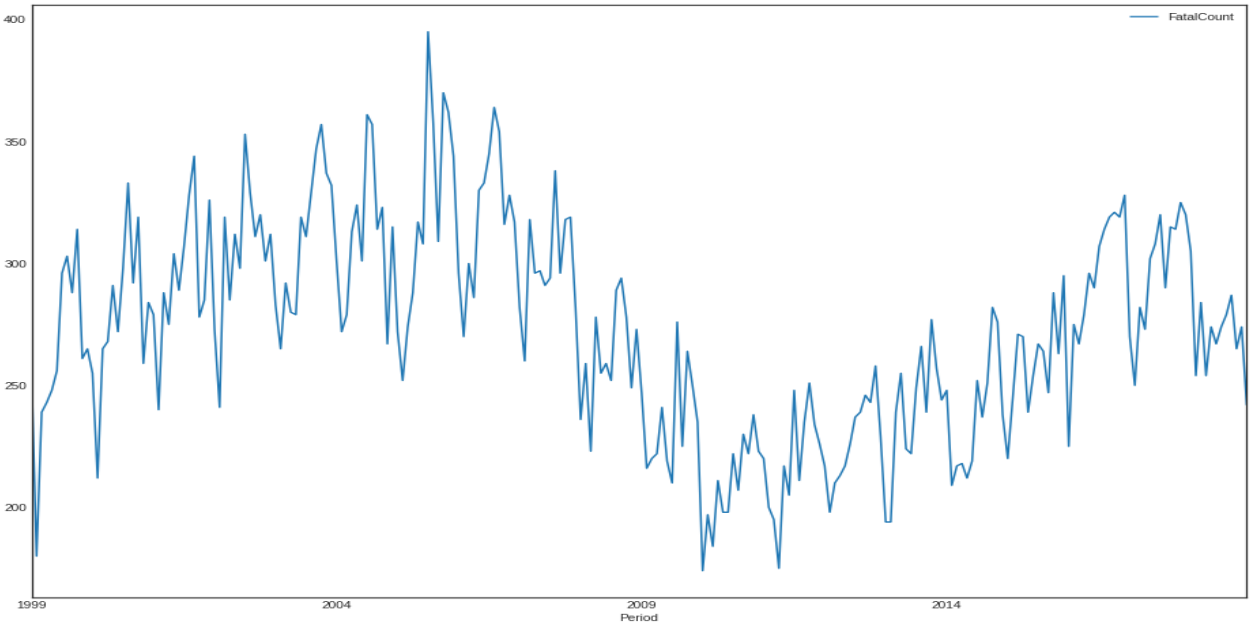
```
[ ]  ts.index
```

```
DatetimeIndex(['1999-01-01', '1999-02-01', '1999-03-01', '1999-04-01',
               '1999-05-01', '1999-06-01', '1999-07-01', '1999-08-01',
               '1999-09-01', '1999-10-01',
               ...
               '2018-03-01', '2018-04-01', '2018-05-01', '2018-06-01',
               '2018-07-01', '2018-08-01', '2018-09-01', '2018-10-01',
               '2018-11-01', '2018-12-01'],
              dtype='datetime64[ns]', name='Period', length=240, freq=None)
```

Note, 1999-01-01 represents the whole January in 1999.

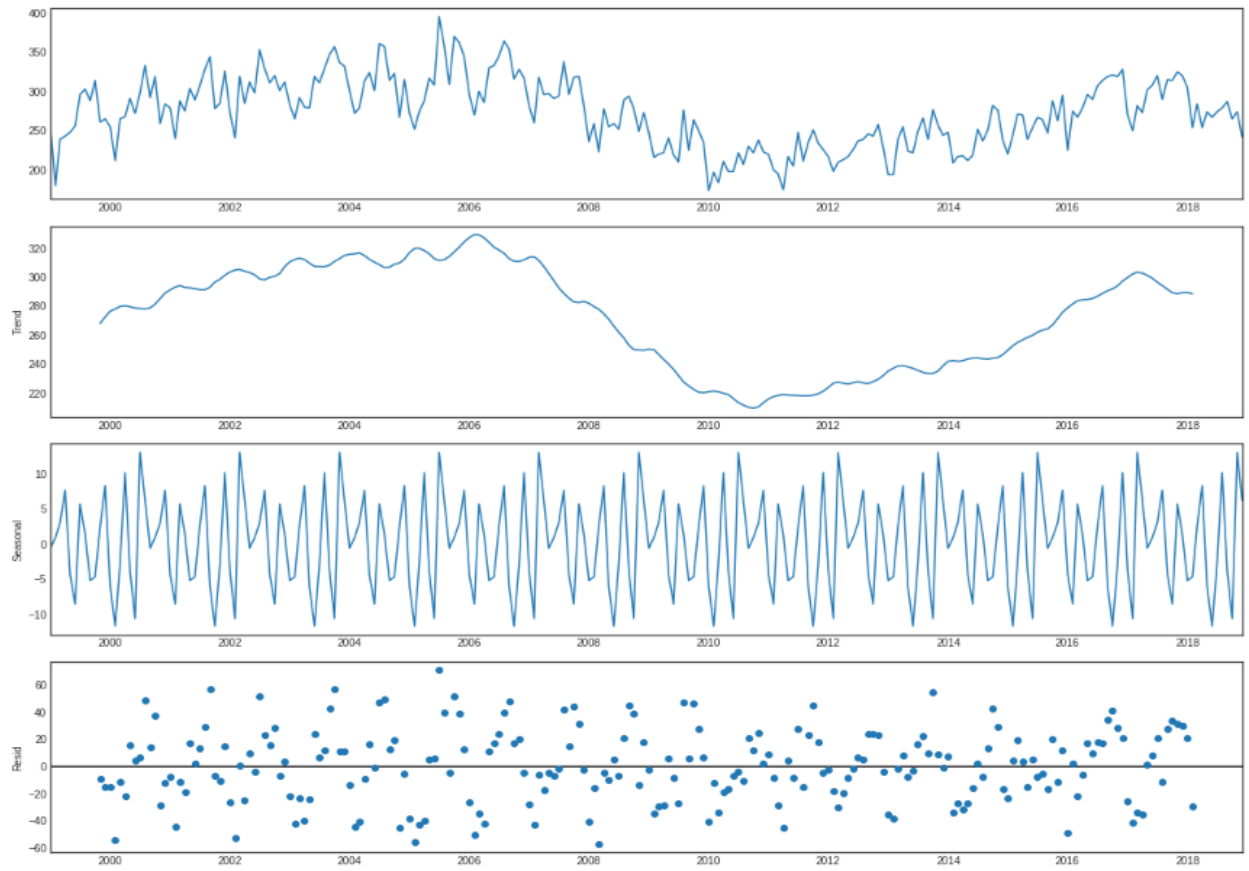
```
ts.plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6414b618d0>
```



As we can see, there are some seasonal and cyclic patterns according to the definition of https://robjhyndman.com/hyndsight/cyclicts/.

Let us check the decomposition now.



It is a good decomposition. As we can see there is clear seasonality, randomized residuals, and a clear cycle.

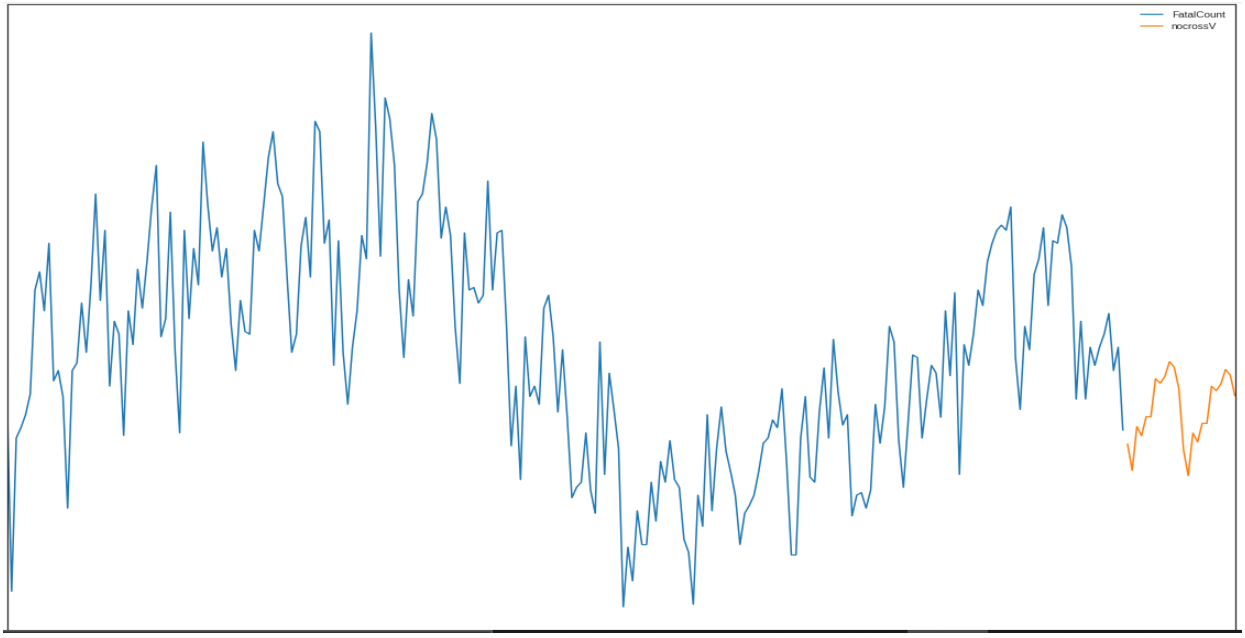
7.2: Build the model

```
 > Is the data stationary ?
Test statistic = -1.484
P-value = 0.541
Critical values :
        1%: -3.459884913337196 - The data is not stationary with 99% confidence
        5%: -2.8745310704320794 - The data is not stationary with 95% confidence
        10%: -2.573693840082908 - The data is not stationary with 90% confidence
```

The data is not stationary, so we need to consider ‘d’ in our ARIMA model.

```
#Try to use a cool Api to fit the model.
#We set stationary  as False, we use adf test. We set the number of period in each season as 12, and we use ocsb for seasonal unit root test.
m1=pm.auto_arima(ts,start_p=0,start_q=0,max_p=3,max_q=3,m=12,start_P=0,start_Q=0,max_P=3,max_Q=3,stationary=False,test='adf',seasonal_test='ocsb',
                 error_action='ignore',information_criterion='aic',suppress_warnings=True,stepwise=True)
```

This api seems running fast with an algorithm called stepwise, we can increase the range of p,d,q.



After plotting the graph, we feel that this graph is very weird. It for sure captures seasonality but we think the trend is not convincing. We feel like the predicted value should have a decrease trend. We are aware that the model does not really do the test-train split. And thus, the aic was based on the entire train set. So, we decide to implement a special train test split for time series called TimeSeriesSplit. It is more like cross-validation.

7.3: Build the model with cross-validation

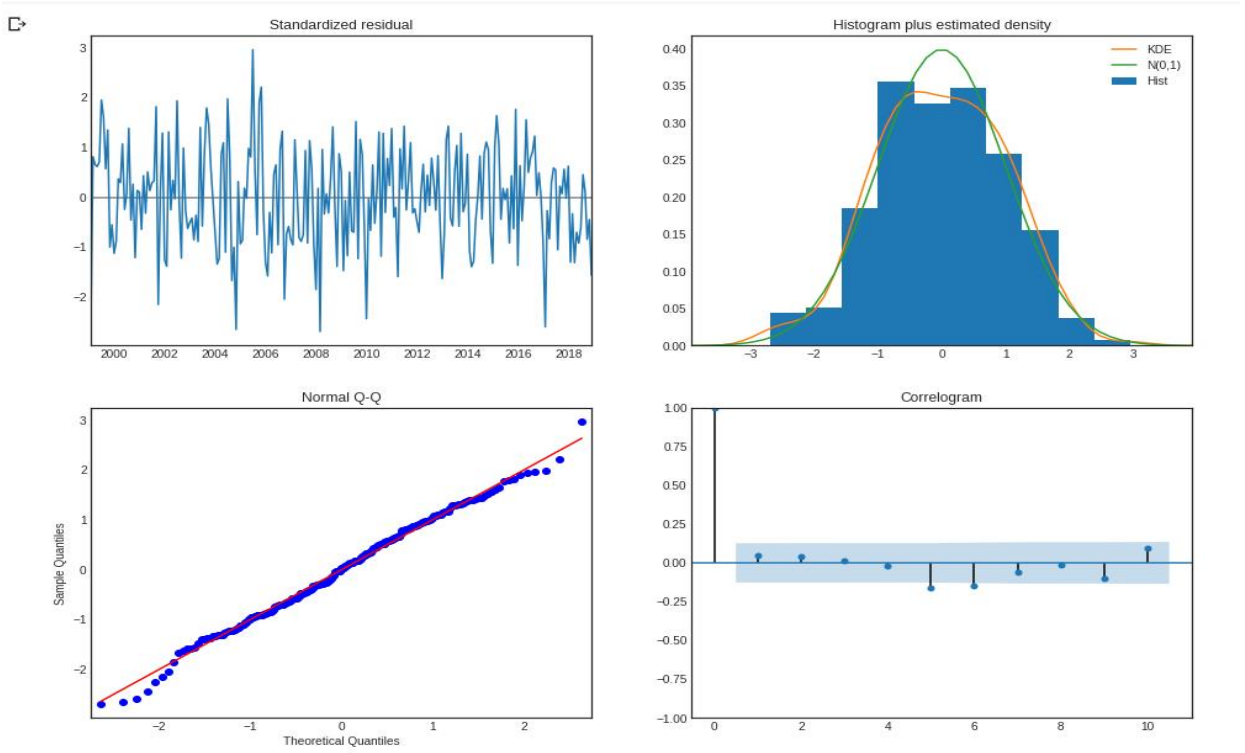
We first split the data into 5 kinds.

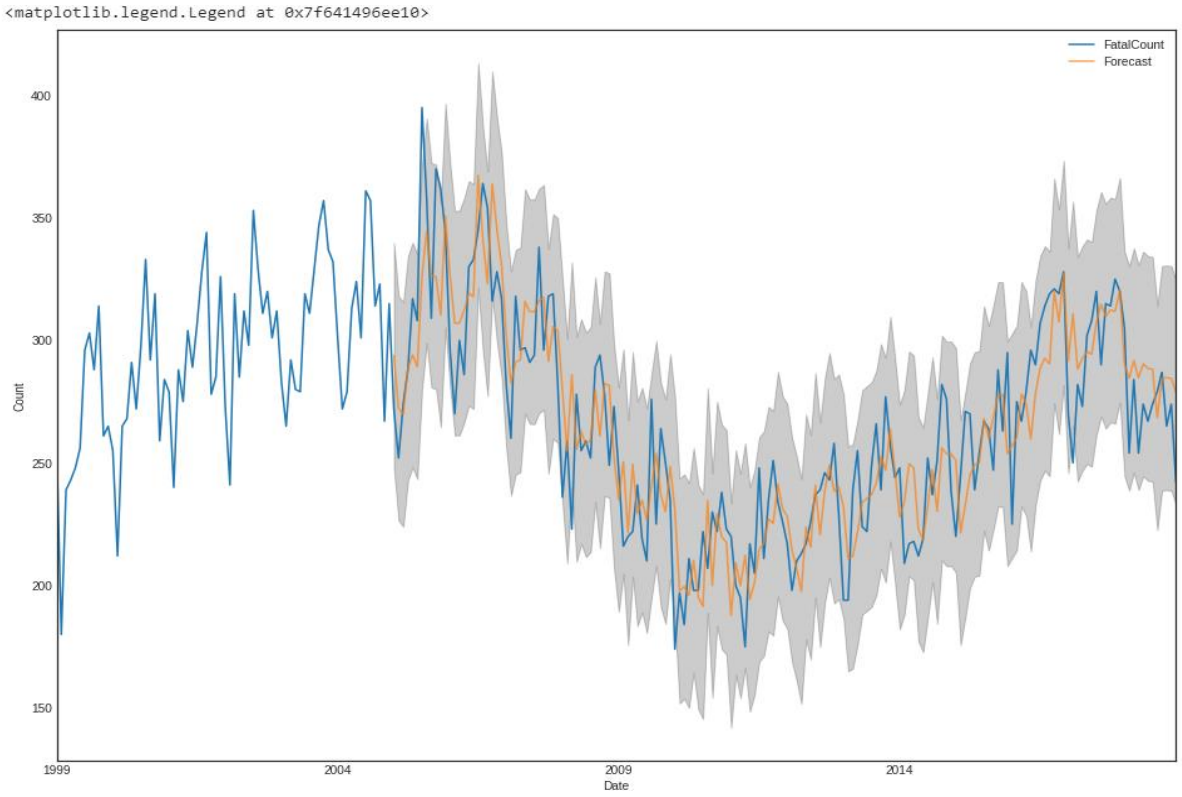
We loop through each combination of hyperparameters and in each loop, we train the model using each of the 5 kinds of splitting.

Remember we use the training group to train and we use the test group to test. We get the aic for the test group and we find the average of aic for each combination of hyperparameters.

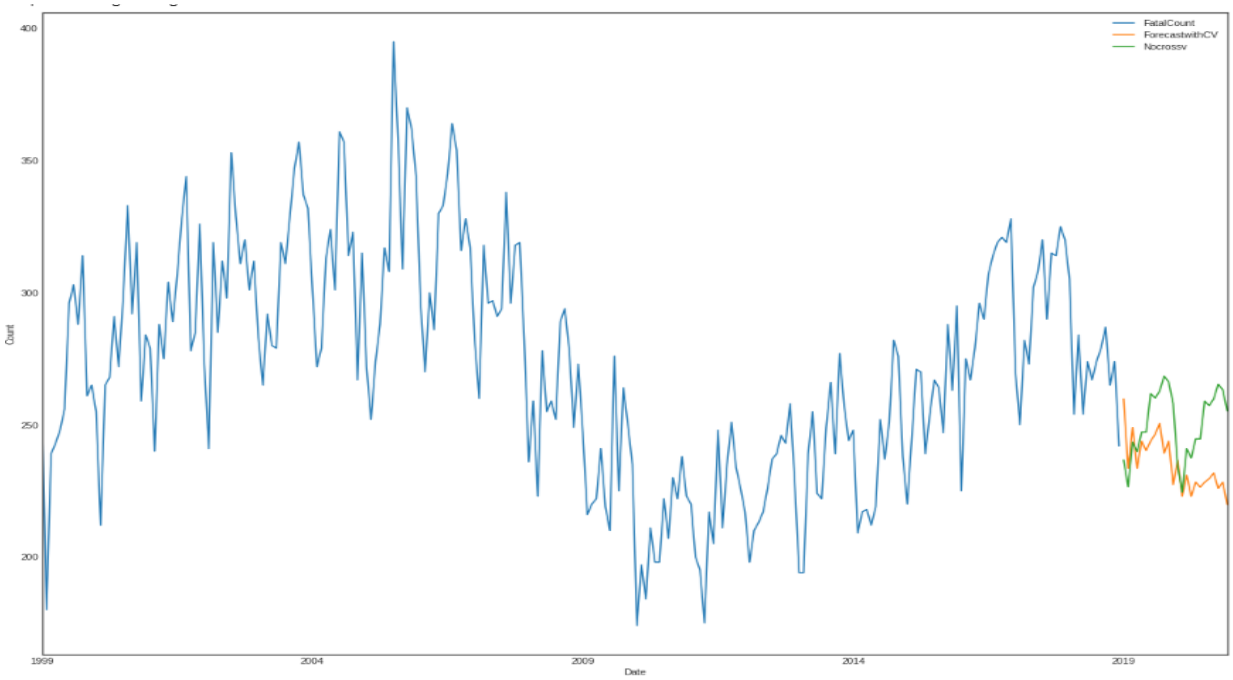
The calculation of aic is complicated. After fitting the model, we need to find the 5 arrays of coefficients and to count non-zero ones. And this is the k for the aic.

```
y = ts['FatalCount']
n=5
tss = TimeSeriesSplit(n_splits = n)
warnings.filterwarnings("ignore") # specify to ignore warning messages
model_output = pd.DataFrame(columns = ['parameters', 'averageaicForagroupofParameter'])
a = "bob"
b = "bob"
c = 1.0
for param in pdq:
  for param_seasonal in seasonal_pdq:
    averageaicForagroupofParameter=0
    for train_index, test_index in tss.split(y):
      y_train, y_test = y.iloc[train_index], y.iloc[test_index]

      mod = sm.tsa.statespace.SARIMAX(y.iloc[train_index],
                                      order=param,
                                      seasonal_order=param_seasonal,
                                  enforce_stationarity=True,
                                  initialization='approximate_diffuse'

                                      )
      results = mod.fit()

      pred_uc = results.get_forecast(steps=len(y.iloc[test_index]))
      pree=pred_uc.predicted_mean
      count=0
      for z in range(0,len(results.polynomial_ar)):
          if results.polynomial_ar[z]!=0:
            count+=1
      for p in range(0,len(results.polynomial_ma)):
          if results.polynomial_ma[p]!=0:
            count+=1
      for q in range(0,len(results.polynomial_seasonal_ar)):
          if results.polynomial_seasonal_ar[q]!=0:
            count+=1
      for h in range(0,len(results.polynomial_seasonal_ma)):
          if results.polynomial_seasonal_ma[h]!=0:
            count+=1
      for j in range(0,len(results.polynomial_trend)):
          if results.polynomial_trend[j]!=0:
            count+=1

      rss = sum((pree.values-y.iloc[test_index].values)**2)
      aiceach= len(y.iloc[test_index])*np.log(rss/len(y.iloc[test_index])) + 2*(count)
      averageaicForagroupofParameter+=aiceach

    a = param
    b = param_seasonal
    c = averageaicForagroupofParameter/n
    model_output = model_output.append({'parameters': 'ARIMA{}x{}12'.format(a, b),
                                        'averageaicForagroupofParameter': c}, ignore_index = True)
```





As we can see, the model fits the data well. And with cross-validation, the risk of overfitting decreased.



As we can see, the orange, which is the new model, is much better to behave a decreasing trend.

7.4: Predict:

a) Fatal car accident in each month without and with COVID-19

The following is the fatal car accident prediction without COVID-19 in California.

```
2019-01-01    259.84
2019-02-01    233.53
2019-03-01    249.00
2019-04-01    233.53
2019-05-01    243.85
2019-06-01    240.23
2019-07-01    243.85
2019-08-01    246.42
2019-09-01    250.55
2019-10-01    239.20
2019-11-01    243.85
2019-12-01    227.34
2020-01-01    236.54
2020-02-01    222.97
2020-03-01    230.95
2020-04-01    222.97
2020-05-01    228.29
2020-06-01    226.43
2020-07-01    228.29
2020-08-01    229.62
2020-09-01    231.75
2020-10-01    225.89
2020-11-01    228.29
2020-12-01    219.77
Freq: MS, dtype: float64
```

In the previous section, we have seen that the average VMT has been dropped by 68.76% in the California State. We used 0.6876 to multiply it and get the prediction for fatal car accident with COVID-19 in California.

```
[29] #The fatal car accident prediction with COV 19 in California
     pred_uc.predicted_mean[15:24]*(1-0.6876)
```

```
2020-04-01   69.65
2020-05-01   71.32
2020-06-01   70.74
2020-07-01   71.32
2020-08-01   71.73
2020-09-01   72.40
2020-10-01   70.57
2020-11-01   71.32
2020-12-01   68.66
Freq: MS, dtype: float64
```

b) Cumulated people who are saved because of the decreased miles travel

```
[30] #The cumulative amount of people saved from fatal car accident in California
     predictPeoplesaved(11)
```

```
2020-04-01 00:00:00  153.0
2020-05-01 00:00:00  310.0
2020-06-01 00:00:00  466.0
2020-07-01 00:00:00  623.0
2020-08-01 00:00:00  781.0
2020-09-01 00:00:00  940.0
2020-10-01 00:00:00  1096.0
2020-11-01 00:00:00  1252.0
2020-12-01 00:00:00  1404.0
2021-01-01 00:00:00  1558.0
2021-02-01 00:00:00  1708.0
```

Hence, from April to Dec in 2020, 1404 people in California who might die because of a car accident are saved because COVID-19 makes travel by car less.

8. Part IV: Improvement

In the previous section, we assumed that there is a directly proportional relationship between the VMT and fatal car accidents. In this section, we tried to find out whether our assumption is correct.

1. Load the data

observation_date	VMT
1999-01-01	2622075
1999-02-01	2626393
1999-03-01	2632934
1999-04-01	2636009
1999-05-01	2638896
...	...
2018-08-01	3230883
2018-09-01	3233105
2018-10-01	3235599
2018-11-01	3237913
2018-12-01	3240325

2. Try Correlation

```
[29]
      corr, _ = pearsonr(vvm3['VMT'], ts['FatalCount'])
      print('Pearsons correlation: %.3f' % corr)

   Pearsons correlation: 0.024
```

Unfortunately, the overall assumption of having a proportional relationship between VMT and car accident fails.

```
Pearsons correlation: 0.569
Pearsons correlation: 0.683
Pearsons correlation: 0.531
Pearsons correlation: 0.547
Pearsons correlation: 0.858
Pearsons correlation: 0.334
Pearsons correlation: 0.800
Pearsons correlation: 0.052
Pearsons correlation: 0.584
Pearsons correlation: -0.479
Pearsons correlation: -0.057
Pearsons correlation: 0.747
Pearsons correlation: -0.722
Pearsons correlation: 0.824
Pearsons correlation: 0.551
Pearsons correlation: 0.668
Pearsons correlation: 0.636
Pearsons correlation: 0.919
Pearsons correlation: 0.856
Pearsons correlation: -0.338
```

We tried to find out the correlation for each year from 1999 to 2018, and it also fails...

But!... We still have time series regression!

3. Time series regression

The general plan is, to calculate predicted people saved from fatal car accidents after VMT drop in California.

Formula we used:

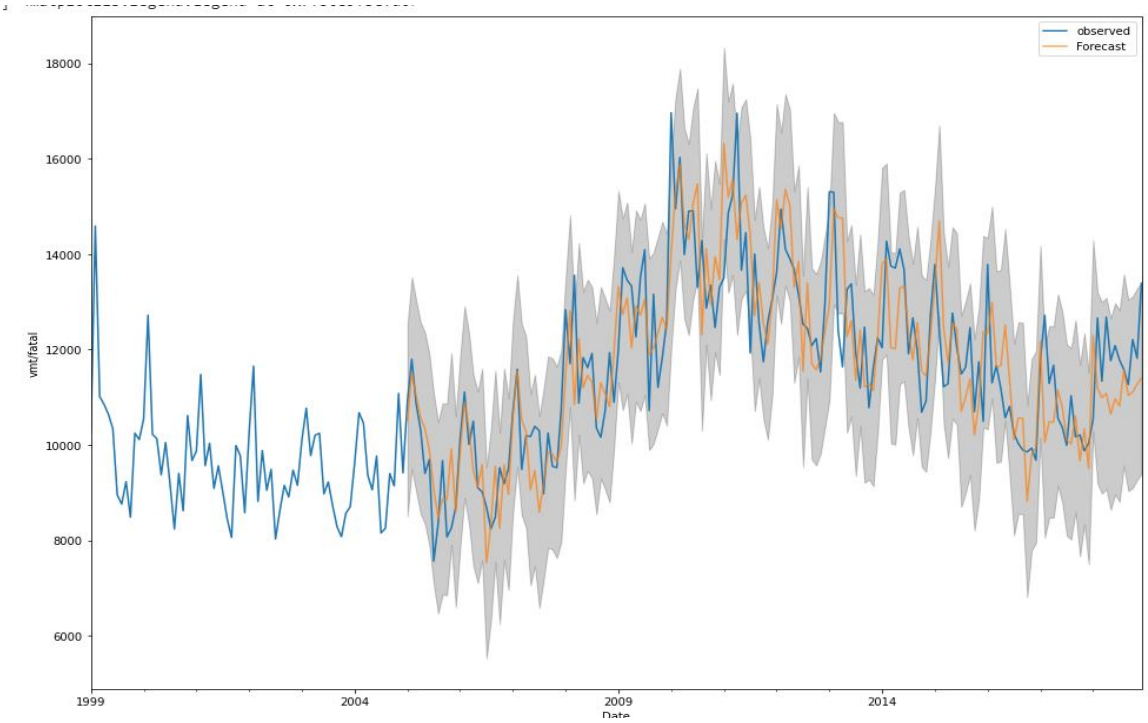
predicted people saved from fatal car accidents after miles drop in California= (predicted (fatal accidents in California/ VMT national)) * ((predicted VMT national) * 0.68) = (1/(predicted(national VMT /fatal accident in California death)))*((predicted VMT national)*0.68).

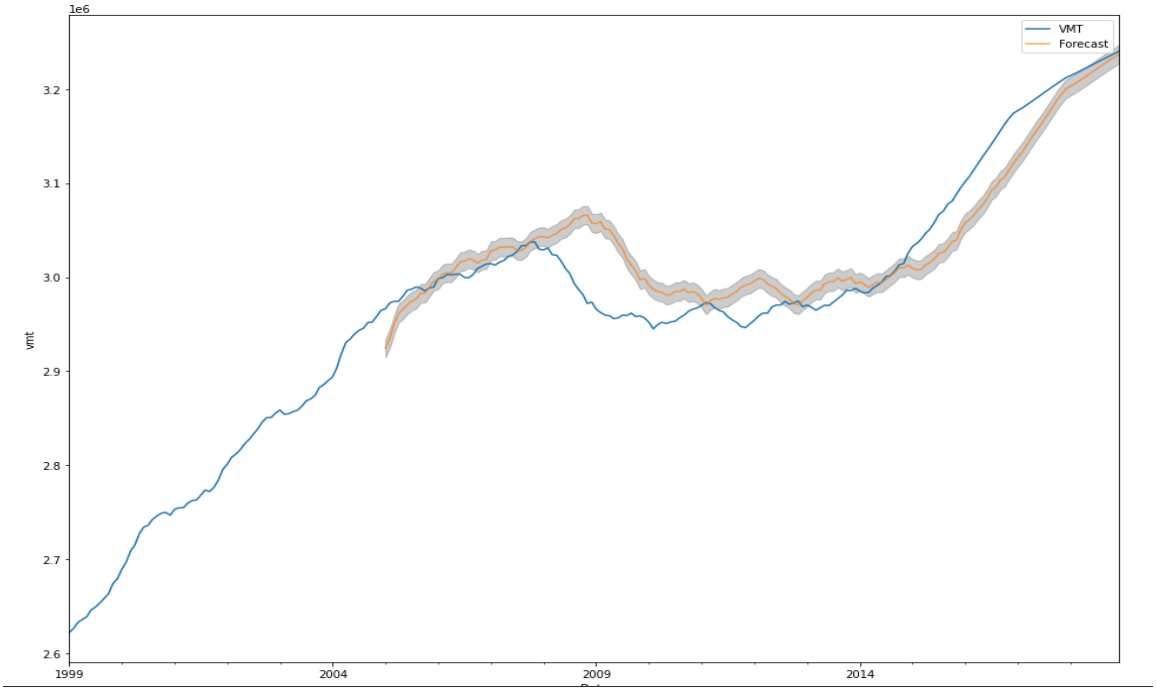
We predicted VMT /death instead of death/ VMT because death/ VMT is too small and causing a bug in Arima.

Another note: Do not feel surprised seeing (in California/national) *(national). The new assumption is that California VMT behaves the same way as the national. So, the denominator canceled out eventually.

Assume this assumption is true, now we are ignoring our previous assumption that VMT and fatal accidents have a directly proportional relationship. The reason is that we directly predict the rate, and then cancel the unit and get the number of fatal car accidents.

We used cross validation again with the same code.






```
2019-01-01    13183.136282
2019-02-01    14777.772307
2019-03-01    13401.981896
2019-04-01    14314.100038
2019-05-01    13310.070731
2019-06-01    13395.842525
2019-07-01    13060.636859
2019-08-01    13421.824218
2019-09-01    12857.986897
2019-10-01    13395.103146
2019-11-01    13027.919451
2019-12-01    13964.077539
2020-01-01    14077.078996
2020-02-01    15911.145707
2020-03-01    14556.316239
2020-04-01    15665.229133
2020-05-01    14700.638511
2020-06-01    14891.027189
2020-07-01    14571.805478
2020-08-01    14678.294411
2020-09-01    14226.880110
2020-10-01    14948.148004
2020-11-01    14569.475932
2020-12-01    15793.221150
Freq: MS, dtype: float64
```

This is the rate: (predicted (VMT national/ fatal accident in California)).

```
2019-01-01    3239342.0
2019-02-01    3241423.0
2019-03-01    3243744.0
2019-04-01    3245964.0
2019-05-01    3248372.0
2019-06-01    3250687.0
2019-07-01    3253066.0
2019-08-01    3255455.0
2019-09-01    3257634.0
2019-10-01    3260089.0
2019-11-01    3262365.0
2019-12-01    3264737.0
2020-01-01    3263665.0
2020-02-01    3265720.0
2020-03-01    3268017.0
2020-04-01    3270212.0
2020-05-01    3272595.0
2020-06-01    3274885.0
2020-07-01    3277239.0
2020-08-01    3279603.0
2020-09-01    3281757.0
2020-10-01    3284187.0
2020-11-01    3286439.0
2020-12-01    3288786.0
Freq: MS, dtype: float64
```

This is the predicted national VMT without COVID-19.

And the following is the final formula:

```
[46] peoplesavedeachmonth=(1/predrateyy)*(prevmtt*0.68)
```

The following is the number of people saved each month in California because of staying at home by COVID-19.

```
peoplesavedeachmonth

2019-01-01    167.088664
2019-02-01    149.154248
2019-03-01    164.583575
2019-04-01    154.201495
2019-05-01    165.956537
2019-06-01    165.011446
2019-07-01    169.370389
2019-08-01    164.933579
2019-09-01    172.281342
2019-10-01    165.497822
2019-11-01    170.281089
2019-12-01    158.980886
2020-01-01    157.652868
2020-02-01    139.568188
2020-03-01    152.665780
2020-04-01    141.954128
2020-05-01    151.378773
2020-06-01    149.547897
2020-07-01    152.933876
2020-08-01    151.933875
2020-09-01    156.857652
2020-10-01    149.399599
2020-11-01    153.387699
2020-12-01    141.603456
Freq: MS, dtype: float64
```

```
[48] sum=0
     for a in range(15,len(peoplesavedeachmonth)):
       sum+=peoplesavedeachmonth[a]
     print(sum)
```

```
1348.9969556678907
```

We sum up from April to Dec of 2020 again, surprisingly, the cumulative people saved is nearly equal to our previous assumption of the direct proportional relationship which is 1404.

The improvement for this is to find the real VMT data for California state.

9. Conclusion

COVID-19 is probably the most widely spread virus the world ever has witnessed. In this study, we tried to find out the current global situation and related statistics. Since the shelter-in-place order implemented to slow down the spread of the virus, it prominently affected all kinds of transportation. We explained how COVID has affected the vehicle miles traveled in the state of California and its consequent effects on fatal car accidents.

The interesting finding of this study is, in California state after shelter-in-place implemented, there was a decrease of 68.76% in the average VMT in the initial span of 18 days. Also, we found a significant decrease in the average VMT for Trinity, Santa Clara, and Los Angeles counties.

Another interesting finding is, because of the reduction in traffic the number of road accidents has reduced, and this could lead to saving at least 1404 people (from April to December 2020) in California who otherwise would have died in a car accident.

In the improvement part of this study, we tried to find out whether our assumption of considering a directly proportional relationship between VMT and fatal car accidents were true, and we found that our assumption was wrong. There is no proportional relationship between the two. Surprisingly, after performing the time-series regression, we got a nearly equal number of cumulative people saved (1349) as our previous assumption (1404).

One further improvement could be to find out the real VMT data for California state which we were not able to find out because of limited time.

10. References

Yogesh Agrawal, Jan 21, 2019, Hypothesis testing in Machine learning using Python, retrieved from:

https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

Katie Dowd, April 30, 2020, California's Modoc County plans to end shelter-in-place order this week, retrieved from:

https://www.sfgate.com/coronavirus/article/california-counties-shelter-in-place-orders-15237258.php

Chengzhi Zhao, Mar 30, Data Visualization for Novel Coronavirus (COVID-19) in Jupyter Notebook with Plotly, retrieved from:

https://medium.com/@chengzhizhao/data-visualization-for-novel-corona-virus-covid-19-in-jupyter-notebook-with-plotly-c8d3300265b4

Kevin Sablan, April 23, 2020, when stay-at-home orders are set to end in all 58 California counties, retrieved from:

https://www.ocregister.com/2020/04/23/when-stay-at-home-orders-are-set-to-end-in-all-58-california-counties/

Stackoverflow.com, SARIMAX python np.linalg.linalg.LinAlgError: LU decomposition error, retrieved from:

https://stackoverflow.com/questions/54136280/sarimax-python-np-linalg-linalg-linalgerror-lu-decomposition-error

pyramid. arima.auto_arima, retrieved from:

http://alkaline-ml.com/pmdarima/0.9.0/modules/generated/pyramid.arima.auto_arima.html

Federal Information Processing System (FIPS) Codes for States and Counties, retrieved from:

https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt

Getting Started with Plotly in Python, retrieved from:

https://plotly.com/python/getting-started/?utm_source=mailchimp-jan-2015&utm_medium=email&utm_campaign=generalemail-jan2015&utm_term=bubble-chart