IBM DATA SCIENCE CAPSTONE PROJECT

YAXUAN XUE

JUNE 29, 2020

1. Introduction

I have been longing to visit Orlando, Florida for a while. As we all know, Orlando has been named the most visited destination in the United States mainly due to its theme parks. And among those attractions there, I want to explore the Universal Studios particularly with my friend, who is a big fan of Harry Potter. However, due to this unexpected pandemic, we got no other choices but to stay at home to be safe and healthy this summer. But we have faith that everything will be alright as long as we stand together and fight together. Thus, I would like to use this project as a guidance for our future trip to Orlando.

In this project, I will use Foursquare to choose our ideal hotels that are close to our destination—Universal Studios, and at the same time has a relatively good rating. And then, I will make a list of nearby restaurants that are located walking-distance from our selected hotel. Finally, because I am curious about when this pandemic can be under control, at least within the United States, so that we can eventually start our trip. I will build an ARIMA model using the active cases data of United States and perform forecasting.

Audience:

Although this project is mainly designed for personal uses, but anyone who wants to visit the amusement park or the city, especially who won't have a vehicle at the time of their visiting and expects to stay as close as possible to the attraction may also find the report helpful. Besides, this report can be interesting and may bring some hope to people who had plans that got delayed due to the COVID-19 and who still intends to implement plans as soon as the epidemic is over.

2. Data preparation

To perform this analysis, we need to know:

1) address of the Universal Studios Florida:

6000 Universal Blvd, Orlando, FL

2) active COVID-19 cases in the United States

source: https://www.coronanet-project.org/download

The COVID-19 data collected from the coronanet-project website includes worldwide data and policies concerning the COVID-19. As for the United States of America, the data covers from Jan.22nd, 2020 to Jun. 3rd,2020.

After collecting the data from the website, I have performed data cleaning, removing duplicates, and selecting the contents that are associated to my project. Given only total positive cases, total deaths, and total recovered cases, I have calculated the total active cases using:

Total Active Cases = Total Positive Cases - Total Deaths - Total Recovered

3. Methodology

3.1 Accommodation

3.1.1 Getting latitude and longitude coordinates

Using Geopy Client, a python library that helps us getting the latitude and longitude of the Universal Studios.

```
address = '6000 Universal Blvd, Orlando, FL'
geolocator = Nominatim(user_agent="foursquare_agent")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print(latitude, longitude)
```

```
28.471878949999997 -81.47121064527349
```

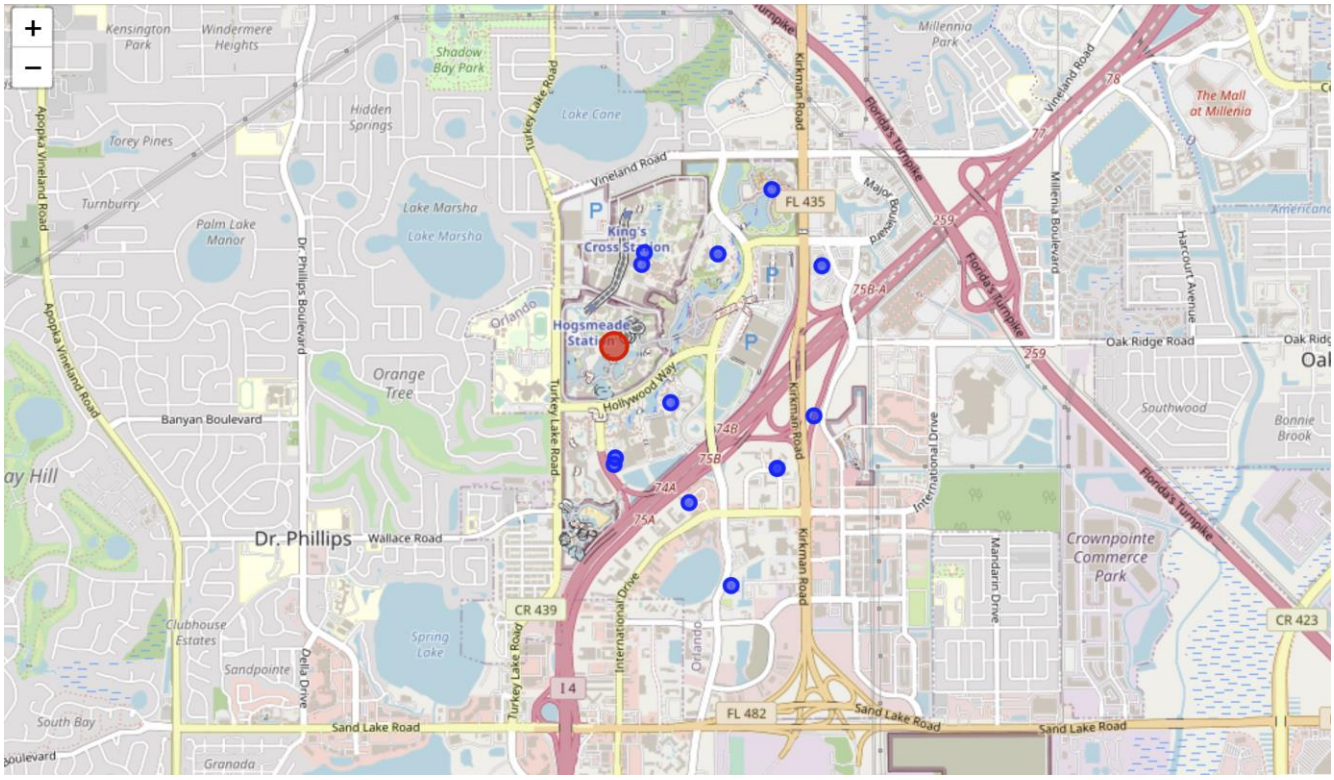
3.1.2 Searching nearby hotels

Considering that we will not have a vehicle during the trip, we would like to stay as close as possible to the park. Here, I used Foursquare API to search nearby venues that are within 1.5 km radius from the studios and have been categorized as “hotel”. I got a dataset in JSON format that contains 13 nearby hotels. Then I transformed venues into a data frame as shown below. For comparison purposes, I have sorted the data frame according to its distance from the park in ascending order.

	name	categories	address	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	country	formattedAddress
26	Loews Royal Pacific Resort	Hotel	6300 Hollywood Way	28.468639	-81.467503	[{'label': 'display', 'lat': 28.46863856649586...	511	32819	US	Orlando	FL	United States	[6300 Hollywood Way, Orlando, FL 32819, United...
10	Pricilla Hotel	Hotel	Universal Studios	28.476559	-81.469380	[{'label': 'display', 'lat': 28.47655939385056...	550	32819	US	Orlando	FL	United States	[Universal Studios (New York), Orlando, FL 328...
15	Empire Hotel	Hotel	Universal Studios	28.477204	-81.469171	[{'label': 'display', 'lat': 28.47720439346763...	625	32819	US	Orlando	FL	United States	[Universal Studios, Orlando, FL 32819, United ...
17	Lowes Adventura Hotel At Universal Studios	Hotel	6725 Adventure Way	28.465466	-81.471115	[{'label': 'display', 'lat': 28.46546554565429...	713	32819	US	Orlando	FL	United States	[6725 Adventure Way, Orlando, FL 32819, United...
7	Universal's Aventura Hotel	Hotel	6725 Adventure Way	28.465076	-81.471188	[{'label': 'display', 'lat': 28.46507579036141...	757	32819	US	Orlando	FL	United States	[6725 Adventure Way, Orlando, FL 32819, United...
18	Club 7 at the Hard Rock Hotel	Hotel	NaN	28.477155	-81.464361	[{'label': 'display', 'lat': 28.47715509313276...	891	NaN	US	NaN	Florida	United States	[Florida, United States]
2	Orlando Continental Plaza Hotel	Hotel	6825 Visitors Cir	28.462913	-81.466296	[{'label': 'display', 'lat': 28.46291303914135...	1107	32819	US	Orlando	FL	United States	[6825 Visitors Cir, Orlando, FL 32819, United ...
20	The Lord Hotel	Hotel	5859 American Way	28.464859	-81.460503	[{'label': 'display', 'lat': 28.46485900878906...	1307	32819	US	Orlando	FL	United States	[5859 American Way, Orlando, FL 32819, United ...
21	The Blue Palm Hotel	Hotel	5859 American Way	28.464859	-81.460503	[{'label': 'entrance', 'lat': 28.464809, 'lng'...	1307	32819	US	Orlando	FL	United States	[5859 American Way, Orlando, FL 32819, United ...
24	Loewe Hotel	Hotel	NaN	28.467865	-81.458107	[{'label': 'display', 'lat': 28.46786511718431...	1357	32819	US	Orlando	FL	United States	[Orlando, FL 32819, United States]
29	Holiday Inn & Suites Across From Universal Orl...	Hotel	5905 S Kirkman Rd	28.476468	-81.457653	[{'label': 'display', 'lat': 28.476468, 'lng':...	1421	32819	US	Orlando	FL	United States	[5905 S Kirkman Rd (at Major Blvd), Orlando, F...
1	Loews Portofino Bay Hotel at Universal Orlando	Hotel	5601 Universal Blvd	28.480846	-81.460843	[{'label': 'display', 'lat': 28.48084591771423...	1423	32819	US	Orlando	FL	United States	[5601 Universal Blvd, Orlando, FL 32819, Unite...
6	Best Western Orlando Gateway Hotel	Hotel	7299 Universal Blvd	28.458103	-81.463500	[{'label': 'display', 'lat': 28.4581031, 'lng'...	1709	32819	US	Orlando	FL	United States	[7299 Universal Blvd, Orlando, FL 32819, Unite...

3.1.3 Visualization

I visualized my data via folium in a Leaflet map.



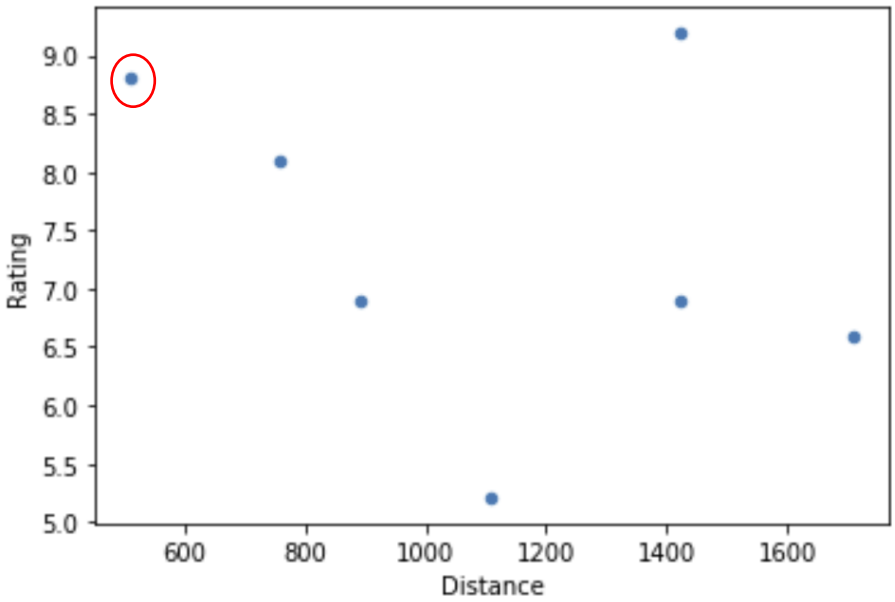
3.1.4 Comparing the ratings

With venue_id of these hotels, I made venue detail calls on Foursquare, receiving the following ratings.

	Name	Rating	Distance
0	Loews Portofino Bay Hotel at Universal Orlando	9.2	1423
1	Loews Royal Pacific Resort	8.8	511
2	Universal's Aventura Hotel	8.1	757
3	Club 7 at the Hard Rock Hotel	6.9	891
4	Holiday Inn & Suites Across From Universal Orl...	6.9	1421
5	Best Western Orlando Gateway Hotel	6.6	1709
6	Orlando Continental Plaza Hotel	5.2	1107

Not all hotels that appeared in my searching have ratings, so I have just collected data for hotels that do have ratings. Also, as for those hotels that were not rated, I will ignore them because I do not want to risk having bad experience there that may potentially ruin my vacation.

For easier comparison, I drew a scatter plot of distance against rating. Theoretically, without considering other factors, the closer the point is to the upper left corner of the plot, the better choice. Because that means the hotel enjoys a short distance from the park and an excellent rating.



From the plot, the point in red circle which represents ***Loews Royal Pacific Resort*** seems to be the most ideal hotel that meets our requirements.

3.1.5 Exploring restaurants near the hotel

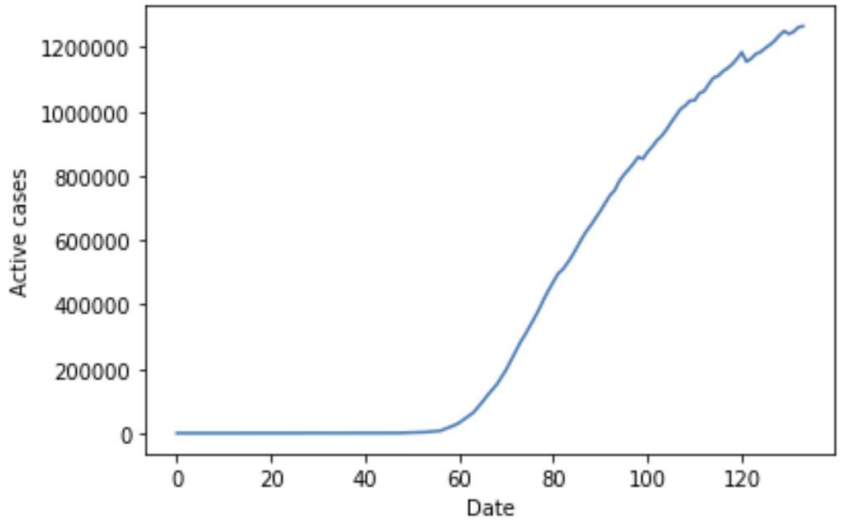
Having decided which hotel to live in, I searched again using Foursquare API to find restaurants that are near our hotel.

	name	categories	address	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	country	formattedAddress
10	Vivo Italian Kitchen	Italian Restaurant	6000 Universal Blvd	28.473063	-81.465647	[{'label': 'display', 'lat': 28.47306265781803...	524	32819	US	Orlando	FL	United States	[6000 Universal Blvd (at Universal CityWalk), ...
7	Bob Marley: A Tribute To Freedom	Caribbean Restaurant	6000 Universal Blvd	28.472875	-81.465123	[{'label': 'display', 'lat': 28.47287499588204...	525	32819	US	Orlando	FL	United States	[6000 Universal Blvd (at Universal Citywalk), ...
9	Circus McGurkus Cafe Stoo-pendous	Diner	Seuss Landing	28.473163	-81.469803	[{'label': 'display', 'lat': 28.47316347107087...	551	32819	US	Orlando	FL	United States	[Seuss Landing (at Islands of Adventure), Orla...
13	Thunder Falls Terrace	BBQ Joint	Jurassic Park	28.469765	-81.473314	[{'label': 'display', 'lat': 28.46976489750437...	582	32819	US	Orlando	FL	United States	[Jurassic Park (Islands of Adventure), Orlando...
14	Mythos	Greek Restaurant	The Lost Continent	28.472640	-81.471439	[{'label': 'display', 'lat': 28.47264026289861...	588	32819	US	Orlando	FL	United States	[The Lost Continent (at Islands of Adventure),...
8	Oshio Korean BBQ	Korean Restaurant	6800 Visitors Cir	28.463251	-81.465498	[{'label': 'display', 'lat': 28.46325100820970...	631	32819	US	Orlando	FL	United States	[6800 Visitors Cir, Orlando, FL 32819, United ...
4	Las Brisas Restaurant	Food	6233 International Dr	28.462166	-81.467087	[{'label': 'entrance', 'lat': 28.461708, 'lng'...	721	32819	US	Orlando	FL	United States	[6233 International Dr, Orlando, FL 32819, Uni...
0	Via Brasil Brazilian Restaurant	Brazilian Restaurant	6315 International Dr	28.461344	-81.467443	[{'label': 'display', 'lat': 28.46134391842936...	812	32819	US	Orlando	FL	United States	[6315 International Dr, Orlando, FL 32819, Uni...
5	Kohola Reef Restaurant	BBQ Joint	Wave Village	28.462716	-81.472433	[{'label': 'display', 'lat': 28.46271588727033...	816	32819	US	Orlando	FL	United States	[Wave Village (Universal's Volcano Bay), Orlan...
11	Tavistock Restaurants	Building	7680 Universal Blvd	28.461786	-81.463047	[{'label': 'display', 'lat': 28.46178633363962...	878	32819	US	Orlando	FL	United States	[7680 Universal Blvd, Orlando, FL 32819, Unite...
1	Louie's Italian Restaurant	Italian Restaurant	New York	28.477348	-81.469196	[{'label': 'display', 'lat': 28.47734786754995...	983	32819	US	Orlando	FL	United States	[New York (Universal Studios Florida), Orlando...
12	The Palm Orlando	Steakhouse	5800 Universal Blvd	28.477117	-81.464620	[{'label': 'display', 'lat': 28.47711656455560...	984	32819	US	Orlando	FL	United States	[5800 Universal Blvd, Orlando, FL 32819, Unite...
2	Thai Silk Restaurant	Thai Restaurant	6803 S Kirkman Rd	28.463811	-81.457950	[{'label': 'display', 'lat': 28.46381058901147...	1078	32819	US	Orlando	FL	United States	[6803 S Kirkman Rd (entrance on Grand National...
6	Tabla Restaurant	Indian Restaurant	5847 Grand National Dr	28.472176	-81.455630	[{'label': 'display', 'lat': 28.4721764, 'lng'...	1226	32819	US	Orlando	FL	United States	[5847 Grand National Dr (Kirkman Road), Orland...
3	La Palma Restaurant	Mexican Restaurant	5545 International Dr	28.463289	-81.455559	[{'label': 'display', 'lat': 28.46328926086425...	1311	32819	US	Orlando	FL	United States	[5545 International Dr, Orlando, FL 32819, Uni...

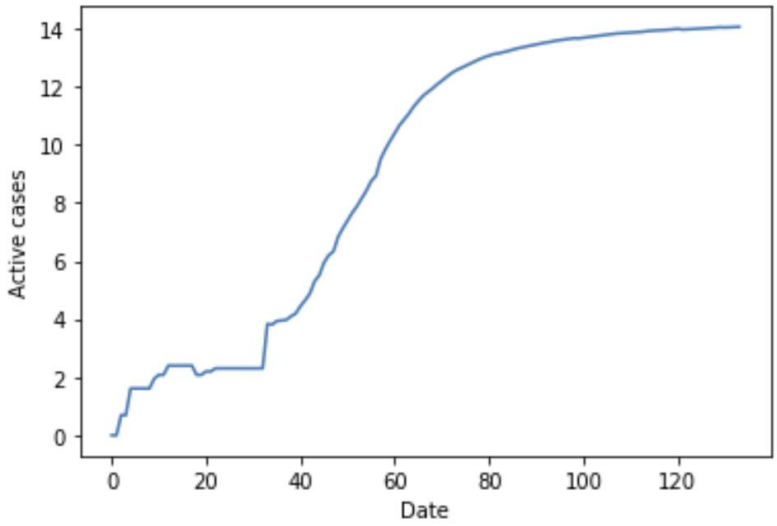
3.2 COVID-19 active cases study

3.2.1 ADF test

To predict when this pandemic will be under control and when it will be safe to travel around the country again, in this part, I am establishing an ARIMA model of total active cases in the United States. To start with, I used matplotlib module in Python to draw the plot of the series and statsmodels module to run ADF test of daily total active cases to further see whether this time series is stationary or not.



From the plot above, it can be clearly seen that the series has a trend and it is not stationary. Thus, I took log of the original series and drew a new plot.



We can see that the log of active cases series has been flattened a little bit but still has an obvious trend. We do an ADF test to prove that the log of active cases series is not stationary either.

```
ADF Statistic: -1.645296370982438
p-value: 0.4594946462270914
Critical Values:
        1%: -3.4816817173418295
        5%: -2.8840418343195267
        10%: -2.578770059171598
```

The p-value is greater than 0.05, it is not stationary. To address this problem, we need to take first difference of log(active cases) and perform ADF test.

```
ADF Statistic: -4.192485632973042
p-value: 0.000678466536687841
Critical Values:
        1%: -3.4816817173418295
        5%: -2.8840418343195267
        10%: -2.578770059171598
```

We can see that the first difference of log(active cases) is stationary. We can then use this series to run ARMA regression.

3.2.2 ARIMA Regression

After choosing the best order for the ARIMA model, I have decided to stick with (2,0,1) and add time as a regressor in the model. In this section, I used ARIMA library from the statsmodels module to run the regression. Here are my regression results:

ARMA Model Results

Dep. Variable:	log_active	No. Observations:	133
Model:	ARMA(2, 1)	Log Likelihood	41.984
Method:	css-mle	S.D. of innovations	0.176
Date:	Wed, 01 Jul 2020	AIC	-71.968
Time:	00:37:58	BIC	-54.626
Sample:	01-23-2020 - 06-03-2020	HQIC	-64.921

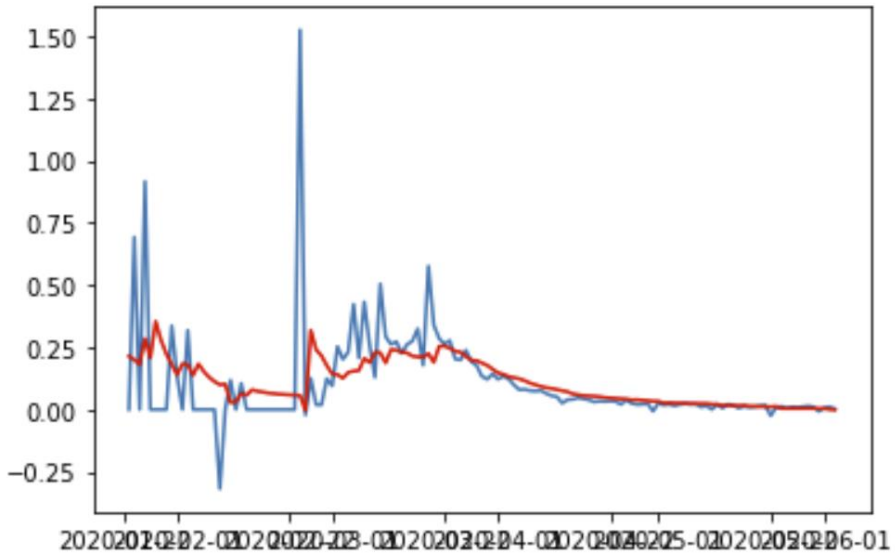
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2158	0.072	3.003	0.003	0.075	0.357
time	-0.0016	0.001	-1.772	0.076	-0.003	0.000
ar.L1.log_active	0.7034	0.127	5.532	0.000	0.454	0.953
ar.L2.log_active	0.2011	0.095	2.115	0.034	0.015	0.387
ma.L1.log_active	-0.7422	0.103	-7.239	0.000	-0.943	-0.541

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0851	+0.0000j	1.0851	0.0000
AR.2	-4.5828	+0.0000j	4.5828	0.5000
MA.1	1.3474	+0.0000j	1.3474	0.0000

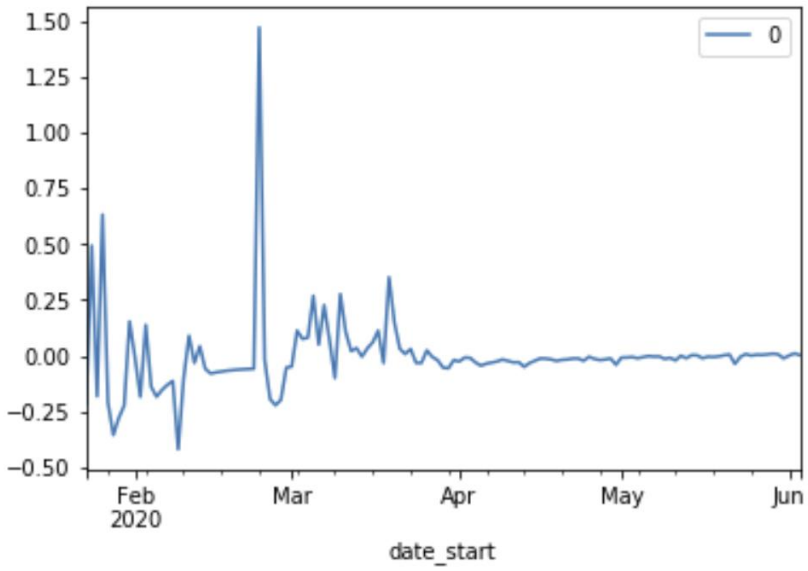
If we set the confidence level to 10%, all regressors significantly influence the independent variable.

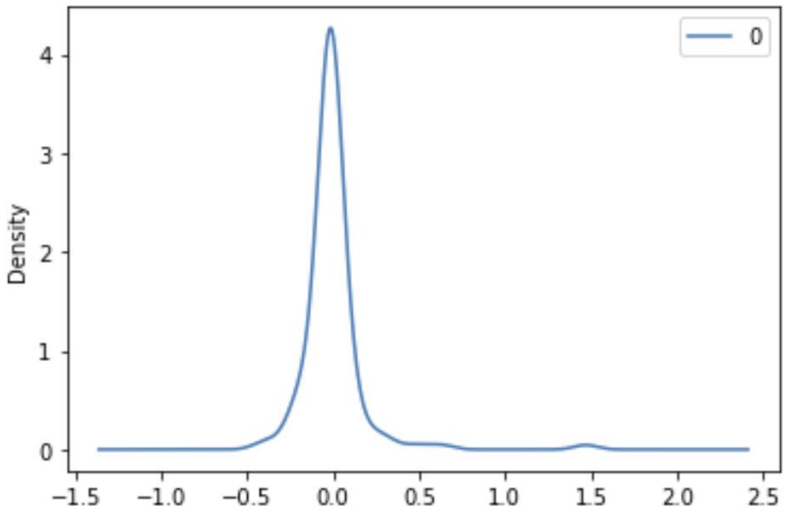
3.2.3 Fitting the model



From the plot above, the blue line is the first differenced log of active cases series, and the red line represents fitted values. Especially for the latest values, where shows less volatility, the model fitted pretty well.

In order to test whether this ARIMA model is qualified, we still need to check the residuals of the model. If the model is a good fit, its residuals should be close to white noise.





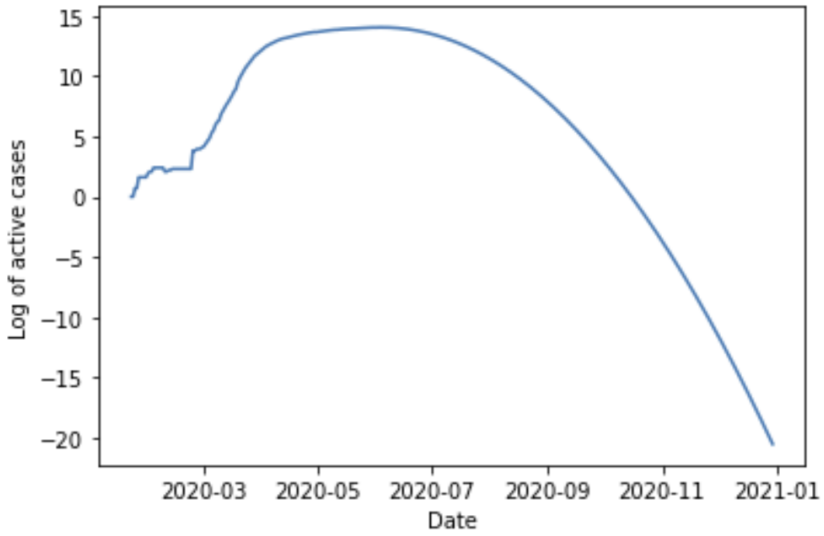
From these two plots, the mean of the residuals is really close to zero and there is no obvious pattern of the residuals, I tend to believe it is white noise. Thus, the model is a qualified model.

3.2.4 Forecasting

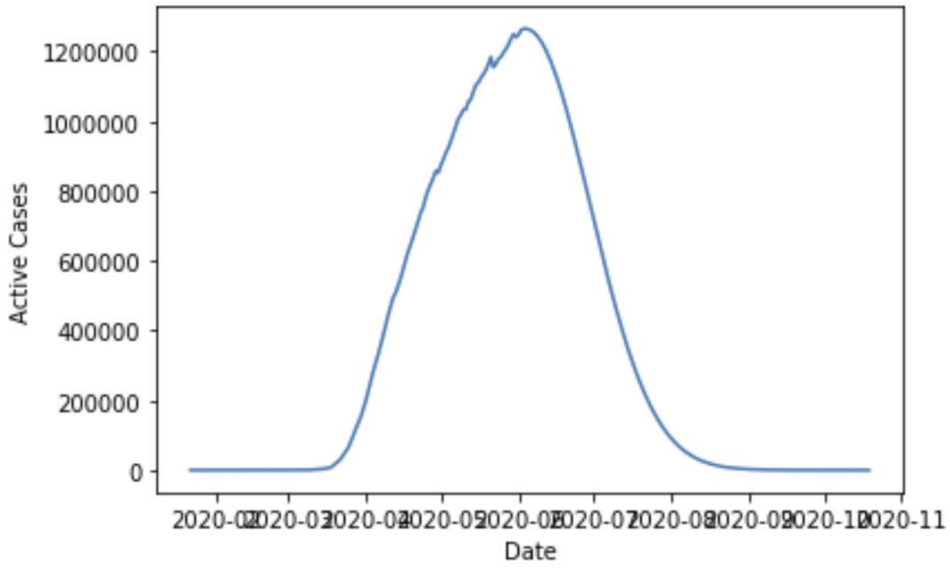
Since everything is well prepared, I am finally able to conduct forecasting. I have predicted values from June 4th,2020 to December 31st,2020 to see if it's possible to defeat the coronavirus within this year. Here are some forecasting values of the first difference of log (active cases).

```
2020-06-04    0.000197
2020-06-05   -0.001049
2020-06-06   -0.002565
2020-06-07   -0.004035
2020-06-08   -0.005526
                 ...
2020-12-27   -0.326478
2020-12-28   -0.328073
2020-12-29   -0.329668
2020-12-30   -0.331263
2020-12-31   -0.332858
```

Because log of a variable can be interpreted as growth rate of the variable, then first difference of log of the variable can be understand as the change of the growth rate at one point of time compared to growth at the point before this point. Therefore, when the first difference of log (active cases) is negative, it means the growth rate (log of active cases) begins to decrease, which means the total active cases are still increasing but at a slower growth rate.



With the first difference of log(active cases) predictions, we can calculate the expected log of active cases as the graph shown above and its corresponding total active cases as shown below.



4. Results

4.1 Where to live and where to eat

Thanks to Foursquare, I found our ideal hotel for the vacation. ***Loews Royal Pacific Resort*** seems to be the most suitable hotel for my friend and I to stay in. And around the hotel, there are 15 restaurants that serve cuisines from all over the world.

4.2 When is it safe to travel

If nothing unexpected happens and everything goes smoothly as right now, from the ARIMA model I built, the active cases of the U.S. should come to zero in late October.

date_start	Active
2020-10-05	8
2020-10-06	7
2020-10-07	6
2020-10-08	5
2020-10-09	4
2020-10-10	3
2020-10-11	3
2020-10-12	2
2020-10-13	2
2020-10-14	1
2020-10-15	1
2020-10-16	1
2020-10-17	1
2020-10-18	1
2020-10-19	0

5. Discussion

The limitation of this project is that I have only considered the effects of time and active cases itself when I was studying the COVID-19 active cases. Particularly, policies such as national/local response efforts and physical distancing requirements are vital to the spread of the epidemic, but these issues are hard to quantify, thus I did not study those effects.

6. Conclusion

By using data collected in Foursquare, now I have some ideas about my future trip to Orlando, such as which hotel to stay in so that we can be extremely close to the Universal Studios and enjoy excellent services, and which restaurants to go to that has cuisines from all over the world. I believe this report can also be a good small guidance to people who have the same interests and plans as I do.

With the help of several Python libraries, I have built an ARIMA model to predict when the active cases in the United States can be zero and when it may be safe to travel around the country again. If everything goes as I expected, according to the model, the situation should be alleviated by the end of October.

Link to Github:

https://github.com/IrisXue96/Coursera_Capstone/blob/master/IBM-Capstone-FinalProject.ipynb