Time Series Analytics Project

CIS 4680 - Advanced Data Analytics  
*by: Calvin Truong*

**U.S. COVID-19 Confirmed Cases**

**Introduction**

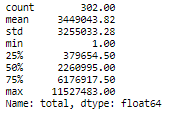
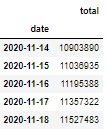
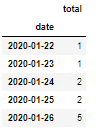
Imagine if tomorrow was your last day to socialize with all your friends and you had to wear a mask wherever you go in public. All the restaurants, gyms, theatres, and barbershops have closed, and everyone had to quarantine themselves at home. Well at the beginning of 2020, an epidemic known as the coronavirus disease or COVID-19 was introduced and broke out at a rapid pace. This virus has successfully spread and infected millions of people around the world and some countries are still currently unable to recover.

This report will be focused on the confirmed cases of COVID-19, specifically in the United States. This report will try to predict the number of cases by the end of the year 2020 assuming that there will be no vaccines available until at least the following year. The report will answer the following questions:

1. Will COVID-19 cases be expected to increase or decrease by the end of the year (Dec 31, 2020)?
2. What is the percentage of COVID-19 cases in the fall season (Sep 22 – Nov 18)? What factors might affect this percentage?
3. Does the previous data of COVID-19 cases affect the number of future cases?

**Data**

To answer the questions, we must first find a dataset that contains an up-to-date total number of COVID-19 cases. While scrolling through the Internet, I was able to find a diverse time series dataset from the *data.world* website. It consists of all the confirmed cases around the world ranging from January 22, 2020 to November 18, 2020. However, I will be focusing on only the number of cases in the United States. So, I have filtered the dataset into a separate excel file so that it contains the following columns: **Country (US only), Date, and Total.** Afterward, I dropped the **Country** column to get rid of the redundancy and it was ready to be imported for the cleaning process.



I used the time series decomposition to determine if the dataset contains any trends, seasonality, and other noise components.

**Background**

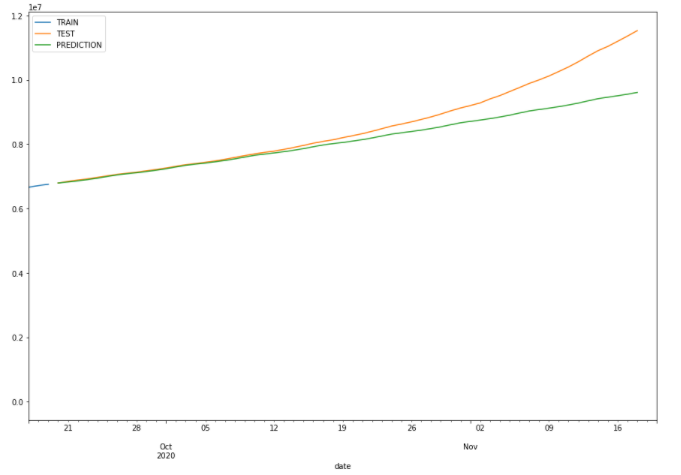
The SARS-CoV-2 disease also known as COVID-19 was declared as a pandemic on February 11, 2020 by the World Health Organization. It first appeared in Wuhan China within bats and made its way to the Asian wet supermarkets. It easily spread throughout China as people came into close contact with one another. Since the infectious disease does a good job hiding along with the common flu and cold, it eventually circulated the rest of the world as people continued traveling and not staying hygienic.

Some symptoms of COVID-19 are fever, shortness of breath, coughs, headaches, sore throats, congestion, fatigue, and inability to taste (CDC.gov). The symptoms of the common flu and COVID-19 are very similar to each other and can be easily mistaken without proper testing. COVID-19 spreads through the respiratory system and can infect people when they inhale the virus particles through their nose, mouth, and lungs. However, the main reason why COVID-19 is dangerous because people contracted with the disease become asymptomatic and will not show any symptoms until after 2 to 14 days. As soon as people realize they are positive for COVID-19, it will already be too late. They may have passed on the virus to other people and the same process repeats.

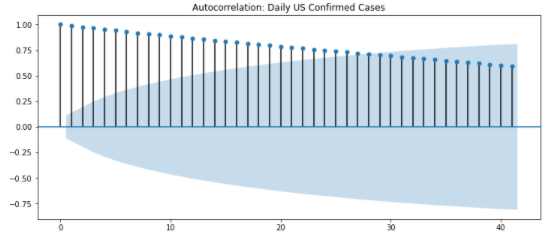
COVID-19 has drastically affected the daily lives of people. For most cases, many negative impacts came out of this pandemic such as increased unemployment rates, family financial distress, company shutdowns, decrease in economic growth, and death of loved ones. Thousands of people are getting tested every single day and hospitals have been filled with COVID-19 patients. And because the disease is so new and always mutating, researchers are having a difficult time developing a vaccination to cure COVID-19. In addition, due to the long process of testing, it is expected for the vaccine to be fully developed after 2020. So as of now, government officials highly encourage people to wear masks, stay at least 6 feet away from others to avoid the spread, and wash their hands frequently for 20 seconds. People should also quarantine themselves and only travel if necessary. Doing all of this, for the time being, will hopefully decrease the total number of confirmed cases in the United States until the vaccination can be distributed to the people.

**Methodology and Data Analysis**

For this report, I used forecasting with the Holt-Winters method to train (80%) and test (20%) of the dataset to predict the estimated total number of confirmed cases in the US that will occur in the future. I fitted the model using a seasonal period of 7 days and evaluated the model against the test set. The test values range from “9/20/2020” to “11/18/2020”. This will allow me to collect *predicted test* values to compare with the *original* *test* values. I also used the Holt-Winters method to show a visual representation of the test prediction and analyzed the differences between the two. Based on the chart, the predicted test values underestimate the original test values because the model did not expect to increase as much as it did during that time range. I proceeded with the same method to predict the confirmed cases for the next 43 days ranging from “11/19/2020” to “12/31/2020” and created the final model. (y-axis: 1.0\*e^7 = 10,000,000 cases)

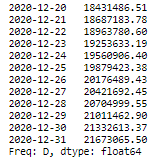
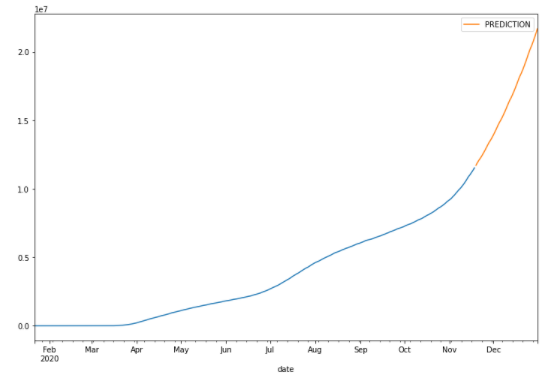


Another method I used was the Exponentially Weighted Moving Average (EWMA) to find the correlation trend of the data. I started with a 7-day simple moving average (SMA) and a 14-day SMA and compared them with the *original total* values. After that, I implemented the autocorrelation function (ACF) to show the change in correlation over time. I had to incorporate another data frame with all the *original total* values. Then I tested the biased, unbiased, autocorrelation, and partial autocorrelation and plotted the number of lags in the correlation.

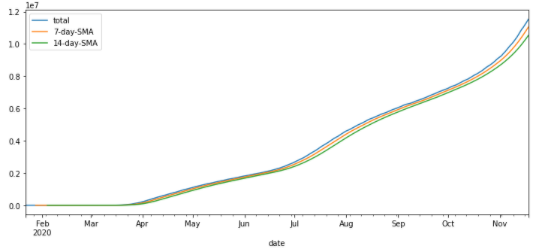


**Findings and Results**

After implementing the Holt-Winters Method, I was able to forecast the future of COVID-19 cases from November 19, 2020 to December 31, 2020. Without any data relating to these time periods, I predicted that by the end of the year, there will be an increase of about 21,673,066 total confirmed cases in the United States. This is an 85% increase since November 18, 2020 if a vaccine has not been distributed in time. (y-axis: 1.0\*e^7 = 10,000,000 cases)

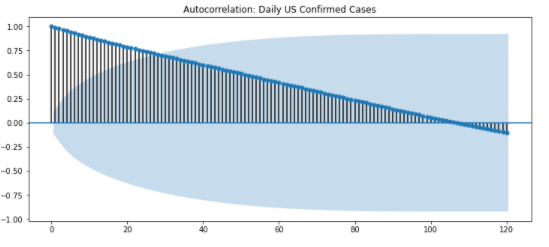


The EWMA tells us the statistic moving mean of the data. We can conclude that using a weekly (7 days) period and a two-week (14 days) period show very similar positive trends as the original total number of confirmed cases.



The percentage of COVID-19 cases in the fall season is 67.47%.

The previous data of COVID-19 cases does affect the number of future cases. According to the ACF plots. The autocorrelation of daily US confirmed cases is gradually decreasing. By 110 lags, the autocorrelation reaches 0, and lags after 110 becomes negative.



**Discussion of Results and Impacts**

From the time-series analysis, we can expect that there will be a massive increase of about 21 million COVID-19 cases by the end of the year. This is because as of now, there is no way to eliminate COVID-19 completely since vaccination has not been established by the government. The prediction of an 85% increase in confirmed cases after November 18 shows how quickly the pandemic is spreading in the United States within a short period of time. Looking at the EWMA plots, there is a positive trend and seasonality for the number of cases. This tells us that the current data weighs as much as the previous data and will continue to rise in the future. Knowing the expected number of COVID-19 cases by the end of the year will benefit and remind the United States community to be more cautious about spreading the disease.

In the time range of September 22 to November 18, the percentage of COVID-19 confirmed cases is 67.47%. An increase of more than half the number of cases in the previous months is affected due to the following factors: seasonality, government policies, and the asymptomatic spread from young adults. As the weather gets colder, people are more likely to catch the flu and become sick. Unfortunately, since COVID-19 has very similar symptoms as the flu, it is difficult to determine whether the person has contracted COVID-19 unless the individual gets tested. Also, since the fall season has many upcoming holidays such as Halloween, Thanksgiving, and Christmas, there are reports of people attending social gatherings to celebrate and therefore have a higher chance of spreading the virus around. During the time range, the government has reopened the United States to boost economic growth, allowing restaurants, gyms, and barbershops to remain open which attracts many people to come together and spread COVID-19. Also, young adults who are asymptomatic can easily spread the virus to older adults who have a weaker immunity system since they do not realize that they have COVID-19 until it is already too late.

According to the ACF model, there is a straight gradual decline and a large number of lags before the ACF values drop off. This concludes that the previous data of COVID-19 cases does affect the number of future cases since the autocorrelation is decreasing. This makes sense because COVID-19 was able to travel all around the world at a fast rate and as more people become infected, it is a lot harder to stop the spread.

**Conclusion**

The COVID-19 pandemic became an existential threat to people’s daily lives in 2020. Due to its complex and non-linear escalation, COVID-19 seems like it is not going away anytime soon. Without a vaccine, confirmed cases will continue to increase and more deaths will occur daily. However, this pandemic has helped us in numerous ways. It has helped people become very hygienic in terms of washing their hands for 20 seconds and avoid the spread of germs by cleaning public areas more often. The innovation of technology has rapidly increased to best fit the new lifestyles. It has also affected the government system of being mindful for the middle and poor class people by funding small businesses and providing unemployment benefits.

But most importantly, COVID-19 has made people realize the importance of building genuine relationships between friends and family. We have all been extremely busy “living our life” that we sometimes forget our loved ones. With COVID-19 going on, life has slowed us down, but this allows us more time to stay connected with people virtually. As of now, it is greatly encouraged to stay quarantined until the spread slows, and vaccination is fully distributed and effective.

**Cited Source**

COVID-19 Time Series Data - dataset by shad. (2020, March 25). Retrieved December 02, 2020, from https://data.world/shad/covid-19-time-series-data

Coronavirus (COVID-19) frequently asked questions. (n.d.). Retrieved December 02, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/faq.html

Thomas, V. (2020, November 24). 8 Positive Impacts of COVID-19 Pandemic. Retrieved December 02, 2020, from https://surveysparrow.com/blog/positive-impacts-to-look-forward-to-post-covid-19-pandemic/