COVID-19 COMPARISON BETWEEN UNITED STATES AND SOUTH KOREA

Saul Arguelles, Raul Mancilla & Maria Patricia Santos

Department of Information Systems, California State University

Los Angeles

**1.Abstract**

COVID-19 has impacted everyone’s lives this past year, so we think it is both interesting and helpful for everyone to see the comparison especially with the growing cases we’ve been having in the United States. In this report, We will see the analyzation of data using Azure ML. This is where we will see the model evaluation and data comparison of total deaths/ cases between the US and South Korea. We will also explain how we have managed to gather / extract the data that we need for this project in Azure

On the second portion, we will see the graph comparison of the confirmed, recovered and death cases between US/ South Korea by using Elastic Search. In addition, we will also show the data comparison in Geo Mapping through Elasticsearch. Through these data anaylzation of datasets, we were able to see a bigger picture as to how the data can be beneficial in order to know what countries and organizations can do in order to beat the pandemic. Furthermore, through using applications such as Azure and Elastic Search, it made it more feasible to get the data in timely matter, especially for having Big Data as your source of information.

**2. Introduction**

Since the COVID-19 outbreak started and evolved into a pandemic, the entire world relied on data coming from organizations to understand and analyze the information. These data gathered are more essential than ever as these are the main basis to determine vast decisions amongst nations. For instance, the government are heavily relying on data to properly distribute funds, supplies and more depending on where it is needed the most. Another instance are health care organizations, where the information is necessary to understand what better know their patients needs in demands, as well as to know what equipment, medicine and exact care is needed.

In this report, we will tackle on the comparison of COVID-19 cases between South Korea and the United States. The two countries have been chosen as the researchers found it very interesting to have two data from both ends of the world. In addition, the researchers wanted to see if there are vast differences amongst the data that were able to gather and how the data reflects on the current conditions of each country and whether there are similarities and differences on the numbers. Through this research, we will also analyze and examine what both countries have done that may have reflect the data that they have, and we will be able to determine what procedures have worked well to flatten the curve.

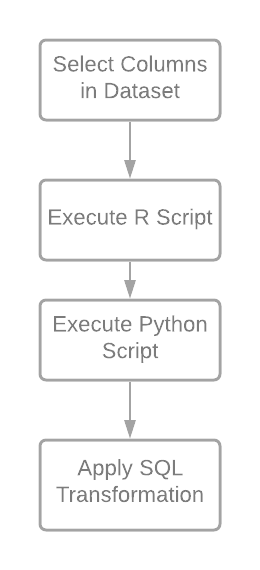
3. Related Work

There is similar work related to our research that could be beneficial. One article talks about the fatality rate related to COVID-19 in areas such as China, Germany, Italy, South Korea, Spain, and the United States. In addition, they have gathered data on the cumulative number of diagnosed infections and deaths that are age specific.[[1]](#footnote-1) Another article that somehow correlates to our research analyzes the data by using the testing data and explained through the daily variation in testing. This research used New York City and Los Angeles data sets that also records data by episode date. Their report reflects COVID- 19 incidence and mortality data to provide information to the public. [[2]](#footnote-2)

What makes our research paper different from the other two is that we used Big Data using Spark on Cloud Computing to analyze the trends and to filter out what is needed within our report. In addition, we have used applications such as Azure ML and Elasticsearch for such to be able to compare datasets and to analyze the data that we are working on.

4.Background

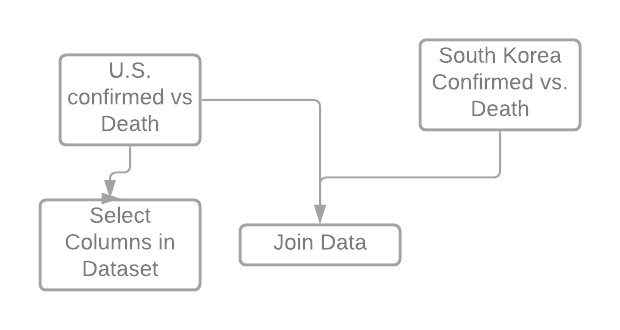
Since there are a lot of data evolving around COVID-19, there are a lot of sources where we can find datasets amongst all countries. The challenge initially was to only find datasets that have United States and South Korea, and by having the same data for comparison. We were able to find data through Community Dashboard project which included the datasets we are looking for which are the geomapping and the data that includes the confirmed, recoverd and dead cases according to the data set.

 In addition, we were able to locate datasets which had a monthly index of number of cases, deaths for both South Korea and the United States through Our Wold in Data, Coronavirus Pandemic Data Explorer. Through these datasets we were able to analyze the data and see a better picture of the data to further explain the trends by using both Elastic Search and Azure ML.

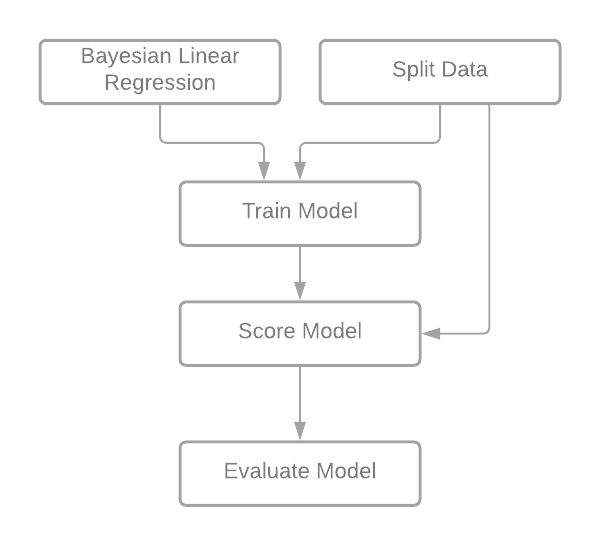
**5. Our Work**

**5.1 Data Analysis using Azure ML**

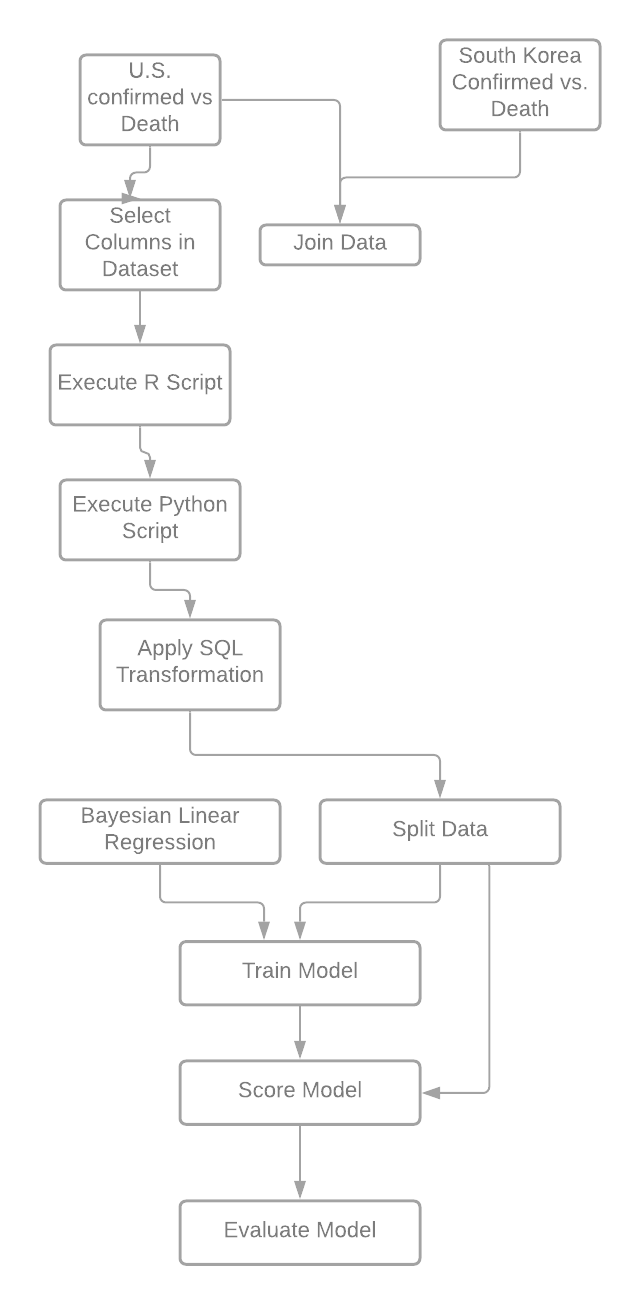
The first Azure experiment we did was to get the dataset of information about Covid-19 from the US and be able to determine how many deaths there were on a certain day from a selected period in time. We needed to focus on just the death column in this dataset so we used the select column in dataset function to display the columns we needed. After we executed R Script and Python Script to Apply the SQL Transformation. [[3]](#footnote-3) When we visualized it, we noticed the only column being displayed was the one we wanted to display the death column and from what it shows the United States number keep increasing over time. We then applied these same steps to the dataset of South Korea and seen that they are far lower number and not increasing over time.



Another Azure experiment we did was to get the dataset of information about Covid-19 from both South Korea and the US and compare between the two the number of deaths related to Covid-19. We needed to focus on the Dead column in both datasets. So, we joined the two using the Join function and selected the correct columns in the column selected. When we visualized the experiment, we were able to see compared to each other. What we saw is that South Korea had a more stable rate of death per day then the U.S where it kept going up especially in the past 3 months.

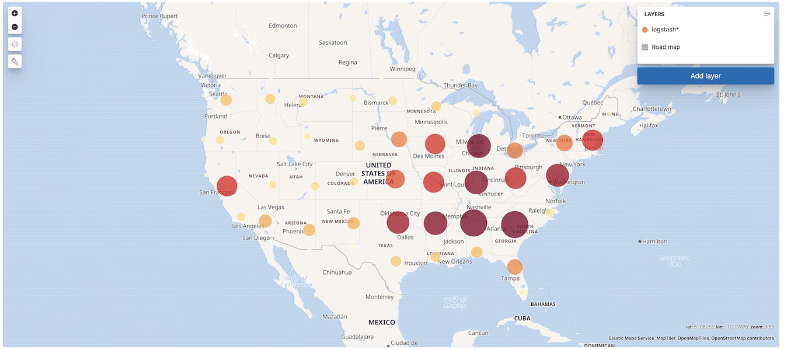
The last experiment we conducted was to try and score the model of both datasets and evaluate the score. We wanted to predict the number of deaths on a given day based on the dataset. We began this by splitting the data and using a train function to score and evaluate the model. When we visualized the results, we saw I believe to be a low number in the relative squared error and the value was 1.09 which is a good indicator of a better predictive results.

Those were the experiments we did using Azure and based on those we can see how much better South Korea is doing now compared to the US.



**5.2 Data Analysis using Elastic Search**

In addition to using Azure for analysis, we have also used Elastic Search to index the data and filter out the information that’s needed. Elastic is an open-source search and analytics engine for all type of data, from maps to geographic visuals to structured data in order to show a visualization in real time. It works by using a variety of sources including logs, system metric and web applications to process data and re create distribution to back up summaries of reports. We tried to show the COVID-19 cases in real time now from confirmed, recovered and death tools to create a result to show the audience. [[4]](#footnote-4)



First, we created an index pattern, then added a layer on the data.Then, we have added the index layer as the confirmed cases in South Korea. Once done, we went to Kibana and visualized the map.

Based on the information on the map, we can see which locations have the surge in cases as well as the surge of deaths in the United States. Here we can also see based on the color scheme (the darker it is, the higher the cases are). In addition, we can also see the surge of cases in South Korea by seeing the data map with cities with the surges.

Based on the analysis on Elastic, the data map is showing that the About 12,000 cases in the start of the year for America. This number skyrockets within these past three months having millions of confirmed cases just for certain states. The top three states with cases include

California 1.35 million, Texas 1.3 and Florida reaching over a million as well.

**6. Conclusion**

In conclusion, through the data that we have seen, South Korea, in the beginning of the data, had a surge of cases in comparison to the United States which had minimal amount of cases. But as time went on, South Korea’s cases although it fluctuates, they have maintained and controlled the total deaths in their country.

In contrast, the United States had minimal cases and have surged through the middle, and it is continuing to rise. Both number of cases and total deaths have increased, and we have not maintained or controlled the amount of deaths.

This data shows that South Korea have been able to control and they have ‘flatten the curve’ while the US has not. Through this data, it could help the United States see and understand that South Korea’s methods are working and they’ll be able to see what could be the most efficient solution in order to be able to flatten the curve in the US.

Through this project, we have learned how helpful both programs are in terms of using big data and being able to filter out what is needed. It also helps us get a better picture when comparing data and both programs made it easier for us to quickly get the data needed. In addition, the visual representations of the data, made it a lot easier for us to analyze the data further.

Being able to get these data easily, the governent and organizations will be able to respond on solutions and response faster which will help us in beating the pandemic that we are currently facing. This will also help healthcare organizations allocate the r In return, it will help us get to the point of going back to normal at a faster pace.

### 7. References

[1] Dudel, Christian, Riffe, Tim, Acosta, Enrique, van Raalte, Alyson, Strozza, Cosmo, & Myrskyla, Mikko. (2020). *Monitoring trends and differences in COVID-19 case-fatality rates using decomposition methods: Contributions of age structure and age-specific fatality.* PloS One, 15(9), e0238904–e0238904. <https://doi.org/10.1371/journal.pone.0238904>

[2] Bergman, Aviv, Sella, Yehonatan, Agre, Peter, & Casadevall, Arturo. (2020). *Oscillations in U.S. COVID-19 Incidence and Mortality Data Reflect Diagnostic and Reporting Factors.* mSystems, 5(4), e00544–20. <https://doi.org/10.1128/mSystems.00544-20>

[3] *The Community COVID-19 Dashboard Project*. (n.d.). Retrieved December 07, 2020, from <https://community.logz.io/covid-19/>

[4] *Coronavirus Pandemic Data Explorer.* (n.d.). Retrieved December 07, 2020, from <https://ourworldindata.org/coronavirus-data-explorer?tab=table>

[5] *Elastic Search* <https://866ef3848e704701ba380af88f52a9a2.us-west1.gcp.cloud.es.io:9243/app/dashboards#/view/a59a9230-3828-11eb-ad0b-1fcae05fc14a?_g=(filters%3A!()%2CrefreshInterval%3A(pause%3A!t%2Cvalue%3A0)%2Ctime%3A(from%3A'2015-05-18T07%3A00%3A00.000Z'%2Cto%3A'2020-05-20T07%3A00%3A00.000Z'))>

[6] *Github* https://github.com/msanto13/desktop-tutorial.git

1. Dudel, Christian, Riffe, Tim, Acosta, Enrique, van Raalte, Alyson, Strozza, Cosmo, & Myrskyla, Mikko. *Monitoring trends and differences in COVID-19 case-fatality rates using decomposition methods: Contributions of age structure and age-specific fatality.(2020).* [↑](#footnote-ref-1)
2. Bergman, Aviv, Sella, Yehonatan, Agre, Peter, & Casadevall, Arturo. (2020). *Oscillations in U.S. COVID-19 Incidence and Mortality Data Reflect Diagnostic and Reporting Factors.* [↑](#footnote-ref-2)
3. *The Community COVID-19 Dashboard Project*. (n.d.). [↑](#footnote-ref-3)
4. *Coronavirus Pandemic Data Explorer.* [↑](#footnote-ref-4)