UiA University of Agder

Health Problems Research Through Big Data Analytics

Abdelrhman Adel Zaher, Email Alteer888.love@gmail.com,
Dan LUO, Email luodan@bupt.edu.cn,
Maonan WANG, Email wangmaonan@bupt.edu.cn,
Mohamed Yahya Jabokji, Email m.jubokji@gmail.com

Data Science Application - IS 216

Date: May 15, 2020

Abstract

In this project we want to learn how to use data science application in working with big data in medicine and healthcare. Our problem is how to make people healthier using data science. Healthcare is one the business fields with the highest "big data" potential. A lot of medical doctors, especially in Norway, and scientists in medicine are doing a lot of research collecting data from patients. But they don't know what to do with all the collected material, and that is where we come in. We have various questions related to Corona virus outbreak in this study, and to answer those questions we will analyze some combined datasets using some tools like Kaggle, python and data analysis mythologies to find out correlation between the different variables.

As a result, we found that there is correlation between many variables like countries with higher GDP and aging rate has higher active cases than other countries, and he correlation coefficient between 'SickRate' and GDP is relatively large. Finally, we did a regression analysis on USA to predict the virus outbreak, the result is if the status quo is maintained, the number of active cases will continue to increase.

Contents

List of Figures

List of Tables

1 Introduction

We are in the technology age, almost every single person generates a huge amount of data through mobile phones, computers, sensors and wherever there is a signal, healthcare data is among them (Raghupathi and Raghupathi, 2014). The extreme challenge in healthcare is how to collect and analyze data to make people healthier. Data analytics can help us to understand new diseases, predict the outcomes and to understand people's habits that decrease or improve their health (Asri et al., 2015).

As we reached the year 2020, still many people die from various disease such as cancer, heart disease, diabetes and the new Corona virus. Half of the 56.9 million deaths worldwide in 2016 were due to the top 10 causes of death and Ischaemic heart disease and stroke are among them. Lung cancer caused 1.7 million deaths in 2016. These diseases have remained the main cause of death globally for the last 15 years (WHO, 2018b).

Later in 2019 **COVID 19** was discovered in China and has since spread globally, resulting in the coronavirus pandemic. Due to 29th April 2020 the virus has infected 3.17 million people, resulting in more than 224,000 deaths' around the world in just 5 months [1]. This virus has now already had a huge impact on the whole world's economy (Anderson et al., 2020), and it slos brings a lot of data that can be analyzed. And there has been a lot of research in this area (Fang et al., 2020; Zhou et al., 2020).

Global Burden of Disease (GBD) study, is an initiative to systematically analyze the causes of death in all over the world. This study helps to guide government, donor institutions, and private sectors through the use of datasets and data analytics (Forbes, 2015).

Yet we need more improvement in healthcare section to reduce the wrong habits that make people vulnerable to many diseases, and here comes our role as data scientists.

In this work, we will try to use data to solve the real problems in health. To be specific, we will focus on analyze the data about **COVID 19**, and we will try to visualize the data. Because we think the charst can provide researchers, public health authorities, and the general public with a user-friendly tool to track the outbreak as it unfolds, as the paper (Dey et al., 2020; Dong et al., 2020) said. The following is a general idea of our analysis. First of all, we will analyze the global trend, and the

[1] https://en.wikipedia.org/wiki/Coronavirus_disease_2019

change in the number of infected people. This mainly gives an overall analysis and some basic data visualization. Then we will analyze the relationship between the number of cases and the country, each country's GDP and Aging. Finally, we will try to predict the number of infected people and make a judgment on the future situation.

The rest of the paper is organised as follows: Section 2 is the background part of the article. In this part, we will discuss the problems of our respective operations. These problems are all related to health. At the same time, we will give the existing solutions to these problems. Section 3 gives solutions to the different problems proposed in section 2. At the same time, this part clarifies the problems that this paper want to focus on, and gives the questions we want to answer later, and give the way about how to answer them. The Methodology and dataset used in this work will be introduced in section 4.3, the methods are included linear regression, calculate correlation coefficient, etc. At the same time, a general step of this analysis will be introduced in section 4.3 too. The final result will be displayed in section 5. This results are divided into three parts, the first part is the basic data visualization, the second part is analyzing with other dataset together, the last part we will try to make some prediction. We will try to solve the problems raised in the section 3 in this part too. In section 6, we answer the questions raised in section 3, and give the detailed explanation. Finally, section 7 gives the contribution of each member of the group.

2 Background

In this chapter, we discussed the multiple problems about health. All these problems are examined as individuals. Furthermore, through existing research articles in key research areas, we have studied whether someone has proposed solutions to these problems we mentioned in this section.

2.1 Problems from Individual Assignments

2.1.1 The Health Problem about Heart

Heart is one of the most significant organs in blood circulatory systems for human. There are many elements can make some problems to heart. For example, smoking, poor eating methodology, poor quality food and insufficient sleep, etc (Keerthana, 2017). All these small negligences can lead to a major threat, that is heart disease. According to the World Health Organization (WHO) statistics in 2016, heart disease is the No.1 killer (Organization et al., 2016). Millions of people die because of the heart disease, and large number of people suffer from heart disease every year. Therefore, prediction of heart disease plays a crucial role for the treatment. In addition, if we can predict heart disease in the early stage, a lot of patient deaths can be prevented. And also, a more accurate and efficient treatment way can be provided (Sivagowry et al., 2013).

2.1.2 The Health Problem about Stroke

Stroke is a leading cause of death worldwide and is a major cause of acquired disability and loss of productive life-years (Feigin et al., 2016). Frist-time incidence of stroke occurs almost 17 million times a year worldwide which means that every two seconds someone in the world will have a stroke for the first time (Thrift et al., 2017). Approximately 30% of the stroke survivors will experience a recurrent stroke or mini-stroke (Mohan et al., 2011). Currently, approximately 3 to 4% of total health care expenditures in Western countries are spent on stroke (van Eeden et al., 2016). The socioeconomic burden of stoke and aging populations are now recognized as a prominent public health issue worldwide (Scalzo et al., 2015).

2.1.3 The Health Problem about Cancer (Lung Cancer)

As we reached the year 2020, still many people die from various disease and cancer one of the deadliest diseases to the mankind. Someone will die every three and half minutes in the United States by lung cancer, accounting for 25 percent of cancer deaths. Survivability rate from lung cancer has been increased according to State of Lung Cancer, but yet what still needs to be done to prevent more deaths and to save more lives[2].

Lung cancer is the most common cancer type around the world, and it is the most common cause of death from cancer with 20% of all cancer types deaths. About 2 million new cases were recorded in 2018[3].

2.1.4 The Health Problem about Liver Failure

Liver failure is life threatening condition and one of the most dangerous diseases that demands urgent medical care. It is difficult to detect at first and it happens when large parts of the liver become damaged and the liver cannot work anymore. It can be either acute or chronic (Bernal et al., 2010) [4]. A variety causes associated with liver failure such as drinking too much of alcohol, toxin, certain prescription medicines, some herbal supplements, etc[5]. All these causes can lead to liver failure. According to the World Health organization (WHO) the liver cancer is leading cause of death worldwide, 783000 deaths in 2018 of liver cancer. This number of people died because of liver cancer and this number may rise every year (WHO, 2018a). Therefore, prediction of liver failure disease plays a crucial role for the treatment. In addition, if we can predict heart disease in the early stage, a lot of patient deaths can be prevented. And a more accurate and efficient treatment way can be provided (Wilson, 2005) [6].

[2] https://www.ascopost.com/news/november-2019/2019-state-of-lung-cancer-report-released/

[3] https://www.wcrf.org/dietandcancer/cancer-trends/lung-cancer-statistics

[4] https://www.webmd.com/digestive-disorders/digestive-diseases-liver-failure#1

[5] https://www.healthline.com/health/hepatic-failure#types

[6] https://www.aacr.org/professionals/blog/liver-cancer-rising/

2.1.5 The Health Problem about Coronavirus

Today we live with one of dangerous diseases that money people die every day because of it. All medical organizations work to prevent spreading this disease and slow down transmission by find a good plan that people can follow and make people well informed about the COVID_19 virus until to find a treatments for this virus[7].

All countries in the world has been followed different ways to stop transmission by closing down the country and do not allow people to leave the house without a very good reason. Some of countries has succeed in decreasing the number and in other countries the number worsened (thelocal, 2020).

According to the World Health organization (WHO) the Coronavirus is leading cause of death worldwide, 287525 confirmed deaths of Coronavirus and 4179479 confirmed cases until 13 May 2020[8]. Therefore, prediction of Coronavirus cases plays a crucial role for finding the most effective plans to apply and reducing the number of people with this disease. In addition, if we can predict Coronavirus cases in the early stage, a lot of patient deaths can be prevented.

2.2 Existing Solutions

In this chapter we will discuss the existing solutions regarding healthcare and how increase survivability from a disease or to prevent it. Many solutions are existing, so we will focus on the main ones.

2.2.1 Existing Solutions to Heart Health

Nowadays, the diagnosis of heart disease in the early stage is still a challenging problem for medical industry. And the treatment cost of heart disease is not affordable by most of the patients, especially the people living in developing countries. At the same time, there are huge amount of healthcare data which are not used effectively. Discovery of hidden patterns and relationships often goes unexploited (Medhekar et al., 2013).

[7] https://www.who.int/health-topics/coronavirus#tab=tab_1

[8] https://www.who.int/emergencies/diseases/novel-coronavirus-2019

There are lots of studies focusing on using big data analytics to analyze and predict heart disease (Chadha et al., 2016). Medhekar and Bote have used the big data analytics to develop an intelligent Heart Disease Prediction System. This system uses Naive Bayes algorithm and it also uses a smoothing technique (Medhekar et al., 2013). Jabbar and Deekshatulu have used K-nearest neighbor and genetic algorithm based on big data to predict heart disease (Jabbar et al., 2015). Dewan and Sharma have discussed various kinds of techniques based on big data for developing a heart disease prediction system. They have used Backpropagation Algorithm as the classification technique for the targeted system (Dewan and Sharma, 2015).

There are two advantages using big data mining techniques to analyze heart disease:

- Advanced data mining techniques is helpful for earlier diagnosis of heart disease. If we can determine the heart disease in the early stage, then this would enhance medical care. And it can also reduce the patient costs.
- At the same time, we can also use data mining techniques to discover the relationships between heart disease and various elements. Knowing which elements easily lead to heart disease can help us improve our lifestyle and keep healthy.

2.2.2 Existing Solutions to Stroke Health

Understanding clinical causation of stroke is critical for patient treatments. There are many predictors associated with stroke diagnosis, medical history, hyperlipidemia, obesity, diabetes mellitus (Black et al., 2015; Hayden et al., 2015). However, there are no single factor would make a definite diagnosis yet (Ni et al., 2018). Physician review of patients' complicated medical records remains the gold-standard method of ascertaining stroke diagnosis which is labor intensive and expensive (Black et al., 2015). Differences about best practices may take years to settle via clinical studies and then to circulated to clinical practice (Alexander and Wang, 2017; Wang and Alexander, 2016). This situation is expected to change rapidly as new sources of healthcare data become increasingly relevant (Tresp et al., 2016). Big data analytics has the potential to improve care, save lives and lower costs and using data analytics in stroke care will build evidence and insights based on the 'real world' (Nishimura et al., 2016).

In stroke care, the use of data analytics has received considerable attention as an important source for creating new evidence (Barocas et al., 2017).

Studying large datasets of patient features, outcomes of treatments and their cost can help identify the most clinically effective treatment to apply thereby influencing healthcare provider behavior (Wang and Alexander, 2016). Data mining, visualization are technologies of big data which will help decision-making through mining the voluminous datasets for information and providing different perspectives (Ghadge et al., 2015). In particular, using data analytics in stroke care will enable the development of stroke phenotypes that can leads to a better understanding of stroke etiology (Ni et al., 2018).

2.2.3 Existing Solutions to Reduce Risk of Lung Cancer

Many factors can lead to develop lung cancer among the human like smoking which is the most common and leading risk factor with 80% of lung cancer deaths. Air pollution can lead to lung cancer especially near heavily trafficked roads. This factor is less than smoking, but yet some researcher estimated that 5% of lung cancer deaths due to air pollution (cancer society, 2019; Lubin and Blot, 1984; Malhotra et al., 2016).

Diet and nutrition are crucial when it comes to cancer. Red meat, saturated fat and alcohol can increase the risk of many cancer types. In the other hand consuming fruit and vegetables reduce the risk of cancer even among the smokers[9].

Some people who get lung cancer do not have any clear risk factors. Although we know how to prevent most lung cancers, at this time we don't know how to prevent all of them (cancer society, 2019).

Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Its cure rate and prediction depends mainly on the early detection and diagnosis of the disease. One of the most common forms of medical malpractices globally is an error in diagnosis. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system (Krishnaiah et al., 2013).

[9] https://www.wcrf.org/dietandcancer/exposures

2.2.4 Existing Solutions to Liver Failure

Today the big Data play great role to find problems in all fields in the life, but the most problem is how we can create value of these collected data to solve problem. Recent studies highlight of big data in offering identifying patients with liver disease and the new methods enable the development of population health algorithms [10].

The big data that are available of liver failure can help to find new values by filter them then make decisions that will help to innovate for reducing the spread of these diseases. The challenge to use the big data to find more causes and increase individual awareness about the factors that cause this disease ultimately create more effective methods to reduce spreading of the liver failure and by prevention. That will lead to innovative solutions for liver failure. **For example**, by using data mining techniques to find the relationships between liver failure and various elements that can lead to know which elements can lead to liver failure and that help us improve our lifestyle. In addition, it can help us for earlier diagnosis of liver failure. Determining the liver failure in early stage can enhance medical care and reduce the patient costs.

2.2.5 Existing Solution for Corona Virus

Until the moment of writing this study, no cure for COVID-19 has been discovered. Therefor many countries have taken strict measurements to lower the spreading of the virus as possible. Here are the major measurements that been taken in many countries:

- Students from all public educational institutions (secondary and higher education) were sent home by March 13, 2020 provisionally for two weeks.
- All public employees who do not perform critical functions (police, health care ect.) are required to work from home.
- Limited use of public transport.
- Closure of restaurants, bars, etc.
- Ban of traveling.
- Turning some hospitals into specialized Coronavirus hospitals and bring all the cases to those hospitals.

[10] http://medicine.buffalo.edu/news_and_events/news/2018/09/talal-liver-care-9076.html

3 The Proposed Solutions

In this chapter, we will discuss different solutions to the different problems we mentioned in the previous chapter. After that, we will focus on the topic about coronavirus, and formed research questions, and developed plans for how to do research on this problem. About the Corona Virus solution, we will analyze it in detail later (in section 4.3 and in section 5), so we will not give the solutions in this section.

3.1 Solutions from Individual Assignments

3.1.1 Heart Disease

The problem described in 2.1.1 focuses on heart disease and how to use big data mining techniques to analyze and predict heart disease. We plan to answer the three questions:

- Are some variables having positive or negative correlations on heart disease?
- Is a person got heart disease?
- Among the many attributes, which attributes are the important ones that cause heart disease?

To solve the problems, we decide to use the dataset named "Heart Disease UCI"[11] to solve the questions mentioned above. This dataset is a part of the "Cleveland Heart Disease Data" (the part obtained from the V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation) and is collected by 303 individuals who have heart disease. After that, we will use python, especially the library named sklearn (Pedregosa et al., 2011) and pytorch (Adam et al., 2017), to analyze the dataset.

- For the first question, we want to analyze the relationship between heart disease and each attribute in the dataset, such as the relationship between age and heart disease. After we calculate the relationship between different attributes, we can answer the question that check whether older people are more likely to get heart disease than youngs. Furthermore, which age group is more likely to be affected. We calculate the Pearson correlation coefficient for each variable in the dataset to analyze the relationships.

[11] https://www.kaggle.com/ronitf/heart-disease-uci

- For the second question, we will use the random forest (Liaw et al., 2002) or the neural network to predict heart disease.
- For the third question, we will use the LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) to give explanations of the model we create for prediction. According to the explanations, we can find which variable has an essential effect on the result of the prediction.

3.1.2 Stroke Disease

The problem described in 2.1.2 focuses on introducing the situation of stroke and the importance of using data analytic to analyze the cause of troke. Four questions are proposed:

- Is there any significant correlation between the influencing factors?
- What is the mainly influencing factors of stroke among 10 indicates collected in the dataset?
- Based on the analyzing, what can we predict about stroke based on given indicates of the individual?
- Based on the answer to the above questions, what solutions exist and can be applied to help people from getting stroke?

To answer those questions, we found the stoke dataset on Kaggle [12]. The dataset contains 43401 records and 10 variables: gender, age, hypertension (whether the patient has hypertension, 0 for no hypertension, 1 for suffering from hypertension), heart disease (whether the patient has heart disease), marital status, type of occupation, area type of residence (urban or rural), average Glucose level, body mass index, smoking status. Not all variables are analyzed the same but all play an important part. Then we utilize Tableau to as a tool for data analysis and visualization. The reason is that Tableau is a common tool used in data analysis which is simple to study, easy and effective.

And I want to do the two following analysis:

- **Profiling** is an unsupervised method. I will use it to describe the natural affecting relationship between the 10 variables and getting stroke.
- **Classifying** is a supervised method. The target on this task in to predict whether the person with determined variables values has stroke or not. Theoretically, prediction based on all 10 given variables will generate the highest accuracy. In order to further research what is the main influencing factors, I also predict the stroke based on single or several combined

[12] https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data

variables. Based on the prediction accuracy the affecting relationship is verified and the variables which given the highest accuracy are the main influencing factors.

3.1.3 Lung Cancer

The problem described in 2.1.3 focuses on increasing survivability from lung cancer and preventions. Survivability and preventions can be increased by following a good fruit diet and increase physical activities. Limiting exposure to cancer causing agents such as asbestos. People with a healthy lifestyle reduces risk of chronic diseases and people with less focus on a healthy lifestyle have a higher risk of developing a number of chronic diseases.

3.1.4 Liver Failure

The problem described in 2.1.4 focuses on liver failure and how to use big data mining techniques to analyze liver failure. We plan to answer the four questions:

- Which age group is most exposed to liver failure?
- Which factors that may lead to this disease?
- Does education affect avoidance of liver failure?
- Are women more exposed to liver failure than men?

To answer these questions, we decide to use the Primary Source of dataset is "Kaggle.com", one of the world's largest data science community and it is labeled as public dataset. This means that available dataset can be shared publicly.

Types of data that will be used in this study are variables such as (age, gender, wight, height, education, $\cdots$) as well as context, based data obtained from "Acute Liver Failure" dataset[13].

There are many tools to analyze data that can be used such as PowerBI, Tableau, SAS and so on. However, I have chosen in this study Tableau as the primary tool in order to analyze the data. The reason to use this tool is that it provides features for visualization and predictions and offering fast data analysis that help see and understand the data.

[13] https://www.kaggle.com/rahul121/acute-liver-failure

There are various methods for data analysis field such as descriptive analysis, regression, cluster, and so on, largely based on two core areas:

- quantitative data analysis
- data analysis methods in qualitative research.

Since this case focuses on which age is most exposed to liver failure and the factors that have influence having liver failure, it will use descriptive analysis, predictive analysis.

Then, we want to achieve exposed outcomes after the different types of analysis be performed on the used data set. By using the research questions and the datasets as well as different variables chosen for this case study, it will be alluded for possible outcomes and values based on this study.

The aim of this study to find factors that lead to liver failure and which age group is most exposed to have this type of diseases "liver failure". After preforming predictive analysis and predictive analysis. We could study the average age of people with liver failure and what the common factors that people with this disease have.

By summarizing the data we have with some graphs and charts that can allow we to come up with some simple conclusions, for instance if we consider that we have charts with to variables weight and maximum blood pressure we could come up with simple results like, for example how many patients have high weight and how many patients have high degree in Maximum Blood Pressure. This data visualization in Tableau will illustrate the reasons and factors that cause liver failure. By applying predictive analysis, Tableau, based on given datasets will show the probability of a person suffering the liver failure disease.

3.2 The Chosen Problem

In this report, we intended to choose COVID-19 as our main problem to analyze. This problem has affected nearly all the people on the earth in one way or another. That's why it is the most important problem to work on. Studies and researches revealed that there is no hope to find a cure or vaccine for the virus before at least one year from now.

3.3 Research Questions

Beside the questions that we might have during the analysis, we have chosen four main questions to focus on basically:

- What is the overall trend of virus development. How many stages is divided into? What are the reasons for these stages?
- Is there a relationship between country and region in spreading or death from COVID-19?
- Is there any correlation between population average age, GDP, population and death rate, sick rate from COVID-19?
- What can we predict about further number of cases?
- What kind of solution can be taken and how it can solve the problem?

Making of questions are based on dataset that we have, and how we can use it to answer these questions. By answering the questions, we could predict many changes in the future and to find potential solution.

3.4 How We Plan to Answer the Research Questions

To answer the questions we will analyze the dataset using python and its packages. These different packages can help us answer the research questions. Based on the chosen methods we will proceed to acquire knowledge related to these. This will eventually lead us closer to answer our research questions. For example, we will use numpy and pandas for data preprocessing, we will discuss these in section 4.4.1. We will use matplotlib to draw static images, the results are shown in section 5.1 and in section 5.2. We will use plotly to draw dynamic charts, the results are shown in section 5.1 too. Finally, we will use sklearn to do some prediction, the results are shown in section 5.3.

4 Methodology

This section covers tools and technologies, sample data, methods of analysis, as well as the details of analysis.

4.1 Tools and technologies

The tools we have used to conduct this analysis include **python** and different **packages** in python. For better display, we uploaded the python code and results on kaggle.

Then we introduce the python packages we actually use in this assignment. We use **numpy** and **pandas** for data processing, including data import, data normalization, and missing value processing. We use **plotly** to draw dynamic graphs, showing the changes in the number of people who are diagnosed with a virus. We use **matplotlib** to draw static graphs, such as drawing fitted curves. We use **sklearn** for linear regression and predicting the number of infected people (Garreta and Moncecchi, 2013).

4.2 Sample data

The data used include datasets obtained from the "The World Bank"[14] and the "Github"[15]. These data include the number of new coronaviruses diagnosed, the number of deaths and the number of cures per day in each country. At the same time, we also use the dataset includes aging data, population data and GDP data for each country. The following will introduce these datasets in detail, and give sample data.

4.2.1 Time Series Data about COVID-19

This dataset is gathered by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), and is uploaded on their github page [16]. Also, Supported by ESRI Living Atlas Team

[14] https://www.worldbank.org/

[15] https://github.com/

[16] https://github.com/CSSEGISandData/COVID-19

and the Johns Hopkins University Applied Physics Lab (JHU APL).

This dataset includes three time series tables are for the global **confirmed cases**, **recovered cases** and **deaths**. All data is read in from the daily case report. Australia, Canada and China in these three table in these three table are reported at the province/state level. Dependencies of the Netherlands, the UK, France and Denmark in these three table are listed under the province/state level. The US and other countries in these three table are at the country level. The three tables are named as **time_series_covid19_confirmed_global.csv**, **time_series_covid19_deaths_global.csv**, and **time_series_covid19_recovered_global.csv**, respectively.

The variables in these tables include the name of the province/state, the name of the country, longitude, latitude, real-time number of infected people. We use longitude and latitude for the visualization on the map, we will introduce this later. The following tabel 1 is the example data in **time_series_covid19_confirmed_global.csv**:

	Data				
Country/Region	Province/State	Lat	Long	1/22/20 ···	5/11/20
Australia	New South Wales	-33.8688	151.2093	0 ···	3053
Canada	Manitoba	53.7609	-98.8139	0 ···	289
China	Hubei	30.9756	112.2707	444 ···	68134

Table 1: The Example Data in Confirmed Cases Table

4.2.2 GDP per capita (current US$)

This dataset is gathered by World Bank, and is published on the World Bank's website[17]. In this data set, it includes GDP per capita for each country in each year. However, in this work, we will use the latest data. The following table 2 is the example data in this dataset.

[17] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Country/Region	Year	Value
Australia	2018	57373.7
Canada	2018	46233.0
China	2018	9770.8

Table 2: The Example Data about GDP per capita

4.2.3 Population ages 65 and above (% of total population)

This dataset is also gathered by World Bank, and is published on the World Bank's website[18]. In this data set, it includes the percentage of population ages 65 and above for each country in each year. Also, in this work, we will use the latest data. In this dataset, the latest year is 2018. The following table 3 is the example data in this dataset.

Country/Region	Year	Value
Australia	2018	16
Canada	2018	17
China	2018	11

Table 3: The Example Data about Population ages 65 and above (% of total population)

4.2.4 Population per Country

This dataset is also gathered by World Bank, and is published on the World Bank's website[19]. In this data set, it includes the population for each country in each year. In this work, we will use the latest data, 2018. The following table 4 is the example data in this dataset.

4.3 Methods for Analyzing

In this section, we introduce several methods which help us solve our research questions.

[18] https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?end=2018&most_recent_value_desc=true&start=2018&view=map&year=2018

[19] https://data.worldbank.org/indicator/sp.pop.totl?end=2018&start=2018

Country/Region	Year	Value (Thousands)
Australia	2018	24992.37
Canada	2018	37058.86
China	2018	1392730.00

Table 4: The Example Data about Population for each Country

4.3.1 Logarithmic Scale

A **logarithmic scale** (or log scale) is a way of displaying numerical data over a very wide range of values in a compact way, typically the largest numbers in the data are hundreds or even thousands of times larger than the smallest numbers. Such a scale is nonlinear: the numbers 10 and 20, and 90 and 100, are not the same distance apart on a log scale. Rather, the numbers 10 and 100, and 100 and 1000 are equally spaced. Thus moving a set distance along the scale means the number has been multiplied by 10 (or some other fixed factor)[20].

Often exponential growth curves are displayed on a log scale, otherwise they would increase too quickly to fit within a small graph. For example, the log scale is often used in financial area (Feigenbaum, 2001; Sornette et al., 2001). In our work, the number of cases in the world increase very fast. In particular, when the United States began to conduct inspections, the number of people who are infected rises exponentially, so we use log scale here in some of our charts.

4.3.2 Sunburst Chart

Sunburst Chart is also known as Ring Chart, Multi-level Pie Chart, and Radial Treemap. This chart is typically used to visualize hierarchical data structures. In our work, our dataset includes country and continent information, this constitutes a hierarchical relationship.

A Sunburst Chart consists of an inner circle surrounded by rings of deeper hierarchy levels. The angle of each segment is either proportional to a value or divided equally under its parent node. All segments in Sunburst Charts may be colored according to which category or hierarchy level they belong to[21].

[20] https://en.wikipedia.org/wiki/Logarithmic_scale

[21] https://www.anychart.com/chartopedia/chart-type/sunburst-chart/

It is likely that the Sunburst Chart type of data visualization was developed to accommodate subunits of the Pie Chart's primary segments (the earliest known example dates back to 1801, and it can be found in William Playfair's Statistical Breviary) (Playfair, 2005; Spence, 2005).

4.3.3 Geographical Scatter Plot

Geographical scatter plot use different colors, shading or symbols within predefined areas to show the values of a particular quantity in those areas. This kind of data representation is often used to map data collected for areal units, such as states or countries. In this work, we want to analyze the number of infected people in various countries, so it's helpful that we use map to help us visualize the data. We will discuss the results of geographical scatter plot in section 5.1.5.

4.3.4 Pearson correlation coefficient

In statistics, the Pearson correlation coefficient (PCC) is a statistic that measures linear correlation between two variables X and Y. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s and for which the mathematical formula was derived and published by Auguste Bravais in 1844(Galton, 1886). The Pearson correlation coefficient has a value between $+1$ and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. It can be calculated by formula 1:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

In our work, we will use Pearson correlation coefficient to analyze the relationship between variables. To be specific, we want to know whether the GDP or aging is related to the number of cases. The analysis and results are shown in section 5.2.2.

4.3.5 Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. The form of the linear model looks like the formula 2

$$y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n \tag{2}$$

We often use least squares approach to fit the linear regression models, that is using least squares approach to solve the β in formula 2. Nowadays, linear regression plays an important role in the field of artificial intelligence such as machine learning. The linear regression algorithm is one of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties (Bishop, 2006; Muller and Guido, 2017).

In this work, we will use linear regression to predict the number of confirmed cases. At the same time, we plan to use linear regression to verify the relationship between different variables, like the relationship between prevalence and GDP or Aging. The linear regression is used in section 5.2.3 and in section 5.3.

4.4 Processes for Analyzing

In this section, we give general steps for our data analyzing. At the same time, we will give the detailed introduction about what we do in each step.

4.4.1 Data Pre-processing

In this work, we plan to focus on the entire country, but the original dataset is subdivided into provincial/state (as shown in table 1), we need to merge them first. We use **group_by** in **pandas** to combine the data, and use different operations on different variables. For example, for number of cases in each day, we sum them up. That is, the number of cases in different provinces of a country, we sum them up to get the total number for that country.

Then we creat a new data table. The original data only contains the three basic datasets, namely, confirmed, deaths and recovery. We create a new data table here to indicate the number of active people, that is, the people who are still sick. We use the formula 3 to calculate the data:

$$Active = Confirmed - Deaths - Recovery \tag{3}$$

Finally, in order to analyze the relationship between the number of cases and the information in countries, we add new variables to the table. That is, to integrate with other datasets, like **Population per Country** and **GDP per capita**.

We add new variables to each table, includung the continent, the GDP per capita of this country, the aging of this country and the population of this country. The following tabel 5 is the example data of our final data table:

	Data								
Country/Region	Lat	Long	1/22/20	···	5/11/20	continent	GDP	Population (Thousands)	Aging
Australia	-33.8	151.2	0	···	3053	Oceania	57373	24992	16
Canada	53.76	-98.8	0	···	289	America	46233	37058	17
China	30.9	112.2	444	···	68134	Asia	9770	1392730	11

Table 5: The Example Data in Final Confirmed Cases Table

4.4.2 Basic Data Visualization

In this part, we do the basic visual analysis to have a better understanding of the dataset we use. Firstly, we analyze the changing trend of the number of infected people around the world, and then in order to get a deeper understanding, we start from the perspective of continents. Next, we do the visualization from the perspective of the country, we look at the the number of active cases in each country per day. At the same time, we also plotted interactive sunburst chart and dynamic geographic scatter plot to better show the results. The results are shown in section 5.1.

4.4.3 Linked with Other Datasets

In this part, we will combine the number of diagnoses with a country's GDP and Aging and other variables for analysis. First of all, we plot the scatter chart with variables like, aging, gdp and confirm cases to roughly determine their relationship. Then we will find the Pearson correlation coefficient between various variables. Finally we analyze the variables with greater correlation, and then use linear regression to test the result. The results are shown in section 5.2.

4.4.4 Predict the Number of Active Cases

In this part, we plan to use regression to predict the number of infected people (including active cases, confirmed cases, and recovered cases). We will predict the number of active cases in the United States, because the US is currently the country with the largest number of infections. And we can use the same method (regression) in other countries.

However, in this part we only make simple predictions without considering some external factors, such as the implementation of new policies, etc. So the final result may be different from the actual one, we can only use it as a reference. The results are shown in section 5.3.

5 Results

In this chapter we present the findings from our analysis. Our findings are presented both visually and writing. The different figures of analysis are from python. We have uploaded the python file to kaggle, and it's easy for us to view some dynamic pictures online, and the link is added in the appendix A.2.

Note that in the various analyzes there are missing some countries in some of the graphic presentations. This is due to small sample sizes. If all countries are displayed at the same time, the result will be very messy.

In this part, it is mainly divided into three parts, which are basic visual display in section 5.1, joint analysis with other dataset in section (advanced data visualization) 5.2, and using regression to predict the number of infected people in section 5.3.

5.1 Basic Data Visualization - Understanding Data Set

The purpose of this part makes us have a better understanding of the dataset. We use visual methods to help us have a deeper understanding of the data.

5.1.1 Worldwide Corona Virus Cases

First, we analyze the changing trend of the number of infected people around the world. To be specific, we analyze the cumulative number of active people and the number of confirmed people every day. We use log scale on the y-axis (this method is introduced in section 4.3.1), this is because the the number rises too fast later. The figure 1 shows the final output:

We can see that there are three time periods in the figure 1:

- 1/12 to 2/12: at first, the confirmed/active cases increase very rapidly.
- 2/12 to 3/12: then, during this period, confirmed cases grows slowly, and the active cases even decreases.
- 3/12 to - : however, after 3/12, the confirmed/active cases increase rapidly again.

We will discuss the reasons for this results in the later visualization.

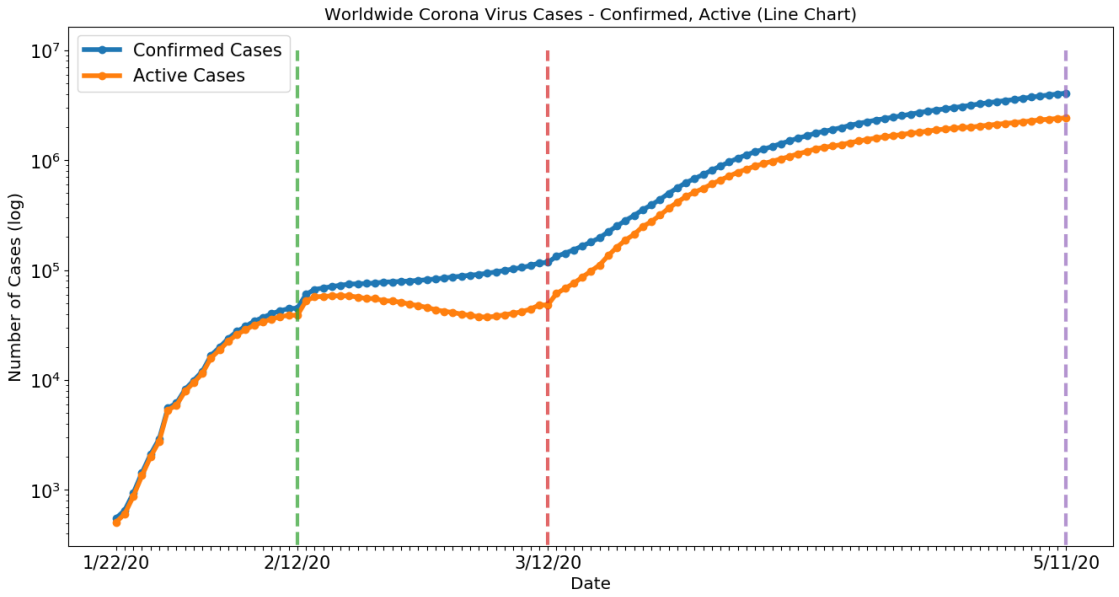

Figure 1: Worldwide Corona Virus Cases - Confirmed, Active (Line Chart)

5.1.2 Corona Virus Cases in Each Continent

Then, we analyze the changing trend of the number of infected people in each continent. Figure 2 shows the active cases in different continents per day.

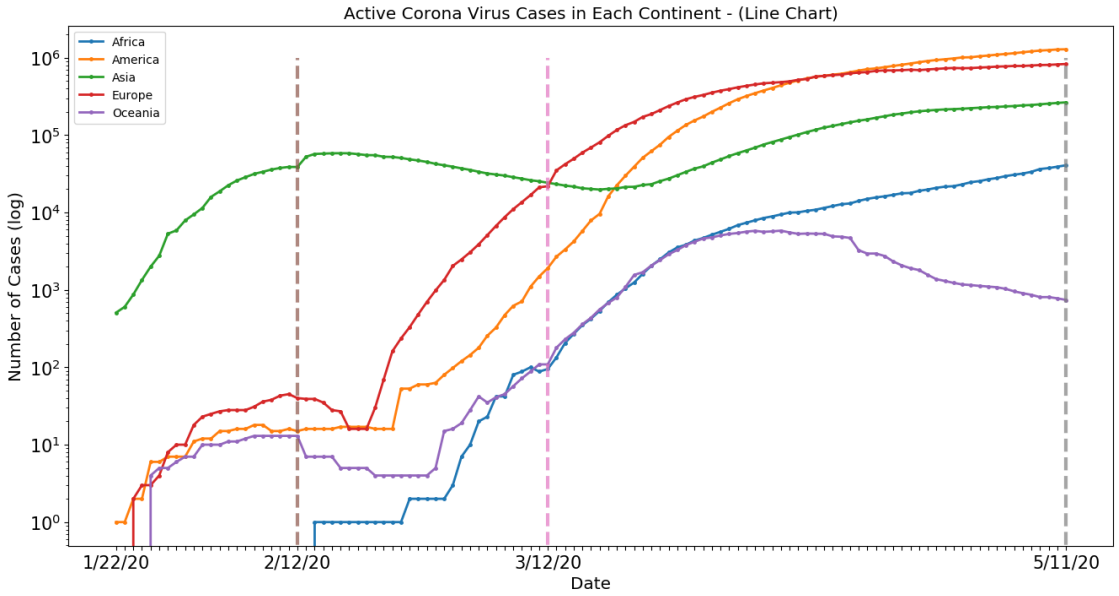

Figure 2: Active Corona Virus Cases in Each Continent - (Line Chart)

From figure 2 we can see that:

- The green curve indicates Asia. At first, the active cases rises fast in Asia, but the number of infected people began to decline after reaching the high point on 2/12.
- In Europe and Americas, the number of active cases has risen from 2/12, and the upward momentum is strong.
- The number of infected people in Asia has started to rise again after 3/12.

We will explain the reason for the above results when we analyze from the national level later. Note that the y-axis here is log scale too. We put results of the confirmed cases of each continent in appendix A, the figure 13.

We can also use a bar chart to represent the number of active cases in each continent. We made a dynamic bar chart, which clearly shows the change process of the number active cases on each continent every day. You can try interact on my kaggle's homepage[22]. As shown in figure 3, this is a bar of active cases on each continent on May 11, 2020.

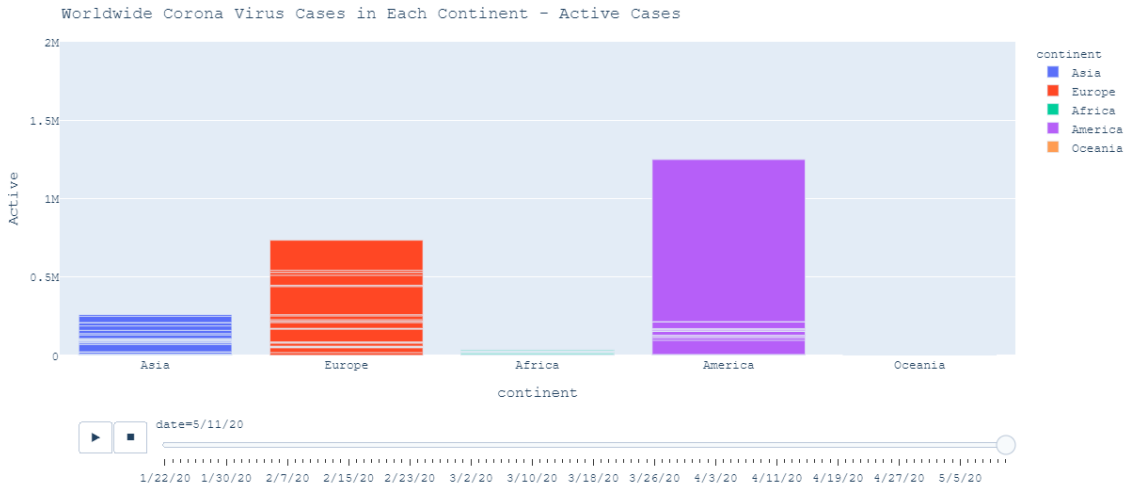

Figure 3: Active Corona Virus Cases in Each Continent - (Dynamic bar chart)

5.1.3 Corona Virus Cases in Each Country

Finally, we look at the the number of active cases in each country per day. In order to have a good visualization, we only select the countries with a larger number of active cases. Figure 4 shows the result of the active cases in each country:

From figure 4 we can see that:

[22] https://www.kaggle.com/maonanwang/data-analysis-on-coronavirus

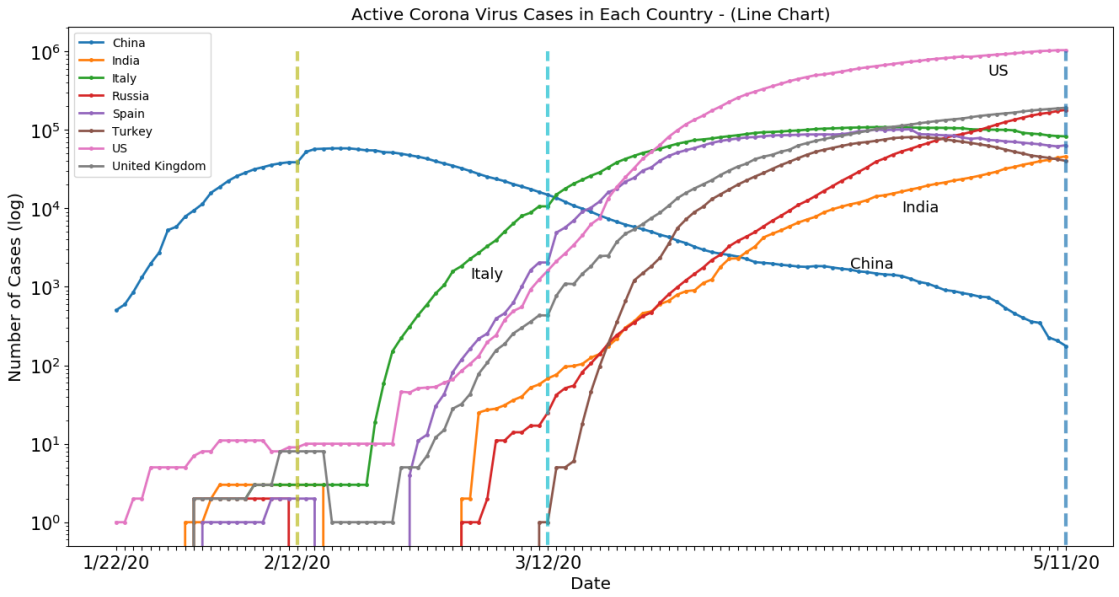

Figure 4: Active Corona Virus Cases in Each Country - (Line Chart)

- At the beginning, the number of infected people rose rapidly from China in Asia, and reached the peak on 2/12. Then, starting from 2/12, the number of infected countries in Europe and in Americas began to rise rapidly, especially in the United States.
- It can be seen that the number of infected people in US is currently the largest.
- At the same time, starting from the 3/12, the cases in other Asian countries, such as India and Turkey, are also increasing, resulting in the number of infections in Asia rising again in the later time period.

This can explain the results in Figure 2 and in Figure 1. The y-axis in Figure 4 is log scale too. We put results of the confirmed cases in each Country in appendix A, the figure 14.

5.1.4 Using the Interactive Sunburst Chart to Display More Details

In order to show the relationship between continents and countries better, we have also made Sunburst Chart (this method is introduced in section 4.3.2), and this chart is interacted. You can try interact on my kaggle's homepage[22]. The figure 5 shows the active cases in each country and continent.

The inner circles mean the continents, and the outer circles mean countries. The angle of each segment means the number of active cases. And this chart can be interactive. For example, we can

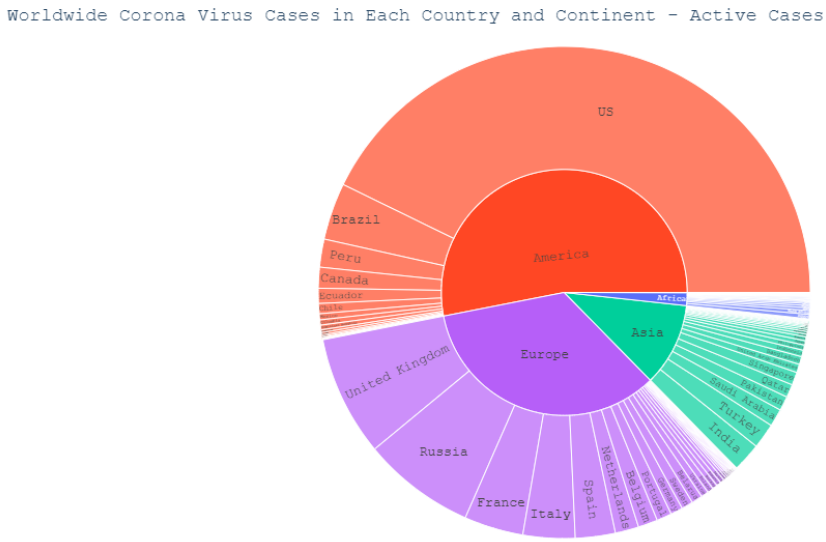

Figure 5: Worldwide Corona Virus Cases in Each Country and Continent - Active Cases

click on a continent, then we can see the detailed information of this continent. If we put the mouse on the country, we can see the detailed information of a country. As shown in the figure 6, we click on Europe, and put the mouse on Norway, then we can see the number of active cases in Norway.

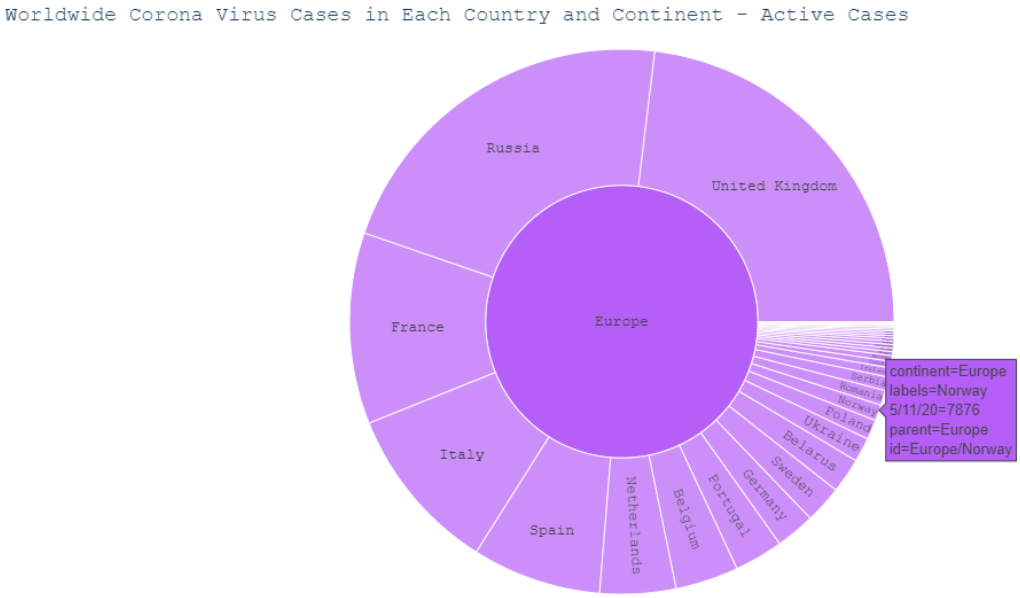

Figure 6: Active Corona Virus Cases in Europe in Norway

5.1.5 Scatter Plots on Maps

Because our data has the variables about country, we can combine the data with the map to make it more intuitive. We draw the distribution of active cases around the world on the day of 5/11. We use different colors to represent different Continents, and the size of the circle represents the number of active cases. The figure 7 shows the number of active cases on 5/11.

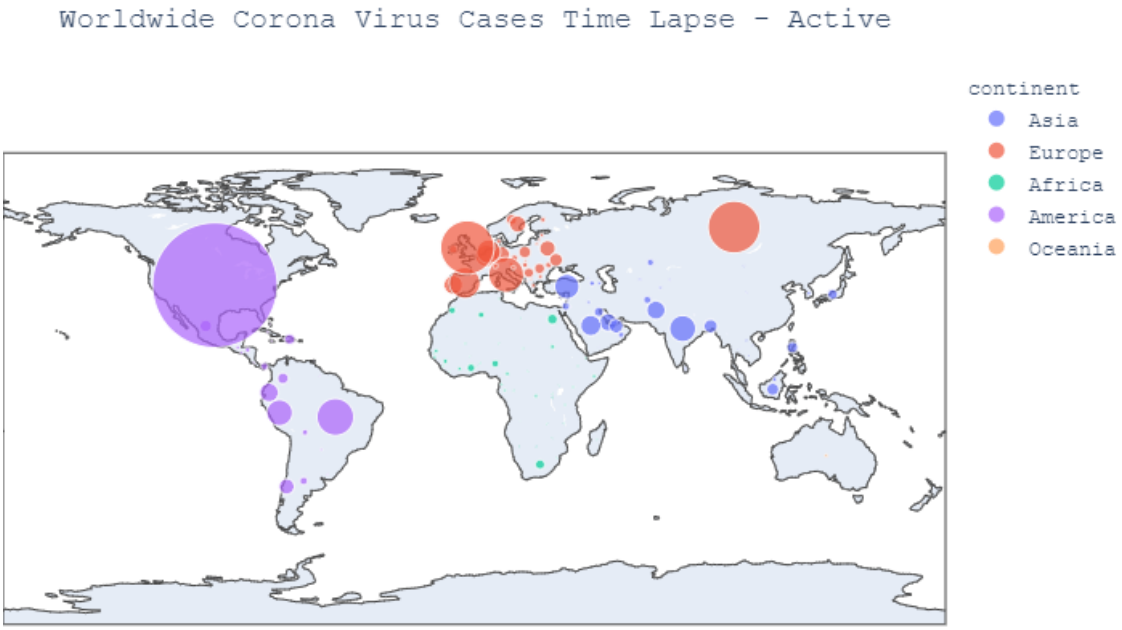

Figure 7: Geographical Scatter Plot on May 11

We can see from figure 7 that the circle in US is big, it means there are many active cases in the United States now.

Similarly, we can also make it as a dynamic graph, showing the changes of each day. You can see the dynamic scatter plot on maps on my kaggle's homepage[22]. We put one of the screenshots in the appendix, in section A, the figure 15.

5.2 Advanced Data Visualization - Linked with Other Datasets

In section 5.1, we only use the dataset about COVID-19, and we just do the basic data visualization. In this section, we will analyze the relationship between GDP, Aging, Population of each country

and COVID-19 data.

5.2.1 Intuitive Analysis of Mixed Data

For the mixed data, we first visualize it. We plot the chart with variables aging, gdp and confirm cases. In the figure 8:

- x axis represents gdp;
- y axis represents aging;
- the size of the circle represents the number of confirmed cases in each country;
- the color represents the continent;

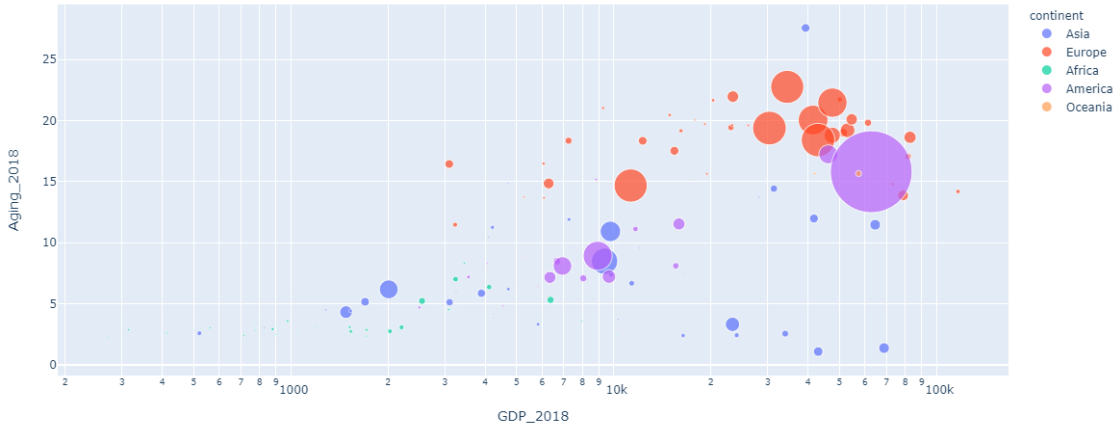

Figure 8: Intuitive Analysis of Mixed Data

From figure 8 we can see that the size of the circle in the upper right corner is larger. It means that in countries with larger GDP and aging, the number of confirmed cases is also higher. This may be because the coronavirus is more infectious to older people, and the transportation in the countries with high GDP is more convenient, so the virus spreads relatively fast.

5.2.2 Correlation Coefficient between Variables

In order to further analyze the relationship between variables, we calculate the correlation coefficient. Before we calculate the correlation coefficient, We add two new variables, that is 'SickRate' and 'DeathRate'. These two new variables are calculated as shown in the formula 4:

$$\begin{cases} SickRate = \dfrac{Confirmed}{Population} \\ \\ DeathRate = \dfrac{Death}{Confirmed} \end{cases} \tag{4}$$

After adding these two new variables, we calculate the correlation coefficient matrix between the variables 'GDP', 'Population', 'Aging', 'Confirmed cases', 'Death cases', 'Recovered cases', 'Active cases', 'DeathRate', 'SickRate'. The correlation coefficient matrix is shown in figure 9:

Figure 9: Correlation Coefficient Matrix of Mixed Data

As shown in figure 9, we can see that the correlation coefficient between 'SickRate' and GDP is high. It is 0.8. This means 'SickRate' and GDP have a strong positive correlation. We will analyze this in the next section 5.2.3.

5.2.3 Relationship Between Prevalence and GDP, Aging

As we can see from the correlation coefficient matrix above, the correlation coefficient between 'SickRate' and GDP is relatively large, so we draw a scatter chart of these two variables. As shown in figure 10:

- The x-axis represents the "GDP per capita (current US$)";
- The y-axis represents the "SickRate", the "SickRate" can be calculatied by formula 4;
- Each circle represent one country;
- The color represent the continent;
- The size of the circle represents the number of confirmed cases;

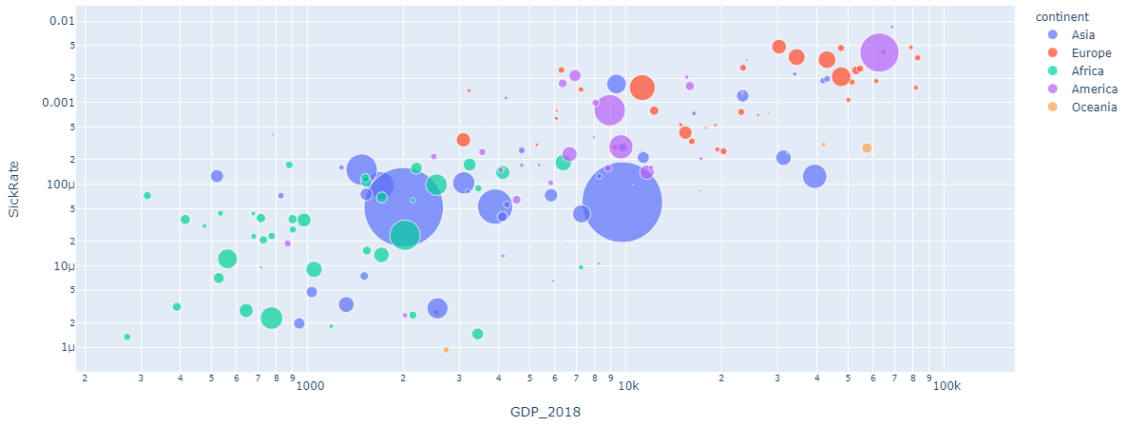

Figure 10: Scatter Chart of 'SickRate' and GDP

It is easy to find a linear relationship between these two variables in figure 10. As shown in figure 11, we use linear regression and plot the linear fitted curve with the scatter plot togerther. From figure 11, we can see that:

- Countries with higher GDP have higher prevalence;
- Most African countries (green spots), have low GDP, and have low prevalence;
- Most Europe countries (orange spots), the GDP in those countries are relatively high, the prevalence are also relatively high.

At the same time, we will calculate R-squared to evaluate the quality of the fit. In this case, the R^2 is 0.54, this model is moderate.

5.3 Predict the Number of Active Cases in US

Finally, we want to use regression to predict the number of active cases in US. We use the United States as an example, because now the United States has the largest number of active cases. And we can use the same method (the regression) on other countries. But there are many factors that affect the spread of the virus, we only consider our existing dataset here, and use the simplest regression

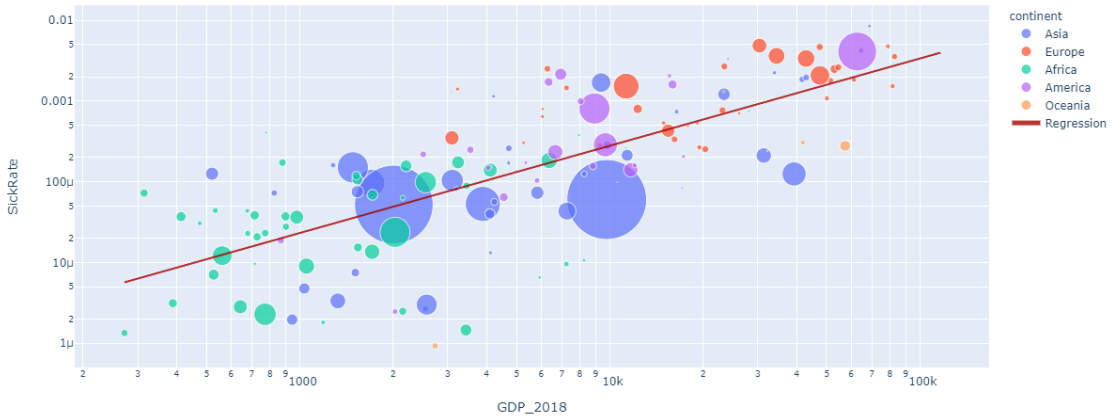

Figure 11: Scatter Chart of 'SickRate' and GDP with Linear Regression

to make predictions. So the results of predictions may not be very accurate, we can only use it as a reference. The predicted results are shown in Figure 12, we can see that if the status quo is maintained, the number of active cases will continue to increase.

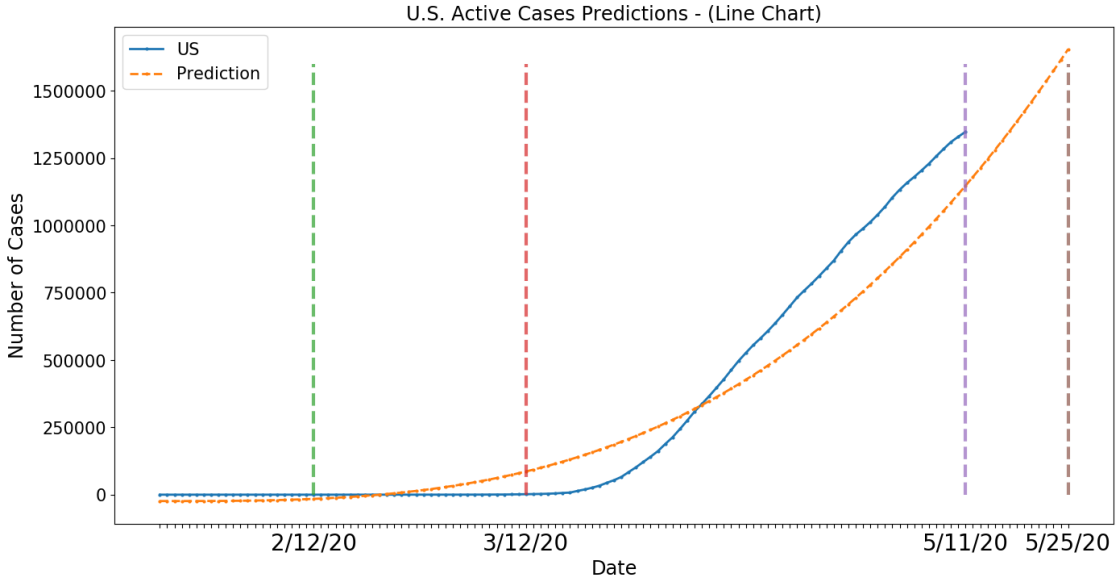

Figure 12: Forecast of Active Cases

6 Discussion

The results represented in the above section were generated by using the python and its different packages. In the above study, it was mainly to answer the 5 questions raised in the section 3.3. In this section, the research questions will be answered and other important issues related to the project will be discussed.

6.1 What is the overall trend of virus development. How many stages is divided into? What are the reasons for these stages?

For the first research question, we think there are three periods as the figure 1 shown.

- The first period is happened from 22/01/2020 to 12/02/2020. In this period time, the confirmed/active cases increase rapidly.
- The second period is happened from 12/02/2020 to 12/03/2020. And in this period time, slow growth in confirmed cases, and active cases decreases.
- the third period from 12/03/2020 to 11/05/2020. In this period time, the confirmed/active cases increase rapidly again.

At first, we were confused about the second time period because it seemed that the situation began to improve. But actuallty the situation began to deteriorate from the 3/12. In order to get further understanding of this situation, we started analyze from the perspective of continents and countries. The results are shown in figure 2 and figure 4.

As can be seen from these figures, in the first time period, the growth rate is very fast because China has a large number of infected people. From the second stage, China's situation has started to improve, but the number of infected countries in Europe has begun to increase rapidly. However, it is just the beginning, the number of infected people is not very large, so the number of cases is declining in the second time period. In the final stage, the number of infected people in the United States and European countries has begun to increase rapidly, so it has started to rise rapidly again.

6.2 Is there a relationship between country and region in spreading or death from COVID-19?

For this question, we believe that there is relationship between the country and region in spreading. It means the spread of the virus is related to the country. It can be seen from the figure 8 that the higher the per capita GDP, the more the aging countries, the higher the prevalence. We think this may be because the elderly are more susceptible to virus infection. And the transportation in these countries is more developed, so the virus will spread faster.

In the Figure 5 and 6 we find that Europe and America have better rates based off corona virus cases than other continents.

6.3 Is there any correlation between population average age, GDP, population and death rate, sick rate from COVID-19?

Yes, as we discussed in the above question, there is relationship between the country and region in spreading. For further analysis, we calculated the correlation coefficients between various variables, the result is shown in figure 9. We can see that the correlation coefficient between 'GDP' and 'SickRate' is relatively large. This means that countries with higher GDP have higher prevalence. To verify this, we plotted a scatterplot between them and used linear fitting to plot the relationship between them, the result is shown in figure 11. It can be seen that the higher the per capita GDP, the higher the prevalence rate.

6.4 What can we predict about further number of cases?

For this problem, we have used linear regression in the section 5.3 to make a simple prediction to the confirmed cases in US. As we said in that section, this prediction is very simple and we only consider our existing dataset here. But we still think if the government does not take some harsher measures, the number of cases will rise in next few weeks.

6.5 What kind of solution can be taken and how it can solve the problem?

For this question, we can say at present apply the same strategies which China and South Korea have done, for example:

- Turning some hospitals into specialized Coronavirus hospitals and bring all the cases to those hospitals, a lot of beds are needed now (Remuzzi and Remuzzi, 2020);
- Make working and studying from home and Ban transportation for at least 2 weeks;
- In addition to efficient testing, tracing, isolating and quarantining.

7 Individual Contribution

Each member of our group has participated with different types of knowledge. This chapter represents what each individual has contributed with in this report.

To organize our group work **Mohamed Jabokji** has taken the role as a project organizer. Every member is responsible for his own work. In addition Mohamed Jabokji has contributed with helping the other members formulate and structure their text.

Maonan WANG has been in charge of analyzing the datasets using Kaggle website and python programing language. Maonan WANG took charge of analyzing and researching the datasets. Every data representation which is used throughout the report has been produced by Maonan WANG.

Dan LUO has written a several parts of this report, as well as she was in charge of organizing the assignment. In addition to the writing she has done a lot of research and double checking all the sources.

Abdelrhman Adel Zaher have contributed in the writing of several parts of the report. While also proofreading and restructuring parts of the report, as well as double checking all the sources with Dan LUO.

We put the contribution by Chapter in appendix A.3, in table 6.

References

Adam, Paszke, Gross Sam, Chintala Soumith, Chanan Gregory, Yang Edward, D Zachary, Lin Zeming, Desmaison Alban, Antiga Luca, and Lerer Adam, "Automatic differentiation in PyTorch," in "Proceedings of Neural Information Processing Systems" 2017.

Alexander, Cheryl Ann and Lidong Wang, "Big data analytics in heart attack prediction," *J Nurs Care*, 2017, *6* (393), 2167–1168.

Anderson, Roy M, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?," *The Lancet*, 2020, *395* (10228), 931–934.

Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel, "Big data in healthcare: Challenges and opportunities," in "2015 International Conference on Cloud Technologies and Applications (CloudTech)" IEEE 2015, pp. 1–7.

Barocas, Solon, Elizabeth Bradley, Vasant Honavar, and Foster Provost, "Big data, data science, and civil rights," *arXiv preprint arXiv:1706.03102*, 2017.

Bernal, William, Georg Auzinger, Anil Dhawan, and Julia Wendon, "Acute liver failure," *The Lancet*, 2010, *376* (9736), 190–201.

Bishop, Christopher M, *Pattern recognition and machine learning*, springer, 2006.

Black, Michael, Wenzhi Wang, and Wei Wang, "Ischemic stroke: From next generation sequencing and GWAS to community genomics?," *Omics: a journal of integrative biology*, 2015, *19* (8), 451–460.

cancer society, American, "Lung Cancer Risk Factors," https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html 2019.

Chadha, Ritika, Shubhankar Mayank, Anurag Vardhan, and Tribikram Pradhan, "Application of data mining techniques on heart disease prediction: a survey," in "Emerging research in computing, information, communication and applications," Springer, 2016, pp. 413–426.

Dewan, Ankita and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," in "2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)" IEEE 2015, pp. 704–706.

Dey, Samrat K, Md Mahbubur Rahman, Umme R Siddiqi, and Arpita Howlader, "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach," *Journal of medical virology*, 2020, *92* (6), 632–638.

Dong, Ensheng, Hongru Du, and Lauren Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet infectious diseases*, 2020.

Fang, Lei, George Karakiulakis, and Michael Roth, "Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?," *The Lancet. Respiratory Medicine*, 2020.

Feigenbaum, James A, "A statistical analysis of log-periodic precursors to financial crashes," *Quantitative Finance*, 2001, *1*, 346–360.

Feigin, Valery L, Gregory A Roth, Mohsen Naghavi, Priya Parmar, Rita Krishnamurthi, Sumeet Chugh, George A Mensah, Bo Norrving, Ivy Shiue, Marie Ng et al., "Global burden of stroke and risk factors in 188 countries, during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013," *The Lancet Neurology*, 2016, *15* (9), 913–924.

Forbes, "How Big Data Can Make People Healthier In Emerging Markets," https://www.forbes.com/sites/techonomy/2015/07/30/how-big-data-can-make-people-healthier-in-emerging-markets/#34dc5f864c15 2015.

Galton, Francis, "Regression towards mediocrity in hereditary stature.," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 1886, *15*, 246–263.

Garreta, Raul and Guillermo Moncecchi, *Learning scikit-learn: machine learning in python*, Packt Publishing Ltd, 2013.

Ghadge, Prajakta, Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh, "Intelligent heart attack prediction system using big data," *International Journal of Recent Research in Mathematics Computer Science and Information Technology*, 2015, *2* (2), 73–77.

Hayden, Derek T, Niamh Hannon, Elizabeth Callaly, Danielle Ní Chróinín, Gillian Horgan, Lorraine Kyne, Joseph Duggan, Eamon Dolan, Killian O'Rourke, David Williams et al., "Rates and Determinants of 5-Year Outcomes After Atrial Fibrillation–Related Stroke: A Population Study," *Stroke*, 2015, *46* (12), 3488–3493.

Jabbar, M Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *arXiv preprint arXiv:1508.02061*, 2015.

Keerthana, TK, "Heart Disease Prediction System using Data Mining Method," *International Journal of Engineering Trends and Technology (IJETT)–Volume*, 2017, *47*.

Krishnaiah, V, G Narsimha, and Dr N Subhash Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," *International Journal of Computer Science and Information Technologies*, 2013, *4* (1), 39–45.

Liaw, Andy, Matthew Wiener et al., "Classification and regression by randomForest," *R news*, 2002, *2* (3), 18–22.

Lubin, Jay H and William J Blot, "Assessment of lung cancer risk factors by histologic category," *JNCI: Journal of the National Cancer Institute*, 1984, *73* (2), 383–389.

Lundberg, Scott M and Su-In Lee, "A unified approach to interpreting model predictions," in "Advances in neural information processing systems" 2017, pp. 4765–4774.

Malhotra, Jyoti, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, and Paolo Boffetta, "Risk factors for lung cancer worldwide," *European Respiratory Journal*, 2016, *48* (3), 889–902.

Medhekar, Dhanashree S, Mayur P Bote, and Shruti D Deshmukh, "Heart disease prediction system using naive Bayes," *Int. J. Enhanced Res. Sci. Technol. Eng*, 2013, *2* (3).

Mohan, Keerthi M, Charles DA Wolfe, Anthony G Rudd, Peter U Heuschmann, Peter L Kolominsky-Rabas, and Andrew P Grieve, "Risk and cumulative risk of stroke recurrence: a systematic review and meta-analysis," *Stroke*, 2011, *42* (5), 1489–1494.

Muller, Andreas C and Sarah Guido, *Introduction to machine learning with Python: a guide for data scientists*, O'Reilly Media, 2017.

Ni, Yizhao, Kathleen Alwell, Charles J Moomaw, Daniel Woo, Opeolu Adeoye, Matthew L Flaherty, Simona Ferioli, Jason Mackey, Felipe De Los Rios La Rosa, Sharyl Martini et al., "Towards phenotyping stroke: Leveraging data from a large-scale epidemiological study to detect stroke diagnosis," *PloS one*, 2018, *13* (2).

Nishimura, Ataru, Kunihiro Nishimura, Akiko Kada, Koji Iihara, J-ASPECT study group et al., "Status and future perspectives of utilizing big data in neurosurgical and stroke research," *Neurologia medico-chirurgica*, 2016, *56* (11), 655–663.

Organization, World Health et al., "New initiative launched to tackle cardiovascular disease, the world's number one killer global hearts 2016 [cited 2016; Available from: http://www. who. int/cardiovascular_diseases/en/]," 2016.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011, *12*, 2825–2830.

Playfair, William, *Playfair's commercial and political atlas and statistical breviary*, Cambridge University Press, 2005.

Raghupathi, Wullianallur and Viju Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, 2014, *2* (1), 3.

Remuzzi, Andrea and Giuseppe Remuzzi, "COVID-19 and Italy: what next?," *The Lancet*, 2020.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in "Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining" 2016, pp. 1135–1144.

Scalzo, Fabien, May Nour, and David S Liebeskind, "Data science of stroke imaging and enlightenment of the penumbra," *Frontiers in neurology*, 2015, *6*, 8.

Sivagowry, S, M Durairaj, and A Persia, "An empirical study on applying data mining techniques for the analysis and prediction of heart disease," in "2013 International Conference on Information Communication and Embedded Systems (ICICES)" IEEE 2013, pp. 265–270.

Sornette, Didier, Anders Johansen et al., "Significance of log-periodic precursors to financial crashes," *Quantitative Finance*, 2001, *1* (4), 452–471.

Spence, Ian, "No humble pie: The origins and usage of a statistical chart," *Journal of Educational and Behavioral Statistics*, 2005, *30* (4), 353–368.

thelocal, "Coronavirus across Europe: An inside view as countries plot a path back to normal life," https://www.who.int/news-room/fact-sheets/detail/cancer 2020.

Thrift, Amanda G, Tharshanah Thayabaranathan, George Howard, Virginia J Howard, Peter M Rothwell, Valery L Feigin, Bo Norrving, Geoffrey A Donnan, and Dominique A Cadilhac, "Global stroke statistics," *International Journal of Stroke*, 2017, *12* (1), 13–32.

Tresp, Volker, J Marc Overhage, Markus Bundschus, Shahrooz Rabizadeh, Peter A Fasching, and Shipeng Yu, "Going digital: a survey on digitalization and large-scale data analytics in healthcare," *Proceedings of the IEEE*, 2016, *104* (11), 2180–2206.

van Eeden, M, GAPG van Mastrigt, SMAA Evers, EPM van Raak, GAM Driessen, and CM van Heugten, "The economic impact of mental healthcare consumption before and after stroke in a cohort of stroke patients in the Netherlands: a record linkage study," *BMC health services research*, 2016, *16* (1), 688.

Wang, Lidong and Cheryl Ann Alexander, "Stroke Care and the Role of Big Data in Healthcare and Stroke," *Rehabilitation Sciences*, 2016, *1* (1), 16–24.

WHO, World Health Organization, "Coronavirus across Europe: An inside view as countries plot a path back to normal life," https://www.who.int/news-room/fact-sheets/detail/cancer 2018.

__, "The top 10 causes of death," https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death 2018.

Wilson, Jennifer Fisher, "Liver cancer on the rise," *Annals of internal medicine*, 2005, *142* (12_Part_1), 1029–1032.

Zhou, Fei, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying

Gu et al., "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *The lancet*, 2020.

A Appendix

A.1 Some Results of Data Visualization

Figure 13 is the line chart about the trend of confirmed corona virus sases in each continent per day.

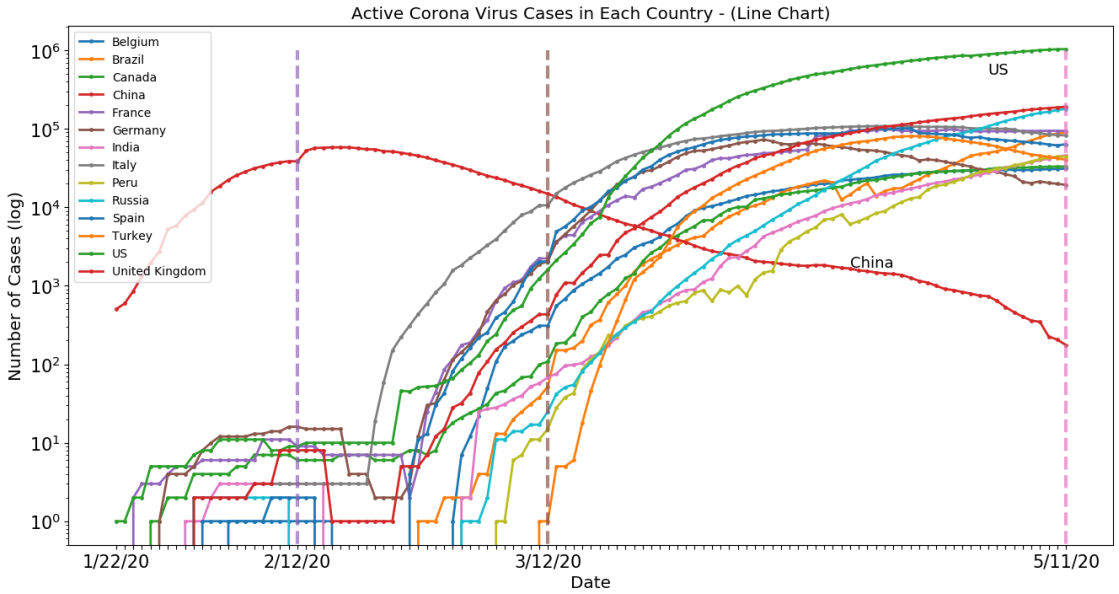

Figure 13: Confirmed Corona Virus Cases in Each Continent - (Line Chart)

Figure 14 is the line chart about the trend of confirmed corona virus sases in each country per day.

Figure 15 is the screenshot of the dynamic geographical scatter plot. This is a screenshot of May 1, 2020, the bottom of the figure 15 is the timeline.

A.2 The Link to Python File

We have uploaded the python file to kaggle, and the link is `https://www.kaggle.com/maonanwang/data-analysis-on-coronavirus/`.

A.3 The contribution by Chapter

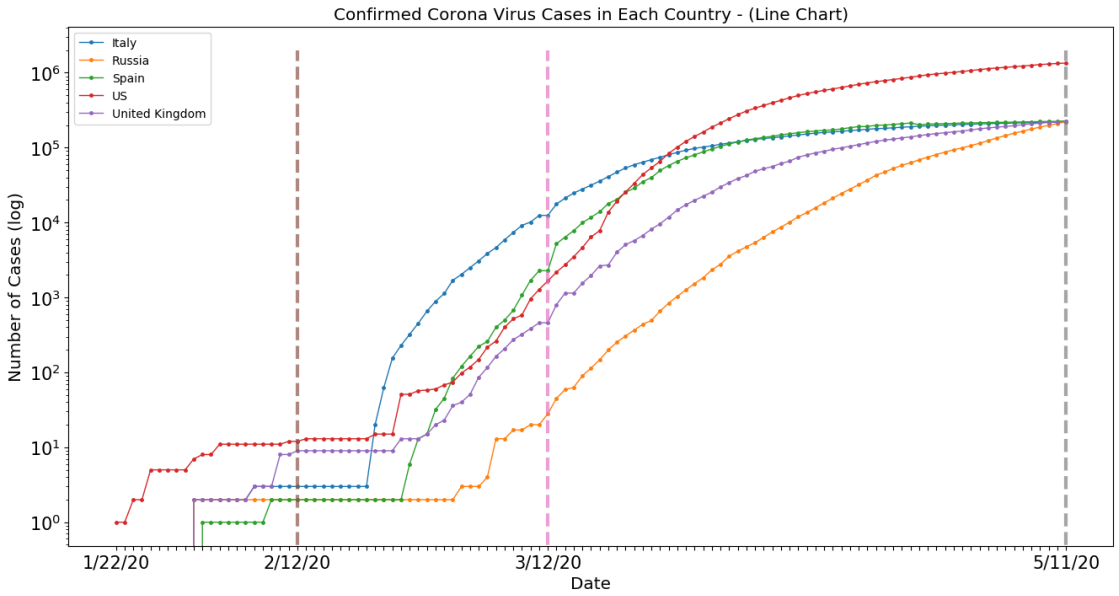

Figure 14: Confirmed Corona Virus Cases in Each Country - (Line Chart)

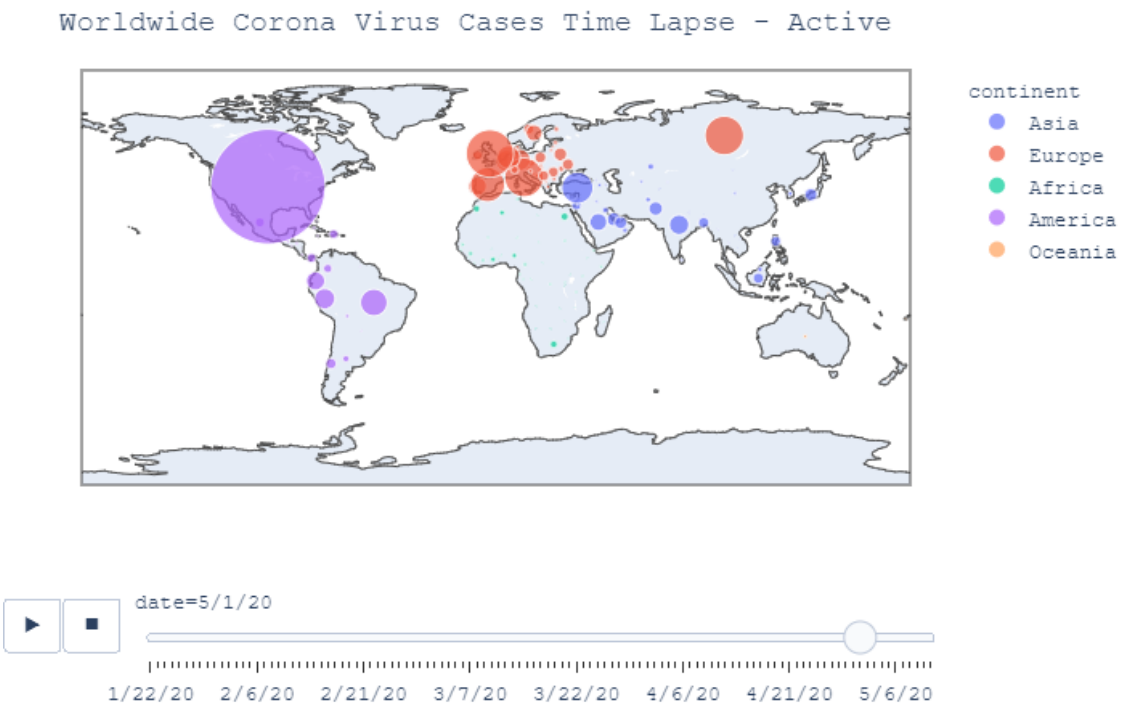

Figure 15: Worldwide Corona Virus Cases Time Lapse - Active Cases

Table 6: The contribution by Chapter

Chapter	Writer & Contributor
Abstract	Mohamed Jabokji & Abdelrhman Adel Zaher
1, Introduction	Abdelrhman Adel Zaher & WANG Maonan
2, Background	
2.1, Problems from Individual Assignments	
2.1.1, The Health Problem about Heart,	Maonan WANG
2.1.2, The Health Problem about Stroke	Dan LUO
2.1.3, The Health Problem about Cancer (Lung Cancer)	Mohamed Jabokji
2.1.4, The Health Problem about Liver Failure	Abdelrhman Adel Zaher
2.1.5, The Health Problem about Coronavirus	Mohamed Jabokji
2.2, Existing Solutions	
2.2.1, Existing Solutions to Heart Health	Maonan WANG
2.2.2, Existing Solutions to Stroke Health	Dan LUO
2.2.3, Existing Solutions to Reduce Risk of Lung Cancer	Mohamed Jabokji
2.2.4, Existing Solutions to Liver Failure	Abdelrhman Adel Zaher
2.2.5, Existing Solution for Corona Virus	Mohamed Jabokji
3, The Proposed Solutions	
3.1, Solutions from Individual Assignments	
3.1.1, Heart Disease	Maonan WANG
3.1.2, Stroke Disease	Dan LUO
3.1.3, Lung Cancer	Mohamed Jabokji
3.1.4, Liver Failure	Abdelrhman Adel Zaher
3.2, The Chosen Problem	Mohamed Jabokji
3.3, Research Questions	Mohamed Jabokji
3.4, How We Plan to Answer the Research Questions	Mohamed Jabokji & WANG Maonan
4.3, Methods for Analyzing	
4.1, Tools and technologies	Dan LUO & Maonan WANG
4.2, Sample data	
4.2.1, Time Series Data about COVID-19	Maonan WANG

Chapter	Writer & Contributor
4.2.2, GDP per capita (current US$)	Dan LUO
4.2.3, Population ages 65 and above (% of total population)	Dan LUO
4.2.4, Population per Country	Dan LUO
4.3, Methods for Analyzing	
4.3.1, Logarithmic Scale	Maonan WANG & Abdelrhman Adel Zaher
4.3.2, Sunburst Chart	Abdelrhman Adel Zaher
4.3.3, Geographical Scatter Plot	Maonan WANG & Abdelrhman Adel Zaher
4.3.4, Pearson correlation coefficient	Maonan WANG & Mohamed Jabokji
4.3.5, Linear Regression	Maonan WANG & Dan LUO
4.4, Processes for Analyzing	
4.4.1, Data Pre-processing	Dan LUO
4.4.2, Basic Data Visualization	Dan LUO
4.4.3, Linked with Other Datasets	Dan LUO
4.4.4, Predict the Number of Active Cases	Dan LUO
5, Results	
5.1, Basic Data Visualization - Understanding Data Set	
5.1.1, Worldwide Corona Virus Cases	Maonan WANG & Abdelrhman Adel Zaher
5.1.2, Corona Virus Cases in Each Continent	Maonan WANG & Abdelrhman Adel Zaher
5.1.3, Corona Virus Cases in Each Country	Maonan WANG & Abdelrhman Adel Zaher
5.1.4, Using the Interactive Sunburst Chart to Display More Details	Maonan WANG & Mohamed Jabokji
5.1.5, Scatter Plots on Maps	Maonan WANG & Mohamed Jabokji
5.2, Advanced Data Visualization - Linked with Other Datasets	
5.2.1, Intuitive Analysis of Mixed Data	Maonan WANG

Chapter	Writer & Contributor
5.2.2, Correlation Coefficient between Variables	Maonan WANG
5.2.3, Relationship Between Prevalence and GDP, Aging	Maonan WANG
5.3, Predict the Number of Active Cases in US	Maonan WANG & Dan LUO
6, Discussion	Abdelrhman Adel Zaher
7, Individual Contribution	Maonan WANG & Mohamed Jabokji