STAT1005 Essential skills for undergraduates: Foundations of Data Science/ STAT1015 Introduction to data science

Group Project

Chapter 1: Introduction

The outbreak of novel coronavirus disease 2019 (COVID-19) in more than 250 countries is threatening peoples' health around the globe. The virus has brought day-to-day life to a grinding halt, affecting every industry in one way or another. Governments and scientists are trying to understand the factors affecting the spread of the virus as well as the best ways to counteract its spread. Our project aims to use the power of data science to tackle these issues in two parts:

1. **How does COVID-19 spread?**
    a. Environmental factors - the impact of several major natural environmental factors (such as temperature, humidity, wind speed, air quality) on the spread of COVID-19
    b. Social and cultural factors - the relationship between population, gender, race, and usage of masks on the confirmed cases and deaths in various countries

2. **What is the best way to counteract the spread of COVID-19?**
    a. Personal protective equipment - the relationship between various types of personal protective equipment and the spread of the virus
    b. Health policies - the effectiveness of various health policies (school suspension, work from home, free testing and restrictions on gathering) on the spread of the virus

The objectives of the project are:

- To analyze the natural environmental factors that have major impacts on the spread of COVID-19 through correlation analysis, and predict the development of the pandemic based on the changing trend of environmental factors through the regression analysis
- To analyze the correlation between the social and cultural factors on the spread of virus and the impact to the global in different countries
- To study the effectiveness of various personal protective equipment (PPE) on slowing down the spread of the virus through correlation analysis, and to predict the most important PPE for curbing further outbreaks
- To study the effectiveness of various health policies including closure policy, restriction on public gathering (social distancing) and testing policy, on slowing down the spread of the virus through descriptive, correlation and linear regression analysis, and evaluate the most effective measure for curbing the outbreaks.

Chapter 2: Backgrounds and objectives

The various objectives of our projects are:

Objective 1: To analyze the meteorological factors that have major impacts on the spread of COVID-19 through correlation analysis and predict the development of the pandemic based on the environmental factors

In objective1, we explore the correlation between meteorological factors and the spread of COVID-19. As we know, meteorological factors are closely related to human life, and human beings experience meteorological changes every day. According to the research on the SARS virus done by Chinese Academy of Sciences (2003) and also the currently published essays on the correlation between meteorological factors and the new coronavirus (Eslami & Jalili, 2020), we find that temperature and humidity played an important role in the spread of the SARS virus, and similar effects occur as for the interaction of some meteorological factors (such as wind speed, sunshine, etc.) with COVID-19. Therefore, it is very necessary to study the correlation between meteorological factors and the spread of COVID-19.

Based on the existing study done by Eslami and Jalili (2020), we can propose that temperature, humidity, wind speed, and duration of sunshine are four variable meteorological factors that are closely related to the spread of COVID-19. In the following chapters, we will conduct correlation analysis to explore the relationship between these four variables and the spread of COVID-19 and use regression model to show and predict the change in the severity of the pandemic during seasonal changes.

Objective 2: To analyze the correlation between the social-cultural determinants and the spread of COVID-19

Even though the spread of the virus shows various patterns in different countries, the social and cultural factors play the utmost role in influencing the current spread of the virus worldwide. First and foremost, the population is one of the factors as there is a possibility of close contact across the population with those infected owing to the mode of transmission is through close contact (George M (2020)). Next, by studying the role of age in transmission, I am able to show that that which age group is more susceptible to the disease (Davies, N.G., Klepac, P., Liu, Y. *e*t al (2020)). The other factor is gender which mortality rate for male to female ratio is being investigated to analyze which gender contributes to a higher COVID-19 mortality rate (Christina P. Tadiri, et.al (2020)). There is evidence that some racial and ethnic groups are at a greater risk of being infected. Hence race is considered a cultural factor (Baligh R. Yehia, et.al (2020)). Since wearing a mask is a key measure to suppress the transmission of the virus, I will study the relationship between the usage of the mask and the total confirmed cases (Holger J Schuenemann, et.al (2020)).

By having statistical data, we are able to have a greater understanding of the strength between the two variables to check whether they are significantly correlated to each other. Hence, by running some analysis, we are able to make a hypothesis on whether social and cultural factors would show a striking trend towards greater infection rates.

Objective 3: To study the effectiveness of various personal protective equipment (PPE) on slowing down the spread of the virus through correlation analysis, and to predict the most important PPE for curbing further outbreaks

Research has shown that a number of personal protective equipment such as masks ((Chan et al., 2020), (Eikenberry et al., 2020), (Li, Liu, Li, Qian, & Dai, 2020)), full body PPE suits ((Smereka & Szarpak, 2020), (Liu et al., 2020)) and sanitizers (Mahmood et al., 2020) are effective to varying degrees in controlling the virus. However, from a governmental standpoint, it may be important to know which particular PPE equipment is the most important to prioritize during the time of a major outbreak, or during a shortage of funding.

This objective aims to study the correlation between the growth of the virus and the distribution of various equipment within the 50 counties of the US state of California. Descriptive analytics will first be used to study the overall situation of the coronavirus over time in each of the 50 counties and also to study the general trends of the distribution of products such as masks, respirators, testing kits and sanitizers. Regression and correlation analysis will then be used to identify the products that played the biggest role in the decrease of the virus (i.e. most negative coefficient of correlation) on a county-wise basis.

Objective 4: To study the effectiveness of various health policies including closure policy, restriction on public gathering (social distancing) and testing policy on slowing down the spread of the virus through correlation analysis, and to predict the most effective measure for curbing outbreaks.

This analysis would include the comparison of numbers of cases confirmed in different countries/regions that has implemented health policies- closure (R., Prof, S., PhD, H., J., J., C., R. (2020, April 06)), restriction on public gathering (A., & M. (2020, September 06)) and testing policy (Carl Heneghan, Dr. Elizabeth Spencer; MMedSci, & Tom Jefferson (2020, August 27)) in different natures (e.g. compulsory or recommended) (Hsiang, et al., 2020). And it would be shown by correlation and regression models.

By testing the correlation between them, we can evaluate the effects and variations of these measures on controlling the spread of the virus. Therefore, we can conclude which health policy is the most effective. The finding can further facilitate the

government’s control of the pandemics and help the world recover from the dilemma soon.

The descriptive analysis would be used in introducing the content of these health policies and some implementation details; while the correlation and regression models would be used in showing the relationships between the number of cases confirmed and the effectiveness of these health policies.

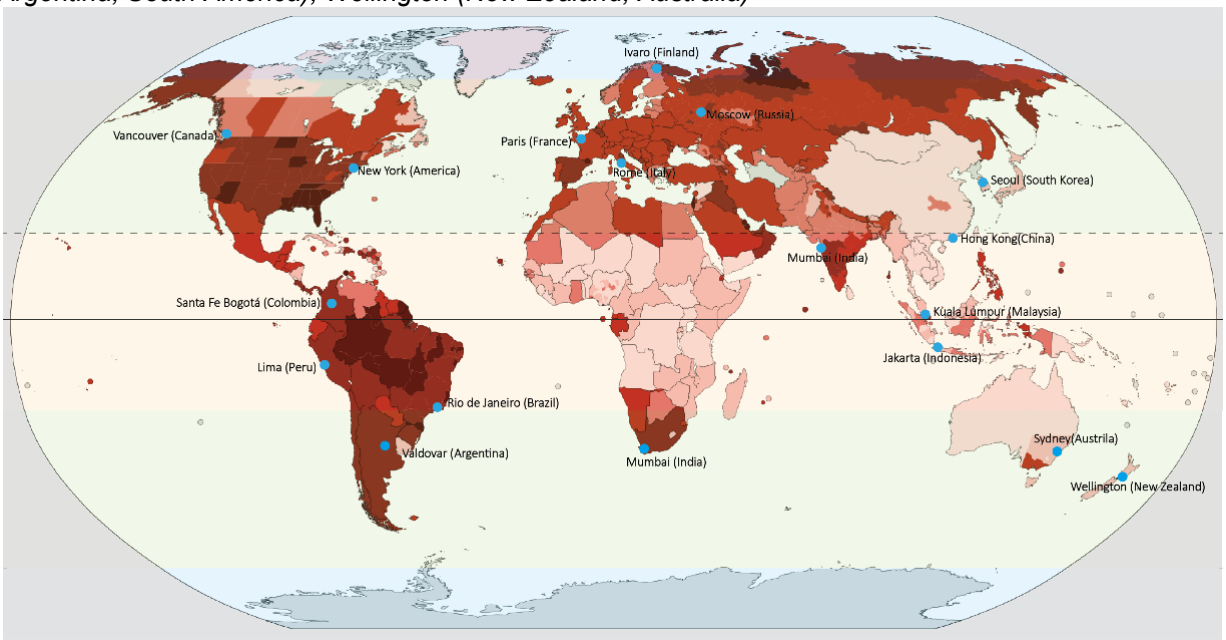
Chapter 3: Data science models

Objective 1:

Data sources, data cleansing and pre-processing

As mentioned above, in order to reflect the trend of the pandemic situation with various meteorological indicators comprehensively and objectively, we select sample cities from around the world, 18 in total. They are distributed in different hemispheres, different continents, different temperature zones, and different countries (in order to make the data results more obvious, the selected cities are located in countries with more confirmed cases):

***Northern Frigid Zone**: Ivalo (Finland, Europe)*
***North Temperate Zone**: Rome (Italy, Europe), New York (United States, North America), Moscow (Russia, Europe), Quebec (Canada, North America), Paris (France, Europe), Seoul (South Korea, Asia)*
***Tropical Zone**: Hong Kong (China, Asia), Mumbai (India, Asia), Kuala Lumpur (Malaysia, Asia), Santa Fe Bogotá (Colombia, South America), Manila (Indonesia, Asia), Rio de Janeiro (Brazil, South America), Lima (Peru, South America)*
***South Temperate Zone**: Cape Town (South Africa, Africa), Melbourne (Australia, Australia), Córdoba (Argentina, South America), Wellington (New Zealand, Australia)*



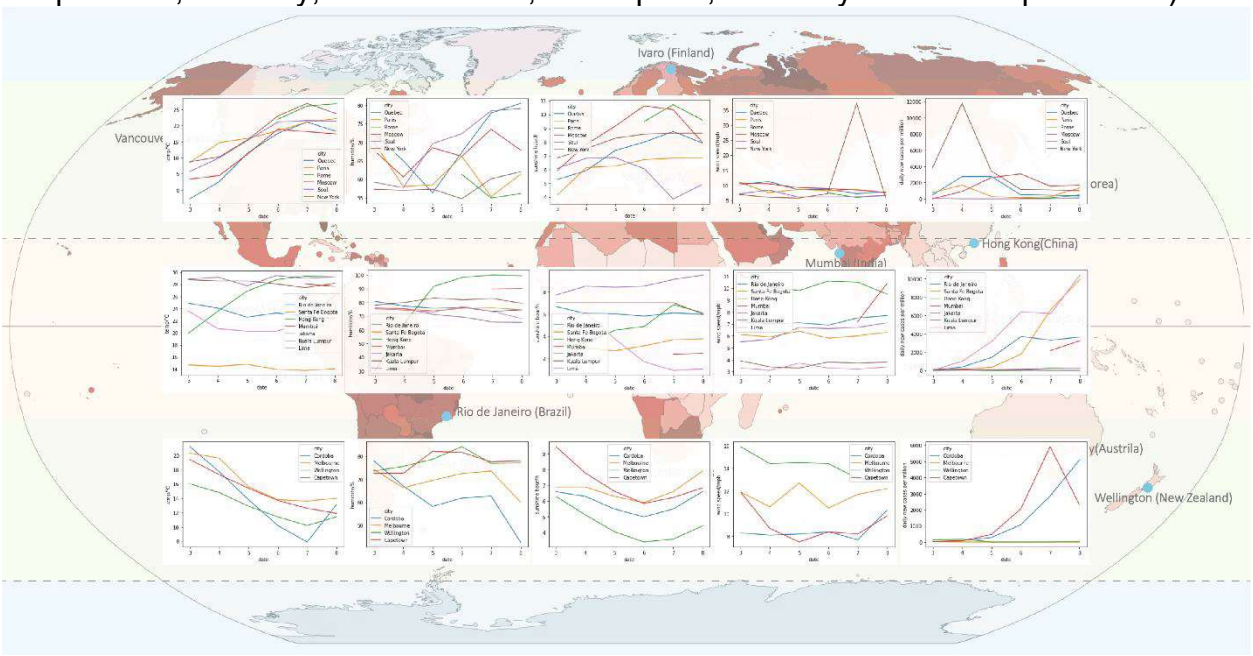
On the *Weather Underground* website, we found the historical weather of each city, and extract their daily average temperature, humidity, and wind speed data from March to August. Due to limited data, we cannot find a record of the daily sunshine duration, so we use the monthly average sunshine duration instead. In the Kaggle database, we find the data of confirmed cases in each country or region (partially specific to cities), and get the data of daily new cases by subtracting the data from the previous day in the excel table. It is worth noting that due to the different population bases of various countries/cities, the scale of confirmed cases will be affected. Therefore, in objective 1, when we present the spread of COVID-19, we use the number of new cases per million people instead.

After importing the data to Python, drop the useless columns "total confirmed cases", "daily new cases", and "temp/℉". In addition, we have also summarized the monthly average data of the four indicators in each city, so that the data can be more concise and summarized in case comparing among groups is needed.

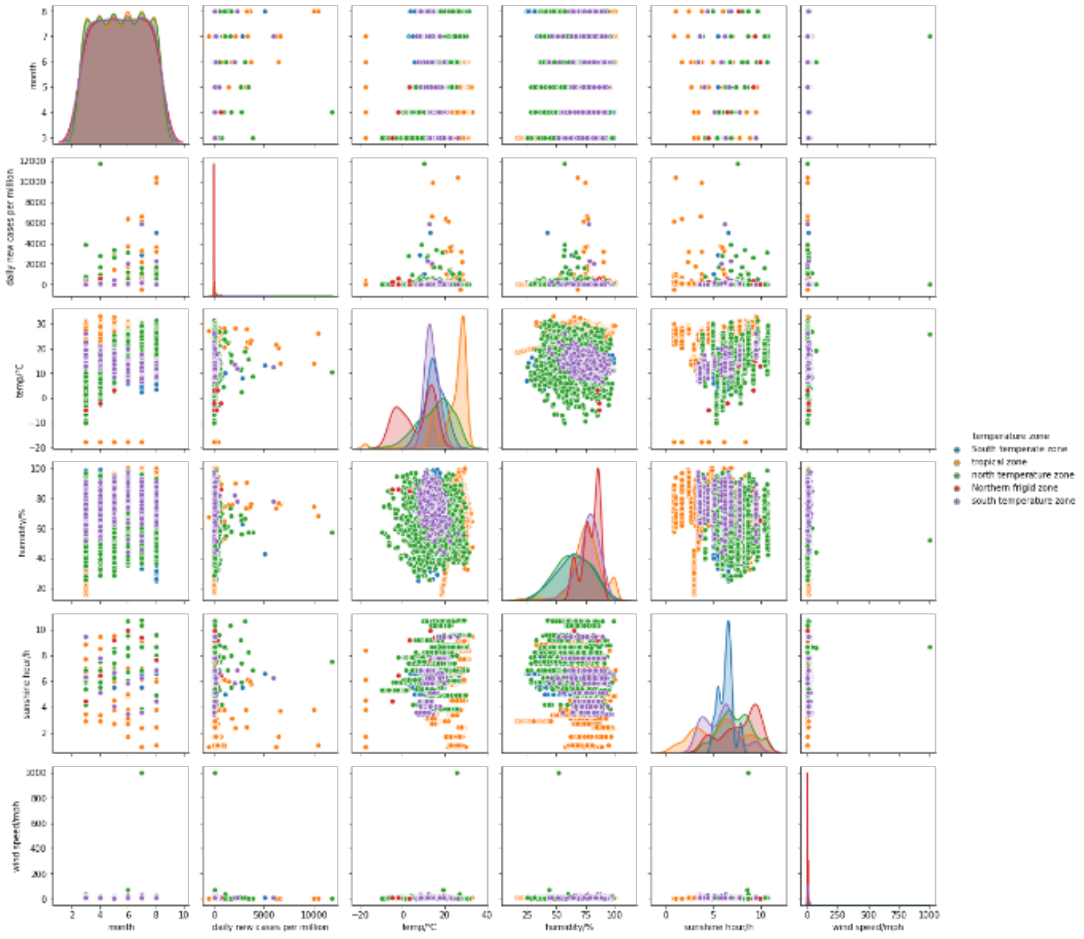
The collated data is as follows: https://docs.google.com/spreadsheets/d/1jBIjiwLqqrd-WpzH65_uxV5qRqRpAwuoacK50980KA0/edit?usp=sharing

Model design

When performing descriptive analysis, we first visualized the changes of each variable over time (cities are grouped by temperature zones, and the y-axis from left to right is temperature, humidity, sunshine time, wind speed, and daily new cases per million):

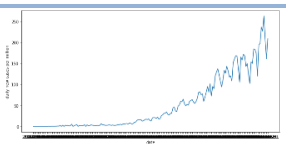
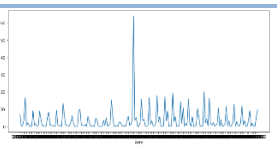
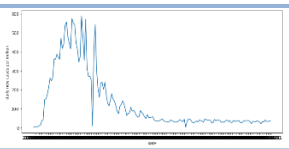
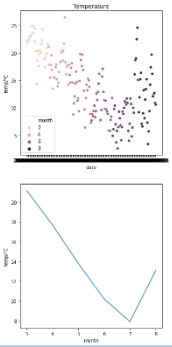
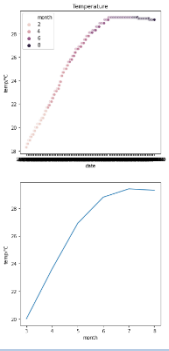
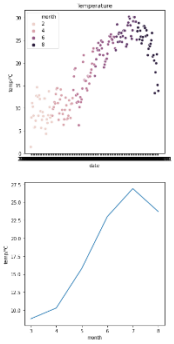
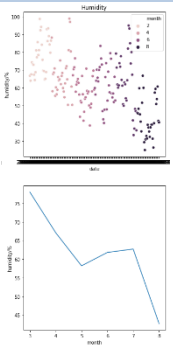
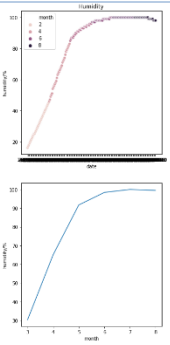
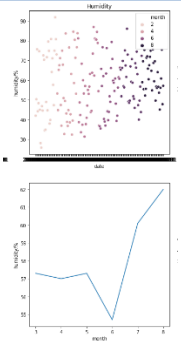
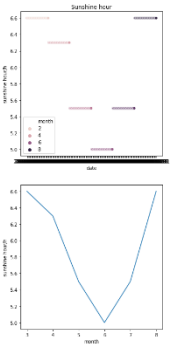
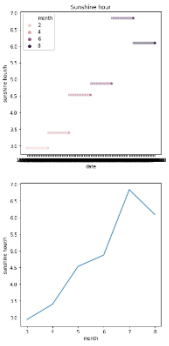
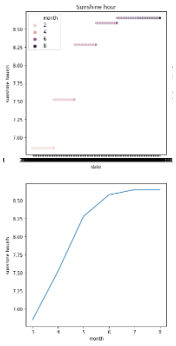
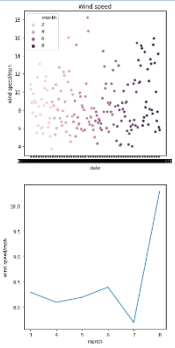
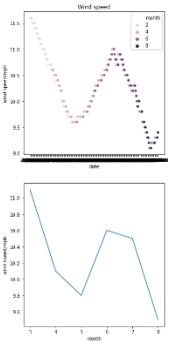
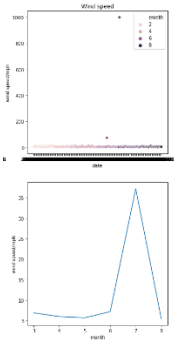


Also, we use the pair plot and heatmap to initially show the changes and relationships between these variables.

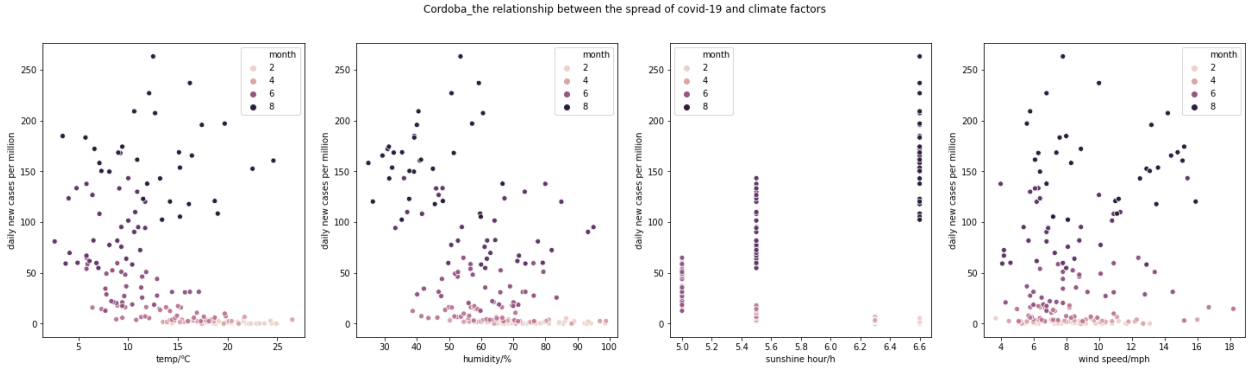


As can be seen from the figures, the data is mixed, and the correlation between weather factors and COVID is very weak. It seems the observing and analyzing of all 18 cities together does not make too much sense. This may be because different countries have different outbreak bases, different population densities, or different health policies during the pandemic, or maybe this is because they are in different climate zones. In the next, I will first conduct a descriptive analysis of each city one by one, hoping to find some common grounds.

Take City Córdoba, Hong Kong, and Now York City as examples: Cordoba of Argentina is located in the southern temperate zone, while Hong Kong is in the tropical zone, and New York in the north temperature zone.

	Cordoba	Hong Kong	New York
Daily new cases per million			
Temperature			
Humidity			
Sunshine hour			
Wind speed			

Since we want to study how do environmental factors affect the spread of covid-19, so we set four meteorological indices as input variables and new confirmed cases per million (DNCCM) as the output variable. To conduct correlation analysis, first we plot the change of DNCCM by each input variable. Take Cordoba city as the example:



To test how significant the correlations are, we set a series of hypothesis according to the plots above, such as "H0: There is a correlation between temperature and the spread of COVID-19 in Cordoba; Ha: There is NO correlation between temperature and the spread of COVID-19 in Cordoba." And then calculate the correlation coefficient R value and the p-value at 95% significant level. In the prediction part, we will set four input variables as the original input variables to build a multiple linear regression model for each city. By comparing the adjusted-r^2, p-value and error value, we can select a more appropriate regression model, and hopefully these regression models can help to predict the severity of COVID-19 with weather forecasting.

Objective 2:

Data sources, data cleansing and pre-processing

Data collected are:

1. Population – population density (people per sq.km of land area)
2. Gender – proportion of male and female in the population and in the total confirmed deaths
3. Age – age group of total confirmed deaths in a country
4. Race – proportion of race in a population and in the total confirmed deaths
5. Percentage of Mask Usage

Countries selection for data collection are based on the population density from each continent. 6 countries are being chosen from Europe, Asia, North America, South America and 3 countries are being chosen from Oceania and Africa.

The timeline for the data chosen are from 1st March 2020 to 31st August 2020.

Europe – Hungary, France, Finland, UK, Italy, Belgium
Asia – Mongolia, China, South Korea, Malaysia, India, Indonesia
North America – Haiti, Nicaragua, Panama, USA, Canada, Puerto Rico
South America – Uruguay, Bolivia, Brazil, Peru, Argentina, Ecuador
Oceania – Papua New Guinea, Australia, New Zealand
Africa – South Africa, Mauritius, Niger

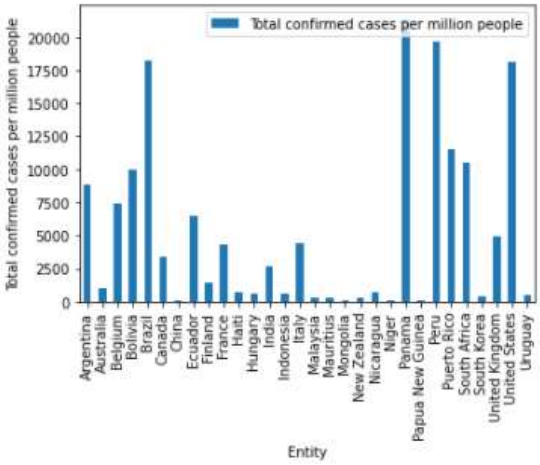
All countries will be present in the population factor for analysis. However, only certain countries are present for the factors such as gender, age, race and mask usage as data sources are limited

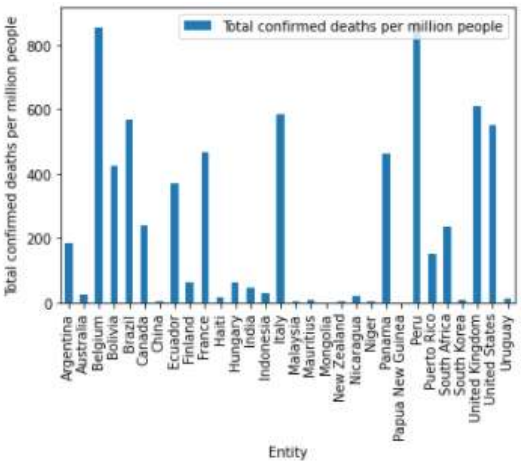
Steps for data cleaning and pre-processing:

1. Drop unnecessary columns which are not related to the model
2. Figure out columns with null values and remove them
3. Insert new columns to the data
4. Calculations for ratio by Microsoft Excel

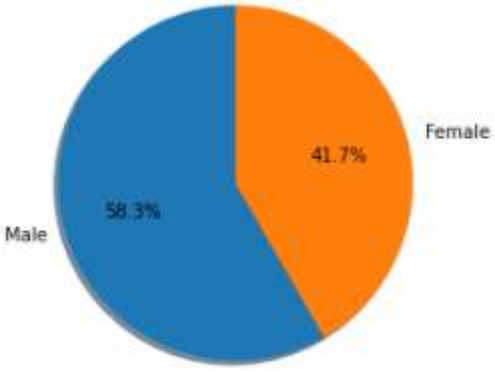
The data is available here:
https://drive.google.com/drive/folders/13Hig8WaQ_xuI5M2mSnYXmiwwWaEqk2Y4?usp=sharing

Model design

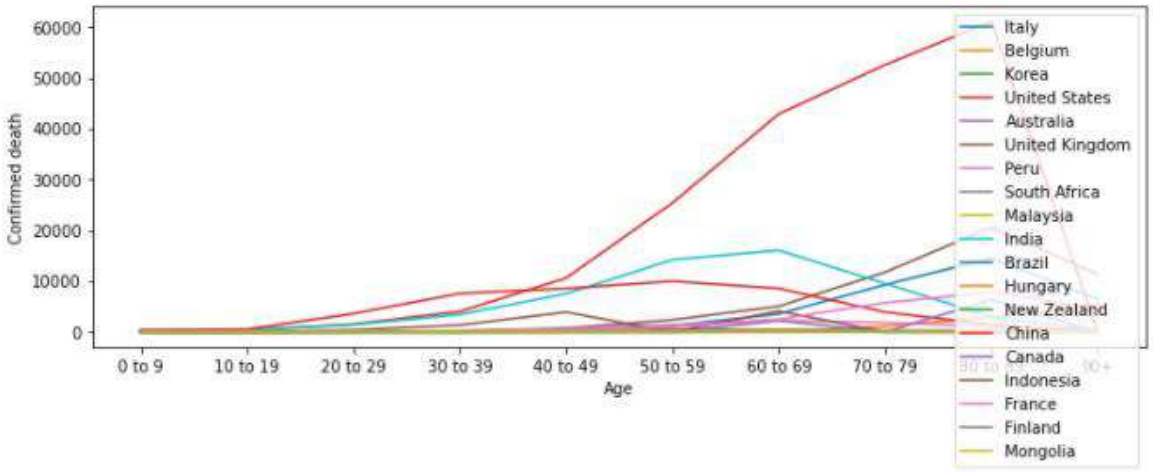




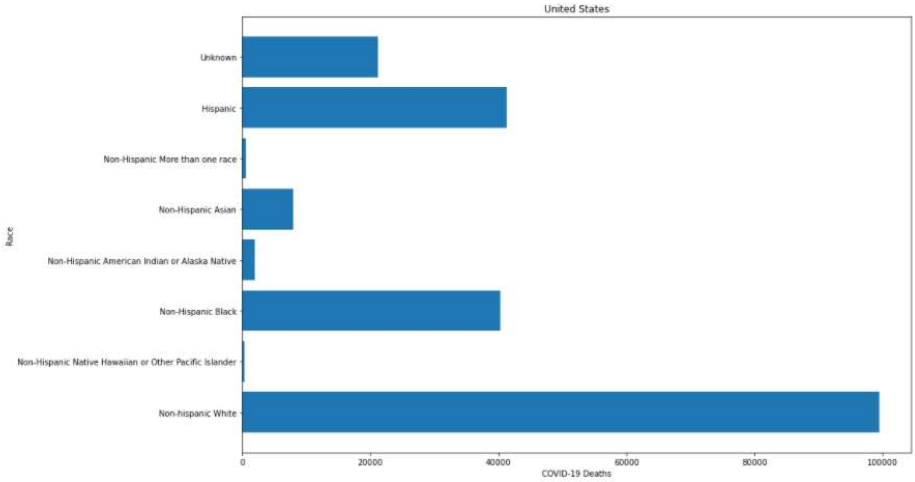
x = entity, y = total confirmed cases per million people
X = entity, y = total confirmed deaths per million people
By performing bar chart, we are able to visualize Panama shows the greatest total confirmed cases per million people while Peru shows the greatest total deaths per million people from March to August



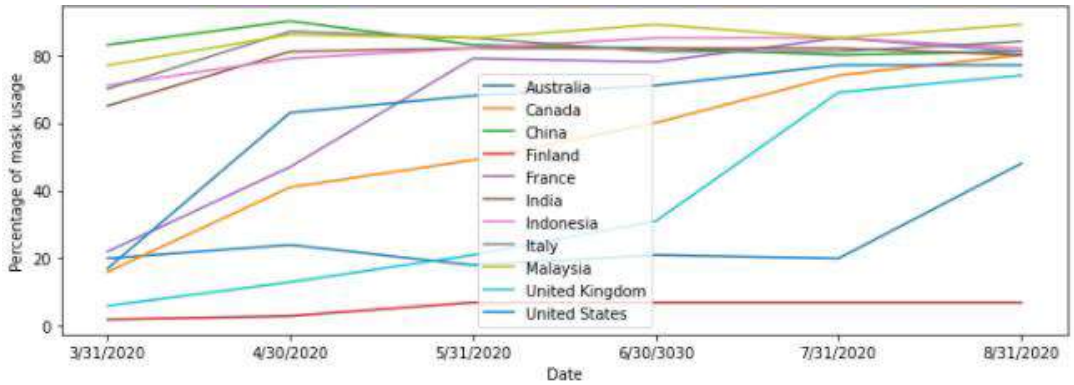
Pie chart shows the male possess a larger proportion in total confirmed death across all countries



x = age group
y = total confirmed deaths
The line chart shows the age 80 to 89 has the largest proportion of total confirmed deaths in all countries except for India which is age 60 to 69

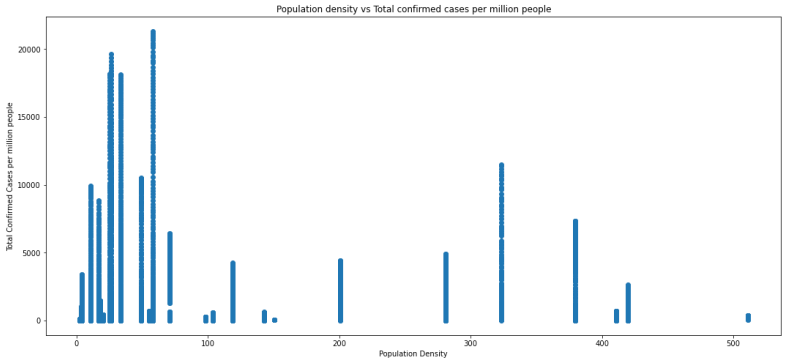


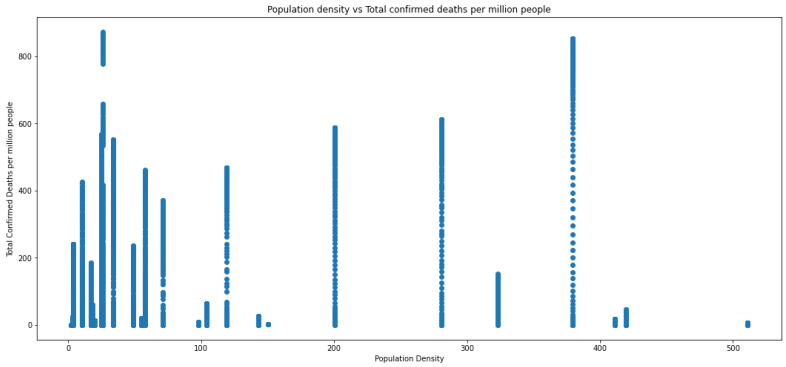
x = confirmed deaths, y = race
By using bar chart, we are able to figure out which race contributes to the highest confirmed deaths in United States

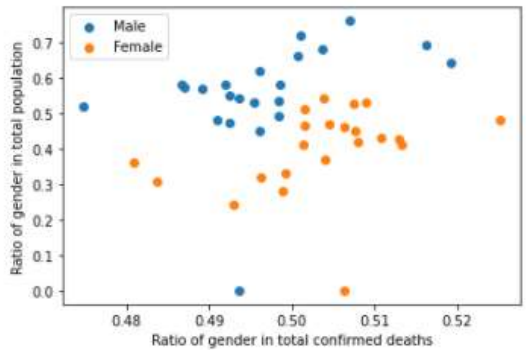


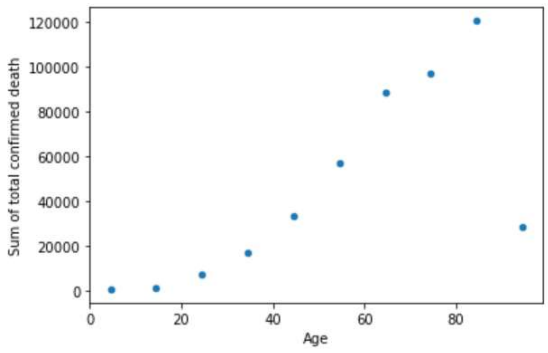
x = date, y = percentage of mask usage

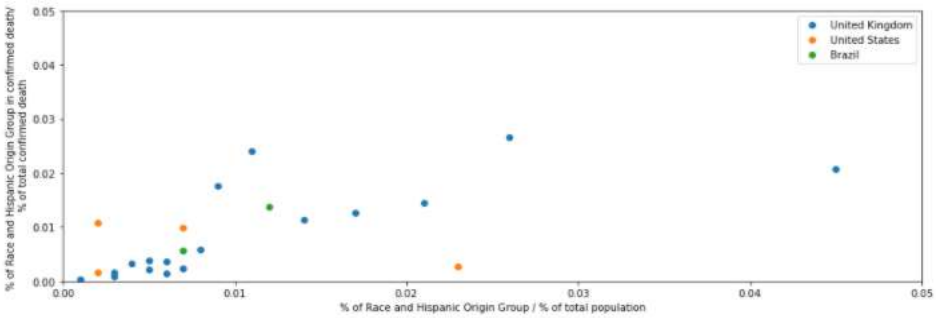
Line chart shows the percentage of mask usage over time for each country. We are able to visualize that China and Malaysia have consistently high percentage of mask usage while Finland has low percentage of mask usage

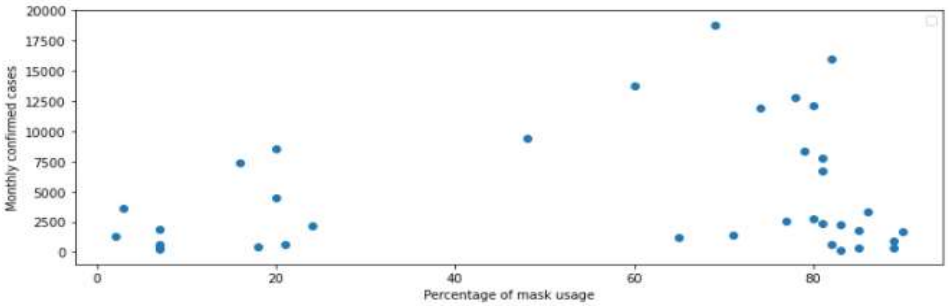












To determine whether both variables are correlated to each other, scatter plot is used to perform the correlation analysis. Firstly, null hypothesis and alternative hypothesis will be set up. Next, r-value and p-value are determined using python. The r-value is can be range from -1 to 1 which indicates that the relationship of the correlation. The p-value is used to show whether the correlation between variables are significant. Lastly, we are able to conclude if the null hypothesis should be accepted or rejected.

The following table shows the x and y variables for the correlation analytics:

Linear regression model is being used for population factor. By figuring out the R-square, Adjusted R-squared and F-statistics, we are able to determine how well does the model fits the data.

	Population	Gender	Age	Race	Percentage of Mask Usage
x - axis	Population density	Ratio of gender in total confirmed deaths	Mean age	Ratio of race and Hispanic origin group in total population	Percentage of mask usage
y - axis	Total confirmed cases / deaths per million people	Ratio of gender in total population	Total confirmed deaths	Ratio of race and Hispanic origin group in total confirmed deaths	Monthly confirmed deaths

Objective 3:

Data sources, data cleansing and pre-processing

Data was gathered from two main sources:

1. COVID-19 PPE Logistics: Data on what kind of products (masks, respirators, testing kits, sanitizers etc.) were exported to which county of California on which date. Published by the California Office of Emergency Services on the California Open Data Portal.
2. COVID-19 Cases: Data on the number of daily and cumulative COVID-19 cases and deaths on a daily basis for each county of California. Published by the California Department of Public Health on the California Open Data Portal.

The following data cleaning and pre-processing steps were applied:

1. Removal of NaN values - NaN values were removed from both CSV files (logistics and cases).
2. Removal of zero values - Some rows in the logistics data sheet were of “zero value” e.g.0 masks exported to XYZ county on 15 July. Such rows were removed.
3. Redistribution of miscellaneous values – Some exports were grouped under broad categories such as “Non-Governmental Entity”, or “State Department”, instead of listing an actual location such as Los Angeles. Such exports were uniformly distributed over all actual counties containing the product on that particular day.
4. Inner join – Finally, the two CSV files were merged through an inner join on the date and counties.

Model design

The exact hypotheses we will be testing are:

Is increased mask/other product import negatively correlated with daily new cases?

Similar hypotheses will be set up for all the other products such as sanitizers we wish to examine. In doing so, we will compare the correlation coefficients for all the products we examine, and finally conclude which product is most negatively correlated (i.e. most effective at preventing) with daily new cases/deaths.

In this procedure, we will first use summary statistics such as mean, median, mode, count, min and max along with basic visualization techniques such as bar and pie charts to first shortlist a number of products to use, as well as some counties in California with high number of cases to restrict our study to.

Then, for each of the shortlisted products and counties, we will use scatter plots and lines of regression as part of the regression-correlation analysis to visualize the

relationship between the increase in imports of the product and the change in cases/deaths.

Finally, the correlation coefficient for each of these products will be computed and compared to determine the most effective product. Here, significance will also be tested under 0.01 significance level:

H_0 = the correlation coefficient is close to 0

H_a = the correlation coefficient is significantly different from 0

In each of the final (i.e. regression-correlation) analyses/visualizations, the independent variable x will be the import quantities of a particular product, while the dependent variable y will be daily new cases/deaths and their respective ratios.

The data is available here:
https://drive.google.com/file/d/1gkpCLb3bg0uRJ_nWfNEiTIFI7AqrNBGL/view?usp=sharing

Objective 4:

Data sources, data cleansing and pre-processing

Data is collected from *The World Data,* includes:

- The number of newly confirmed cases (daily)
- The nature of health policies (closure, social distancing and testing) implemented

Data cleaning and pre-processing works of data are as follows.

1. Drop unnecessary columns from raw data
2. Removal of NaN values
3. Group some columns for better computation
4. Indicate policies' implementation nature with number

Policy/ Nature	no measures imposed	Recommended imposition	Required at some levels	Required at all levels
Closure	0	1	2	3
No. of regions/countries	3	9	6	2
Policy/ Nature	Restrict gathering >1000 people	Restrict gathering between 100-1000 people	Restrict gathering between 10-100 people	Restrict gathering <10 people
Social distancing	1	2	3	4
No. of regions/countries	2	5	3	10
Policy/ Nature	no measures imposed	Test who has symptoms and work in specific industries	Test who has symptoms	Open public testing
Testing	0	1	2	3
No. of regions/countries	3	11	4	2

5. Calculate the means of the confirmed cases of regions with different indicated numbers (done by Microsoft Excel)

The data is available here: https://drive.google.com/drive/folders/1XggTS7nRUMMS1--EUN83RwNG4zY7XxVT?usp=sharing

Model design

	Closure	Social distancing	Testing
x - axis	The indicators of policies nature		
y - axis	Newly confirmed cases		

Chapter 4: Result summary

Objective 1

Hypothesis-testing example (Cordoba):

H0: There is a correlation between temperature and the spread of COVID-19
Ha: There is NO correlation between temperature and the spread of COVID-19
r value = -0.41480365248849477
p value = 4.792476634131689e-09
Since p = 4.79*e-09 is much smaller than 0.05, we conclude that temperature is signific antly negatively correlated with the spread of COVID-19 (r= -0.415; p <0.05).
H0: There is a correlation between humidity and the spread of COVID-19
Ha: There is NO correlation between humidity and the spread of COVID-19
r value = -0.5185427291001568
p value = 4.7091025988357564e-14
Since p = 4.71*e-14 is much smaller than 0.05, we conclude that humidity is significantly negatively correlated with the spread of COVID-19 (r= -0.519; p <0.05).
H0: There is a correlation between sunshine hour and the spread of COVID-19
Ha: There is NO correlation between sunshine hour and the spread of COVID-19
r value = 0.1694218271987471
p value = 0.021497311484511786
Since p = 4.79*e-07 is smaller than 0.05, we conclude that sunshine hour is significantly positively correlated with the spread of COVID-19 (r= 0.169; p <0.05).
H0: There is a correlation between wind speed and the spread of COVID-19
Ha: There is NO correlation between wind speed and the spread of COVID-19
r value = 0.16337655640765406
p value = 0.02669455575225696
Since p = 4.79*e-07 is smaller than 0.05, we conclude that wind speed is significantly po sitively correlate with the spread of COVID-19 (r= 0.163; p <0.05).

On Python, we perform the hypothesis testing like the example shown above, the correlation coefficient and p value between the weather variables of each city and DNCCM are shown in the following table:

	City	Temperature		Humidity		Sunshine Hour		Wind Speed	
		r	p	r	p	r	p	r	p
north	Ivalo	-0.589	1.435e-18	0.553	3.652e-16	-0.361	4.508e-07	0.439	4.338e-10
	Moscow	-0.567	6.065e-17	0.097	~~0.191~~	0.748	1.937e-34	-0.195	0.008
	Quebec	-0.301	3.241e-05	-0.304	2.763e-05	-0.278	0.000	0.127	~~0.085~~
	New York	-0.653	1.471e-22	0.031	~~0.680~~	-0.634	6.516e-21	-0.040	~~0.598~~
	Rome	0.261	0.006	0.055	~~0.569~~	-0.102	~~0.291~~	0.034	~~0.728~~
	Seoul	-0.028	~~0.711~~	0.039	~~0.601~~	-0.196	0.008	-0.020	~~0.790~~
	Paris	-0.058	~~0.434~~	-0.089	~~0.228~~	-0.120	~~0.106~~	-0.061	~~0.413~~
	HK	0.210	0.004	0.150	0.042	0.382	9.457e-08	-0.060	~~0.422~~
	Mumbai	-0.472	7.755e-06	0.345	0.001	-0.643	7.132e-11	0.232	0.036
	Bogota	-0.276	0.000	0.038	~~0.608~~	0.655	7.777e-24	0.031	~~0.673~~
	KL	0.236	0.001	-0.127	~~0.084~~	0.297	4.195e-05	-0.078	~~0.296~~

south	Jakarta	0.068	0.360	-0.598	4.613e-19	0.890	1.975e-63	0.258	0.000
	Rio	-0.275	0.000	0.139	0.060	-0.468	2.427e-11	0.076	0.304
	Wellington	0.237	0.001	0.073	0.326	0.447	2.058e-10	0.0276	0.709
	Cape Town	-0.264	0.000	0.036	0.627	-0.310	2.252e-05	-0.037	0.0625
	Cordoba	-0.415	4.792e-09	-0.518	4.709e-14	0.169	0.021	0.163	0.027
	Melbourne	-0.311	1.757e-05	0.078	0.294	0.058	0.435	0.045	0.543

Then, according to the r and p-value, we established a regression model for each city. Still taking Cordoba as an example, the selection process of the model is as follows:

Model	R^2	Adjusted R^2	F-statistic	Error
DNNCM~ Temp+ Humidity+ Sunshine hour+ Wind speed	0.537	0.526	51.83	0.561
DNNCM~ Temp+ Humidity+ Sunshine hour	**0.534**	**0.526**	**68.66**	**0.561**
DNNCM ~ Temp	0.172	0.168	37.82	1.134
DNNCM ~ Temp^2	0.173			1.133
DNNCM ~ Humidity	0.269	0.265	66.94	1.065
DNNCM ~ Humidity^2	0.343			1.010
DNNCM ~ Sunshine hour	0.029	0.023	5.378	1.228
DNNCM ~ Temp+ Humidity	0.347	0.340	48.08	1.010
DNNCM ~ Temp+ Sunshine hour	0.424	0.418	66.64	0.948
DNNCM ~ Temp+ Sunshine hour	0.31	0.293	38.98	1.044

By comparing adjusted r^2, F-statistic and error value, the final multiple linear regression model we choose for city Cordoba is:
DNCCM = -91.1714 - 7.1659* temperature - 1.3987* humidity + 55.5186* sunshine hour
Other cities
e.g.
Hong Kong: DNCCM= -54.8411 + 3.5328* temperature - 0.4297* humidity
New York: DNCCM= 866.0055 - 8.6569* temperature - 70.6130* sunshine hour

Objective 2

Hypothesis testing result:

Population

The r-value and p-value are -0.0490 and 0.0004 respectively
The hypothesis is set up:
H0: There is a correlation between the total confirmed cases per million people and population density.
Ha: There is NO correlation between the total confirmed cases per million people and population density.

- The p-value is smaller than 0.05, hence we can conclude that the total confirmed cases per million people is significantly negative correlated with population density. The null hypothesis is accepted.

The r-value and p-value are 0.1723 and 9.393 respectively.
The hypothesis is set up:
H0: There is a correlation between the total confirmed deaths per million people and population density.
Ha: There is NO correlation between the total confirmed deaths per million people and population density.

- The p-value is greater than 0.05, hence we can conclude the total confirmed deaths per million people is not correlated with population density. The null hypothesis is rejected.

Gender

The r-value and p-value for male are 0.3779 and 0.0829 respectively.
The hypothesis is set up:
H0: There is a correlation between ratio of male in total confirmed deaths and ratio of male in total population.
Ha: There is NO correlation between ratio of male in total confirmed deaths and ratio of male in total population.

- The p-value is greater than 0.05, hence we can conclude that the ratio of male in total confirmed deaths is not correlated with the ratio of male in total population. The null hypothesis is rejected.

The r-value and p-value for female are 0.3450 and 0.1159 respectively.
The hypothesis is set up:
H0: There is a correlation between ratio of female in total confirmed deaths and ratio of male in total population.
Ha: There is NO correlation between ratio of female in total confirmed deaths and ratio of male in total population.

- The p-value is greater than 0.05, hence we can conclude that the ratio of female in total confirmed deaths is not correlated with the ratio of female in total population. The null hypothesis is rejected.

Age

The r-value and p-value are 0.7498 and 0.0125 respectively.
The hypothesis is set up:
H0: There is a correlation between total confirmed death and age.
Ha: There is NO correlation between total confirmed death and age.

- The p-value is less than 0.05, hence we can conclude that total confirmed death is significantly positively correlated with age. The null hypothesis is accepted.

Race

United Kingdom: r-value = 0.999, p-value = 3.926
United States: r-value = 0.9723, p-value = 5.193e-05
Brazil: r-value = 0.9663, p-value = 0.0074

The hypothesis is set up:
H0: There is a correlation between the ratio of race and Hispanic group in total population and ratio of race and Hispanic group in total confirmed deaths.
Ha: There is NO correlation between the ratio of race and Hispanic group in total population and ratio of race and Hispanic group in total confirmed deaths.

- The p-value for United Kingdom is greater than 0.05, hence we can conclude that the ratio of race and Hispanic group in total population is not correlated with the ratio of race and Hispanic group in total confirmed deaths. The null hypothesis is rejected.

- The p-value for United States and Brazil are smaller than 0.05, hence we can conclude that the ratio of race and Hispanic group in total population is significantly positively correlated with the ratio of race and Hispanic group in total confirmed deaths. The null hypothesis is accepted.

Percentage of Mask Usage

The r-value and p-value are 0.1771 and 0.1548 respectively.
The hypothesis is set up:
H0: There is a correlation between total confirmed cases and percentage of mask usage.
Ha: There is NO correlation between the total confirmed cases and percentage of mask usage.

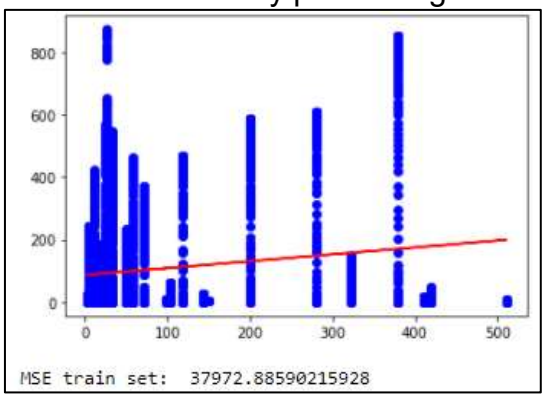
- The p-value is greater than 0.05, hence we can conclude that the total confirmed cases is not correlated with percentage of mask usage. The null hypothesis is rejected.

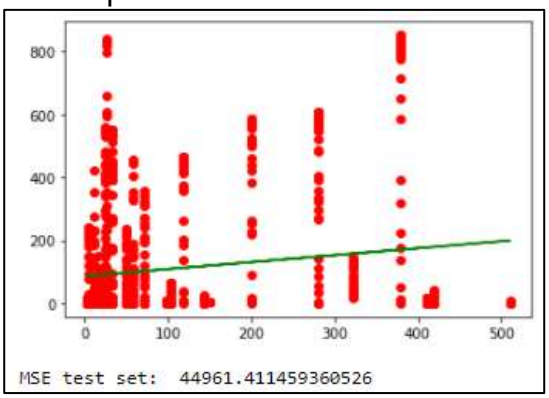
The regression equation for the population factor is:

$$Total\ confirmed\ cases\ per\ million\ people = 2033.2532 - 1.1276\ Population\ Density$$

The R-squared and Adjusted R-squared are 0.002 while F-statistics is only 12.49. This tells us that the model does not fit the data.

We can visualize by performing the train test split:





Based on the plotted graph, even though the correlation is significant, but the data shows no relationship between the two variables. The mean squared error calculated are very large.

Objective 3

Step 1: Descriptive Analytics

Summary:

- The counties with the most number of cases were found to be Los Angeles, Riverside, San Bernardino, Orange and San Diego.
- The products with the maximum number of imports were found to be Surgical Masks, N-95 Respirators, Hand Sanitizers, Examination Gloves and Surgical or Examination Gowns.
- Every product was found to be roughly negatively associated with the number of cases.
- The date range for analysis was selected as July 08 2020 – September 28 2020.

Detailed steps with screenshots:

In the descriptive analytics step, basic tools such as summary statistics and charts were used to narrow down the counties and products, and to get a general feel for the data.

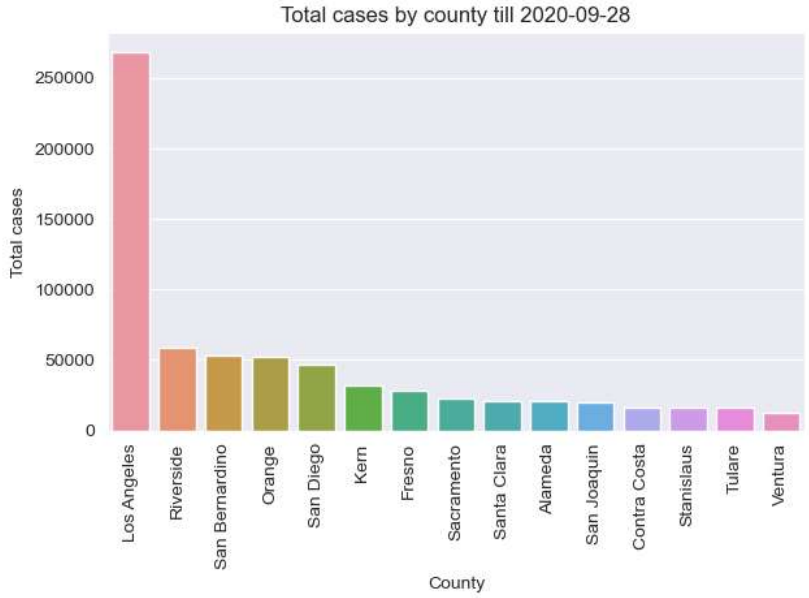
The Python application was first used to generate a list of the various counties, products and date range:

```
The range of dates is:
First date:  2020-06-08 00:00:00
Last date:  2020-09-28 00:00:00

Press ENTER to continue...

The various counties in the data are:
['Alameda', 'Alpine', 'Amador', 'Butte', 'Calaveras', 'Colusa', 'Contra Costa', 'Del Norte',
 'El Dorado', 'Fresno', 'Glenn', 'Humboldt', 'Imperial', 'Inyo', 'Kern', 'Kings', 'Lake', 'Lassen',
 'Los Angeles', 'Madera', 'Marin', 'Mariposa', 'Mendocino', 'Merced', 'Modoc', 'Mono', 'Monterey',
 'Napa', 'Nevada', 'Orange', 'Placer', 'Plumas', 'Riverside', 'Sacramento', 'San Benito',
 'San Bernardino', 'San Diego', 'San Francisco', 'San Joaquin', 'San Luis Obispo', 'San Mateo',
 'Santa Barbara', 'Santa Clara', 'Santa Cruz', 'Shasta', 'Sierra', 'Siskiyou', 'Solano', 'Sonoma',
 'Stanislaus', 'Sutter', 'Tehama', 'Trinity', 'Tulare', 'Tuolumne', 'Ventura', 'Yolo', 'Yuba']

Press ENTER to continue...


The various products in the data are:
['Cloth Masks', 'Coveralls (Hospitals or EMS)', 'Examination Gloves', 'Face Shields (Disposable)',
 'Goggles', 'Hand Sanitizers', 'KN95 Respirators', 'Lab Supplies', 'Medical Equipment',
 'N-95 Respirators', 'Other / None of the above', 'Pharmaceuticals', 'Surgical Masks',
 'Surgical or Examination Gowns', 'Swabs', 'Test Kits', 'Ventilators', 'Viral Testing Media',
 'Wipes', 'Sample Collecting Kits', 'Personnel', 'Shoe Covers', 'Cleaning Supplies', 'Lab Kit',
 'Beds/Cots', 'Body Bags']

Press ENTER to continue...
```

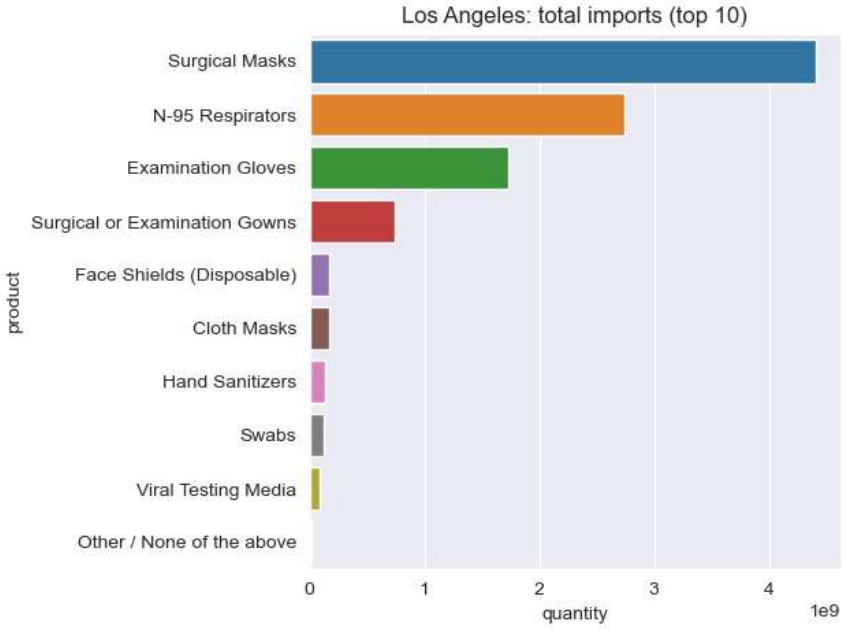
However, it does not make sense to analyze every available county and product. Instead, we may choose to focus our analysis on a few counties with the most number

of cases, and a few products which have the maximum number of imports in these counties. Bar charts were used for this purpose.

First, a bar chart of the total cases in each county helped narrow down the five cases with the most number of cases: **Los Angeles, Riverside, San Bernardino, Orange** and **San Diego.**

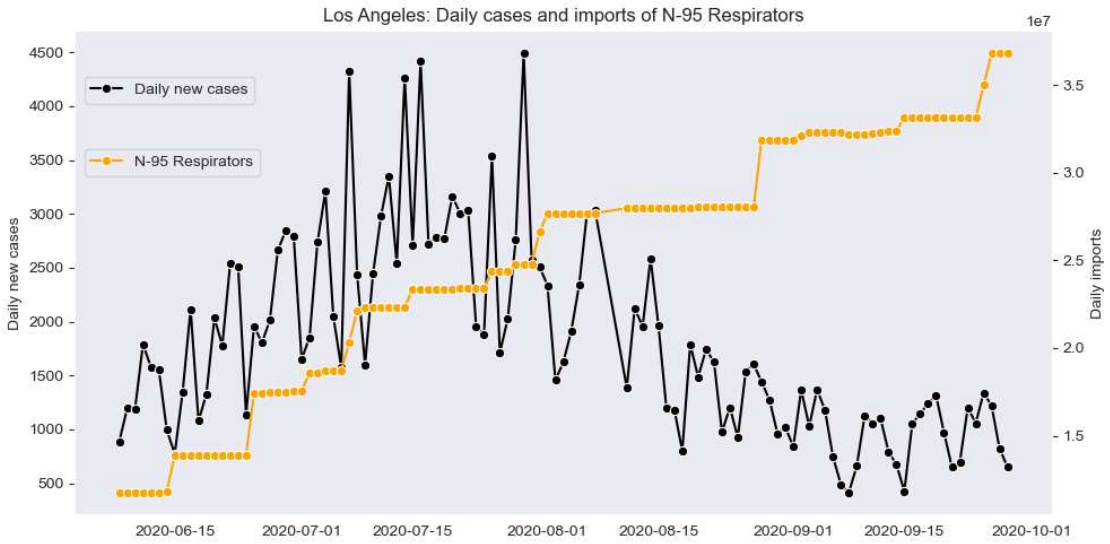


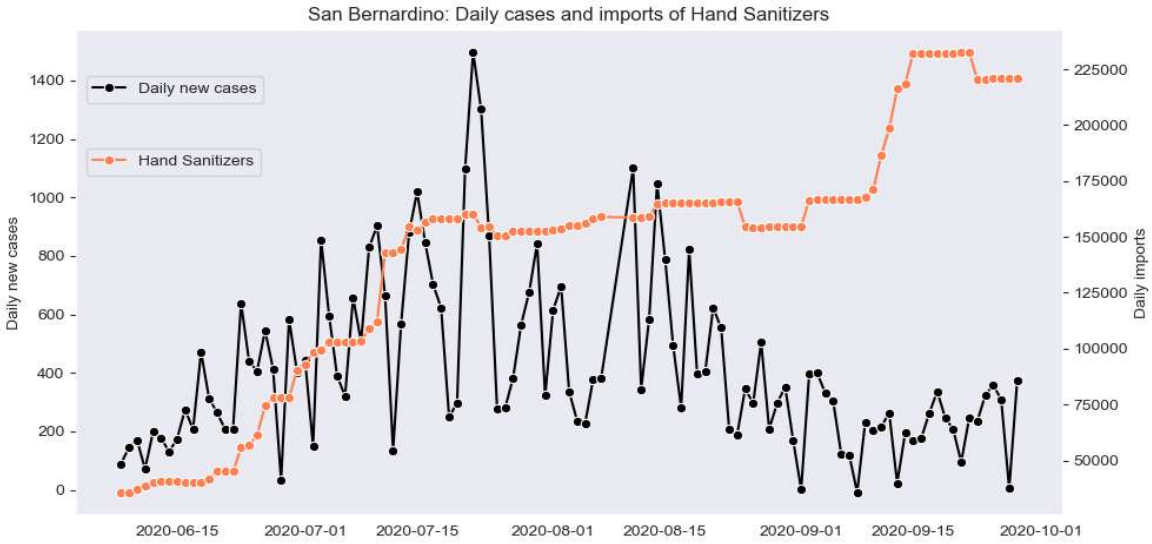
Next, another bar chart was used to filter the products which had the maximum number of imports in each of these counties. One such chart is shown below:



Using such charts for each of the five counties, the five most exported products were found to be **Surgical Masks, N-95 Respirators, Hand Sanitizers, Examination Gloves** and **Surgical or Examination Gowns.**

Next, line plots were used to visualize the general change in cases and imports for each county and product pair. Two such plots are shown below.





These plots served two purposes. First, they showed that each of the products were **roughly negatively associated** with the number of daily new cases. Secondly, they showed that during the first 30 days of this data (June 08 – July 08), the cases and the imports were still reaching a steady state. Thus, it would be inappropriate to analyze this period as the effects of the products cannot be seen immediately. Thus, the date range for analysis was chosen as **July 08 2020** to **September 28 2020.**

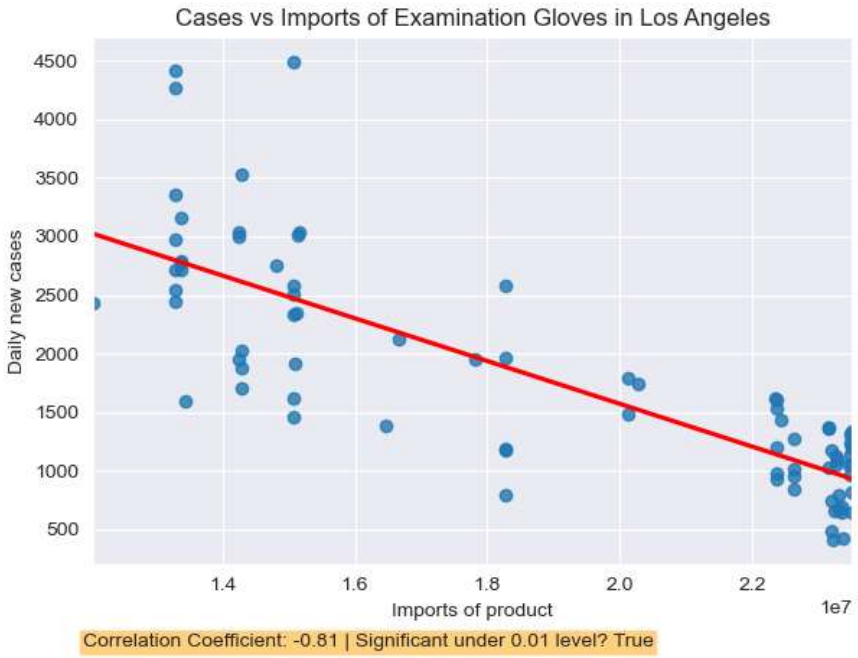
Step 2: Correlation Analysis

Summary:

- All the products were found to be **significantly negatively correlated** with daily number of cases under **0.01 significance level.**
- **Examination gloves** had the most negative coefficient of correlation, followed by **respirators, sanitizers** and **gowns.**

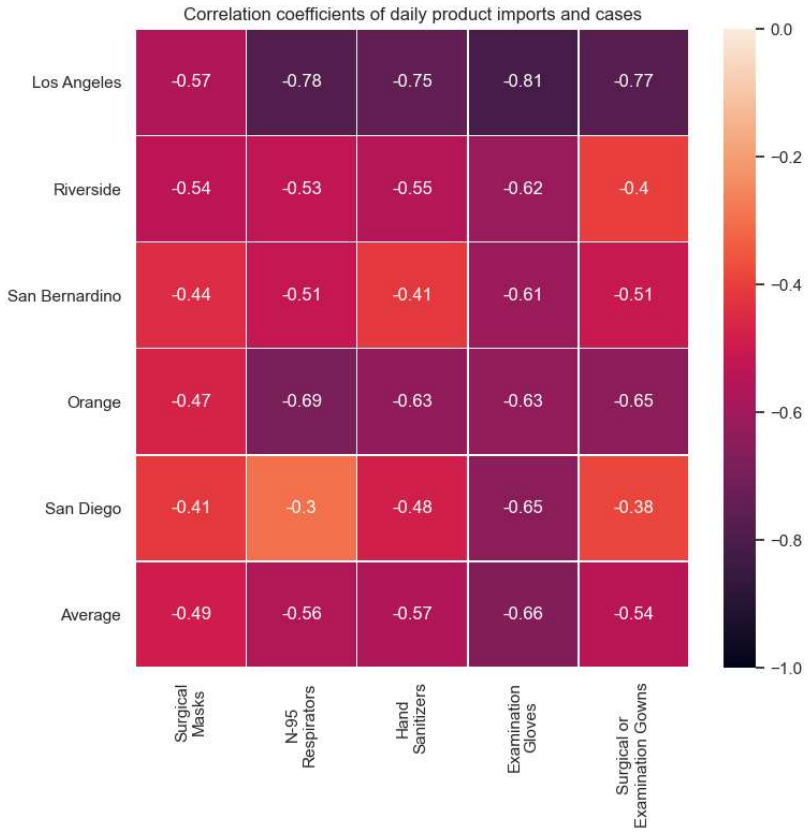
Detailed steps with screenshots:

Scatter plots with lines of best fit were drawn for each of the products in each county. One example (gloves in Los Angeles) is shown below. Furthermore, their correlation coefficients were calculated, and hypothesis testing was conducted under 0.01 significance level.



It was found that all the products were **significantly negatively correlated** with daily new cases.

Next, a heatmap was drawn to compare the various correlation coefficients in an effective manner.



As it can be seen, on average **examination gloves** were found to have the most negative correlation coefficient (i.e. most effective), followed by **sanitizers, gowns** and **respirators. Masks** has the least negative coefficient of correlation on average.

Step 3: Regression Analysis

Summary:

- A linear model consisting of only **imports of Examination Gloves** was found to be the best predictor of **daily new cases.**
- The model was found to be **significant**. The **adjusted R-squared** was 0.658, while the **F-statistic** was 153.34.

Detailed steps:

As Los Angeles has the highest number of cases, it was selected for building the linear regression model. The output variable was the number of **daily new cases**, and the input variables were various combinations of the **five products** discussed in the previous section. Thus, 31 different models were evaluated. The five best models are summarized below:

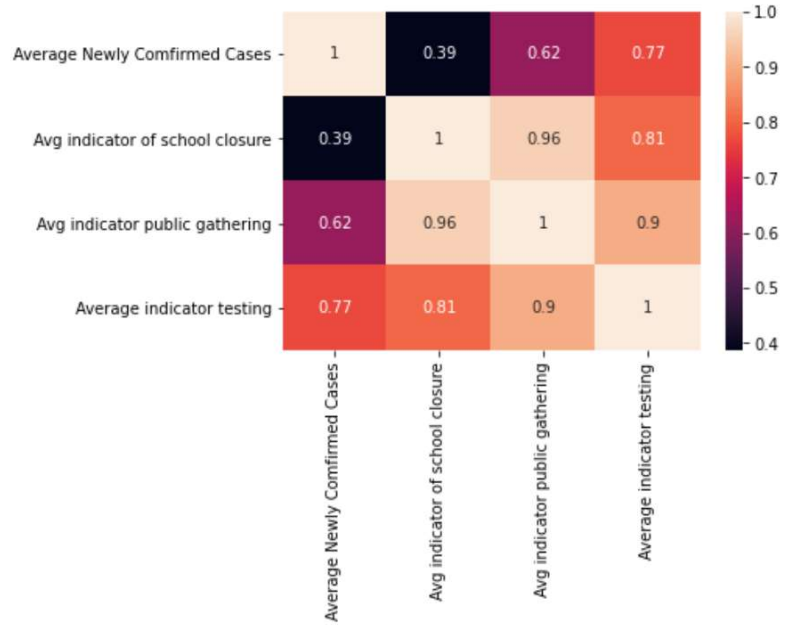
Predictors used	Adjusted R-squared	Error	F-statistic
Examination Gloves	0.658583096	0.315959093	153.38866
N95 Respirators	0.600862731	0.341624878	119.9268942
Surgical/Examination Gowns	0.589889228	0.346289192	114.6308827
Hand Sanitizers	0.556492644	0.36011297	100.1255686
Surgical Masks + Examination Gloves	0.666376693	0.312332042	79.89700378

Thus, the model consisting of only **examination gloves** as the predictor was found to be the most accurate model.

Objective 4

Descriptive Analytics

The brief overview the relationship between variables:



Correlation Analysis

Hypothesis testing results:

- Setting up the hypothesis

H0: There is a correlation between closure policy/ restriction on public gathering/ testing policy and the spread of COVID-19

Ha: There is NO correlation between closure policy/ restriction on public gathering/ testing policy and the spread of COVID-19

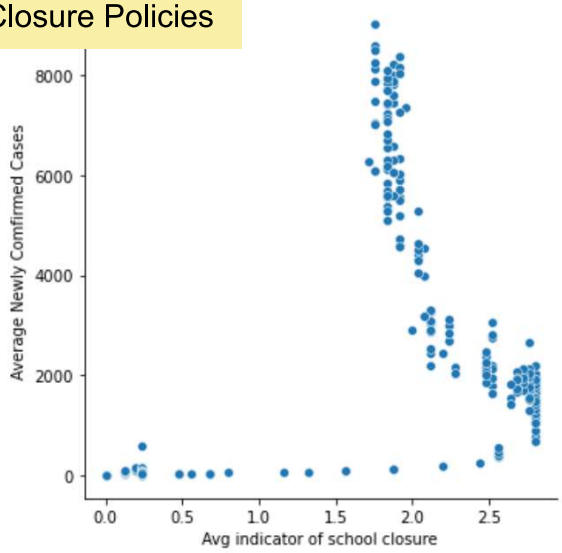
- Calculating the r-value and p-value

Indicators/ Policy	Closure	Restriction on public gathering	Testing
r-value	0.387	0.617	0.768
p-value	3.99^-10	5.06^-27	1.22^-48
relationship	+ve	+ve	+ve

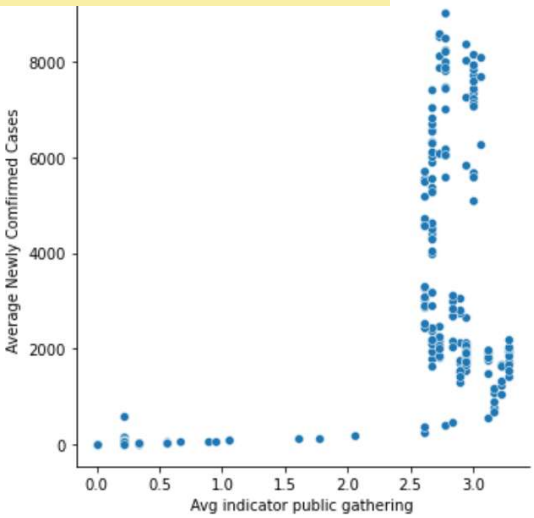
➔ The r-values show the relationship different health policies and spread of the COVID-19. Since the p-values are smaller than 0.05, we can conclude that the health policies are significantly correlated to the spread of the COVID. We accept the null hypothesis.

- Data visualization:

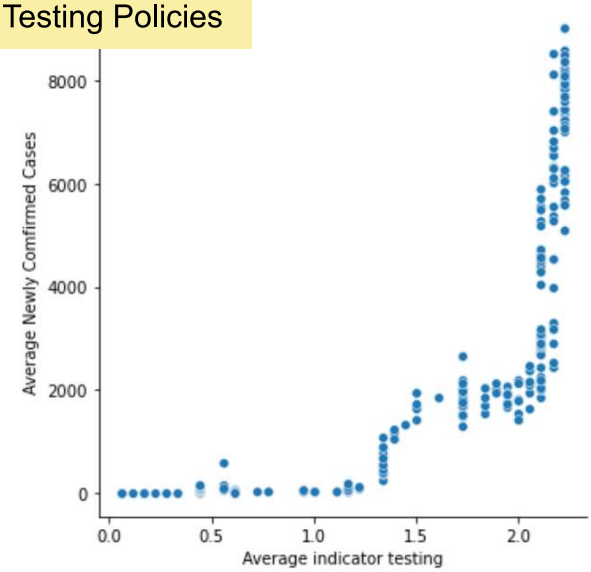
Closure Policies



Restriction on public gathering



Testing Policies



Linear Regression

- Model selection:
  There are 7 possible regression models:
  1. Newly confirmed cases ~ closure policy
  2. Newly confirmed cases ~ restriction on public gathering
  3. Newly confirmed cases ~ testing policy
  4. Newly confirmed cases ~ closure policy + restriction on public gathering
  5. Newly confirmed cases ~ closure policy + testing policy
  6. Newly confirmed cases ~ restriction on public gathering + testing policy
  7. Newly confirmed cases ~ closure policy + restriction on public gathering + testing policy

To select the best model, we would compare the adjust R^2 and the F-statistic. Finally, we found out the best model is the 8. Newly confirmed cases ~ closure policy + restriction on public gathering + testing policy, with R-squared = 0.942; Adj. R-squared = 0.941 and F-statistic = 1309

This tells us that the model quite fit the data.

OLS Regression Results

Dep. Variable:	Average Newly Comfirmed Cases	**R-squared (uncentered):**	0.942
Model:	OLS	**Adj. R-squared (uncentered):**	0.941
Method:	Least Squares	**F-statistic:**	1309.
Date:	Sun, 06 Dec 2020	**Prob (F-statistic):**	7.87e-149
Time:	10:59:11	**Log-Likelihood:**	-2010.8
No. Observations:	244	**AIC:**	4028.
Df Residuals:	241	**BIC:**	4038.
Df Model:	3		
Covariance Type:	nonrobust		

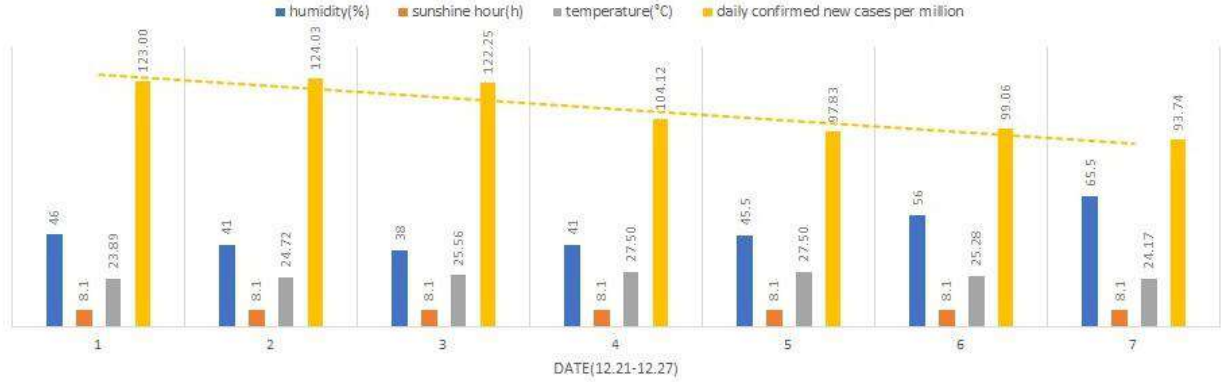
	coef	std err	t	P>|t|	[0.025	0.975]
Avg indicator of school closure	-5780.7762	211.723	-27.303	0.000	-6197.840	-5363.712
Avg indicator public gathering	5041.3644	234.269	21.520	0.000	4579.888	5502.841
Average indicator testing	1322.5296	163.395	8.094	0.000	1000.665	1644.394

Omnibus:	3.916	**Durbin-Watson:**	0.447
Prob(Omnibus):	0.141	**Jarque-Bera (JB):**	3.694
Skew:	0.218	**Prob(JB):**	0.158
Kurtosis:	3.415	**Cond. No.**	19.0

Chapter 5: Data Interpretation and Discussion

Objective 1

For each city, we can obtain its prediction model separately. For example, for Cordoba, the following regression model was selected: DNCCM = -91.1714-7.1659*temperature-1.3987*humidity+55.5186*sunshine hour. Based on this, we can bring in future weather data, make simple predictions on the spread of the pandemic, and adjust the prevention measures in advance when necessary. The trend line below shows the change of the number of daily new confirmed cases per million people according to Cordoba's weather forecast during the week from December 21 to 27, 2020.



From the result, it can be seen that the correlation between four meteorological factors and the pandemic situation in each sample city is very different. In Cordoba City, which we have been using as the example, the four x variables are all related to the spread of COVID-19. However, in some other cities, such as Paris, the result based on existing data shows that at 95% significant level, all four x variables have no relationship with the spread of COVID-19.

In addition, comparing four x variables, temperature and sunshine hour are correlated with the pandemic in most cities, while humidity and wind speed are not correlated at 95% significant level in many cities. Among them, the correlation between humidity and COVID-19 is somehow random and does not show a certain rule for the time being, but once the two are correlated, the correlation is generally strong.

The correlation between temperature and the spread of COVID-19 is generally negative, and from the current results, the higher the latitude, the stronger the correlation (the greater the |r|). Based on this, we can predict that the spread of COVID-19 will be more serious in cold weather. Therefore, countries or regions in such weather conditions should pay more attention to the prevention and control of the pandemic. As can be seen from the charts above, in the summer, pandemic situation in some northern hemisphere countries improves, while in southern hemisphere countries become serious. This should be partly due to temperature changes. However, when the northern hemisphere ushered in winter again, people need to pay more attention again to prevent the pandemic from rebounding.

Objective 2

According to the results, we know that population density is positively correlated to the total confirmed cases but not correlated to the total confirmed deaths. This is because the spread of covid-19 is by airborne transmission. Hence, when the population density in a country is high, people are tended to stay closer causing the increase in total confirmed cases. It is not related to deaths as deaths are normally caused by poor health conditions. Data shows more male are dying from Covid-19 worldwide than female. This may due to male are more likely to smoke and hence causing their lungs to be more vulnerable. Besides, race in United States and Brazil shows positive correlation to the total confirmed deaths. Some people from racial and ethnic minority group have different eating or living conditions which increase the chances of getting infected and facing deaths. Age is positively correlated to the total confirmed deaths. Age 80 to 90 shows the greatest confirmed deaths among all age groups. This is because the elderly immune systems are not as good as the teenagers and adults at reacting to the coronavirus. Moreover, elderly more likely to have poor health conditions such as heart disease, diabetes or kidney disease which further weaken their bodies' abilities to fight with the disease. For the social factor, there is no correlation between the mask usage and confirmed deaths as masks are used for protection of healthy persons or to prevent onward transmission but do not imply that people will not infect with COVID-19.

Objective 3

According to the results, it was found that **all the products discussed** are significantly negatively correlated with the daily new cases. This validates present day research, as the analysis shows that increased usage of such equipment reduces the daily number of new cases. This cements the recommendations made by authorities and governments on recommendations and regulations on wearing masks, and regularly using sanitizers.

Furthermore, the analysis finds that **gloves** are the most negatively correlated product, followed by **respirators** and **gowns.** The likeliest interpretation for this is that these products are typically used as **medical equipment** and mostly worn by frontline medical workers and nurses. This suggests that **adequate protection and preparation** for **medical workers** may result in a **better handling** of the pandemic, as the doctors are able to treat cases more effectively, **thus controlling the spread.** On this basis, we may infer that in times of severe outbreak, when the supply of protective equipment is limited, it may be a wise choice to **prioritize the distribution of protective equipment to frontline workers such as doctors and nurses.**

Lastly, the regression analysis also finds that the most effective predictor of daily new cases is the import quantities of **gloves**. This corroborates the interpretations made in the previous paragraph, as the import of gloves is likely to be influential in controlling the spread of the virus.

Objective 4

Even though health policies and spread of COVID show a positive relationship, however, it does not mean that the health policies are ineffective in fighting the COVID. It is because the governments around the globe started to increase tightness of health measures (i.e. change from recommended to compulsory nature) when there is an increasing number of confirmed cases since they believe such changes can help relieve the seriousness of the COVID spreading. Take Italy as an example, it changed its closure policy only when the number of weekly newly confirmed cases increased 18000. Therefore, increase in tightness of measure associates with increase in newly confirmed cases, and that is also the reason why we can see a positive relationship between tightness of closure policy and the spread of the COVID.

In addition, it does take time to see the effectiveness on controlling the spread of the COVID, therefore, it is understandable that there will not be a sudden drop in newly confirmed cases right after changing the tightness of the health measures. Take Wuhan, China as an example, the newly confirmed cases drop 11.3% after the government change the closure policy to a compulsory level for 76 days. Therefore, longer the period of data collected, better the assessments of the effectiveness of the health policy. Nevertheless, due to limited source of data, we cannot show a more comprehensive assessment and we would like to add it to the future to-do list.

Fortunately, we can still evaluate the effectiveness of closure policy and restriction on public gathering from the graphs. We can see that there is a slightly decreasing slope at the tail part. Therefore, we can still conclude that the closure policy and restriction on public gathering are effective in controlling the spread of the COVID.

Similar case applied to testing policy as well, yet, the graph of it cannot show due to limited data.

Chapter 6: Conclusions

Objective 1

Basically, this part has achieved the goal: The relationship between weather factors and the spread of COVID-19 has been obtained through correlation analysis, and the regression models predicting the future trend of the pandemic (from weather perspective) can be figured out. However, in the process of research, there are two issues: First, the regional nature of the data exists, that is, the pandemic situation of each sample city is personalized. Even at similar latitudes, with similar temperature and other meteorological indicators, different cities have different COVID-19 trends. During the researching process, I tried to group the samples in some ways, but the correlations obtained did not strengthen much. Therefore, the initial idea of getting a global forecasting model may not be in line with the reality form the perspective of weather. The second question is about the regression modeling. The data is somehow scattered, so when improving the model, I tried to use forms other than linear, such as power two form, exponential form, and logarithmic form. When the model has one-variable, the quadratic form is more effective. However, since I lack the knowledge of building a multiple non-linear regression model, the final result is still a multiple linear regression model. In further study, we may consider the establishment of multiple nonlinear regression model.

What's more, as for the reasons behind the strong regional nation, we think it is because the impact of meteorological factors on the spread of COVID-19 is somehow limited, and the correlation is weak. Other influencing factors may have a greater impact, such as the wearing of masks and the implementation of prevention policy, etc. And this is why we take meteorological factors within our first research objective. In the following objectives, we get to know more correlations between COVID-19 and variables in other factors, which can help us understand the spread of the COVID-19 more comprehensively.

Objective 2

Even though we are able to show that population density is correlated to the confirmed cases of Covid-19, the linear regression model is unable to show the plots as the data for population remained unchanged from March to August. Hence, vertical plots can only be obtained. Population data could be recorded daily in order to increase the accuracy of the result. Similarly, confirmed cases and deaths for age, gender and race are limited across countries and they are not updated daily or monthly. This will cause bias in the result as unsimilar timeframe of data is considered. Moreover, measurements for mask usage in the community may contain errors as there is no data on county-level mandates for wearing masks in public. Therefore, different methods are used to build a multi-regression model but could not achieve anything as data obtained is limited. By obtaining precise data with the same timeframe which is updated daily or monthly across all countries, we believe that cultural and social factors can be shown to have a strong relationship to the spread of Covid-19.

Objective 3

Overall, the data analysis has achieved the goals set in the preliminary chapters. It has corroborated existing research by showing that the usage of protective equipment such as masks, gloves, respirators etc is negatively correlated with the number of daily new cases. Furthermore, it has been able to establish a clear “most effective equipment” – examination gloves.

However, the study has a few caveats:

- The analysis assumes that there is a one-to-one correspondence between the **import** of these products, and the **usage** of these products i.e. import data is used as a proxy for usage data. This assumption is made as usage data for equipment is typically not available. While this assumption is *roughly* valid, it may not always hold true e.g. in the case of government stockpiling.
- Much of the contemporary research that has been cited has focused around the effectiveness of **masks** in preventing the spread of the virus. However, this study was not able to establish very strong evidence in favour of the usage of masks. The likely cause for this is that mask import is **overwhelmingly higher** than the other products, thus leading to a perceived *lower* correlation coefficient.

Thus, while this data analysis serves as a good starting point for analyzing and proving the effectiveness of various PPE (personal protective equipment), further research based on scientific experiments and smaller groups is required.

Objective 4

Among the health policies we studied, the effectiveness of closure policy is the most obvious, follows by the restriction on the public gathering and testing policy. Therefore, we can conclude that these health policies are useful in controlling the spread of the COVID.

However, there are some limitations when we carry out the evaluations. For instance, since it may take some times to see the effectiveness of the health policies, while the time period we have studied is rather short, so we may under-state the effectiveness of health policies. With longer time span studied, more comprehensive and accurate conclusions can be made, and we would like to add this into the future to-do-list.

Chapter 7: References

Objective 1:

- Chinese Academy of Sciences. (2003, September 28). Meteorological factors are closely related to the SARS epidemic. Retrieved from http://www.cas.cn/zt/kjzt/fdgx/zjpl/200309/t20030928_1710323.shtml
- Eslami, H., & Jalili, M. Author's initials. (2020). The role of environmental factors to transmission of SARS-CoV-2 (COVID-19). 10.1186/s13568-020-01028-0.

Objective 2:

- European Centre for Disease Prevention and Control (ECDC) (2020) Total Confirmed Deaths due to Covid-19 per million people. Retrieved from https://ourworldindata.org/grapher/covid-19-death-rate-vs-population-density
- National Center for Health Statistics. (2020, December 16). *Deaths involving coronavirus disease 2019 (COVID-19) by race and Hispanic origin group and age, by state | Data | Centers for Disease Control and Prevention*. CDC. https://data.cdc.gov/NCHS/Deaths-involving-coronavirus-disease-2019-COVID-19/ks3g-spdg

Objective 3:

- Chan, J. F., Yuan, S., Zhang, A. J., Poon, V. K., Chan, C. C., Lee, A. C., Yuen, K. (2020). Surgical Mask Partition Reduces the Risk of Noncontact Transmission in a Golden Syrian Hamster Model for Coronavirus Disease 2019 (COVID-19). *Clinical Infectious Diseases*. doi:10.1093/cid/ciaa644
- Eikenberry, S. E., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Gumel, A. B. (2020). To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling, 5*, 293-308. doi:10.1016/j.idm.2020.04.001
- Li, T., Liu, Y., Li, M., Qian, X., & Dai, S. Y. (2020). Mask or no mask for COVID-19: A public health and market study. *Plos One, 15*(8). doi:10.1371/journal.pone.0237691
- Liu, M., Cheng, S., Xu, K., Yang, Y., Zhu, Q., Zhang, H., Xiao, H. (2020). Use of personal protective equipment against coronavirus disease 2019 by healthcare professionals in Wuhan, China: Cross sectional study. *Bmj,* M2195. doi:10.1136/bmj.m2195
- Mahmood, A., Eqan, M., Pervez, S., Alghamdi, H. A., Tabinda, A. B., Yasar, A., Pugazhendhi, A. (2020). COVID-19 and frequent use of hand sanitizers; human health and environmental hazards by exposure pathways. *Science of The Total Environment, 742*, 140561. doi:10.1016/j.scitotenv.2020.140561

- Smereka, J., & Szarpak, L. (2020). The use of personal protective equipment in the COVID-19 pandemic era. *The American Journal of Emergency Medicine, 38*(7), 1529-1530. doi:10.1016/j.ajem.2020.04.028

Objective 4:

- R., Prof, S., PhD, H., J., J., C., R. (2020, April 06). School closure and management practices during coronavirus outbreaks including COVID-19: A rapid systematic review. Retrieved October 29, 2020, from https://www.thelancet.com/journals/lanchi/article/PIIS2352-4642(20)30095-X/fulltext#%20 DOI:https://doi.org/10.1016/S2352-4642(20)30095-X
- A., & M. (2020, September 06). Mass gatherings contributed to early COVID-19 mortality. Retrieved October 29, 2020, from https://voxeu.org/article/mass-gatherings-contributed-early-covid-19-mortality
- Carl Heneghan, Dr. Elizabeth Spencer; MMedSci, & Tom Jefferson (2020, August 27). COVID-19 testing and correlation with infectious virus, cycle thresholds, and analytical sensitivity. Retrieved October 29, 2020, from https://www.cebm.net/study/covid-19-testing-and-correlation-with-infectious-virus-cycle-thresholds-and-analytical-sensitivity/ DOI: https://doi.org/10.1101/2020.08.05.20168963
- Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., . . . Wu, T. (2020, June 08). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. Retrieved October 29, 2020, from https://www.nature.com/articles/s41586-020-2404-8

Chapter 8: Appendix

- Code for objective1: https://drive.google.com/file/d/1H4rv7dmsRmkP2ncG8UrppwvIm50xHxmB/view?usp=sharing
- Code for objective 2: https://drive.google.com/file/d/1pkgTm12DTL2-gpVyeTqtpsHdAQc2R86P/view?usp=sharing
- Code for objective 3: https://drive.google.com/file/d/1OjMXqY8Y5p1uZejI4ijS2XGV_Ibr5CeG/view?usp=sharing
- Code for objective 4: https://drive.google.com/drive/folders/1c4uYPAQ_BGc7PJTFW50aB0jBIug30Dj?usp=sharing