COMP5310 Principle of Data Science
Assignment 1

Student name: Nikheil Malakar
Student ID: nmal0773
Unikey: 490037555

Problem

Covid-19 or SARS-COV-2 was first detected in Wuhan China, has spread around the world. This virus in its first year has infected over 135 million people and killed almost 3 million people around the globe. It has costs people their livelihood and trillions of dollars in economic damage, it has brought humans to a standstill. After a year and more there has been plenty of data collected regarding covid-19 and its cases around the globe. For this project, it has been concluded to perform simulation and test to their spread for different populations and different countries. We will also discuss their Reproduction Number of how we can expect the infection to increase.

There remains mystery on how it spread through the population, but we want to find if measures taken by countries are lower than the Basic Reproduction Number (BRN) of the pathogen. This determines if the trend of infections is on the rise or declining. With the knowledge gained from this data we can assess the strategy and will also be able to predict the number of infections in the future. This will further help governments around the world to decide on policy to curb infection number and effective decision making until the vaccine have arrived. We will also compare the Reproduction number for USA and check effectiveness of the policies. We will compare the reproduction number for USA and the BRN for USA and the covid-19 itself.

By this analysis, the governments around the globe would have data and analysis required for their policy modeling to minimize damage to the economy and minimizing the infection numbers.

Approach

Our approach is very simple, we want to experiment what type of population/countries are more vulnerable to covid-19. Firstly, we will be experimenting with a simulation for a normal population of 500 people with a default size of land mass. Then we will compare what happens if the virus is introduced to a very dense population with same population and then socially distant population with same population size.

We will confirm if our hypothesis is true for countries and regions around the world. For each population we will check their density and covid-19 infections, this will show the correlation between density and spread of the virus. We will further check the correlation with policies and covid-19 cases if we can find a correlation between face-mask policies and reduction of covid-19.

We will further study how the covid-19 should be spreading using their Basic Reproduction Number (BRN) using this we can check how well the countries has been controlling the spread. We will take a country and check their basic reproduction number for that country and compare it will the virus's BRN itself to check the effectiveness of Covid-19.

Data

The data was selected from multiple public sources and also made using simulations.

1. Population Density Dataset: https://www.kaggle.com/tanuprabhu/population-by-country-2020
   Tables:
   Country Population (2020), Yearly Change, Net Change, Density (P/Km), Land Area (Km), Migrants (net), Fert. Rate, Med. Age, Urban Pop %, World Share
   This Dataset of Population Density shows density of the general population for the year 2020 which is when the covid-19 pandemic started. Allowing comparison with our simulation with countries. This comma separated file (CSV) consist of 205 countries which some of them have issues with their recognition with the ISO standards and the ISO dataset we also have downloaded.
   The analysis and graph for population Density is shown in Appendix 1.
2. Covid-19 Dataset: https://ourworldindata.org/coronavirus-source-data
   Tables:
   iso_code, continent, location, date, total_cases, new_cases, new_cases_smoothed, total_deaths and etc.
   This dataset of 18MB+ has daily infection number for every country around the world with cases and death with other information like percentage of both gender getting infected in different countries.
3. World-ISO codes: https://www.kaggle.com/juanumusic/countries-iso-codes
   Tables:
   English short name lower case, Alpha-2 code, Alpha-3 code, Numeric code, ISO 3166-2
   When mapping world map graphs of the world, we can only map them with ISO code for example Afghanistan's ISO code is AFG. This dataset will be used in combination of other datasets. This dataset is taken from the ISO Wikipedia page which is publicly available to everyone.
4. US-Covid-19 Cases: https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset
   Tables:
   date, county, state, fips, cases, deaths
   This dataset records daily infection cases for COVID-19 in the United states for different counties. This dataset also has the number of deaths in the united states. This dataset is frequently updated and has the latest information and data we need.
5. Mask Policy: https://ourworldindata.org/grapher/face-covering-policies-covid
   Tables:
   Entity, Code, Day, facial_coverings
   This shows the number of masks policy for each country to curb the infections for covid-19.

These datasets are different datasets which will be used by combining to analyses what works for covid-19 and what does not.

The main question to be asked is what kind of policies are effective to what type of population. First, we will simulate how Covid-19 will spread in an unconstrained environment for a population of 500. We will discuss how this virus will spread among the population until there will be many immune people and prevent others from getting infected. Typically, Covid-19 will be cured after 14 days therefore we will consider immune with a chance of 40%.

Health Experts has predicted that Covid-19 will likely infect from 40-60% of the world's population so we will check how our model did. The simulation is stored in data/simulation1.csv which is plotted and animated and included in Appendix 6, which shows only 219 infections of 500 people which is on target to the expected 40-60% of the world population. The people infected further get immune to Covid-19 after 14 days with the chance of 40% or they will remain infectious.

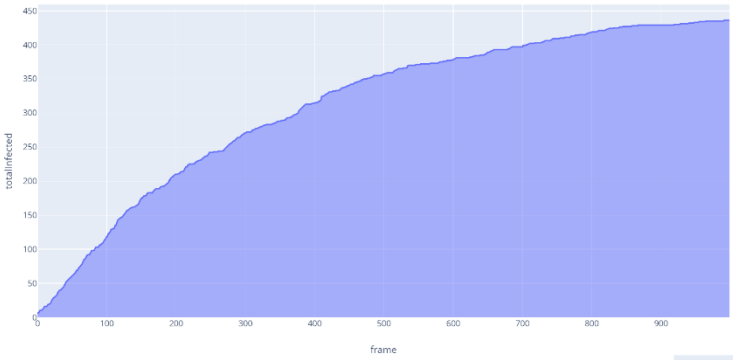
After simulating the infection in the population our data is stored in data/simulation1.csv which has x points, y points, Number of recovered people, type and total infection along with people ids. Which was plotted as an animation and can be seen on Appendix 6. After confirming our simulation is running with no problem with expected results, further study of Dense and Socially distanced population was simulated and kept in

data/simulation1_smallBorder and data/simulation1_largeBorder respectively and their simulation can be seen in Appendix 7 and Appendix 8 respectively.

The ISO dataset consist of ISO codes for each country which is needed to differentiate between countries. It is collected from the official Wikipedia page. This dataset consists both ISO version 2 and ISO version 3 which will be used in mapping countries. In combination of other dataset, we can map with values and ISO codes.

The population by country is a dataset collected in 2020 which records the population's information such as density, land area etc. It also has information which will not be used in the current project is Net change of population with migrant numbers too. OWID Covid-19 dataset and us counties Covid-19 dataset both contains Covid-19 cases per day around the world and US counties. Last dataset being used for this project is face-mask policies which contain the number of mask policies implemented in each country and this dataset too will be used in combination with ISO dataset.

To study our question of how Covid-19 spreads in two different population groups of dense and socially distanced we performed our simulation with different areas therefore a dense population will have a small area meaning less socially distant and the large area will be more socially distant. After simulating for both dense and socially distant populations our result for number of infections were:



The Dense population shows an exponential increase in number of infections within the population and the growth of infection slowly decline after more immune people appear in the population than infected.

Figure 1 – Number of infections of covid-19 for Dense Population

The Socially distant simulation with the same population also shows an increase of covid-19 cases but never hitting the high infections rates as the dense population, due to infected people not infecting and becoming immune. After a while there seems to be low growth of infection number and the showing control of infection in the population.

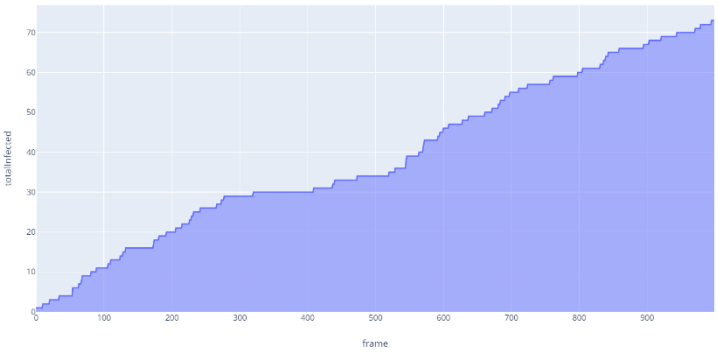


Figure 2 - Number of infections of covid-19 for Socially Distant Population

With the help of our simulation, we concluded that dense populations are vulnerable to Covid-19 compared to socially distanced population. Infected people hardly infect other people and get immune fast due to less people closely distanced to them. We will check this hypothesis by checking the population density of countries and Covid-19 cases for countries.

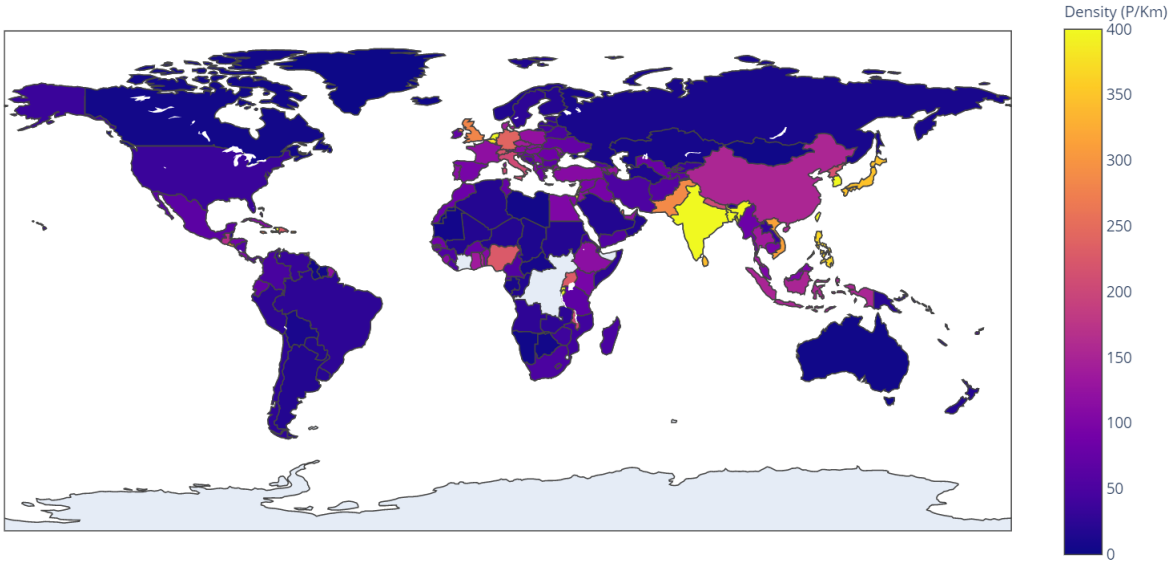


Figure 3 – Population Density of Countries around the world

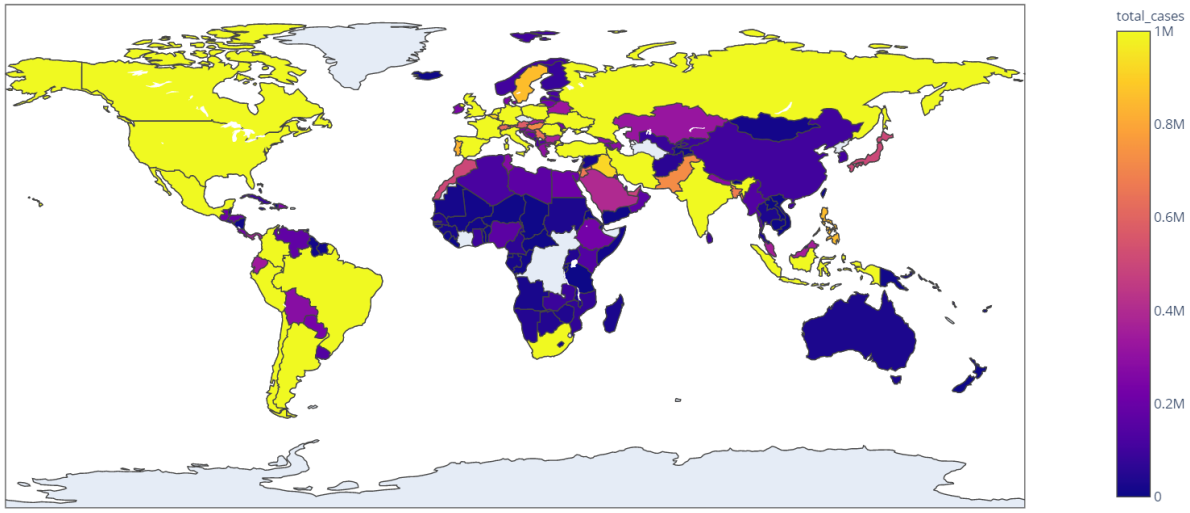


Figure 4 – Covid-19 infections of Countries around the world

From viewing Figure 3 and Figure 4 from Appendix 1 and Appendix 2, we can see that there is very little correlation between covid-19 cases and population density leaving some exceptions like India and Australia. Covid-19 cases not only reflect population density but also government policies that curb infections. For example, USA is not dense but due to bad government policies it has the highest covid-19 cases whereas Australia is not dense too, but the government policies has curbed the covid-19 cases. This can be seen in Appendix 3 that show clear correlation between population density and covid-19 cases which show no correlation between them.

We will also check if face-mask policies help to curb the number of infections in the country.

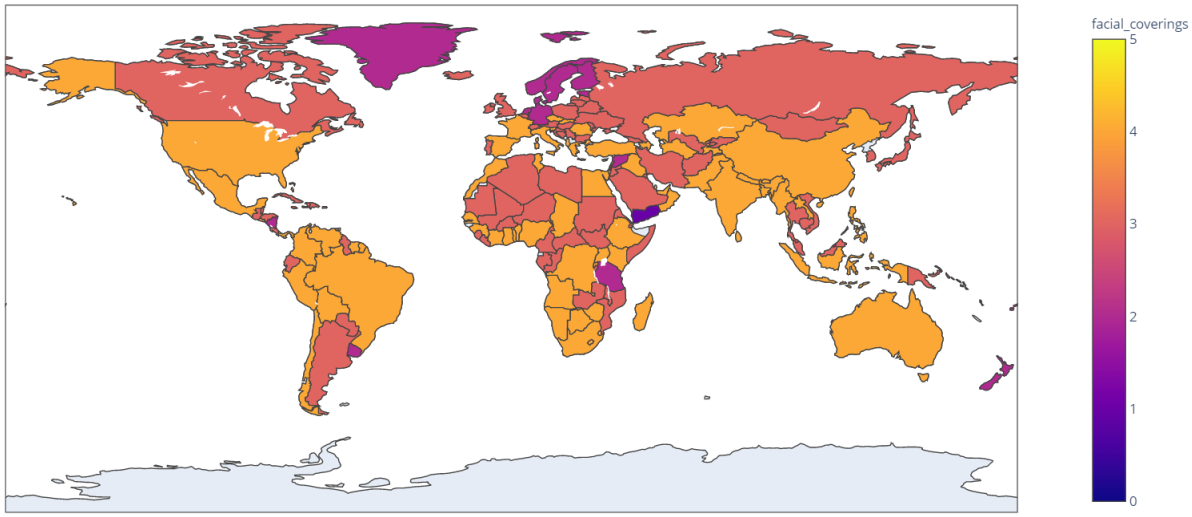


Figure 5 – Face mask policies for different countries around the world

We can see almost all countries have face-mask policies, but this does not reflect the number of infections in the country. From these two studies we can say with confidence that covid-19 infection numbers are a result of multiple factors and I think one of the greatest factors is public cooperation's in face mask wearing and lockdowns.

Finally, we will check how effective the measures are for a given country. Firstly, we calculated the Basic Reproduction Number for the country and compare it to the Basic Reproduction Number for Covid-19 which is 2.2 taken on 1/11/2020 in Wuhan, China.

After 10 infections, covid-19 infections from 1 person should increase exponentially to 1024 which can be seen in Appendix 6. It will result to expected count if there is no restriction or precautions taken by the population which is not the case. We will check the reproduction number for USA for November 11.

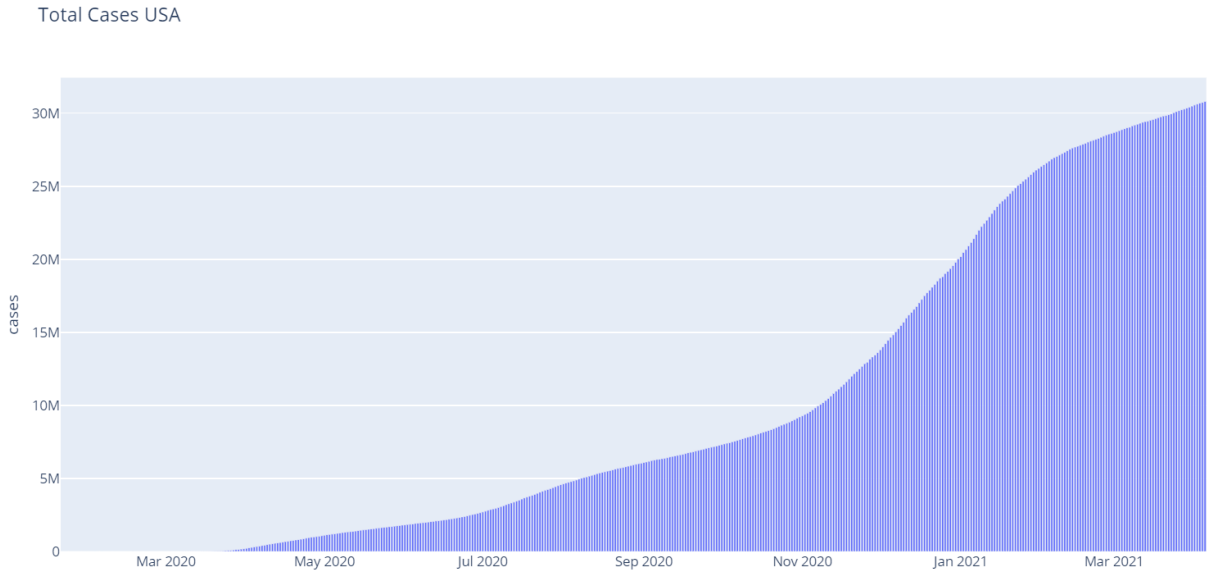


Figure 6 – total USA covid-19 cases daily

From November 11 to November 25 we have calculated the reproduction number is 1.7. Which is a proof that measures for covid-19 is taken but it is not effective and basic reproductive number remains high and growth after November is still exponential.

Proposal for Next stage

For stage of the Project, the objective would be to figure out effective way to lower infection number for different countries and regions and find out specific relationship between other factors and Covid-19 infections. Factors to be in considerations:

1. Lockdown
2. Curfew with time data
3. Vaccinations in the country with the type of vaccine
4. Covid-19 immunity in the population
5. Covid-19 Tests
6. Travel Restriction
7. Etc

All these data is very ambiguous with different types of data coming in and handling with just csv files will be very hard therefore, moving it to a No-SQL database will ease storage problem of data extraction and will be able to new type of columns of data too. Due to so many countries and regions with different dates, the use of NoSQL database would perfect.

After all the data are sorted, we will be able to train our model to predict the number infection with the help of all the data we have collected. There should be a relationship between policies and Covid-19 infections.

References

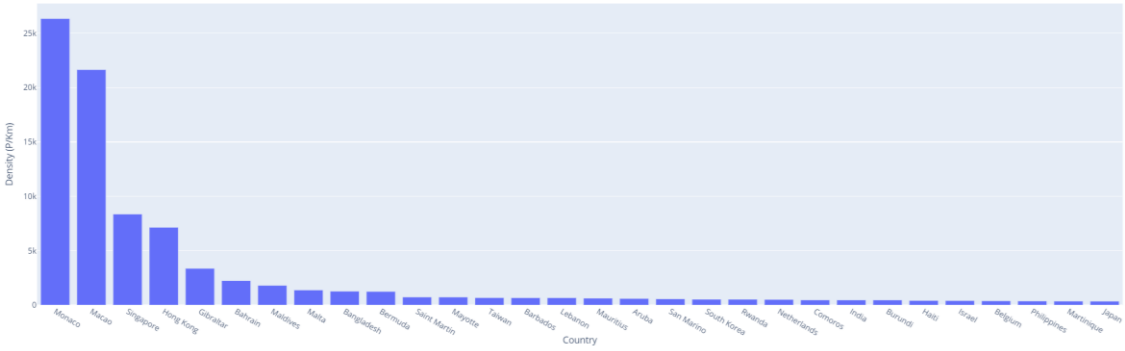
1. Basic Reproduction Number, Medrxiv, viewed April 2021 https://www.medrxiv.org/content/10.1101/2020.09.26.20202010v1.full.pdf
2. Kaggle 2020, Population by Country, Kaggle, viewed April 2021 https://www.kaggle.com/tanuprabhu/population-by-country-2020
3. Our World in Data, Coronavirus Source Data, OWID, viewed April 2021 https://ourworldindata.org/coronavirus-source-data
4. Kaggle 2017, Countries ISO Codes, Kaggle, viewed April 2021 https://www.kaggle.com/juanumusic/countries-iso-codes
5. Kaggle 2021, US counties COVID-19 dataset, Kaggle, viewed April 2021 https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset
6. Our World in Data, Face Covering policies during COVID-19, OWID, viewed April 2021 https://ourworldindata.org/grapher/face-covering-policies-covid

Appendices

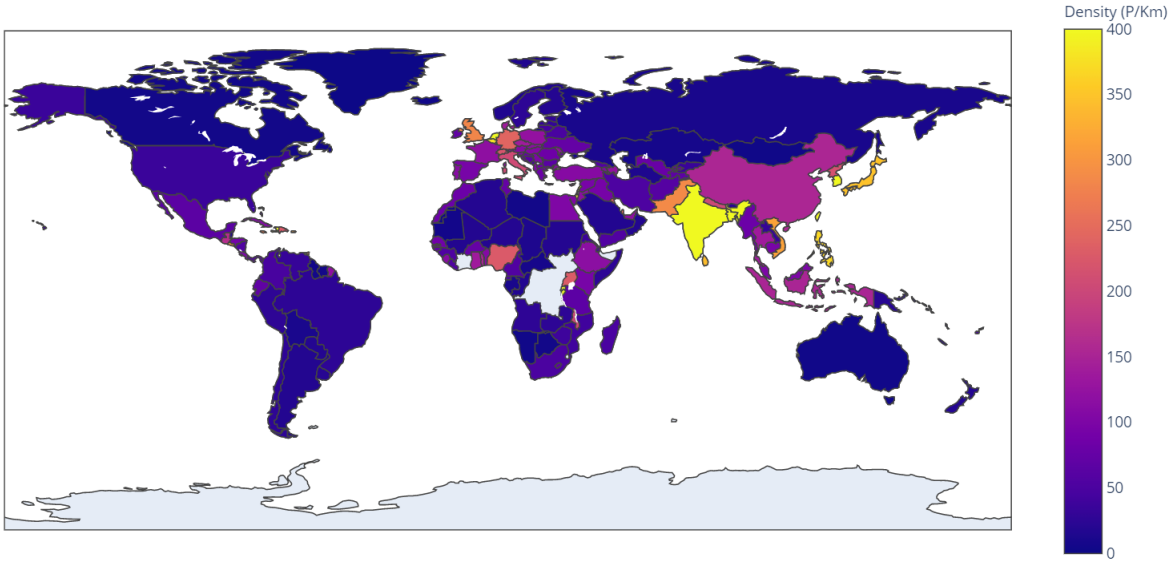
1. Population Density Summary for countries
2. Number of Covid-19 Cases for different countries
3. Correlation between number of covid-19 cases and population density
4. Face masks policy
5. Covid-19 Cases in the United States of America
6. Covid-19 Basic Reproduction Number

Appendix 1- Data Visualization for population Density

The chart shows the highest Population Density around the globe.

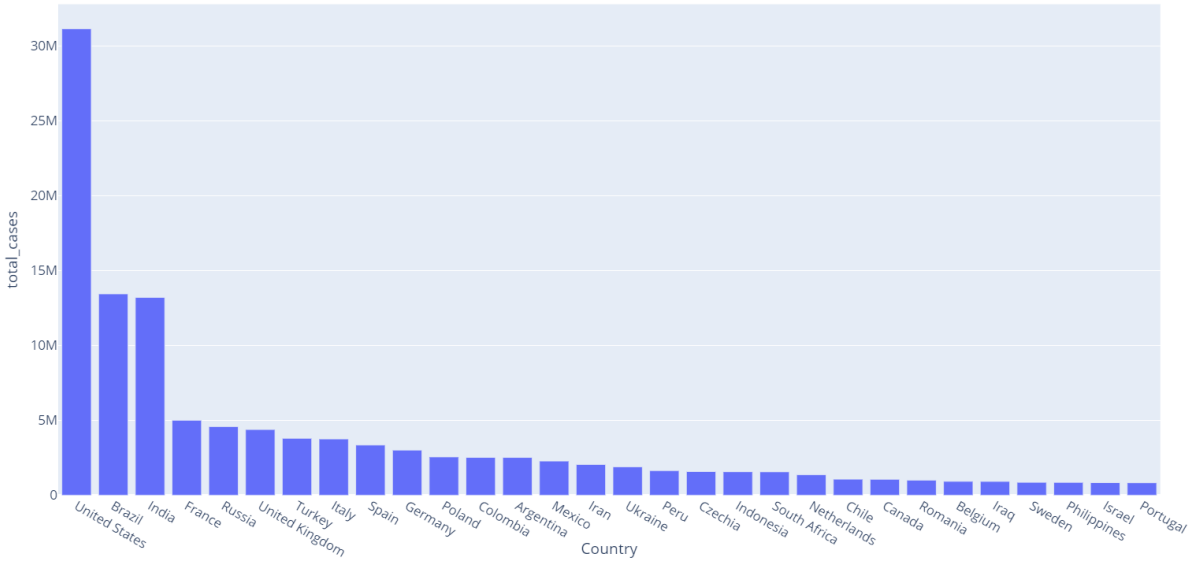


Map Graph for the country's population density around the globe.

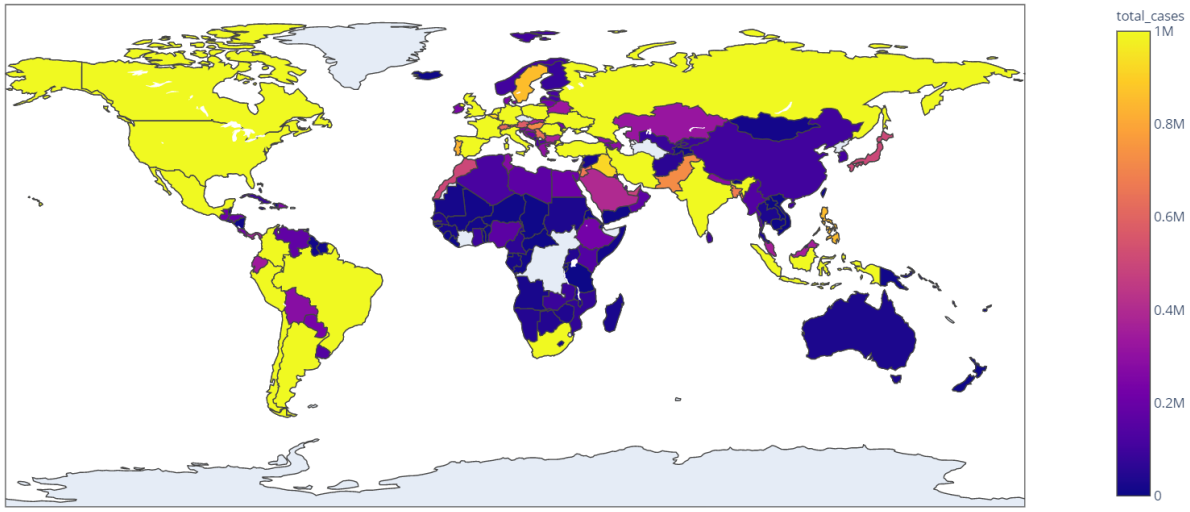


Appendix 2 – Number of Covid-19 Cases for different countries

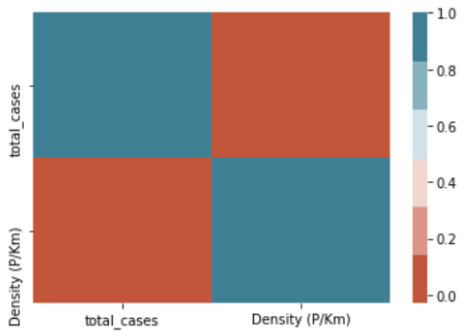
This chart shows the number of covid-19 cases around the globe for different countries.



Map Graph for covid-19 cases of countries around the globe.



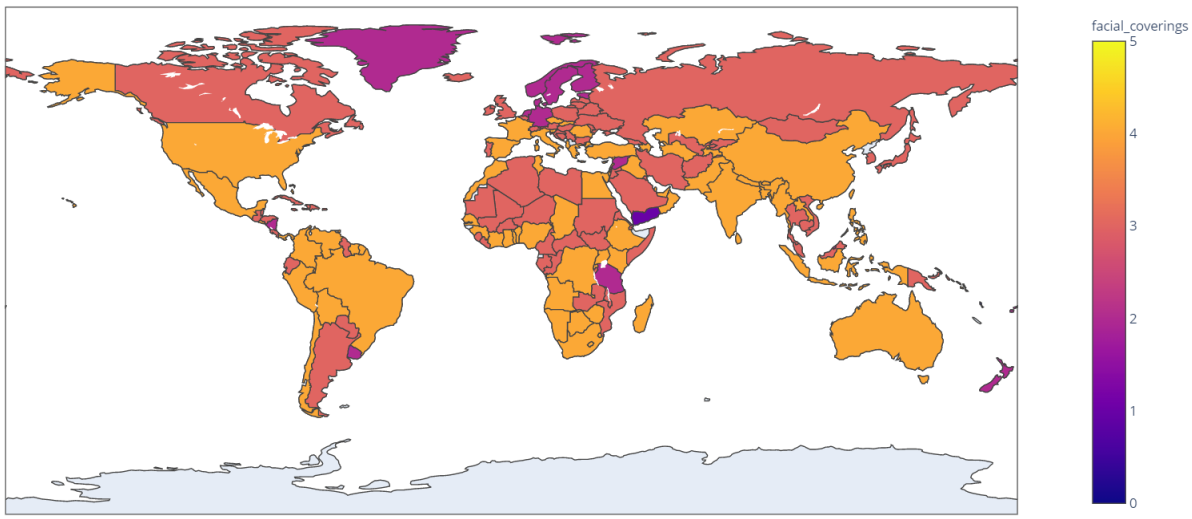
Appendix 3 – Correlation between number of covid-19 cases and population density



This correlation graph shows the correlation between the number of covid-19 cases around the globe.

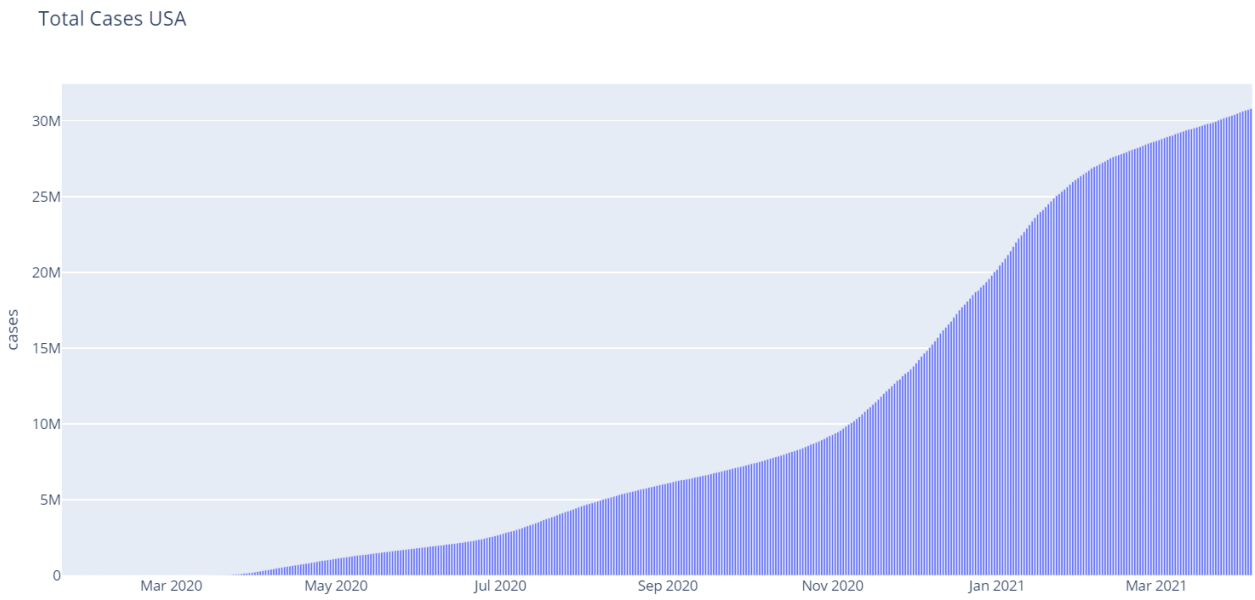
Appendix 4 – Face masks policy

This Map shows the different countries and their number of face masks policies.



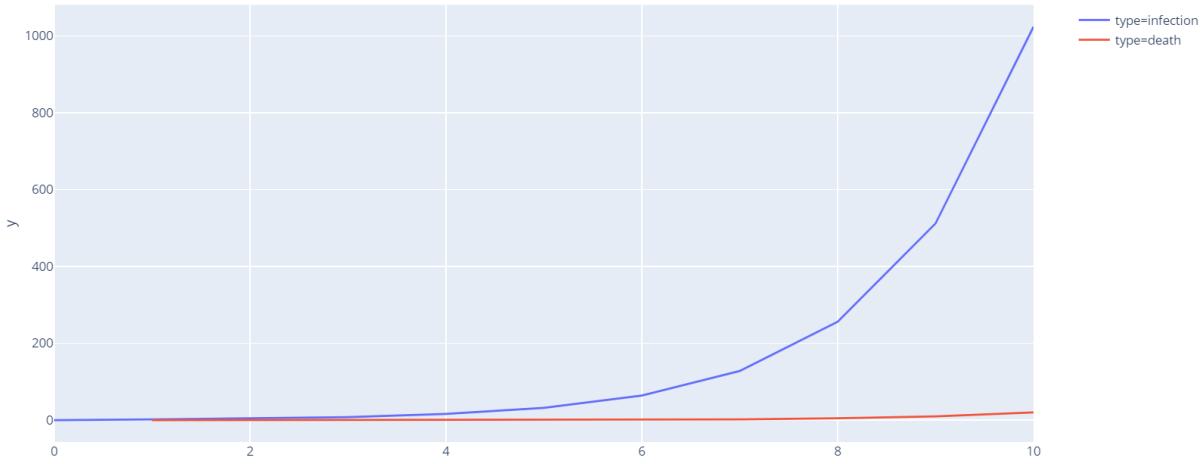
Appendix 5 – Covid-19 Cases in the United States of America

This chart show the increase of total cases of covid-19 in the States.



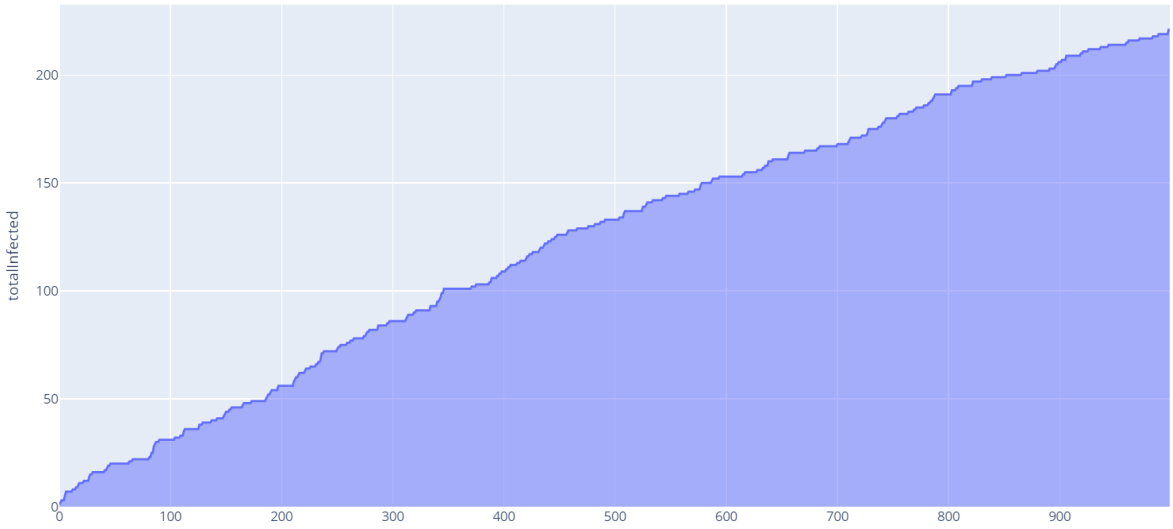
Appendix 6 – Covid-19 Basic Reproduction Number

The increase of covid-19 cases and Covid-19 deaths from a person

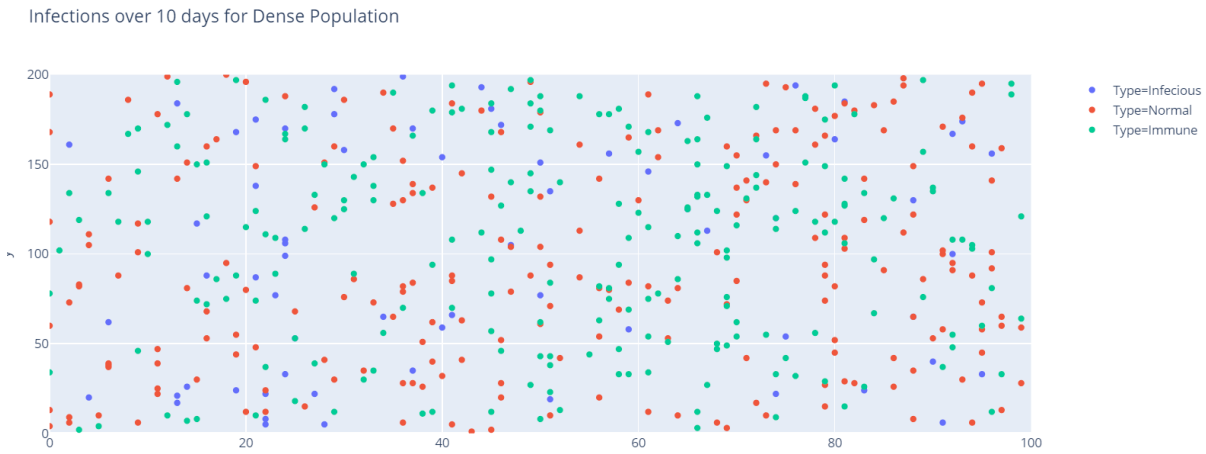


Appendix 6 – Covid-19 Simulation and Results for Normal Population

(video on visuals/NormalPopulation.mp4)
Below is the graph for number of cases for covid-19 in the population

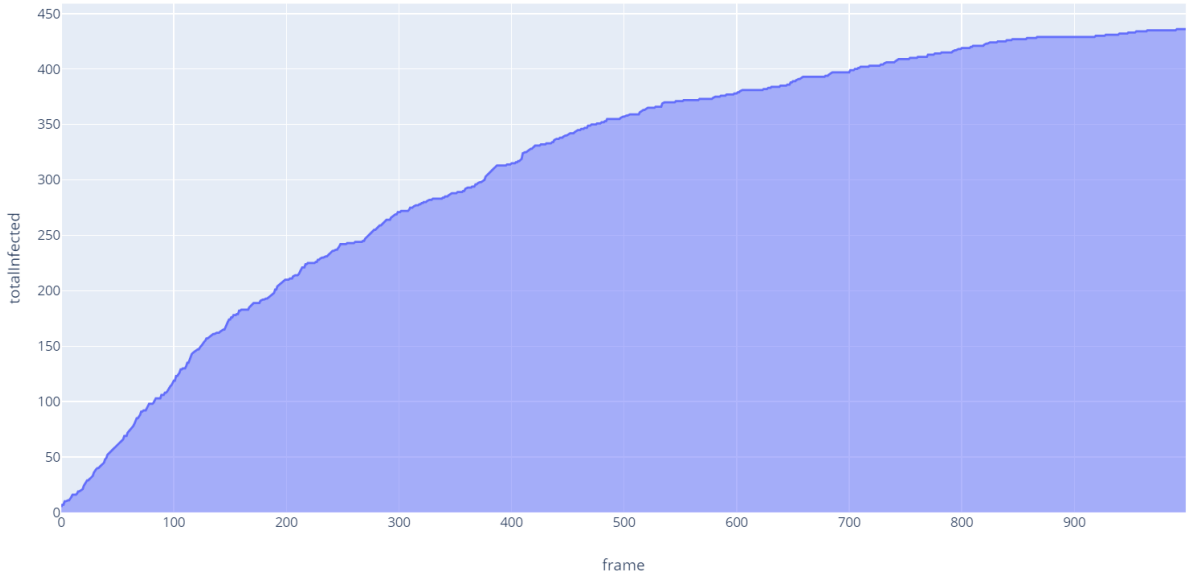


Appendix 7 - Covid-19 Simulation and Results for Dense Population

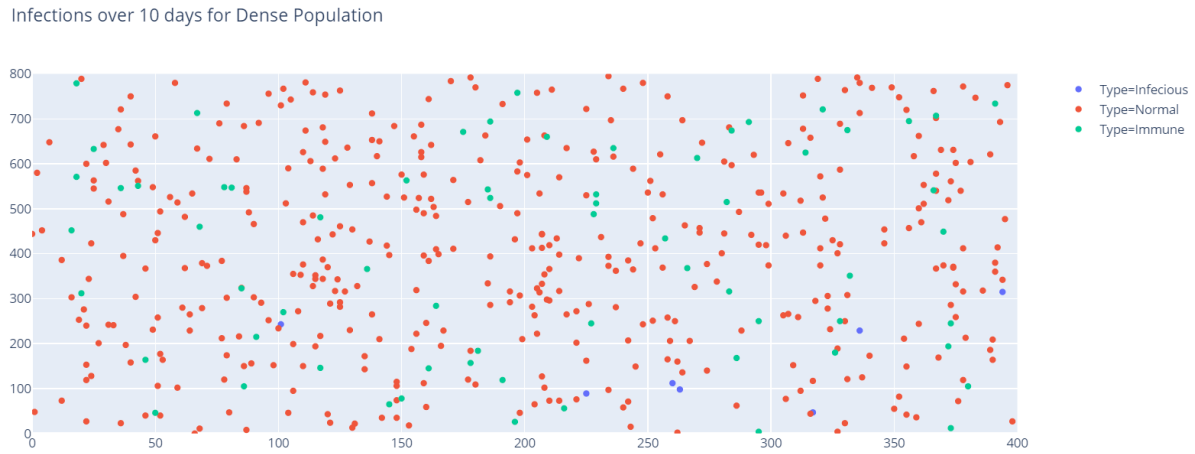


(video on visuals/DensePopulation.mp4)

Below is the graph for number of cases for covid-19 in dense population.



Appendix 8 - Covid-19 Simulation and Results for Socially Distant Population



(video on visuals/SociallyDistantPopulation.mp4)

Below is the graph for number of cases for covid-19 in Socially distant population.