Investigation of Factors Affecting Spread of COVID-19

Ruohan Li, Isobella, Teddy Ong, Zhuolin Wu

Sunday 21st February, 2021

Abstract

This report identifies the factors that affect the severity of the transmission of Covid-19, the disease caused by SARS-CoV-2, and the strength of correlation between these factors and the weekly number of total cases. From the data of 30 European countries, including most of the Schengen Zone, we found that there are 10 factors having relatively significant impacts. Our finding suggests that the percentage of the population aged 70 and above in a country has the most significant influence on the total cases of Covid infection.

1 Non-Technical Executive Summary

Factors we examined in this report that affect the rate of transmission of Covid-19 can be classified into two areas:

1. Social-Economic Indicators: GDP per capita, Human Development Index, Hospital Beds per Thousand, and Life Expectancy
2. Demographic Indicators: Elderly aged 65 and above, Elderly aged 70 and above, Cardiovascular Death Rate, Diabetes Prevalence, Percentage of M/F Smokers

We found that the **elderly aged 70 and above** has the most significant impact on the transmission of Covid-19, measured by the number of total cases each week for each country. This relationship was weaker in the **elderly aged 65 and above**, and even weaker in the **median age of the population**. The shows that not only are the elderly more susceptible to the diesease, the reduced mobility of older age groups has little impact on the correlation, as is expected from an airborne respiratory disease.

Population Density has relatively little impact on the transmission of Covid-19, in contrary to literature on US States [2]. We draw the following conclusions:

1. There is significant difference between the behaviour of residents and their compliance to covid regulation for the two continents.
2. The inter-connected nature of mainland Europe and the Schengen Zone allowed disease to spread regardless of the density of each population centre.

The correlations listed above have a high level of precision, especially when the total cases each week is relatively low. The level of precision decreases as the total cases each week increases, likely due to population dynamics outside our focus of investigation.

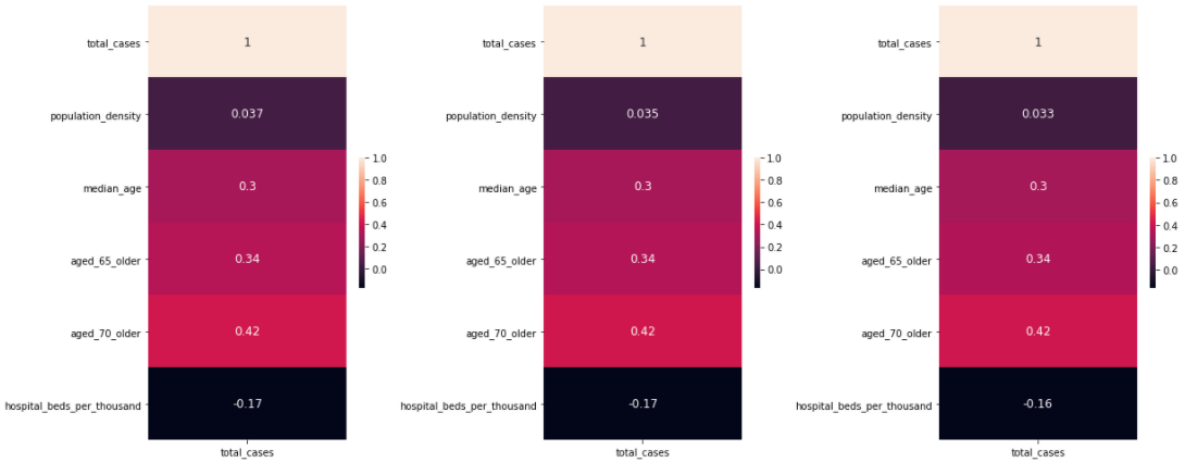


Figure 1: The level of correlation between total cases and the top five factors. Data from the first three weeks of analysis, beginning 9th March 2020.

2 Background Research

We began by identifying factors that affect the pandemic, and researched the indicators that were used by governments to measure the spread of the virus. Data sets were used to create models so we could observe how strongly the factors affected the virus spread. It is important to analyze the impact of each factor to help policy makers in the decision-making process to contain the spread of Covid-19.

2.1 Factors

2.1.1 Age

An older population seemed to naturally slow the contagion due to fewer social contacts. Socioeconomic factors (especially the rates of employment and public transportation usage) [1]. High age groups emerge as significant in affecting the case and mortality rates.

2.1.2 Population Density

COVID-19 infection rate was found to be both positively correlated with population density, and negatively correlated with social isolation rate. Population density emerges as a significant affecting factor for the virus. [2]

2.1.3 Healthcare System

Organizational features of local healthcare systems (particularly the proportion of private healthcare facilities) also seemed to play an important part in the infection's spread. Patients who are not admitted to the hospital and are discharged home from the ED play a crucial role in curbing the spread of COVID-19. (3)

2.1.4 Environmental

Changes in weather (i.e., increase of temperature and humidity as spring and summer months arrive in the Northern Hemisphere) will not necessarily lead to declines in case counts without the implementation of drastic public health interventions. (4) Inversely correlated with both temperature and UV radiation, the UV radiation provided by sunlight potentially contributes to depletion of SARS-CoV-2 infectivity. (5)

2.1.5 Social-economical

Healthcare index, homeless and GDP have little or no impact on pandemic spread and mortality. (6)

2.1.6 Transportation

Airport Traffic is positively related to the rates in the US. (7) The spread of COVID-19 within the expanded metropolitan area of São Paulo, Brazil, is influenced by the availability of highways. (8)

2.1.7 Testing rate

If testing rates are slow, there could be a spread of the virus due to people not being safe and accidentally spreading the virus due to unknowingly having the virus. Contact tracing being incorrectly tracked down, due to a variety of reasons, such as people not using the app properly. (9)

2.1.8 Life Expectancy

While a lower life expectancy doesn't affect the spread of the virus, the virus causes a lower life expectancy. Research found that a 2% infection rate could cause a decrease in life expectancy in countries with high life expectancies, which Europe has quite a few of. Getting COVID for any ages could cause health problems later in life, if not immediately. (10)

2.1.9 Hospital Beds per Thousand

COVID clearly had a major effect on the availability of hospital beds, especially ICU beds with ventilators. The pandemic also proved how unprepared most of the world was to deal with the sudden need for so many ventilators and ICU beds.

2.1.10 Diabetes Prevalence

Having diabetes doesn't increase your chances of getting the virus, but it definitely but you at a higher risk of serious health complications if you have both. Any illness that can affect your heart puts you at more risk of being seriously ill with COVID-19, and diabetes can lead to serious heart problems. (11)

2.1.11 Cardiovascular Death Rate

COVID can severely affect people with underlying health conditions, such as cardiovascular problems, and these cases act contribute to many of the death cases. For example, in Ireland, more than 8 out of 10 of the deaths that happened were to people that had underlying conditions, with chronic heart disease being the most common illness. (12) (13)

2.2 Indicators

2.2.1 R Values

This value is the number of people the virus will be spread to by one infected person (on average) (it is the measure of the contagiousness of the disease). If the R value is low enough, the virus will stop spreading. Restrictions are proven to have lowered the value of the R number: in England for example, successive restrictions brought the value down, but only did full lockdown bring the figure to below one. The R value is calculated using the amount of new cases, it cannot be done by tracking the virus' moving from person to person so scientists work backwards with the data. (14) (15) (16)

2.2.2 Daily Cases

Daily case numbers are a tell of how the virus is spreading. When restrictions are brought into place, the effects they have on the spread of the virus is not shown until weeks later. This needs to be taken into account when looking at the case numbers: new restrictions will not be reflected by the numbers for a while to come.

2.2.3 Death Rate

Daily deaths are not a very good indicator of the spread of the virus, just a good representation of the demographic that the virus has spread to (i.e: the elderly, people with respiratory problems, smokers etc). If there are a lot of deaths, the severity of the situation is shown more seriously: a lot of cases with few deaths is better than with many deaths, as cases can be from teenagers or young adults (people less affected by the virus), and there is opportunity for recovery.

2.2.4 Positivity Rate

The positivity rate for COVID is the percentage of tested positive cases, ((positive cases/total cases) * 100). This figure is used to see what the transmission rate for the virus is, and to see if the number of tests being done is an adequate measure for the amount of infected people. A high positivity rate shows a high level of transmission, and that there are more people with the virus who haven't been tested yet. (17)

2.2.5 Vaccination Rate

Israel has been acting as the country to lead by example when it comes to vaccinating it's people, with over 50% of their population being vaccinated already (with a higher vaccination rate for the elderly). Vaccination rates in the UK are the highest in Europe, with 22% of the vaccines going to the elderly and healthcare workers. These vaccines will not be shown in the new cases daily, as the elderly do not make up for a large percentage of the cases, with that demographic taking extra care to be safe. (18) (19)

3 Technical Exposition

3.1 First Look

Firstly, we noted that the OWID data was the most comprehensive. Out of those, countries in Europe had much less missing data than the countries in other continents. Most of the data regarding the USA was only on the National level, but we believe that analysis of USA must be broken down into individual states, due to the effect of partisan politics on the policies enacted and the strictness of enforcement. This data was difficult to find.

Overall, we decided to narrow our analysis to 30 European countries, which were selected as they were present in both the OWID and the ECDC dataset. This allows us to correspond and merge the data easily. Our principle is to avoid adding non-zero values into missing entries in order to preserve the integrity of the source data.

3.2 Cleaning Data Sets

3.2.1 Start / End Dates

We then noted that each country's dataset started and ended on different dates. From our research, the first wave of COVID impacted Europe around the beginning of March, and indeed most countries began recording around that time. For ease of data manipulation, we truncated all OWID data to begin on the 9th March 2020, Monday and end on 3rd Jan 2021, Sunday. This allowed us to group by weeks and match corresponding entries directly with ECDC data.

3.2.2 Missing Entries in ECDC

We then noticed ECDC data has missing weeks. We decided to fix that by adding the required row with all data being 0. We could reasonably use the average of the neighbouring cells, but at this point we decided not to pollute the data with unnecessary assumptions. This dataset was also missing many entries under Positivity Rate. Furthermore, we hoped that OWID and ECDC data could be merged to fill each other's missing gaps.

3.2.3 Missing Entries in OWID

We plotted the total number of missing entries for each column. Out of 9670 rows, some columns such as Weekly ICU Admissions and Total Handwashing Facilities were completely blank, which made them untenable to use for further investigation. Along with columns which could be easily re-derived if needed (such new tests per thousand), we decided to drop them entirely.

3.3 Comparison of Testing Data to Gauge Reliability

Testing is a key factor to provide context for total cases recorded. As preliminary visualisation, we aggregated OWID by weeks and plotted the number of new tests. We observed that some countries, such as Belgium, had a perfect match between OWID and ECDC data.

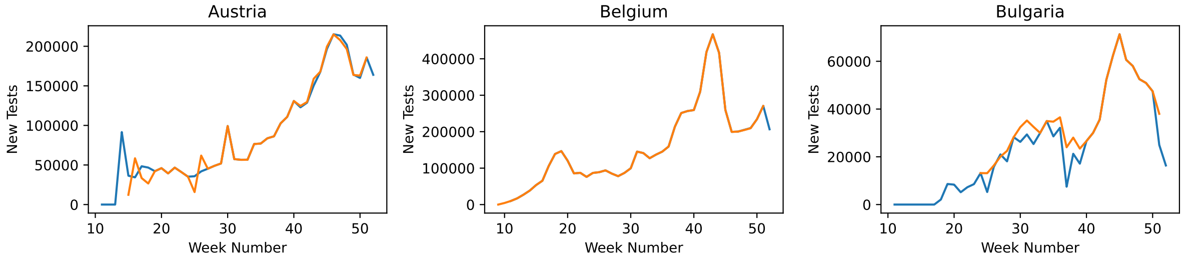


Figure 2: OWID (Blue) vs ECDC (Orange) Testing Data for Austria, Belgium, Bulgaria.

Meanwhile, France systematically under-reports its new tests to ECDC. Germany had not reported any tests under OWID, which makes it necessary to use the data entirely from ECDC. Greece had entire weeks of missing data, causing sharp spikes down to 0. In general, whenever the OWID data is below the ECDC data, it meant that the week had missing entries, subsequently causing the sum to be lower.

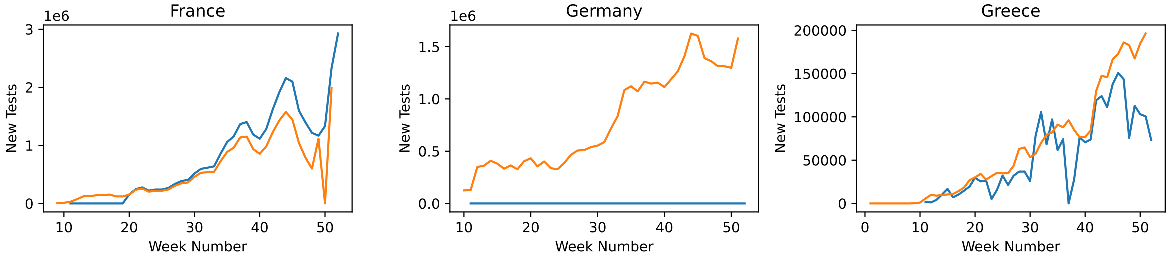


Figure 3: OWID (Blue) vs ECDC (Orange) Testing Data for France, Germany, Greece.

We believe it is justified to merge these two datasets by taking the maximum between OWID and ECDC for each week. By inspection, this will reasonably patch the gaps seen without imposing any external assumptions by interpolation.

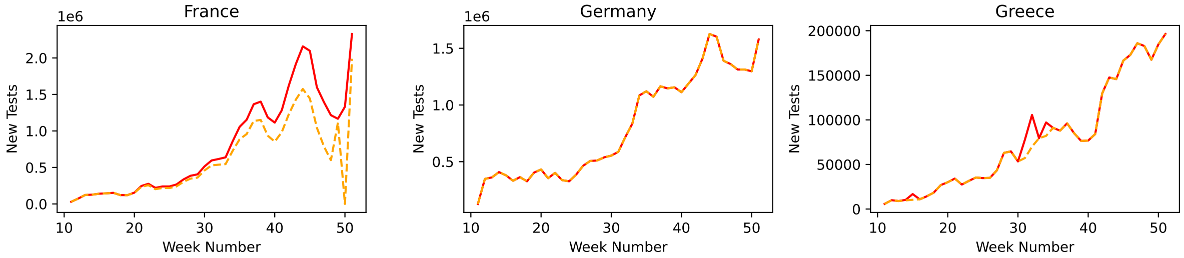


Figure 4: Merged (Red) vs ECDC (Orange) Testing Data for France, Germany, Greece.

We used this testing data to adjust for the total cases reported, by calculating total cases per thousand tests. This will be referred to as ”total cases” here-on.

3.4 Heat Map

We plotted a Heat Map to show the correlation between different factors to the total number of cases of COVID-19. Each graph represents a different week. This shows the relationship between the values of these in each country to their total cases.

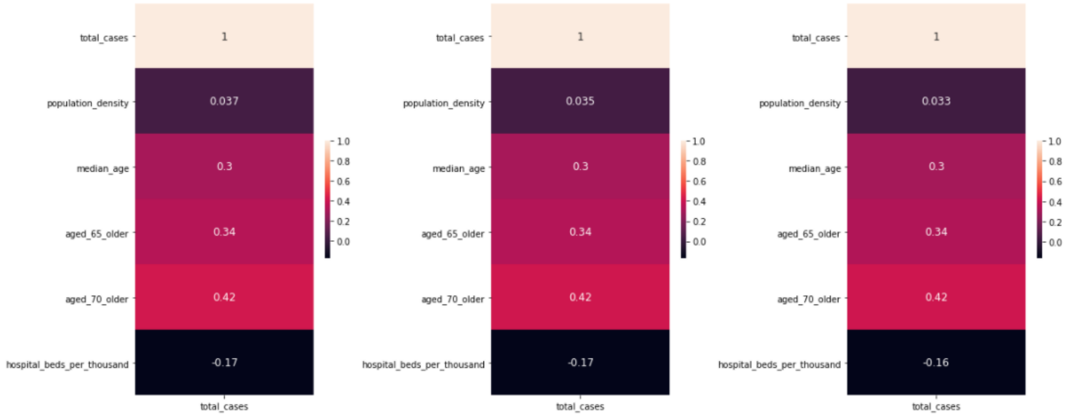


Figure 5: The level of correlation between total cases and the top five factors. Data from the first three weeks of analysis, beginning 9th March 2020.

3.4.1 Interpretation of Heat Map

As seen, the age groups show a strong correlation to total cases, while, surprisingly, population density did not affect total cases. This is in contradiction with literature [2]. We believe this is because: the population density is given by population divided by land area, whereas to have a more detailed insight, we need to look at cities vs suburbs.

Hospital beds per thousand had a slightly negative correlation, showing that while it is slightly better to have more medical facilities, hospitals were generally unable to affect the spread of covid. That is because they only deal with the cases after the disease has already spread.

3.4.2 Evolution of Heat Map Across Time

When we look at the changes to the Heat Map and the correlation numbers across time, we notice that they decrease significantly as we approach Week 40. This corresponded to the start of the "second wave" for many countries in September and October 2020.

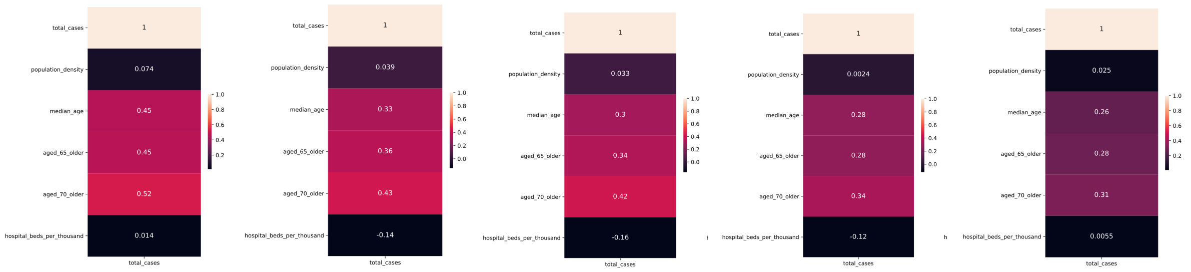


Figure 6: Heat Map from Weeks 11, 18, 25, 32, 39

3.5 Quantifying the Trend

We seek to investigate the trend between the correlation and progression of weeks.

3.5.1 Random Forest Model

A correlation matrix model, a random forest model, and a heat map was fitted using our cleaned up data. The forest and the heat map were used to see what weight the factors had on the COVID cases per day (i.e spread of the virus) and the matrix was used to see the level of accuracy we had.

3.5.2 KNN Matrix Model

We applied the K-Nearest Neighbours machine learning model to improve our results from the Random Forest Model. Moreover, we examined the level of precision for the correlation between these factors and the total cases each week for each country through two machine learning models: 1) random forest; 2) k-nearest neighbours classifier. It appears that there exists a strong level of precision for the correlation especially when the total cases each week is relatively low, although the level of precision for the correlation decreases as the total cases each week are high.

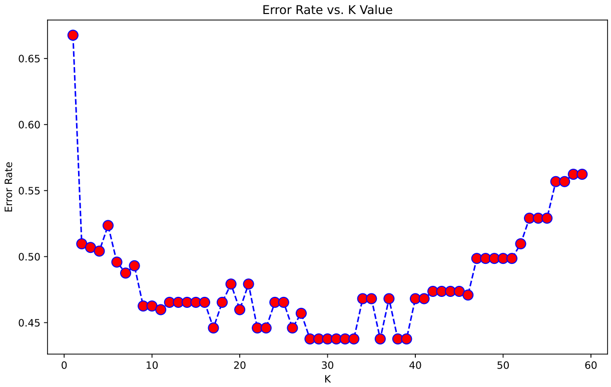
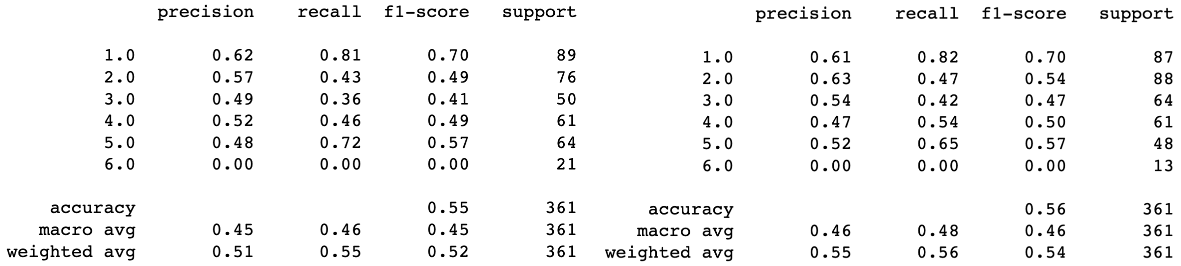


Figure 7: Graph of Error Rate vs K-value.

In particular, the k-value for the KNN model was selected by plotting the graph of error rate against k-value. We then used $k = 24$ for the table above.



```
              precision    recall  f1-score   support                     precision    recall  f1-score   support

         1.0       0.62      0.81      0.70        89              1.0       0.61      0.82      0.70        87
         2.0       0.57      0.43      0.49        76              2.0       0.63      0.47      0.54        88
         3.0       0.49      0.36      0.41        50              3.0       0.54      0.42      0.47        64
         4.0       0.52      0.46      0.49        61              4.0       0.47      0.54      0.50        61
         5.0       0.48      0.72      0.57        64              5.0       0.52      0.65      0.57        48
         6.0       0.00      0.00      0.00        21              6.0       0.00      0.00      0.00        13

    accuracy                           0.55       361         accuracy                           0.56       361
   macro avg       0.45      0.46      0.45       361        macro avg       0.46      0.48      0.46       361
weighted avg       0.51      0.55      0.52       361     weighted avg       0.55      0.56      0.54       361
```

Figure 8: Output of Random Forest Model and Matrix Model, Respectively.

Note: the values 1.0, 2.0, 3.0, 4.0, 5.0, 6.0 refer to Aged 70 and older, Aged 65 and older, Median Age, Population Density, Hospital Beds per Thousand, Testing Rate

3.6 Results

The matrix model showed us that as the weeks progressed, the higher the cases, the lower the accuracy of the results. The conclusion remains that the age groups had the biggest correlation to the case numbers, with the hospital beds having a slightly negative correlation. When we looked at the bar plot of correlation between the population Aged 70 and above with the Total cases, the pattern remains nearly identical throughout all the weeks. Furthermore, the KNN Model had lower error than the Random Forest Model.

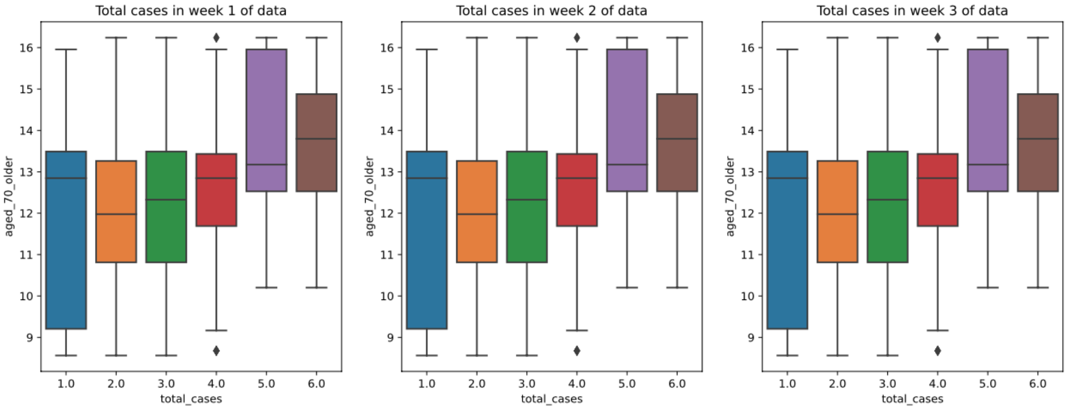


Figure 9: Bar Plot of Correlation Between Aged 70+ and Total Cases, Week 1, 2, 3

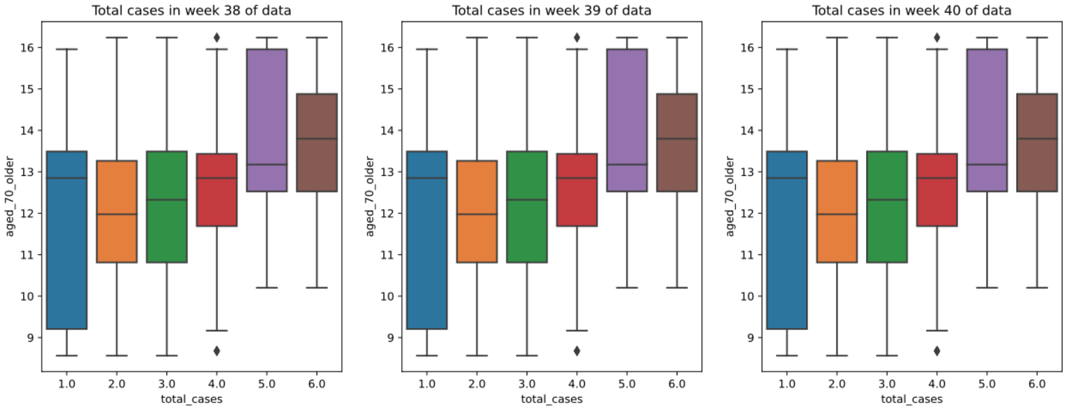


Figure 10: Bar Plot of Correlation Between Aged 70+ and Total Cases, Week 38, 39, 40

This plot shows a society with an older age tends to have a higher number of cases, through the trend of overall increasing median in the distribution

4 Conclusion

Conclusion: A correlation can clearly be seen between the factors, mostly age groups, and the COVID-19 cases. This information could be used by healthcare professionals and governments to implement regulations and focus their energy where it is needed most (in protecting older people) which would lower the numbers of cases, and can be used to predict how the virus will then continue to spread.

References

[1] *Buja, Alessandra, et al.* 'Demographic and Socio-Economic Factors, and Healthcare Resource Indicators Associated with the Rapid Spread of COVID-19 in Northern Italy: An Ecological Study'. PLOS ONE, vol. 15, no. 12, Dec. 2020, p. e0244535. PLoS Journals, doi:10.1371/journal.pone.0244535.

[2] Roy, Satyaki, and Preetam Ghosh. "Factors Affecting COVID-19 Infected and Death Rates Inform Lockdown-Related Policymaking." PLOS ONE, vol. 15, no. 10, Oct. 2020, p. e0241165. PLoS Journals, doi:10.1371/journal.pone.0241165.

[3] James, Melissa K. .. .., et al. "Demographic and Socioeconomic Characteristics of COVID-19 Patients Treated in the Emergency Department of a New York City Hospital." Journal of Community Health, Oct. 2020. Springer Link, doi:10.1007/s10900-020-00937-2.

[4] Poirier, Canelle, et al. "The Role of Environmental Factors on Transmission Rates of the COVID-19 Outbreak: An Initial Assessment in Two Spatial Scales." Scientific Reports, vol. 10, no. 1, Oct. 2020, p. 17002. www.nature.com, doi:10.1038/s41598-020-74089-7.

[5] Nakada, Liane Yuri Kondo, and Rodrigo Custodio Urban. "COVID-19 Pandemic: Environmental and Social Factors Influencing the Spread of SARS-CoV-2 in São Paulo, Brazil." Environmental Science and Pollution Research, Sept. 2020. Springer Link, doi:10.1007/s11356-020-10930-w.

[6] Roy, Satyaki, and Preetam Ghosh. "Factors Affecting COVID-19 Infected and Death Rates Inform Lockdown-Related Policymaking." PLOS ONE, vol. 15, no. 10, Oct. 2020, p. e0241165. PLoS Journals, doi:10.1371/journal.pone.0241165.

[7] Nakada, Liane Yuri Kondo, and Rodrigo Custodio Urban. "COVID-19 Pandemic: Environmental and Social Factors Influencing the Spread of SARS-CoV-2 in São Paulo, Brazil." Environmental Science and Pollution Research, Sept. 2020. Springer Link, doi:10.1007/s11356-020-10930-w.

[8] Lewis, Dyani. "Why Many Countries Failed at COVID Contact-Tracing — but Some Got It Right." Nature, vol. 588, no. 7838, Dec. 2020, pp. 384–87. www.nature.com, doi:10.1038/d41586-020-03518-4.

[9] "COVID-19 May Cause Decrease in Life Expectancy This Year." Healthline, 22 Sept. 2020, https://www.healthline.com/health-news/covid-19-may-cause-decrease-in-life-expectancy-this-year.

[10] Diabetes and COVID-19. https://www2.hse.ie/conditions/coronavirus/diabetes-and-coronavirus.html. Accessed 21 Feb. 2021.

[11] McGreevy, Ronan. "Covid-19: Majority of Fatalities Had Underlying Conditions." The Irish Times, https://www.irishtimes.com/news/ireland/irish-news/covid-19-majority-of-fatalities-had-underlying-conditions-1.4471443. Accessed 21 Feb. 2021.

[12] World Health Organisation, Coronavirus disease 2019 (COVID-19) Situation Report – 51, 11 March 2020.

[13] "Coronavirus: What Is the R Number and How Is It Calculated?" BBC News, 19 Feb. 2021. www.bbc.com, https://www.bbc.com/news/health-52473523.

[14] "Coronavirus: How Infection Rates Are Changing across Europe and What It Means for Ending Lockdown." Sky News, https://news.sky.com/story/coronavirus-how-infection-rates-are-changing-across-europe-and-what-it-means-for-ending-lockdown-11981382. Accessed 21 Feb. 2021. Week 06, 2021. https://covid19-country-overviews.ecdc.europa.eu/. Accessed 21 Feb. 2021.

[15] Vadlamani, Ranjan, and JH Bloomberg School of Public Health. "Page Not Found." Johns Hopkins Bloomberg School of Public Health, https://www.jhsph.eduhttps://www.jhsph.edu/404.html?SEUseVersionObject=9FD79F3CD822C4 Accessed 21 Feb. 2021.

[16] "Strong Decline in Coronavirus across England since January, React Study Shows." BBC News, 18 Feb. 2021. www.bbc.com, https://www.bbc.com/news/health-56098313.

[17] "Europe: COVID-19 Vaccination Rate by Country 2021." Statista, https://www.statista.com/statistics/1196071/covid-19-vaccination-rate-in-europe-by-country/. Accessed 21 Feb. 2021.

[18] Wibbens, Phebo D., et al. "Which COVID Policies Are Most Effective? A Bayesian Analysis of COVID-19 by Jurisdiction." PLOS ONE, vol. 15, no. 12, Dec. 2020, p. e0244177. PLoS Journals, doi:10.1371/journal.pone.0244177.

[19] https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker

Appendix

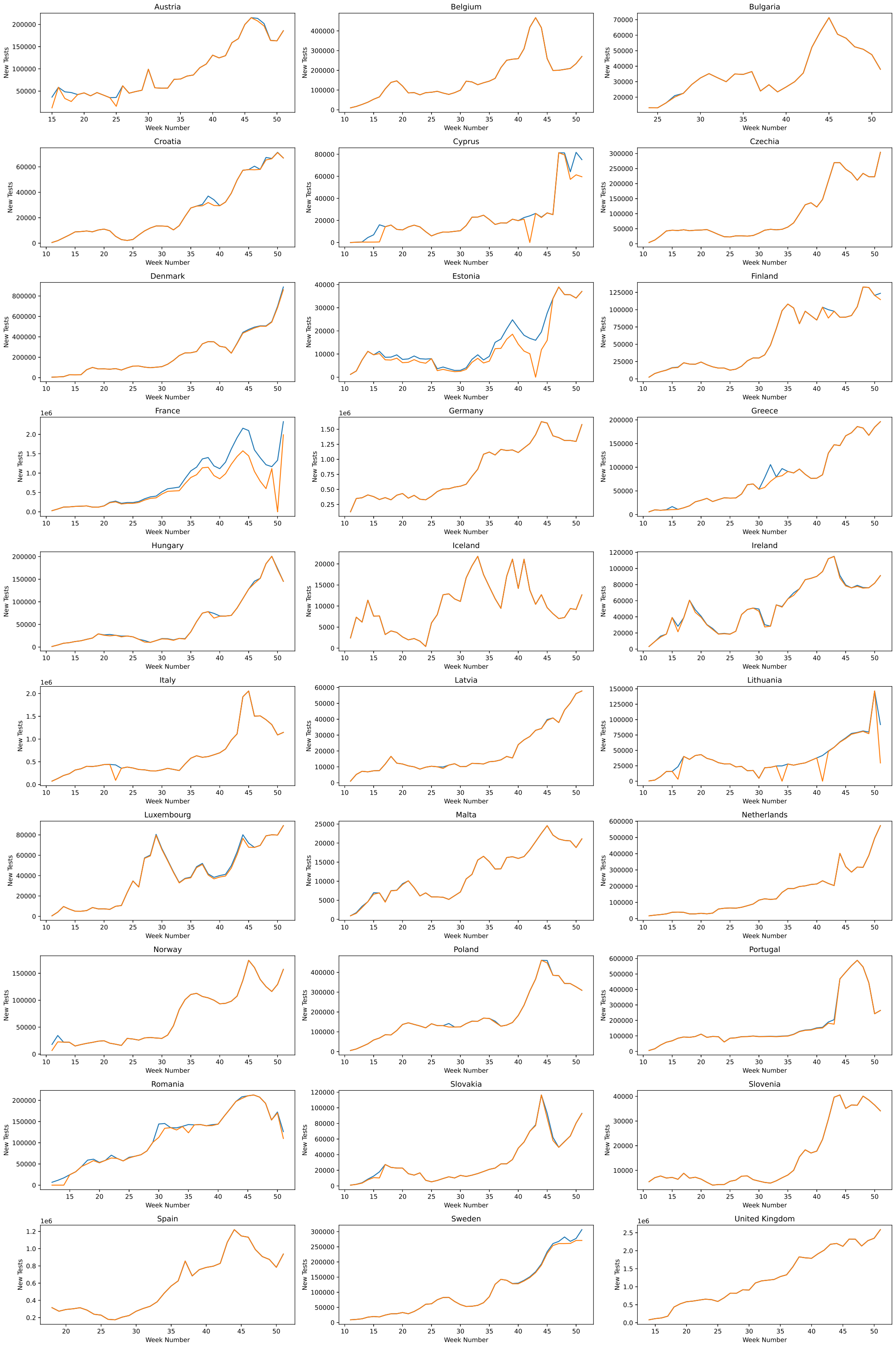


Figure 11: Full Testing Data

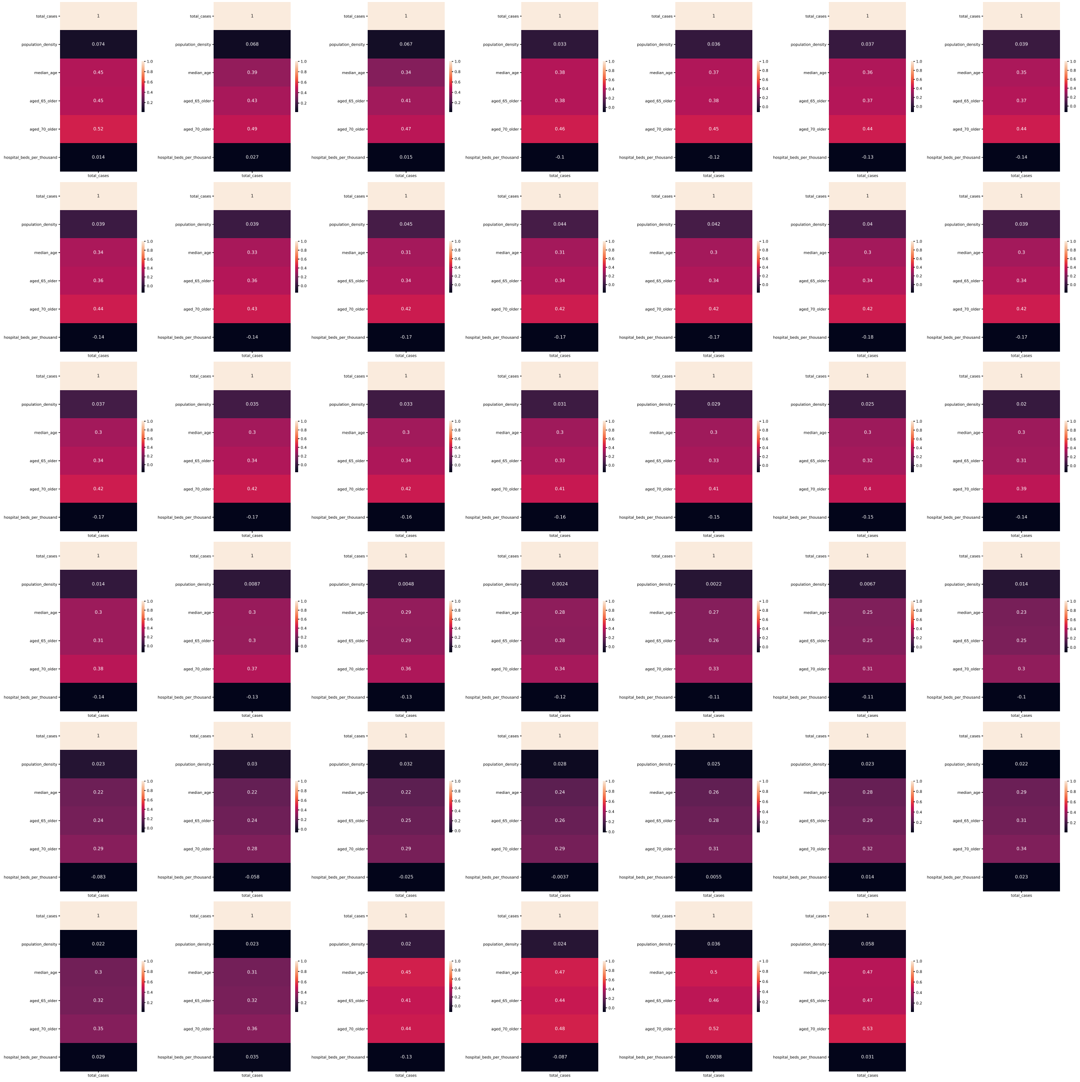


Figure 12: Full Heat Map