Date: June 15, 2020

To: Mr. Daniel Gutierrez

From: Bassam Kaaki

Subject: Data Science Project (Covid-19)

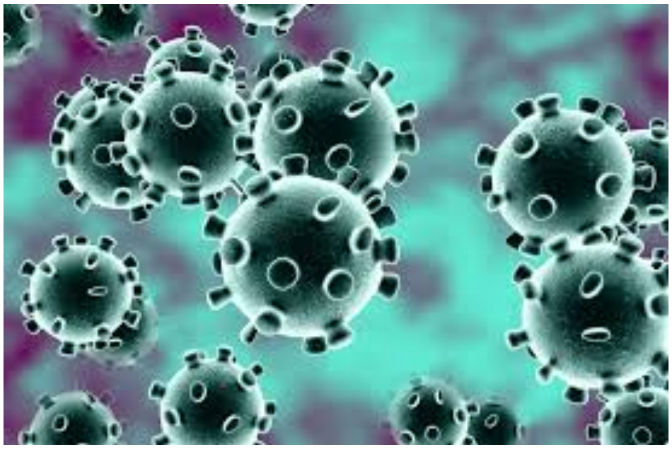


Table of Contents

Project Introduction

The project I am investigating is about the Corona Virus, known as COVID-19. It is regarded as an infectious disease that has mutated from SARS virus into a newly discovered corona virus. People infected with this disease will have mild to moderate respiratory symptoms and can be able to recover without the need for special treatment. However, people with underlying medical issues such as a heart problem, high blood pressure, diabetes, low immunity, cancer and chronic respiratory disease, will develop symptoms that are likely to be more serious and can be fatal. The virus can be passed to others through the spread of droplets of saliva during communication or any discharge from the nose, usually transmitted when a person coughs or sneezes. As of the current date and time, there is no vaccine to counterattack this critical pandemic which has affected the world in terms of deaths, critical cases, the economy and jobs, including a heavy burden on medical staff and equipment.

Project Goal

The goal of this project is to provide proof that testing is an important aspect in curbing the Covid-19 cases that are happening around the world and what type of case transmissions are causing the spread. I was fond of this subject because living in the U.S, I saw that the mandatory lock down issued on March 19, 2020 was not very firmly followed and people seemed to be moving from place to place with ease without considering the importance on the governmental rules and regulations for maintaining the "stay at home order". Even travel was not restricted as it should have been, and flights continued to resume in and out of LA, even when the lock down was put in place. The same is happening in other parts of the world (Kuwait for example where I come from), has been experiencing an increase in Corona virus infections lately after having a steady rate. Although a strict lock down has been put in place, several individuals continue to visit family and friends. Another issue is the open bridge for traveling that was set by the Kuwaiti government, to bring back all its nationals stuck in the US and other parts of the world in Europe and Asia.

Project Hypothesis

My hypothesis will revolve around the following: "**Testing for the corona virus is important for quick identification of cases and transmission type for immediate isolation to**

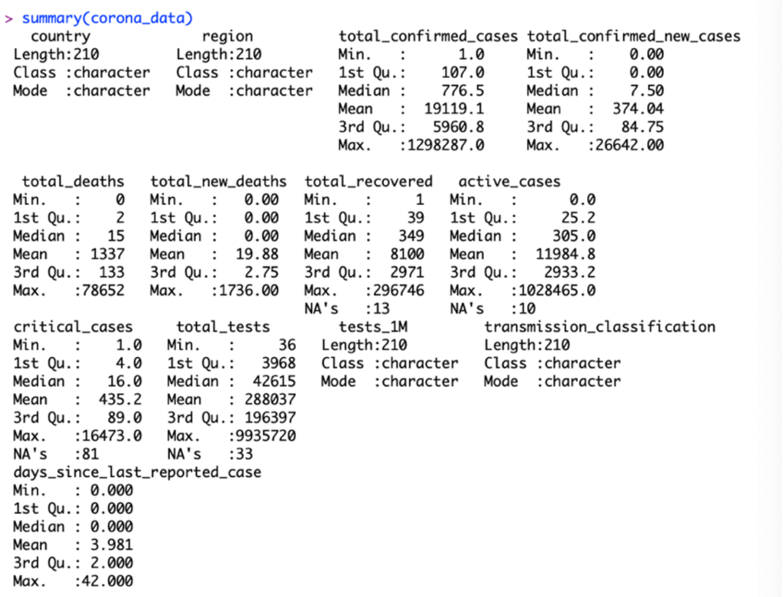
prevent the spread". We need to understand the different case types associated with the infections reported worldwide. There are 4 types of transmission classifications, which will help in identifying the infection types to curb the spread. **Sporadic cases** are considered as imported from travel or locally detected. **Clusters of cases** are infections clustered in time, geographic locations and/or by common exposures. **Community transmissions** are cases that are caused by large infectious outbreaks with the inability to relate to the case through a chain of transmission, large cases from sentinel lab surveillance or multiple unrelated clusters in a country/area/territory. **Pending cases** are those that have not been reported as either sporadic, cluster of cases, or community transmissions since their origin is not known yet.

Data Munging

Dataset Description

Two final dataset files were used as PDF files produced by the World Health Organization. The first one is the global situational report numbered "95" found on the following link: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200424-sitrep-95-covid-19.pdf?sfvrsn=e8065831_4 which was produced on (April 24th 2020) and the second global situation report numbered 113" produced in (May 12, 2020), from the following address also associated to the World Health Organization: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200512-covid-19-sitrep-113.pdf?sfvrsn=feac3b6d_2. This dataset contains country, region, total_confirmed_cases, total_confirmed_new_cases, total_deaths, total_new_deaths, transmission_classification and days_since_last_report columns. Two more additional datasets were used from https://www.worldometers.info/coronavirus one for April 24, 2020 and one for May 12, 2020. This dataset contained additional columns which I wanted to place in both the final datasets I worked with called corona_data_proj for April situation report and corona_data for May's situation report. The columns included are total_recovered, active_cases, critical_cases, total_tests and test_1M. The extraction of this website was needed to be taken immediately for these columns for April 24th report and May 12, 2020, since they don't keep track of all the days reports under separate links to be accessed in the future, as the World Health Organization does. As such the final output were two datasets used for comparison purposes. Most of the cleaning work for the worldometers site was done in excel due to limitations in data not being stored to be reviewed in the future and second due to the nature in WHO's PDF files. However, in the R file, the code is shown where data munging was used in both datasets and how it could have been performed using R on the April dataset where excel cleaning was more dominant.

Before any data munging took place, both the corona_data_proj and corona_data excel sheets required to be imported and transformed from a tibble to a dataframe. The below is an example summary of the corona_data dataset (May 12 report) after the data.frame conversion was made.



```
> summary(corona_data)
   country              region          total_confirmed_cases total_confirmed_new_cases
 Length:210         Length:210         Min.   :      1.0      Min.   :    0.00
 Class :character   Class :character   1st Qu.:    107.0      1st Qu.:    0.00
 Mode  :character   Mode  :character   Median :    776.5      Median :    7.50
                                       Mean   :  19119.1      Mean   :  374.04
                                       3rd Qu.:   5960.8      3rd Qu.:   84.75
                                       Max.   :1298287.0      Max.   :26642.00

  total_deaths   total_new_deaths  total_recovered   active_cases
 Min.   :    0   Min.   :   0.00   Min.   :     1   Min.   :      0.0
 1st Qu.:    2   1st Qu.:   0.00   1st Qu.:    39   1st Qu.:     25.2
 Median :   15   Median :   0.00   Median :   349   Median :    305.0
 Mean   : 1337   Mean   :  19.88   Mean   :  8100   Mean   :  11984.8
 3rd Qu.:  133   3rd Qu.:   2.75   3rd Qu.:  2971   3rd Qu.:   2933.2
 Max.   :78652   Max.   :1736.00   Max.   :296746   Max.   :1028465.0
                                   NA's   :13       NA's   :10
 critical_cases      total_tests        tests_1M          transmission_classification
 Min.   :    1.0   Min.   :     36   Length:210         Length:210
 1st Qu.:    4.0   1st Qu.:   3968   Class :character   Class :character
 Median :   16.0   Median :  42615   Mode  :character   Mode  :character
 Mean   :  435.2   Mean   : 288037
 3rd Qu.:   89.0   3rd Qu.: 196397
 Max.   :16473.0   Max.   :9935720
 NA's   :81        NA's   :33
 days_since_last_reported_case
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 0.000
 Mean   : 3.981
 3rd Qu.: 2.000
 Max.   :42.000
```

Creating categorical variables

In order to start the analysis of the corona_data dataset, some variables needed to be taken care off.

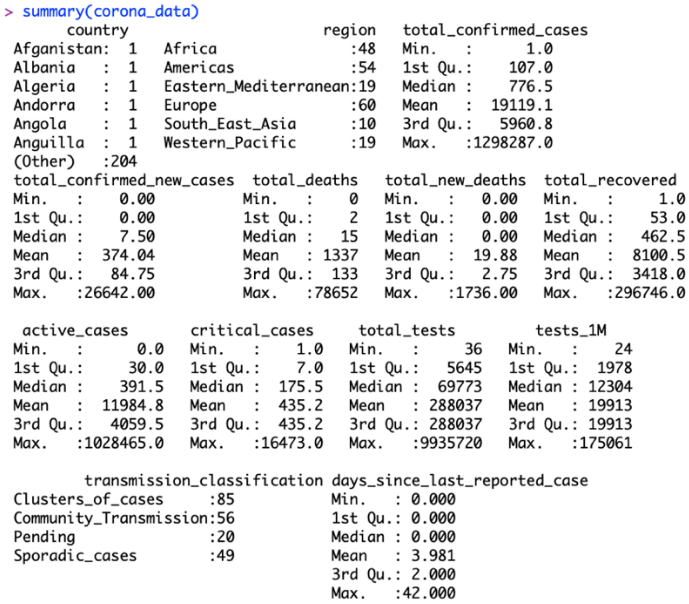
- **country:** Is a character class and needs to be changed to as.factor so we can be able to see all the countries within their groups.
- **region:** Is a character class and need to be changed to as.factor so we can be able to see the different regions.
- **transmission classification:** Is a character class and needs to be changed to as.factor, so we can view the different categories under it.

- **tests_1M:** Is a character class and needs to be changed to numeric so we can be able to work with it in R.

Handling Missing Data

- total_recovered: contained 13 NA's
- active_cases: contained 10 NA's
- critical_cases: contained 81 NA's
- total_tests: contained 33 NA's

To get rid of the NA's, it was not a good idea just to remove them completely from the data set because critical cases alone had 81 NA's. As such a wise idea was to use the means of each of these columns and fill the NA's instead. The final output of the summary after the NA's have been removed and the creation of the categorical variables looks as below:



```
> summary(corona_data)
       country                        region   total_confirmed_cases
 Afganistan:  1   Africa               :48   Min.   :      1.0     
 Albania   :  1   Americas             :54   1st Qu.:    107.0     
 Algeria   :  1   Eastern_Mediterranean:19   Median :    776.5     
 Andorra   :  1   Europe               :60   Mean   :  19119.1     
 Angola    :  1   South_East_Asia      :10   3rd Qu.:   5960.8     
 Anguilla  :  1   Western_Pacific      :19   Max.   :1298287.0     
 (Other)   :204                                                    
 total_confirmed_new_cases  total_deaths    total_new_deaths  total_recovered   
 Min.   :    0.00          Min.   :    0   Min.   :   0.00   Min.   :     1.0  
 1st Qu.:    0.00          1st Qu.:    2   1st Qu.:   0.00   1st Qu.:    53.0  
 Median :    7.50          Median :   15   Median :   0.00   Median :   462.5  
 Mean   :  374.04          Mean   : 1337   Mean   :  19.88   Mean   :  8100.5  
 3rd Qu.:   84.75          3rd Qu.:  133   3rd Qu.:   2.75   3rd Qu.:  3418.0  
 Max.   :26642.00          Max.   :78652   Max.   :1736.00   Max.   :296746.0  

  active_cases       critical_cases      total_tests          tests_1M     
 Min.   :      0.0  Min.   :    1.0   Min.   :     36   Min.   :    24  
 1st Qu.:     30.0  1st Qu.:    7.0   1st Qu.:   5645   1st Qu.:  1978  
 Median :    391.5  Median :  175.5   Median :  69773   Median : 12304  
 Mean   :  11984.8  Mean   :  435.2   Mean   : 288037   Mean   : 19913  
 3rd Qu.:   4059.5  3rd Qu.:  435.2   3rd Qu.: 288037   3rd Qu.: 19913  
 Max.   :1028465.0  Max.   :16473.0   Max.   :9935720   Max.   :175061  

         transmission_classification days_since_last_reported_case
 Clusters_of_cases       :85          Min.   : 0.000               
 Community_Transmission:56            1st Qu.: 0.000               
 Pending                 :20          Median : 0.000               
 Sporadic_cases          :49          Mean   : 3.981               
                                      3rd Qu.: 2.000               
                                      Max.   :42.000               
```

Column name changes

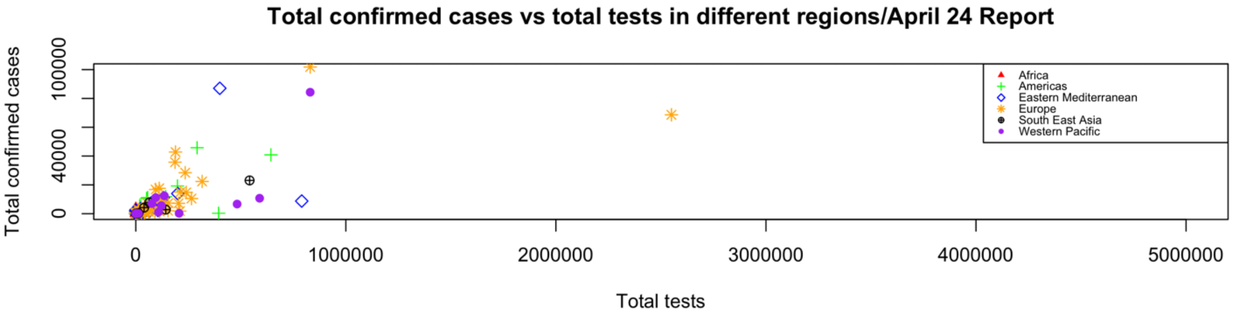
The column names were changed in excel from capital cases to lower cases and separated by an underscore without spacings in between words. The code for the column name changes can be found in the r document incase these columns were to be changed using R language.

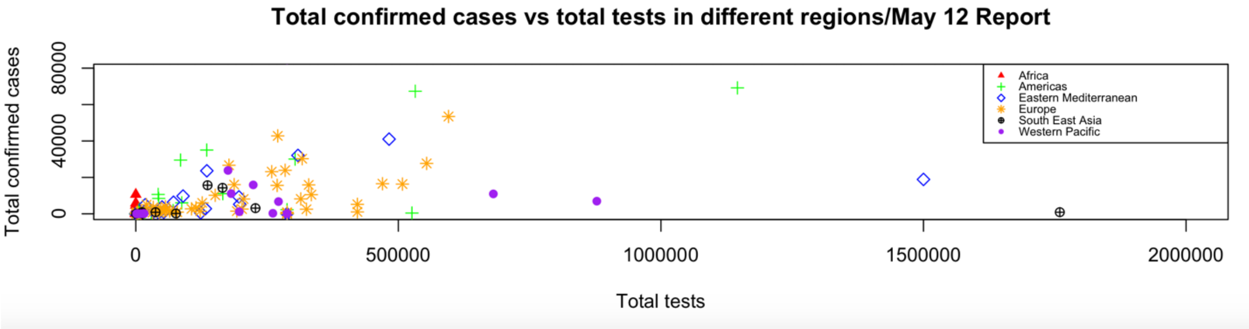
EDA and Statistical analysis

To gain a deep familiarity with the datasets used, exploratory data analysis will be presented below to set the stage for supervised and unsupervised statistical learning.

Total confirmed cases vs Testing in different regions

If we look at the below scatterplot for April 24 report, we can depict that Europe had the highest number of total confirmed cases of around 100,000 with more than 800,000 tests being made. The least region that performed testing was the Western Pacific including South East Asia. In comparison with the May 12 report, we can see that South East Asia, particularly India, increased its testing and had the highest number of tests made worldwide with 1,759,579 tests and 50,000 total confirmed cases. We can also depict that in the African region in both reports there are very few tests made to confirm cases for Covid-19.

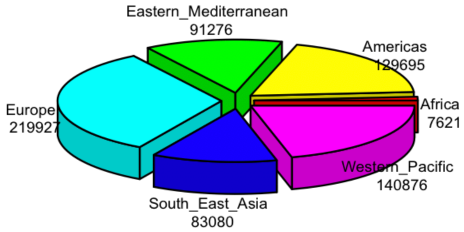




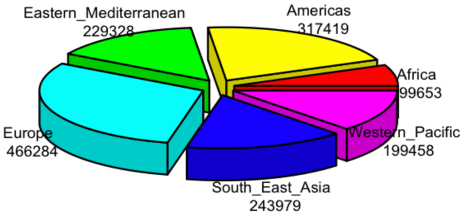
Average total tests in different regions

Most of the testing is taking place in Europe with an average of 219,927 tests. Second place comes the Americas with an average of 129,695 tests. Third comes South East Asia with 83,080 tests. The 4th region for testing is the Eastern Mediterranean with 91,276 tests. In 5th place is the Western Pacific with 140,876 tests and finally the African region with the lowest testing average of 7,621 compared to the rest of the regions. If we compare these results to May 12 Report, we can see that all regions have increased their testing for the virus by the following percentages: Europe [112%], Eastern Mediterranean [151%], Americas [144%], Africa [1,207%], South East Asia [193%] and finally Western Pacific [41.60%].

Average total tests in different regions, April 24 report



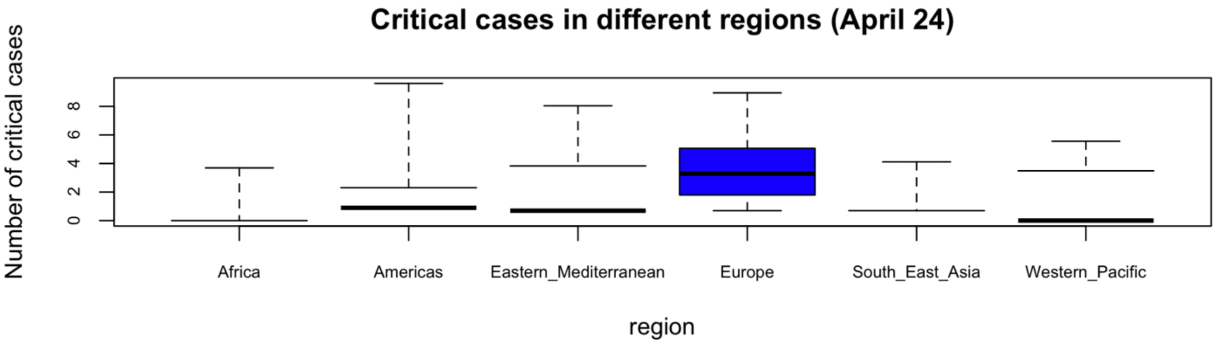
Average total tests in different regions, May 12 report

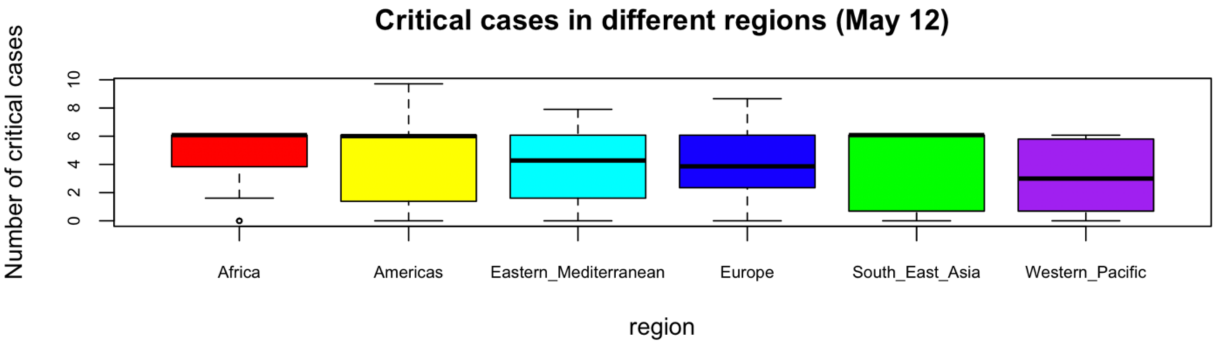


Critical cases in different regions

Taking into consideration the April 24 report, below is a distribution of critical cases in each quartile below. Most of the critical cases in different regions are concentrated in the 4th quartile with around 14,932 critical cases. Most of the critical cases are seen in the Americas and Europe followed by the Eastern Mediterranean regions and the least in Africa as reported.

Quartile1	Quartile2	Quartile3	Quartile4
0	2	36.25	14932





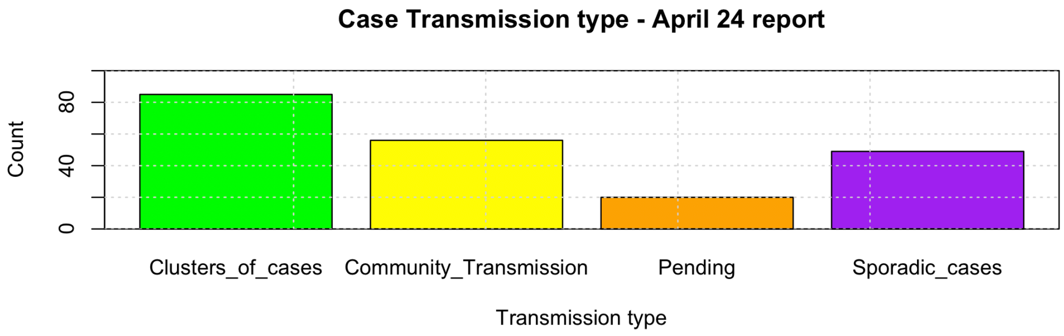
If we look at the critical cases quartiles table for May 12 report for all the regions worldwide, we begin to see a clearer pattern with a wider distribution of cases as the days progressed in the following manner:

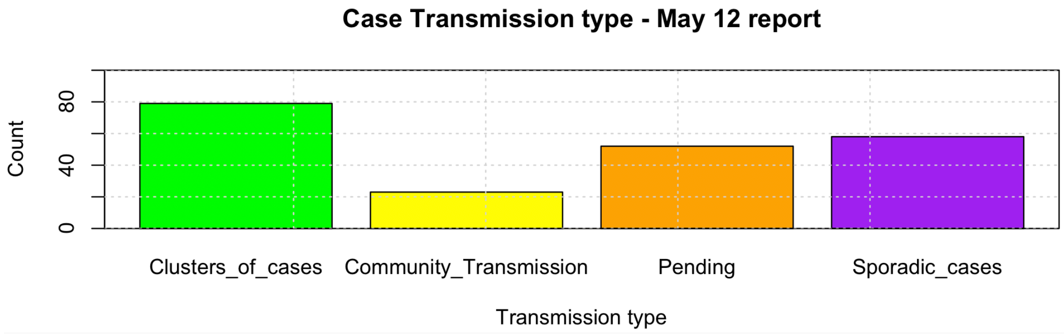
Quartile1	Quartile2	Quartile3	Quartile4
7	175.50	435.2093	16473

Critical cases are more heavily concentrated in the 4th quartile than any other quartile shown. Most of these cases are found in the Americas and Europe and the least in South East Asia and the Western Pacific. The box plot clearly identifies the critical cases from 6,000 – 9,000 in the Americas, in Europe between 6,000 to 8,500 in their 4th quartiles. The lowest median critical cases are found in the Western Pacific. The median critical cases are close for the Americas and South East Asia. It appears as well that critical cases in the Eastern Mediterranean are the third highest between 6,000 – 7,900 cases. South East Asia is interesting in that most of its critical cases are concentrated below the median number of cases for that region.

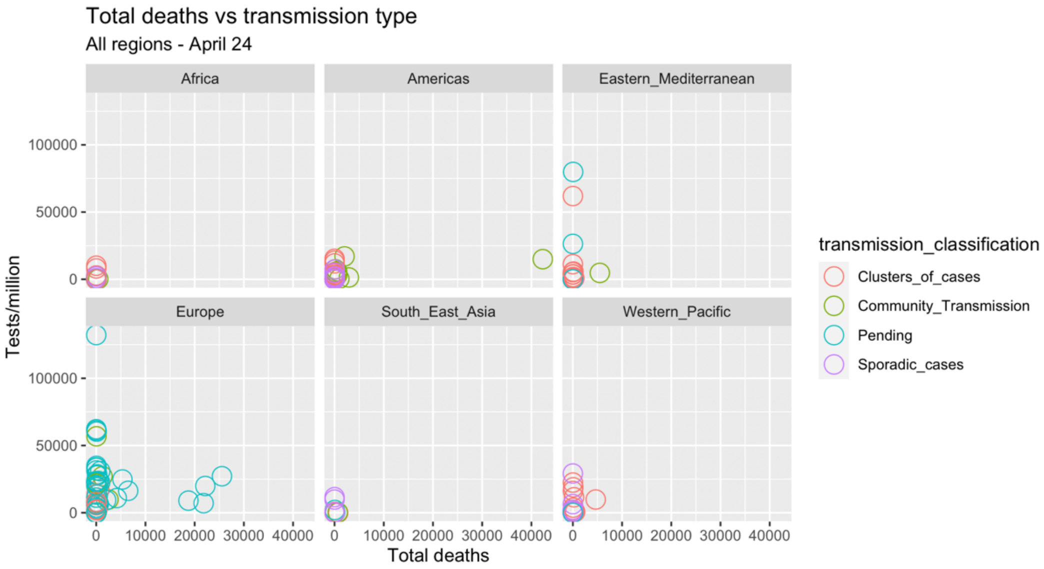
Transmission Cases

In viewing the below bar plots, we can deduce that in both April and May reports the number of cases stated by countries is mostly concentrated in the cluster of cases transmission type. Community transmission has decreased in May's report since it was higher in April's report. Country classifications have increased in May's report from 50 to 60 countries reporting cases widely spread in the sporadic type. Some countries don't report cases in a timely fashion as others do.

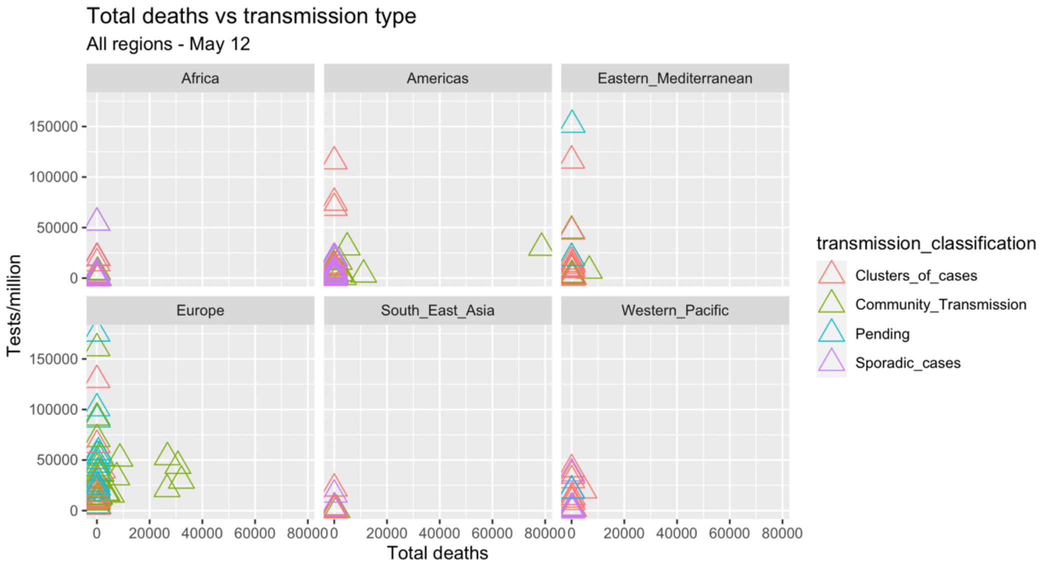

Case Transmission type - April 24 report


Case Transmission type - May 12 report

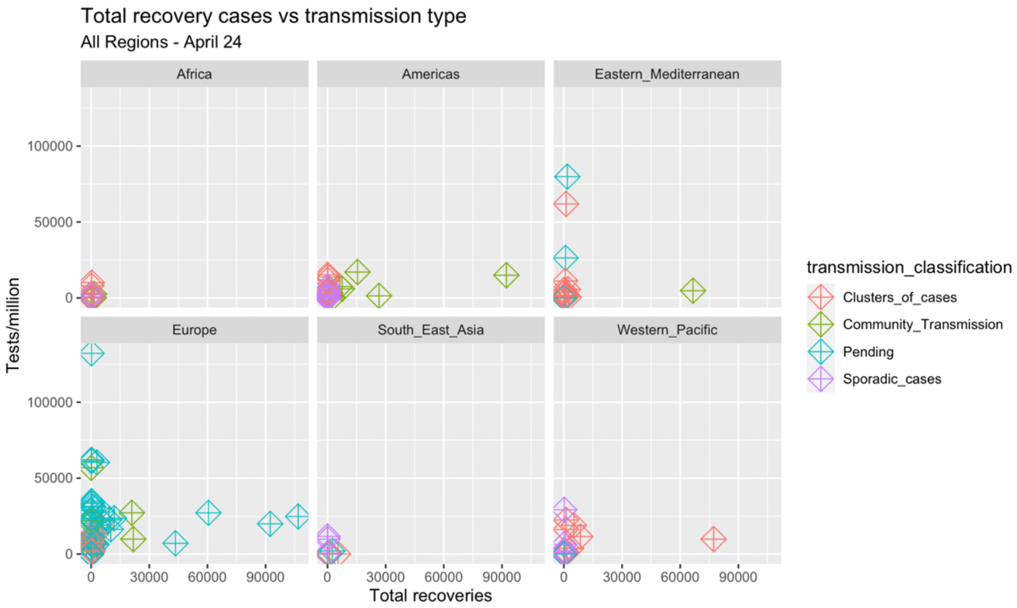
Most of the death cases of which were tested for April 24 report are visible under the pending classification in the European region. In the Americas more of the death cases are classified under community transmission. In Africa, the tested death cases that were reported are classified under clusters of cases. In the Eastern Mediterranean most of the tests made have been classified under pending transmission, however deaths that occurred in that region were mainly a community transmission.



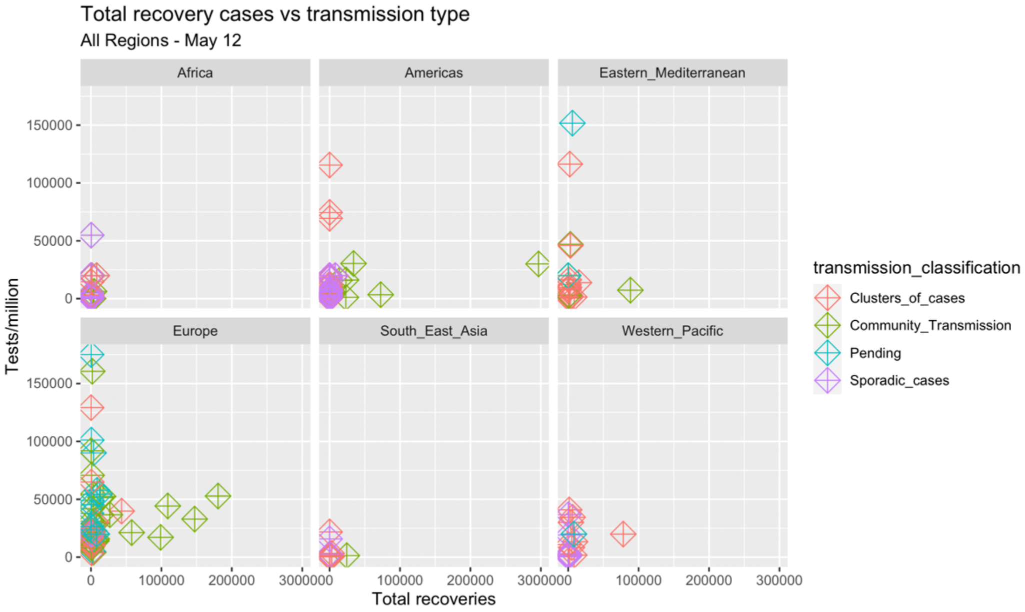
By May 12 report, in Africa the tested cases where deaths were concerned changed from cluster of cases to sporadic cases. Europe changed from pending to community transmission and continues to have the highest numbers of deaths.



As depicted from the April 24 report most of the tested cases that recovered in the European region were classified under the pending cases. In the Americas, most of the total recoveries were classified under the community transmission type. In the Western Pacific most of the tested cases that recovered were classified under clusters of cases.



As we can see from the below graphs for May 12 report Africa test cases that recovered have moved from clusters of cases type to sporadic cases. In Europe however, the recovered tested cases have changed type from pending to community transmission. Most of the regions show more recoveries for their tested corona cases in the community transmission type than any other type.

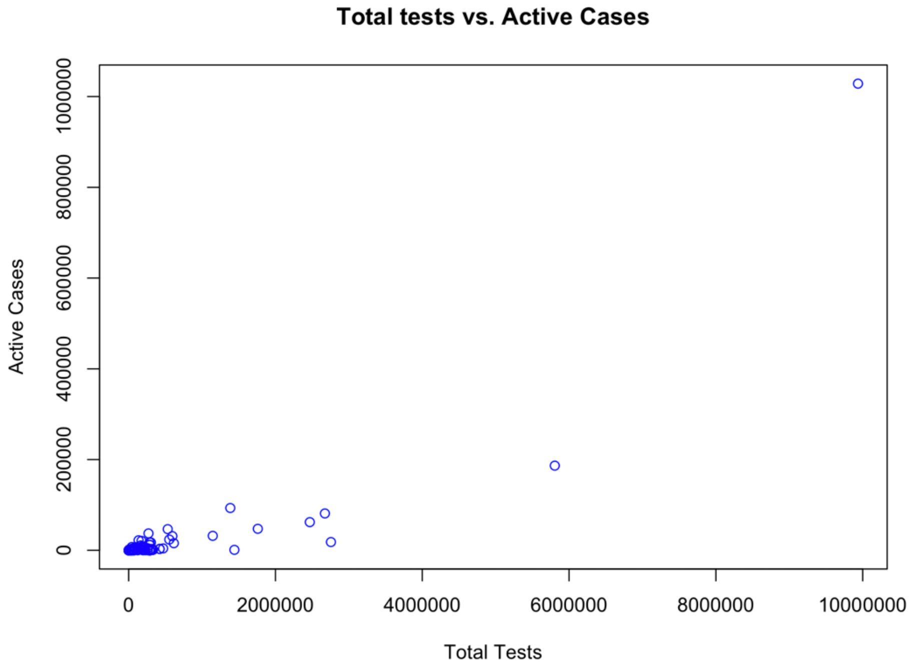


Supervised Machine Learning – Multiple Regression Analysis

Before conducting the multiple linear regression analysis, the May 12 report was chosen, and the data was split into a training set and a testing set. The training set was 60% of the dataset while the testing set was 40% of the dataset. The total number of observations was 210 as such the split was 126 observations for the training set and 84 observations for the testing set.

Relationship between the response variable and predictors

While plotting various predictors against total tests and total_tests1M, the best fitted for predicting was the total_tests variable as shown below.



What are the most important predictors?

The most important predictors are total confirmed new cases, total recovered and critical cases showing 3 stars next to each variable with the lowest pr values. Next important predictors with two stars are total confirmed new cases and total deaths, and with a 1 star follows total new deaths and finally tests_1M. The below is the outcome of the regression model using all the variables to predict total tests as the response variable.

```
Coefficients:
                            Estimate Std. Error t value       Pr(>|t|)    
(Intercept)               96690.1824 28790.7084   3.358       0.001059 ** 
total_confirmed_cases       -10.1942     3.4447  -2.959       0.003732 ** 
total_confirmed_new_cases   443.6722    63.7154   6.963 0.000000000207 ***
total_deaths                 36.5803    11.7361   3.117       0.002300 ** 
total_new_deaths          -3167.4028  1365.2456  -2.320       0.022073 *  
total_recovered              24.7037     3.5620   6.935 0.000000000238 ***
active_cases                 11.5160     3.1759   3.626       0.000428 ***
critical_cases             -312.6989    49.4485  -6.324 0.000000004833 ***
tests_1M                      1.9597     0.8009   2.447       0.015904 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 254100 on 117 degrees of freedom
Multiple R-squared:  0.9498,	Adjusted R-squared:  0.9464 
F-statistic: 276.8 on 8 and 117 DF,  p-value: < 0.00000000000000022
```

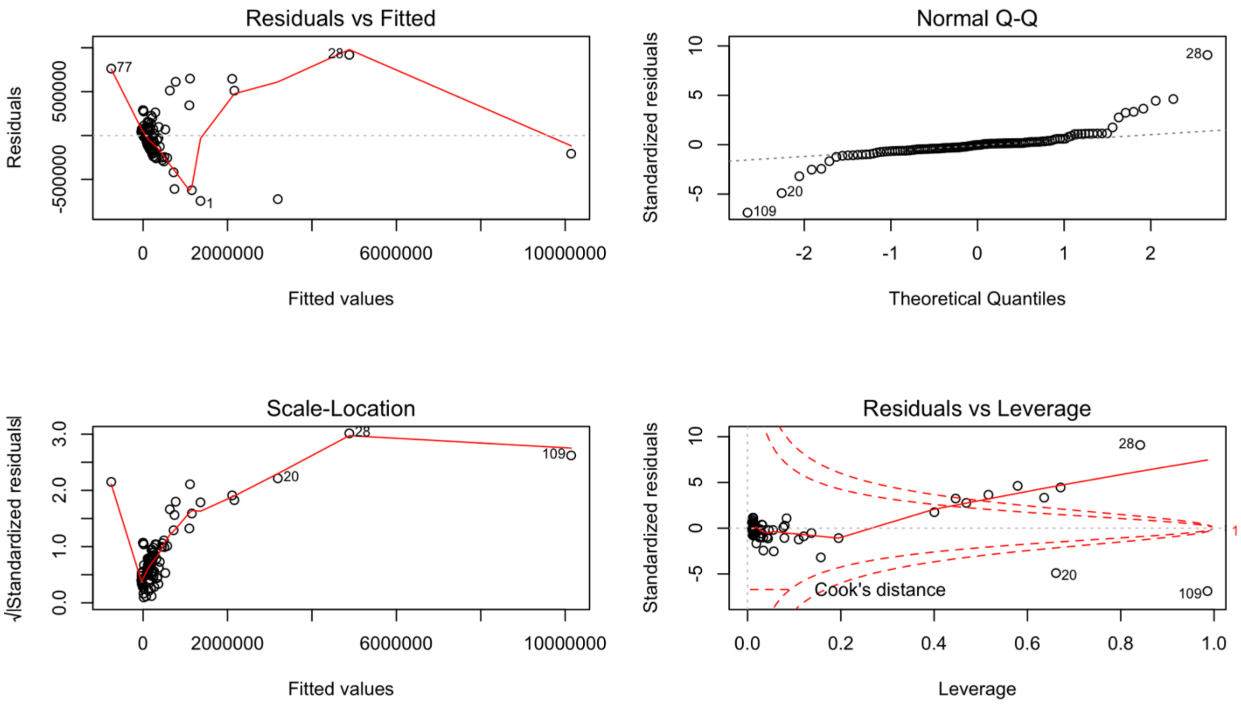
The R-squared above of **0.9498** shows the goodness of the fit of the model using total tests (response variable) to be predicted by the above variables. Next we will use test_corona_data (test set) to predict the response variable total_tests, where the values are already known. In order to check for the accuracy of the prediction of the test set, we will compare the values of total tests with the actual values and we get **0.7122038** which is moderately close to 1. We can say that it is moderately linear related.

Algorithms Performance and Accuracy

Now that we have our multiple regression model formed, we are going to check for the root mean standard error of both the training and test sets. This is to further check how well the model fits the data. From the RMSE calculation performed for the training set on total tests, we get (244903.8/9935720)*100 = **2.464882.** For the RMSE calculation performed on the testing set on total tests we get (235034.2/9935720)*100 = **2.365548**. The RMSE’s for both the training set and testing set are close and corresponds to the average error of 2.46882 + 2.365548/2 = **2.42**, This shows us that the predictor variables using total tests variable (response variable), has an RMSE average of **2.42**, meaning that the observed total tests values deviate from the predicted values approximately 2.42 units in average. The small RMSE number for both the training set and testing set represents a good algorithmic performance and accuracy in the model.

Diagnostic Plots of Multiple Regression Model

- **In Residuals vs Fitted:** The red fitted line is not relatively flat and straight in the residual versus fitted plot. This is due to the outlier numbered 28[Russia], 77[FaroeIslands].
- **QQ plot:** In assessing homoscedasticity, there is a diversion in the tail and head due to the nature of the data in corona cases with countries having higher numbers in certain variables point 28[Russia] and low numbers in other variables point 20[Spain]. However, there is a relatively equal spread around y = 0 showing equal variance if the outliers were not present.
- **In Scale-Location:** Some deviation from the line where we see outliers numbered point 20[Spain], point 28[Russia] and point 109[USA], otherwise most of the data diverged due to the extreme outlier country data in reported numbers in the dataset.
- **In Residuals vs. Leverage:** The measure of distance of the data pull to the mean shows that there are data points crossing the thresholds once again for point 20[Spain], point 28[Russia] and point 109[USA].



To add further, the outlier points have a significant level of shaping the outcome of the dataset analysis. As such no outliers were removed since for further analysis purpose i.e., in the K-means analysis to be conducted below, I could see that the U.S is going to form its own cluster at this point, while Russia and Spain will be together in one cluster.

Unsupervised Machine Learning - Clustering

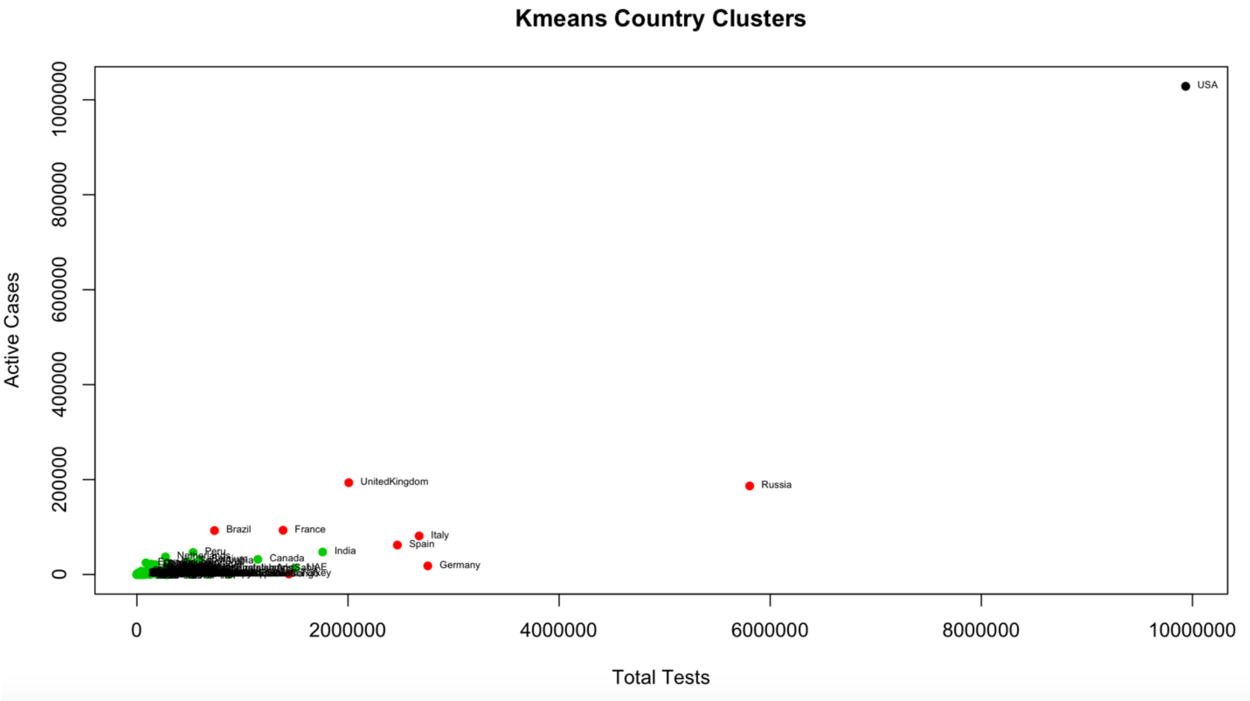
In order to start the k-means clustering analysis, we will remove all factor variables in our case [country], [region] and [transmission_classification]. Due to the nature of the dataset where variables can have either high or low quantitative data, we will normalize these numbers, so all variables have a level playing field so high-level numbers don't dominate the whole show of the clustering analysis. When we normalize the variables the average for each variable becomes 0 and the standard deviation is approximately 1. We will calculate the mean for all variable's columns and similarly for standard deviation. We can then have a normalized dataset to work with.

The total clusters chosen is 3 so we can see all the countries divided into 3 divisions according to distance to one another. The outcome of the cluster analysis gives us the following:

```
K-means clustering with 3 clusters of sizes 1, 9, 200
Cluster means:
  total_confirmed_cases total_confirmed_new_cases total_deaths total_new_deaths
1          13.2292346            12.6847276               11.3805229      13.4047805
2           1.6664970             1.4008130                2.2147359       1.2272729
3          -0.1411385            -0.1264602               -0.1565657      -0.1222512
  total_recovered active_cases critical_cases total_tests    tests_1M
1      9.8038163   13.7438392     11.6591821  11.0329811  0.35953015
2      2.7664225    0.9561076      1.8940393   2.1973941  0.27321850
3     -0.1735081   -0.1117440     -0.1435277  -0.1540476 -0.01409248
  days_since_last_reported_case
1                -0.44283943
2                -0.44283943
3                 0.02214197
(between_SS / total_SS =  70.6 %)
```

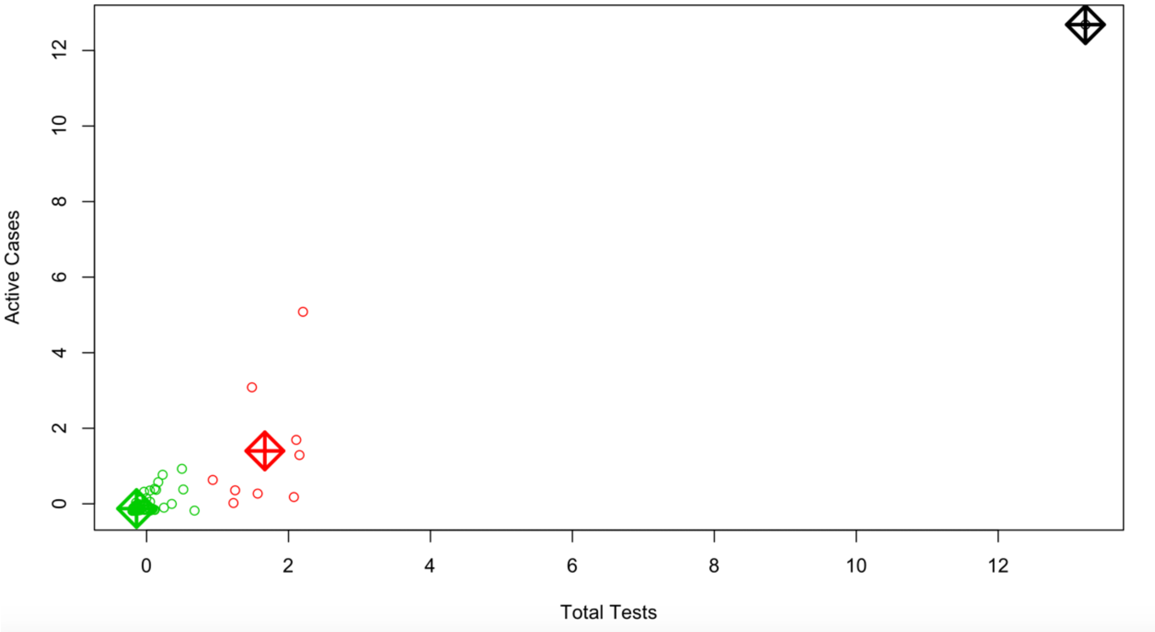
The first cluster has 1 country, the second cluster has 9 countries and the third cluster has 200 countries. The cluster means are all shown for each for the variables associated with the k-means clustering analysis conducted. Within cluster variability for cluster one is [0.0000], second cluster is [146.3867] and third cluster [467.2590]. Each cluster shows the closeness in distance of countries to one another. We can find from the cluster means that the highest number of tests are made in cluster 1 (USA) and the lowest in cluster 3, while most death cases are occurring in cluster 1 (USA) and less in cluster 3 (all other countries). Most of the critical cases are in cluster 1 (USA) while less are in cluster 3. Total confirmed new cases are more as well in cluster 1 and less in cluster 3 (all other countries). Total recoveries are highest in cluster 1 (USA) and second comes cluster 2 (Iran, Peru, Brazil, France, UK, Italy, Spain, Germany and Russia). Least recoveries are occurring in cluster 3 (all other countries).

We will plot active cases variable with total tests variable to see the cluster formations in relations to these variables as an example.

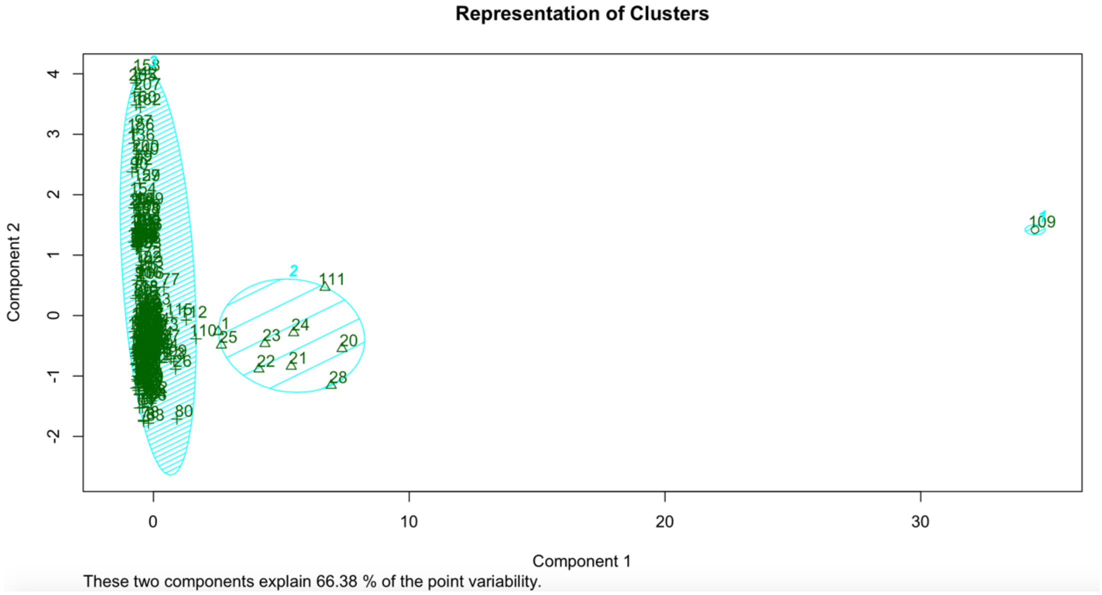


From the k-means cluster graph we can see that the total tests made of around 9,000,000 for the USA confirms active cases of above 1,000,000 (black point). The red points show cluster 2 with the nine countries explained earlier and the third with the 200 countries in green. The distance from Germany for example to USA is very large in terms of the total tests made to active cases and is explained with around 2,800,000 total tests producing around 50,000 active cases.

The below graph will show the centers of each of the three clusters. As mentioned earlier, the USA, which was an outlier in the diagnostic plots, if removed could have resulted in a loss of a cluster and could have resulted in a different analysis outcome. As such, this outlier was not removed because of its importance to the contribution to the k-means analysis made.



In the representation of clusters graph below, the 3 clusters show hardly any overlap between them. Cluster 1 is the USA showing the furthest distance from cluster 1 and cluster 2. As can be seen, clustering is good when between cluster distance is high and within cluster distance is low. In this case we get to see in the cluster centers that there is a good separation, which is a good sign in explaining the different circumstances that are occurring in these different countries in terms of the different variables they are affected by.



Conclusion

The multiple regression study using total tests as the response variable being predicted by other variables such as active cases, total new confirmed cases, death cases etc., showed an R-squared of **0.9498** which provided evidence that testing is a first measure prevention to inhibit disease spread. As more tests are made, more active cases can be detected and treated at an earlier stage to prevent future complications. The following examples show the prediction made from using the coefficients and the trained model to make predictions for the new data (test data) when 500 active and 1000 cases are recorded. It states from the following calculation (coef(lm2)[1] + coef(lm2)[7]*500) = 102,448 (the number of predicted tests made for having the 500 cases are 102,488). If we increase the number of active cases to 1,000, we would require 1,082,061 tests to be made to reach the 1,000 cases mark, coef(lm2)[1] + coef(lm2)[7]*1000 = 1,082,061. This shows a linear relationship being formed, which gives us that more active cases are recorded when more tests are being conducted in countries worldwide.

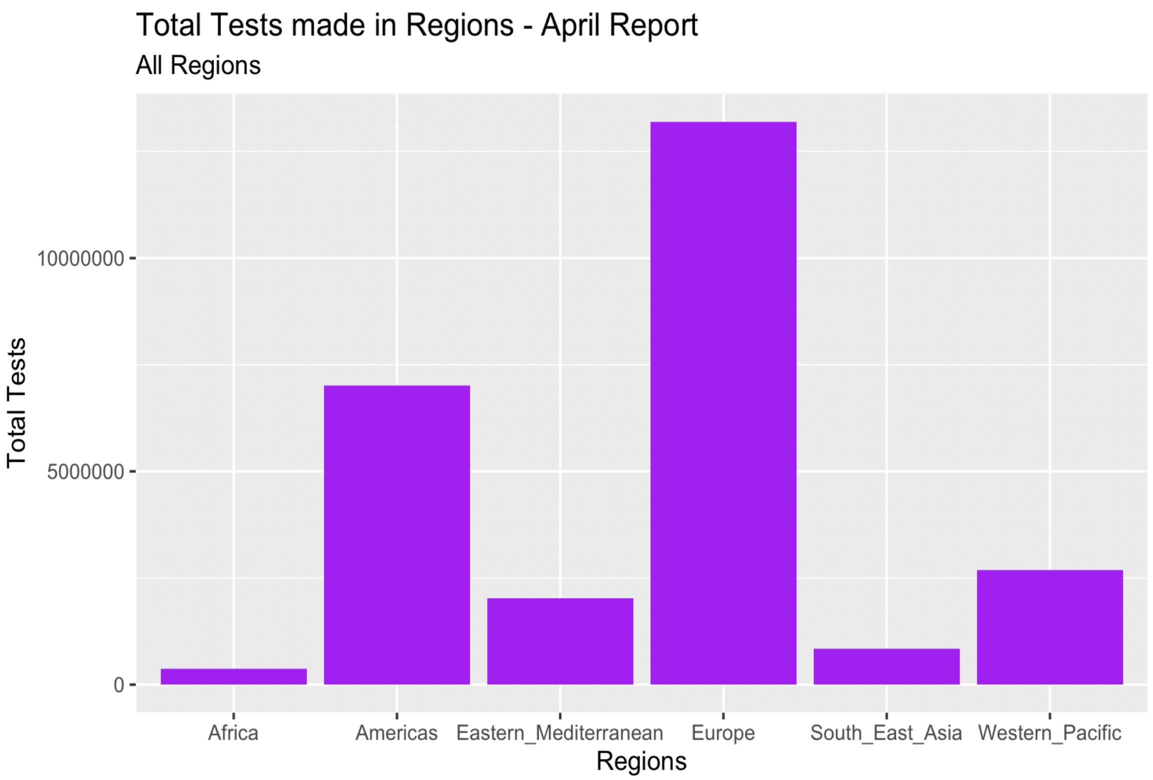
There are five reasons why the diagnostic plots are right skewed. The first reason relates to the fact that there are countries that don’t report any death activities or testing being made to confirm any corona activity, whereas we have the United States with the most tests being made in all regions of the world. The second reason is that some countries experience very low death rates compared to others and regions where low death cases are seen are recorded under the African region, parts of Asia, the Middle East and less populated islands such as those belonging to U.S, Europe and Britain. The third reason is that there are countries with low populations that seem to have lower population rates and less corona activity such as Greenland and Saint Barthelemy to name a few, which record low numbers of cases to the World Health Organization. The fourth reason is that there are high risk countries such as the USA, Russia and the United Kingdom, Brazil, Italy and Spain that record a very high corona activity. The fifth reason is that there are countries that don't wish to share information and are non-transparent such as North Korea, Yemen, Turkmenistan and Iran. As such lack of information from such countries would cause a pull to the right and contribute to a pull in the tails of the diagnostic plots as well.

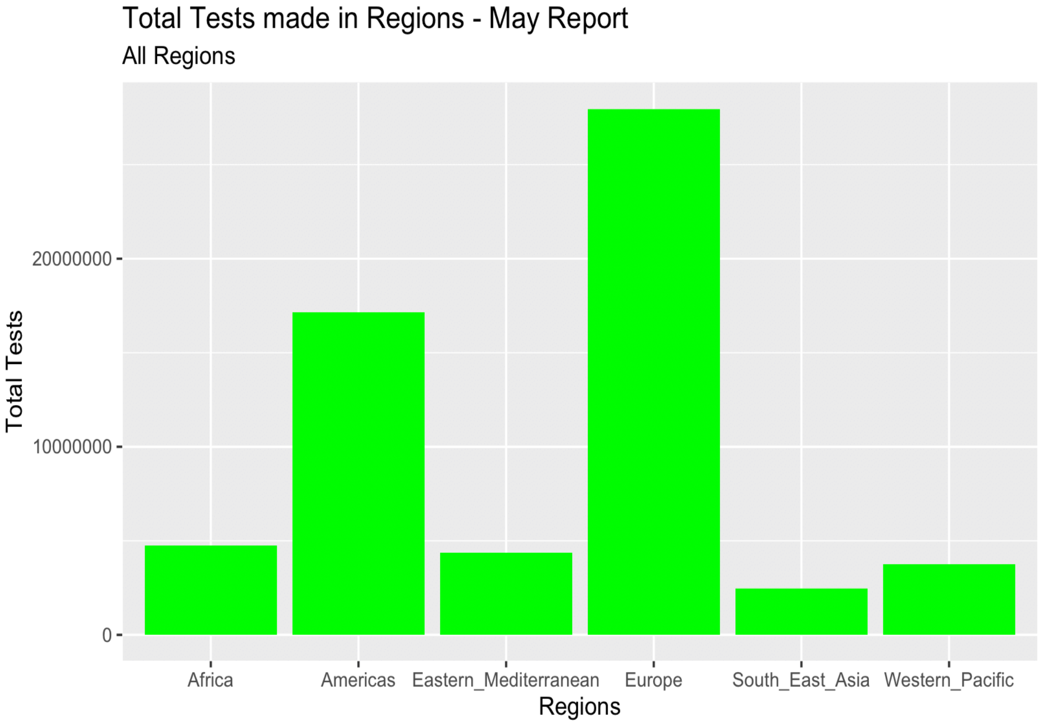
We can also conclude that less tests are made when active cases are low in regions such as the relationship outlined in the K-means clustering. The K-means clustering also showed that there is a relationship with more tests being made and increases in active cases taking place.

In terms of the case transmission types, it is very interesting to see that the most cases come from what is highlighted as "clusters of cases" in both April's report and May's report as well. These clusters of cases are infections that are clustered in time, geographic locations and/or by common exposures. However, sporadic case transmission type has increased in May's report than in April's report due to more tests being made assessed as sporadic cases.

Transmission type seems to influence type of tests made. When tests were performed at a later stage as in the case with Africa in the death's vs transmission type scatterplot earlier, they seem to change as months progress. African transmission cases were marked as clusters of cases in April report and in May's report turned to be sporadic cases as was shown earlier.

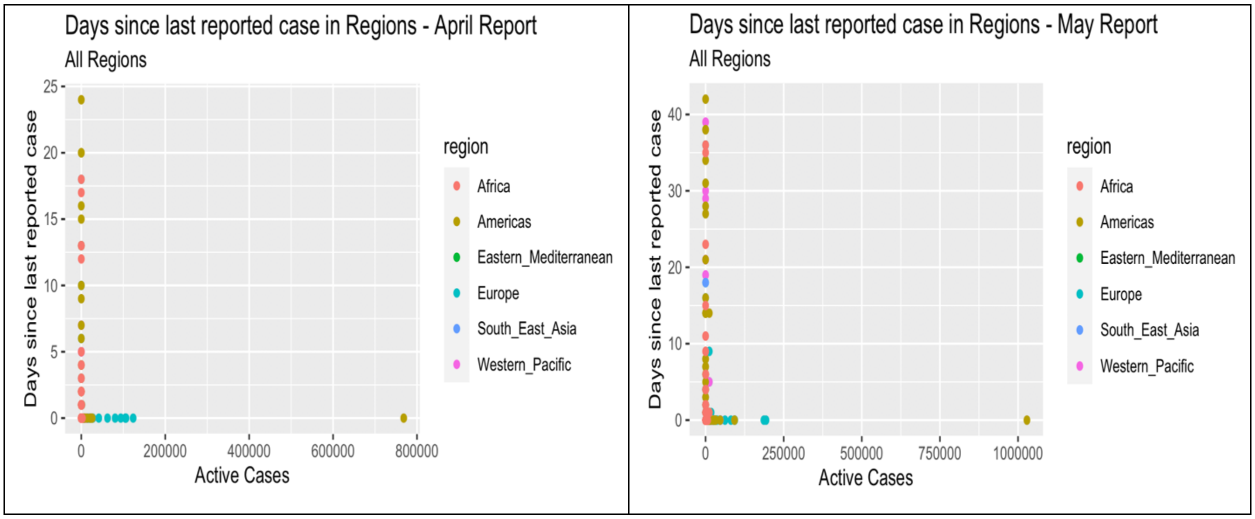
Below are graphs that depict when Africa's transmission type in April was clusters of cases (number of tests made < 200,000 tests) and then with an increase (> 400,000 tests being conducted) changed to sporadic cases.





Europe continues to lead with the number of tests being made, while Africa jumped from last place in April's report to second place in May's report and the Eastern Mediterranean being the third. As the current news is coming in, we are seeing that countries which opened back its economy such as the U.S and China, are being exposed back again to higher numbers of corona detections, which will require further testing by its populations to curb this incredibly unknown virus behavior.

It is also worth mentioning that the number of days since the last report is provided to the World Health Organization by a given country which provides information on the current criticality of its status. Countries situated in the Americas such as Suriname, last reported to WHO 24 days since the April 24th report. This gives 3 options to consider, 1. they either have less reported cases, are waiting to understand the transmission type (pending), or there are no cases to report. In Mays report, in Suriname, there are no actual total confirmed cases to report which explains this lag. As for the European region (Spain as an example) we see that the reporting is continuously updated with 0 days as last day reported in its cases in both reports, meaning there are no lags in days for reporting in Europe (Spain) like the Americas (excluding the U.S which updates on a continual basis).



Limitations in the Covid-19 study

There were minor limitations to having this study finalized. Among the few are listed below:

- Collection of data from different websites (WHO and Wordometers)

- Worldometers data is not kept for future need. Updated daily only and requires manual input to work with.
- File name extension (PDF) difficulty to transform to numerical data using R
- Learning curve of ggplot package.

Future Considerations

- Using “RO” which stands for R naught. It is the number of average cases an infected person will cause in their infectious period.
- Data utilizing time series.