Simulation and Data Visualization Assignment

M Meher Deepthi

Student No: 21115542

ACADEMIC HONESTY AND INTEGRITY: I agree to abide by the expectations as to my conduct, as described in the academic honesty and integrity statement.

Part 1: Analytics

a) Following are the research questions discussed in this paper

Q1: Analyze the development of virus mutations over time. Are there detectable seasonal trends?

Q2: Analyze the effect of covid vaccination on death rates in UK

Q3: Study of various virus mutations across different regions:

i) Total number strains across different countries

ii) Spread of mutations with same ancestry (clade membership) across different continents.

b) Following data sets are used to answer the above questions

Q1: To analyze the development of virus mutations over time in UK the data set '***nextstrain_groups_ecdc_ncov_united-kingdom_metadata.tsv'*** is used. Following columns are used for the analysis:

i) clade_membership
ii) date

Q2: To understand how the covid deaths/cases are influenced with covid vaccination, data from https://coronavirus.data.gov.uk is used. Following fields are used for analysis:

i) Cumulative COVID cases
ii) Cumulative Deaths
iii) Cumulative Vaccination numbers
iv) Date

Q3: To study of various mutations across different regions the following data set has been used '***nextstrain_ncov_open_global_metadata.tsv'.*** The following attributes are taken into account

i) clade_membership
ii) Region (Countries and Continents).

Appropriateness of each data set:

For Q1, 2 variables – clade-membership and date help us understand how each strain of covid mutations 19A, 19B, 20A, 20B, 20C is spread over time (in months) and hence can be used to analyze the relationship between them.

For Q2, number of deaths, covid cases and vaccination numbers are taken into account for plotting the relation between them and understanding how one is affecting the other.

For Q3 the number of covid mutations found in different countries and continents are taken into account. These fields can be used to plot various mutations across regions and thus help us understand them better.

c) Correlation Between Datasets:

The 1st and 3rd data sets 'nextstrain_groups_ecdc_ncov_united-kingdom_metadata.tsv' and nextstrain_ncov_open_global_metadata.tsv have the fields of mutations in UK and across the world they can be correlated with respect to date or with respect to Virus mutations.

The 2nd data set which is taken from https://coronavirus.data.gov.uk contains cumulative Deaths, Cases and Vaccination numbers distributed across time and thus can be cross correlated with time to other datasets.

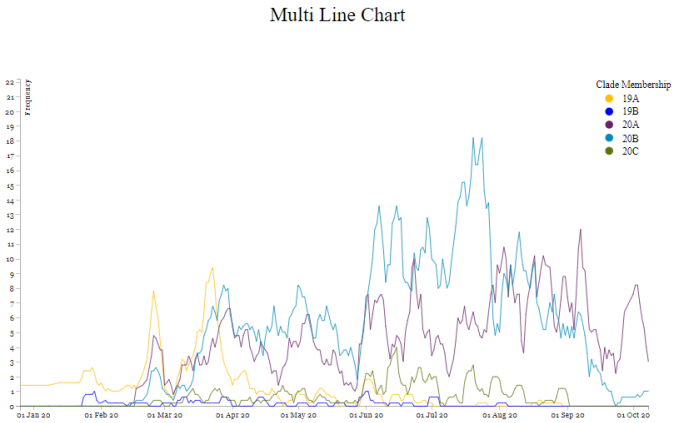
The correlation between the datasets can help us understand relationship between the Covid cases and mutations during different times

Part 2. Design and Discussion

a) Design proposal:

Following design has been proposed for answering each of the research questions:

Q1: Development of virus mutations over time can be plotted with a line graph or 100 % area graph as show in Fig 1 and Fig 2



Frequency of Covid mutations(clade_membership) varying with time

Fig 1 Multi line Chart

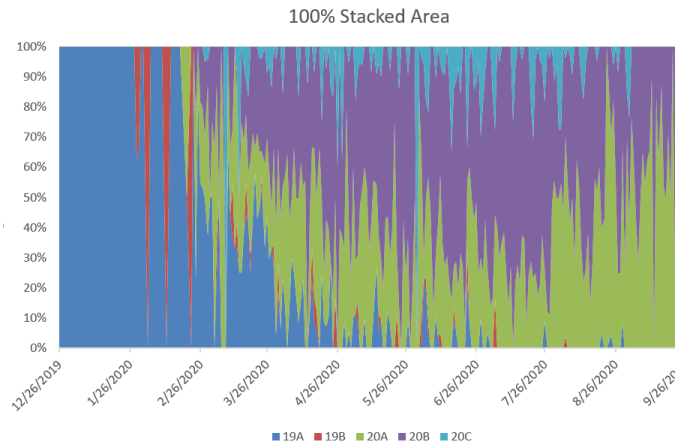


Fig 2 100% stacked area chart

Q2: For this, a combined chart of bar and line graph has been used as shown below.

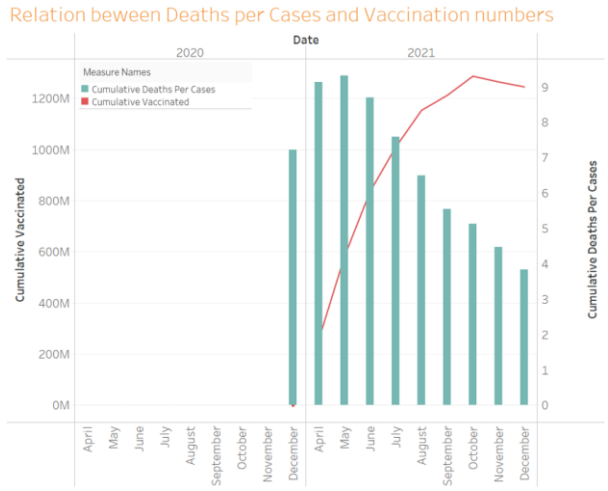


Fig 3 Relation between Deaths per cases and Vaccination numbers

Q3: To answer the first part of this question (Q3-i), the following map plot Fig 4 has been plotted in tableau to understand the number of covid variants detected across different countries.

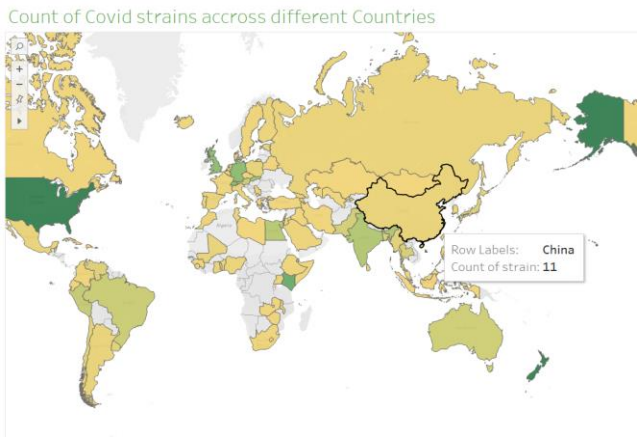


Fig 4: Count of Covid Strains across different countries

And for Q3-ii) following bar chart with proportions is constructed

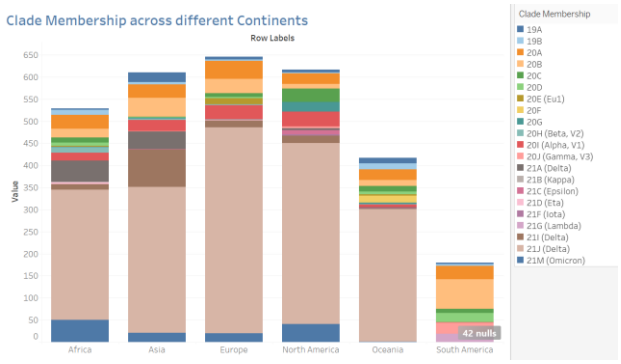


Fig 5: Clade Memberships across different Continents

b. Design Discussion:

Q1: Q1 uses temporal data i.e., variation of categorical variables is plotted with X- axis as time where Categorical variable is Clade Membership (mutations with same ancestry -19A, 19B, 20A, 20B, 20C). These variables are plotted over time in 2 visual formats using Fig1 and Fig2 and thus help in answering Q1. They are plotted in different colours (Hue), as Hue is one of the best ways to visualise categorical variables

Fig1 is a Multi-Line chart which shows the frequency of clade membership with respect to time. Hue is used as a differentiator of Clade Memberships along user interaction that Highlights the line which is selected with the mouse. This helps to point out the exact trends of the selected line and prevents overlapping.

Fig2 is a 100% area chart, this plot converts frequency values on Y-axis to percentage, thus giving a detailed sense of most popular variants present in particular time. Main advantage of 100% area chart over regular area chart is that it eliminates the confusion caused by simple area chart in making trends look very different compared to the original trend. This helps in better representation of data and hence makes it easier to understand the prominence of clade memberships at a given date.

Seasonal Trends: From Fig 1 it is apparent that not all Variant frequencies fluctuated the same but it can be seen that Clade Memberships 20A, 20B and 20C had a significant increase in frequency in June, July and August which might indicate that the rainy season might contribute to increase in Covid Mutations.

Q2 Fig 3: For Q2, analysis has to be done to understand the relationship between Covid-19 Deaths/Cases and Vaccination. The Data set had the counts of cumulative number of deaths, covid cases and vaccinations. But to understand the relation between these two variables (deaths and cases) following variable is calculated.

Derived Variable (Deaths/Cases): Cumulative Deaths/Cases is calculated using python.

Here the Dates are taken as categorical variables and are plotted against each month and year. Chart with 2 Y-axes is taken for this comparison to understand the impact that vaccinations had on Death/Cases.

Two visual channels are used to represent the data. One is change in graph type other is colour, while the spatial positions represent the corresponding values of the Y axis (Cumulative Deaths/Cases or Cumulative Vaccinations) the visual channels help us differentiate both graphs. Cumulative Deaths/Cases are represented with Bar Graph while the Cumulative vaccination numbers are plotted with line graph. By plotting each variable in 2 these two formats along with colour difference, the distinction becomes clear and hence increases the separability of the channels.

From the graph it is apparent that number of Deaths/Cases started decreasing in May and June which is around the same time that vaccination counts were increasing. this can help us understand the relationship between the 2 variables and to further investigate if one has an effect on the other.

Q3 (i) Fig 4: In Fig 4 For the colour scheme, monochromatic colours are used as it is far more effective than using different colours in understanding the ranges or values (increasing or decreasing) of variables. The count of strains across each country is ordered quantitative data. The saturation variable of Hue has been used to differentiate these values. The Graph shows how counts of different strains are distributed across different countries. The Tableau worksheet also has pop up on each country which gives the country name and number of strains which makes it easier for the user to understand the spread of various mutations.

Q3 (ii) Fig 5: In Fig 5 different clade memberships are represented as different colours and different continents are represented by bars of the plot. Hue is used as a differentiator for clade memberships as it is a categorical variable. Fig 5 plot helps us understand the frequency and proportion of each clade membership distributed along the bar graph. The proportions are also in the order of naming, making it even easier. The 2 visual channels here are hue of each clade membership and the Y – axis values which correspond to their frequency value. The value and the mutation name appears as mouse hovers over the bar chart on each of the proportions.

Both the plots help us understand various mutations across different regions making it easier to study their spread.

Part 3. Implementation

a) Data Pre-processing:

The data set used for answering the Q1 is '*nextstrain_groups_ecdc_ncov_united-kingdom_metadata.tsv'*. Following columns are used for the analysis

i) clade_membership
ii) date

the column Clade Membership which represents the mutations with same ancestry is used to represent various mutations of covid. The data is first grouped with respect to date and 5 columns with mutations 19A, 19B, 20A, 20B, 20C are formed with following format.

	A	B	C	D	E	F	G
1	date	19A	19B	20A	20B	20C	
2	7-Mar-20	0	0	2	1	0	
3	8-Mar-20	0	0	0	0	1	
4	9-Mar-20	2	0	1	0	0	
5	10-Mar-20	4	0	2	3	0	
6	11-Mar-20	4	2	2	3	0	
7	12-Mar-20	2	0	4	0	0	
8	13-Mar-20	4	1	5	1	1	
9	14-Mar-20	1	0	1	0	2	
10	15-Mar-20	1	0	2	1	0	
11	16-Mar-20	6	0	5	4	1	
12	17-Mar-20	4	1	2	1	1	
13	18-Mar-20	5	0	2	3	2	
14	19-Mar-20	4	1	5	6	0	
15	20-Mar-20	6	0	4	4	0	
16	21-Mar-20	7	0	1	4	0	
17	22-Mar-20	3	0	2	2	0	
18	23-Mar-20	12	0	5	8	1	
19	24-Mar-20	14	1	2	8	1	
20	25-Mar-20	6	0	5	7	0	

Moving average is taken for each column for every 5 consecutive days to eliminate high frequency fluctuations which gives better visualisation. The data looks like the figure below after the adjustments.

	A	B	C	D	E	F
1	date	19A	19B	20A	20B	20C
2	16-Feb-20	1.4	0.2	0.2	0	0
3	17-Feb-20	1.2	0.2	1.2	0.4	0
4	20-Feb-20	2.2	0.2	1.4	0.4	0
5	22-Feb-20	3.2	0.2	1.8	0.8	0
6	23-Feb-20	4.8	0	2.6	1.4	0
7	24-Feb-20	6.4	0	3.4	2.4	0.2
8	25-Feb-20	7.8	0	4.8	2.4	0.4
9	26-Feb-20	6.8	0	4.6	2.6	0.4
10	27-Feb-20	6	0	4.2	2.4	0.4
11	28-Feb-20	4.4	0	3.8	2	0.4
12	29-Feb-20	2.8	0.2	3.8	1.2	0.2
13	1-Mar-20	1.2	0.2	1.4	1	0
14	2-Mar-20	1	0.2	1.6	0.8	0
15	3-Mar-20	0.6	0.2	1.8	0.8	0
16	4-Mar-20	0.6	0.2	1.4	0.6	0.2
17	5-Mar-20	0.6	0	0.8	0.4	0.2
18	6-Mar-20	1.2	0	1.2	0.8	0.2
19	7-Mar-20	2	0.4	1.4	1.4	0.2
20	8-Mar-20	2.4	0.4	1.8	1.2	0.2

b) WebApp Live:

The following link has the visualization in D3

WebApp Live: https://meherd4.github.io/visualizationCW/

c) Interaction:

The visualization allows user interaction with mouse movement. Line graph is highlighted upon user interaction and the selected line becomes more apparent compared to other lines which makes it easier to distinguish and understand the trend of the selected line properly.

References:

Implementation Code Reference

[1] *https://datawanderings.com/2019/10/28/tutorial-making-a-line-chart-in-d3-js-v-5/*

[2] *https://datawanderings.com/2019/11/01/tutorial-making-an-interactive-line-chart-in-d3-js-v-5/*