DATA423 – Assignment 2

Initial Review and Clean of Dataset

The dataset for Assignment 2 consists of 1 ID variable, 2 categorical variables, 10 numeric variables, 1 outcome variable and 1 observation type variable. The majority of these variables had missing data in different forms and only 4 variables were whole and full of valid variables. Of those which contained missingness, the missing value placeholders identified were as follows:

- -- : POLITICS
- <NA> : POLITICS
- NA : POPULATION, AGE25_PROPTN, AGE_MEDIAN, AGE50_PROPTN, POP_DENSITY, GDP, INFANT_MORT, DOCS, VAX_RATE, HEALTHCARE_COST
- -99 : AGE25_PROPTN, AGE50_PROPTN, DOCS, VAX_RATE

From this view, the decision to hardcode some changes to the missingness was established. The two missing value placeholders in the "POLITICS" column were removed and placed into a new level called "NONE". This choice was made because it kept the meaning of the NA value while establishing that it was still usable data without needing to be placed into the other 4 levels already within the column. Putting them into a new level rather than the "OTHER" level was to establish that it was data that could be in any of the 4 other variables but the information was not present, or not willing to be submitted.

The numeric missing values that were identified as having -99 were changed to NA to prevent any mix up when plotting the data which is present in the EDA graphs associated with the original data set. It was confirmed that these were missing values, rather than valuable outliers when comparing them with the rest of the data available in each column.

Finally, "HEALTHCARE_COST" was clearly associated with "HEALTHCARE_BASIS". This established a need to change the NA values present within the HEALTHCARE_COST variable as every one of them was linked to a "FREE" healthcare basis, and so this missing variable was replaced with 0.

All other missing value placeholders had no clear reasoning behind the missingness and so no other infilling was done to clean the data set. A comparison of the original data set and this initial clean is available with the "Clean Data" check box on the summary page.

Further Review of Dataset and Initial Exploratory Data Analysis

Missing Data Graph:

There are many curious features displayed within this graph, the first most obvious is the large amount of missing data within the AGE_MEDIAN variable, there are 3 main clusters that display the most missingness, but it's generally a widespread with no clear reason behind why the data is missing. POPULATION and POP_DENSITY seem to hold similar results starting with a higher level of missingness which becomes more sporadic before increasing again until a small chunk is missing at the same time in later States/Countries. In other respects, all the other data that has missingness appears to have no other connections and the missingness seems to be primarily of a random dispersion with some occasional matches throughout the variables. In terms of the valid data, having HEALTHCARE_BASIS be complete is an interesting concept and different meanings could be interpreted as to why out of all the other variables there is no missingness. Overall, 83.8% of the raw data is present and this can be improved on with the cleaning as mentioned above and further cleaning associated with the thresholds mentioned below.

Box Plot:

The Box Plots initialise with a view into GDP, DOCS, VAX_RATE and DEATH_RATE, as without standardizing they fall within similar values and because of their context they are interesting to compare. For instance, how GDP and VAX_RATE have similar outliers, with VAX_RATE having only low outliers and GDP having primarily low outliers. This also displays the missing value placeholders present in the raw data with DOCS and VAX_RATE having values close to -99. When excluding POPULATION, POP_DENSITY and HEALTHCARE_COST, the boxplots show several sets of variables that appear similar without centring and scaling. When increasing the IQR multiplier to 5, outliers are still present in GDP, and VAX_RATE. DOCS VAX_RATE, AGE25_PROPTN and AGE50_PROPTN all display the outliers that are associated with the missing value placeholder -99. HEALTHCARE_COST has very high outliers which are still present even at the max IQR multiplier, these values could be associated with the conversation of high healthcare costs around the world or could have other reasons for their value.

Correlation Graph:

All variables are present in the initialisation of the correlation graph, using the Pearson method and the OLO grouping method, two groups are present. AGE50_PROPTN, DOCS and VAX_RATE all have varying levels of dark blue which means that there is a high level of positive correlation present between them. This makes sense because the first wave of vaccinations were given to those over 50 and having a higher amount of Doctors in an area would enable easier access to vaccination. The other group shows a strong positive correlation between INFANT_MORT, DEATH_RATE and POP_DENSITY, which also has good reasoning as COVID was known to spread easier to the more vulnerable and those in a denser area. There is a strong negative association between AGE25_PROPTN and INFANT_MORT. When changing to the Spearman method, a positive association is observed between INFANT_MORT and DOCS and adjusting the grouping keeps the same grouping observed in the first instant.

Mosaic:

The mosaic graph of the raw data compares the two categorical variables available in the data set that have a low enough cardinality. There are no obviously coloured areas in this graph which means that all the variables are evenly balanced throughout the dataset.

Rising Value Chart:

HEALTHCARE_COST shows a very large discontinuity in which the values jump from around 5000 to 1500 which matches the high outliers observed in the boxplots. DOCS, VAX_RATE, AGE25_PROPTN and AGE50_PROPTN all display a large discontinuity that are associated with the missing value placeholder -99. GDP also shows a discontinuity between 20 and 45, and AGE_MEDIAN falls short of the end of the chart due to the large number of missing values it contains. When centred and scaled, DEATH_RATE is the only variable that continues until the end of the graph with a fairly continuous line.

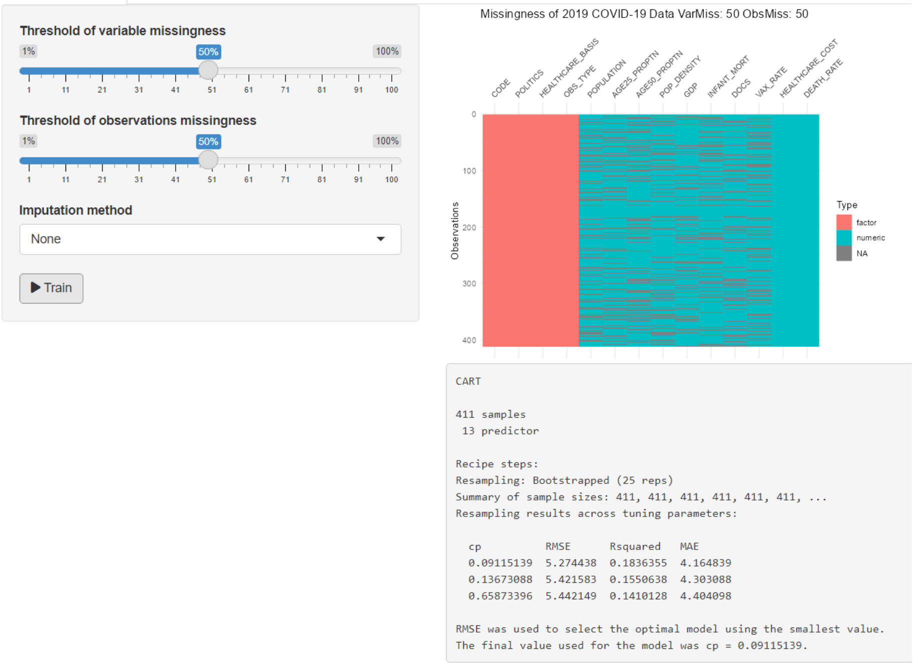
Pairs Chart:

The pairs chart overall shows little correlation between the variables even when coloured by the available categorical variables. Some observations noticed was the strong positive correlation between POP_DENSITY and DEATH_RATE, and the distinct colouring and grouping displayed between DEATH_RATE, HEALTHCARE_COST and POP_DENSITY with colouring by HEALTHCARE_BASIS. Further observation would be easier with a clean data set as the missing values make data interpretation difficult.

Missingness:

Thresholds:

Because of the difficulties in analysing the raw data, cleaning was needed to update the dataset to a more user-friendly view. The first steps of cleaning were detailed above and then thresholds were used to remove any excessively missing variables and then observations. These steps are shown in the series of images below with reasonings.



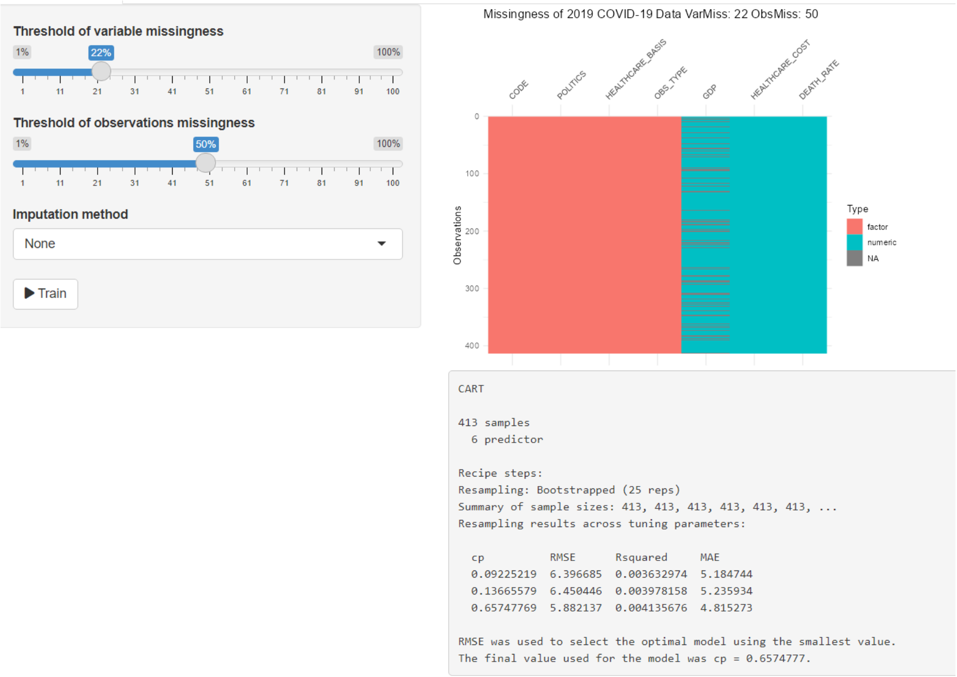
Threshold of variable missingness: 50%

Threshold of observations missingness: 50%

Imputation method: None

```
CART

411 samples
 13 predictor

Recipe steps:
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 411, 411, 411, 411, 411, 411, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.09115139  5.274438  0.1836355  4.164839
  0.13673088  5.421583  0.1550638  4.303088
  0.65873396  5.442149  0.1410128  4.404098

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.09115139.
```

1. *The first step was starting off with a threshold of 50% for both the variables and outliers. This removed 1 variable, the AGE_MEDIAN and none of the observations. This means that all observations are still present in the dataset, and all observations hold at least 50% valid data.*



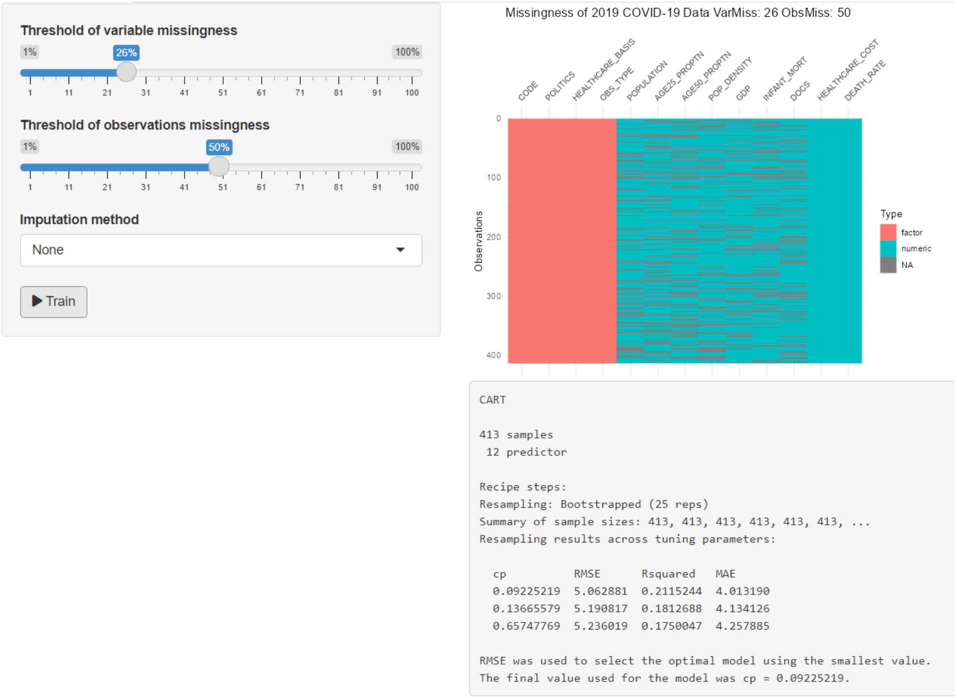
Threshold of variable missingness 22%

Threshold of observations missingness 50%

Imputation method None

```
CART

413 samples
  6 predictor

Recipe steps:
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 413, 413, 413, 413, 413, 413, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared     MAE
  0.09225219  6.396685  0.003632974  5.184744
  0.13665579  6.450446  0.003978158  5.235934
  0.65747769  5.882137  0.004135676  4.815273

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.6574777.
```

2. A threshold of 22% for variable missingness was the point where the RMSE value was increasing and so increasing the threshold would be needed. This also shows that the threshold removed 7 observations which would be detrimental to the analysis of the dataset.



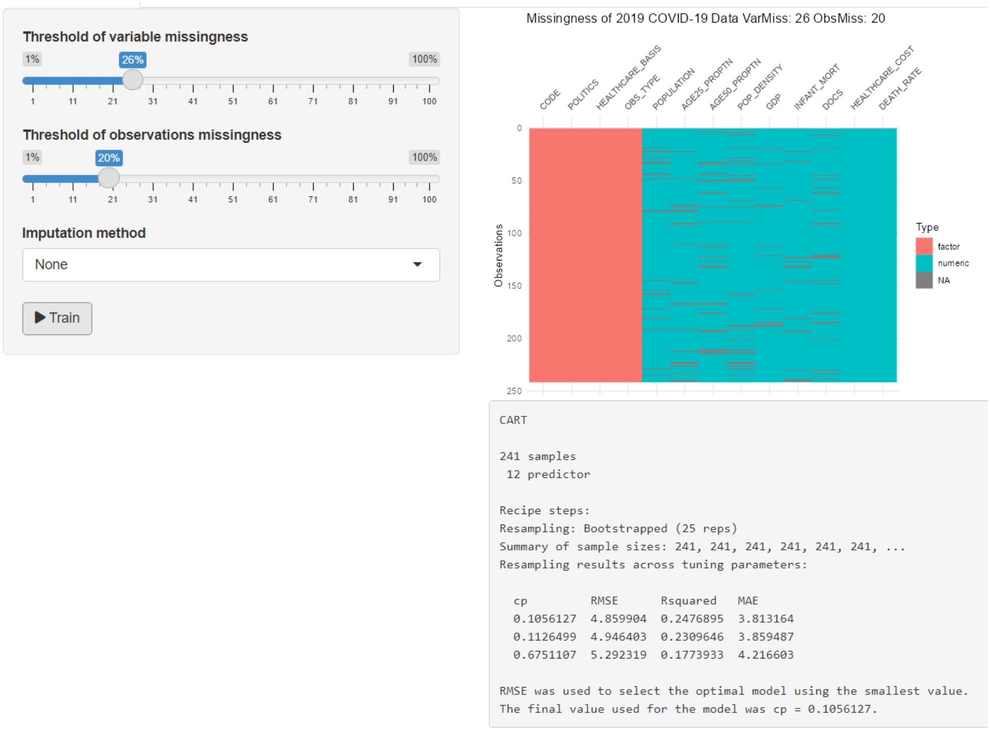
Threshold of variable missingness 26%

Threshold of observations missingness 50%

Imputation method None

```
CART

413 samples
 12 predictor

Recipe steps:
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 413, 413, 413, 413, 413, 413, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.09225219  5.062881  0.2115244  4.013190
  0.13665579  5.190817  0.1812688  4.134126
  0.65747769  5.236019  0.1750047  4.257885

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.09225219.
```

3. Finally, the threshold of 26% allowed for reduction of RMSE without removing a significant amount of variables. A similar process was then undertaken for the threshold of observation missingness.



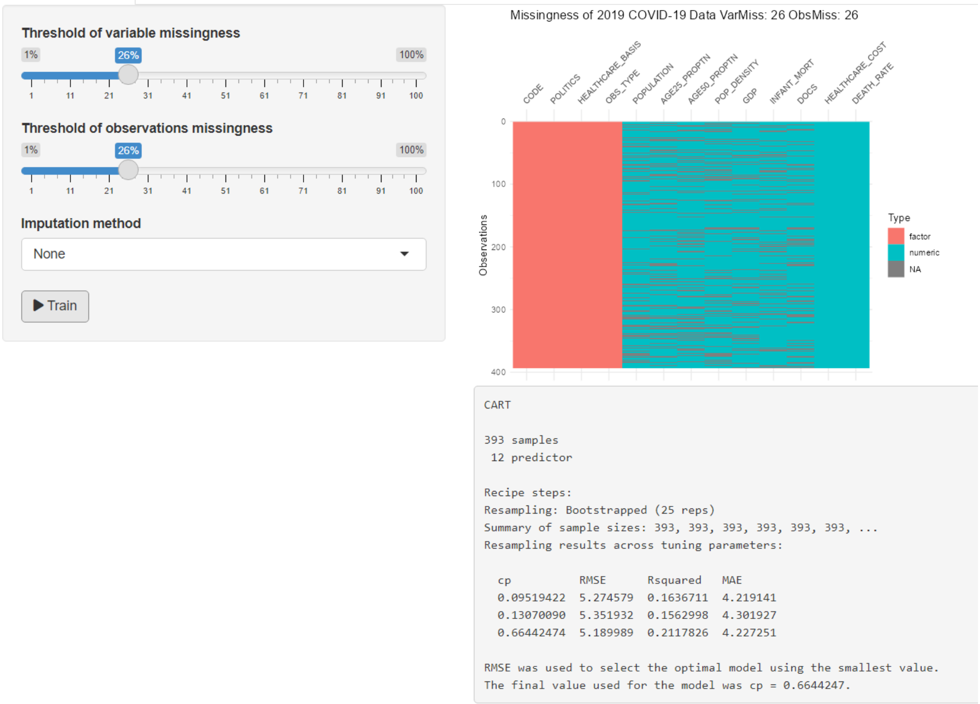
Threshold of variable missingness: 26%

Threshold of observations missingness: 20%

Imputation method: None

```
CART

241 samples
 12 predictor

Recipe steps:
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 241, 241, 241, 241, 241, 241, ...
Resampling results across tuning parameters:

  cp         RMSE      Rsquared   MAE
  0.1056127  4.859904  0.2476895  3.813164
  0.1126499  4.946403  0.2309646  3.859487
  0.6751107  5.292319  0.1773933  4.216603

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.1056127.
```

4. Reducing to 20% was the point where the RMSE increased compared to the previous thresholds. This lost nearly half the observations in the process though and clearly would be another situation that would impede data analysis.



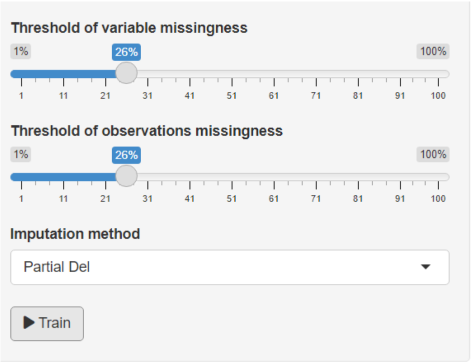
Threshold of variable missingness: 26%

Threshold of observations missingness: 26%

Imputation method: None

```
CART

393 samples
 12 predictor

Recipe steps:
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 393, 393, 393, 393, 393, 393, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.09519422  5.274579  0.1636711  4.219141
  0.13070090  5.351932  0.1562998  4.301927
  0.66442474  5.189989  0.2117826  4.227251

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.6644247.
```

5. Once again, 26% proved to be an acceptable threshold which kept most of the observations while still reducing the RMSE. Further analysis to determine if imputation would help improve this score was performed.



Threshold of variable missingness

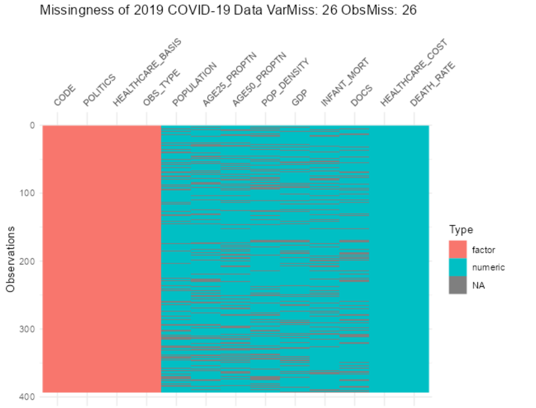
1% 26% 100%

Threshold of observations missingness

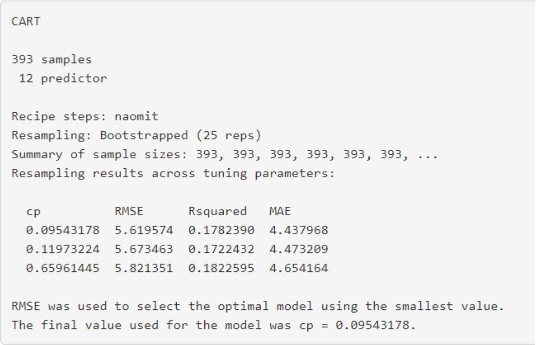
1% 26% 100%

Imputation method

Partial Del

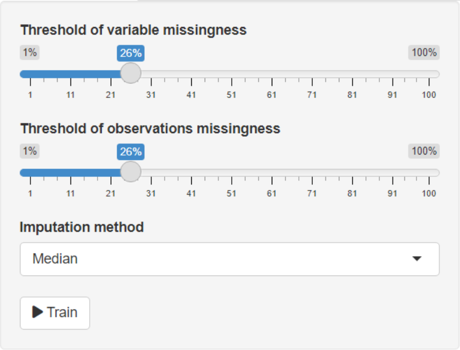


Missingness of 2019 COVID-19 Data VarMiss: 26 ObsMiss: 26



```
CART

393 samples
 12 predictor

Recipe steps: naomit
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 393, 393, 393, 393, 393, 393, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.09543178  5.619574  0.1782390  4.437968
  0.11973224  5.673463  0.1722432  4.473209
  0.65961445  5.821351  0.1822595  4.654164

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.09543178.
```

6. Partial deletion increased the RMSE and due to the missingness pattern observed in the next section, it would not be appropriate to use for this situation.



Threshold of variable missingness

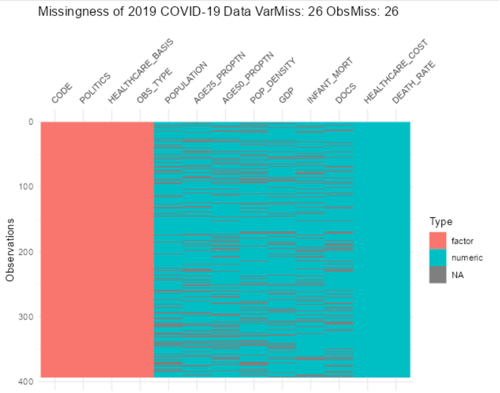
1% 26% 100%

Threshold of observations missingness

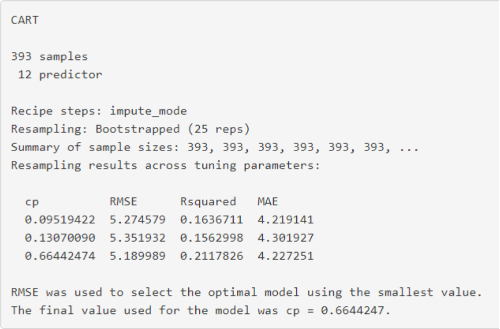
1% 26% 100%

Imputation method

Median

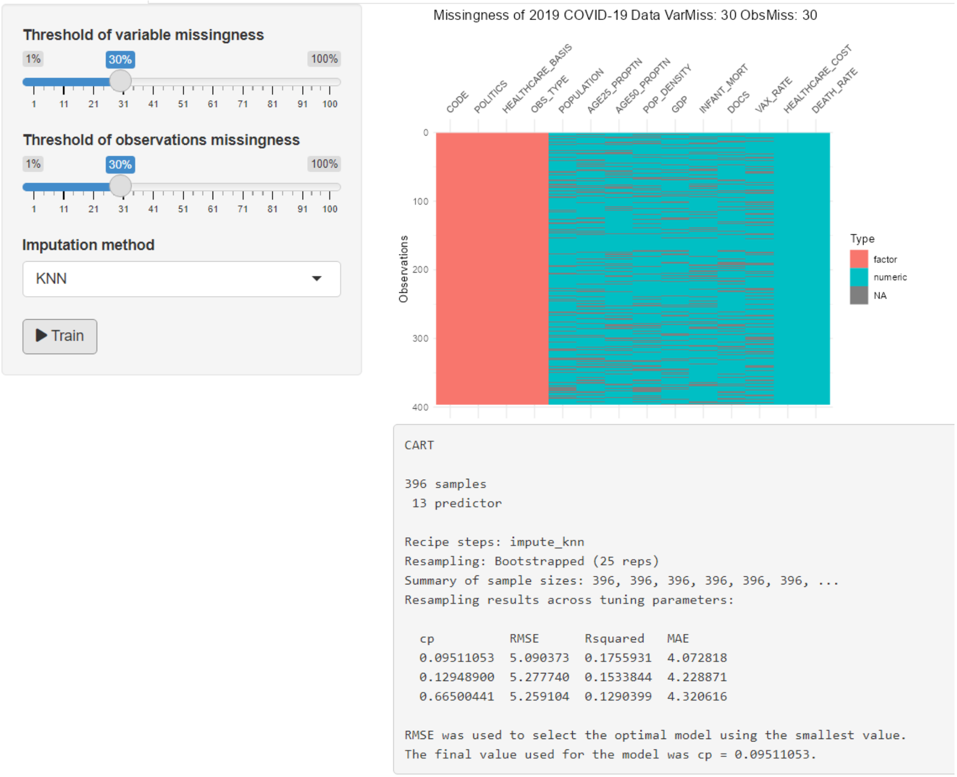


Missingness of 2019 COVID-19 Data VarMiss: 26 ObsMiss: 26



```
CART

393 samples
 12 predictor

Recipe steps: impute_mode
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 393, 393, 393, 393, 393, 393, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.09519422  5.274579  0.1636711  4.219141
  0.13070090  5.351932  0.1562998  4.301927
  0.66442474  5.189989  0.2117826  4.227251

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.6644247.
```

7. Imputation through Median and through KNN provided the same results as having no imputation method. Although applying them it was discovered that increasing the thresholds slightly reduced the RMSE again and included a few more observations.



Threshold of variable missingness: 30%

Threshold of observations missingness: 30%

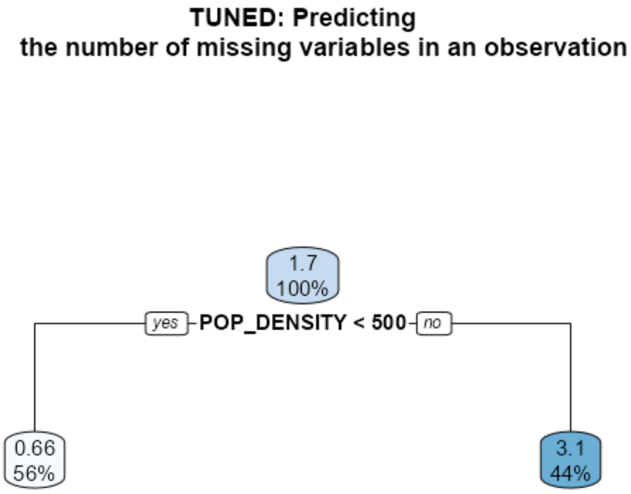
Imputation method: KNN

```
CART

396 samples
 13 predictor

Recipe steps: impute_knn
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 396, 396, 396, 396, 396, 396, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.09511053  5.090373  0.1755931  4.072818
  0.12948900  5.277740  0.1533844  4.228871
  0.66500441  5.259104  0.1290399  4.320616

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.09511053.
```

8. This final image are the thresholds chosen for both variable and observation missingness, 30% and 30% with an RMSE of 5.09 and containing 396 observations and 13 variables. These thresholds were used for all further analysis.

Pattern:

A regression tree was created to determine is the missingness had an observable pattern. Only one variable appeared in the tree with two nodes branching from it. POP_DENSITY seemed to have a correlation with missing variables, with many of the ones greater than 500 appearing to have more missingness in comparison to the POP_DENSITY values less than 500. As there are multiple nodes, it can be argued that there is some partial way to predict observation missingness. This means we can rule out MCAR.



Cleaned Data Charts:

In the majority of the updated charts, the similarity observations still hold true, and the data is easier to interpret. For the Boxplots, its now important to note that HEALTHCARE_COST now has low outliers too which account for the free healthcare in the dataset, and GDP and VAX_RATE aren't as similar to the other variables as previously noted as there aren't missing value placeholders to skew the view of the data.

There are some stronger types of correlation present in the updated graph where a slight negative correlation is now more obvious between VAX_RATE and POPULATION. The groups have also shifted around slightly, but there are still two clear groups to be observed.

The mosaic chart has the Pearsons residuals change slightly but the data has stayed balanced even with excessive missing values removed.

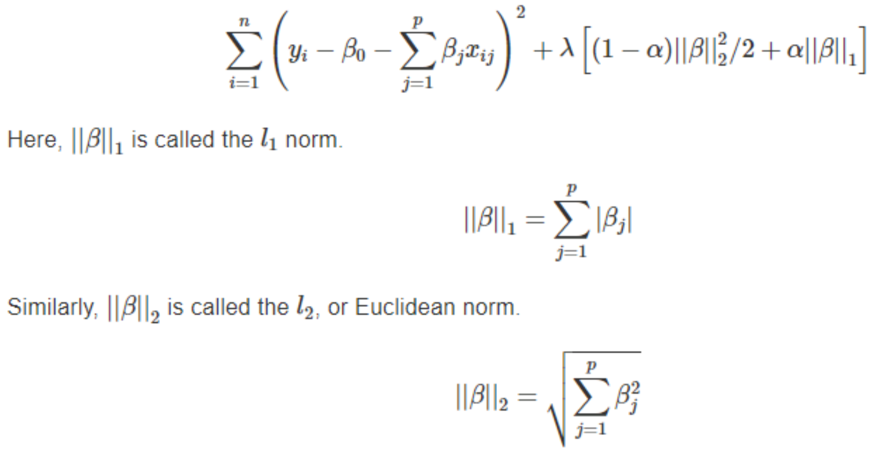
The rising value chart shows the most distinct changes as there aren't as many discontinuities as mentioned before. GDP and VAX_RATE still have an observable discontinuity from around 20 to 50 and HEALTHCARE_COST now has two large discontinuities to account for the adjusted missing values, from 0 to 5000 and then from 5000 to 15000.

The pairs chart is easier to analyse now without the missing variables. DEATH_RATE and POPULATION show a no relationship. The scatter plot has no discernible trend. POP_DENSITY and POPULATION also show a no relationship within their scatter plot. INFANT_MORT and DEATH_RATE show a slight positive relationship as do INFANT_MORT and DOCS and DOCS and DEATH_RATE.

GLMNET Method

Recipe-Based Processing and CARET Training:

As per the data exploration regarding the data missingness threshold, the "recipe" based data processing imputes by KNN on all the predictors with a 5-neighbour limit. The recipe can then regress DEATH_RATE against all other variables when implemented within the CARET train.



$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\left[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1\right]$$

Here, $||\beta||_1$ is called the l_1 norm.

$$||\beta||_1=\sum_{j=1}^{p}|\beta_j|$$

Similarly, $||\beta||_2$ is called the l_2, or Euclidean norm.

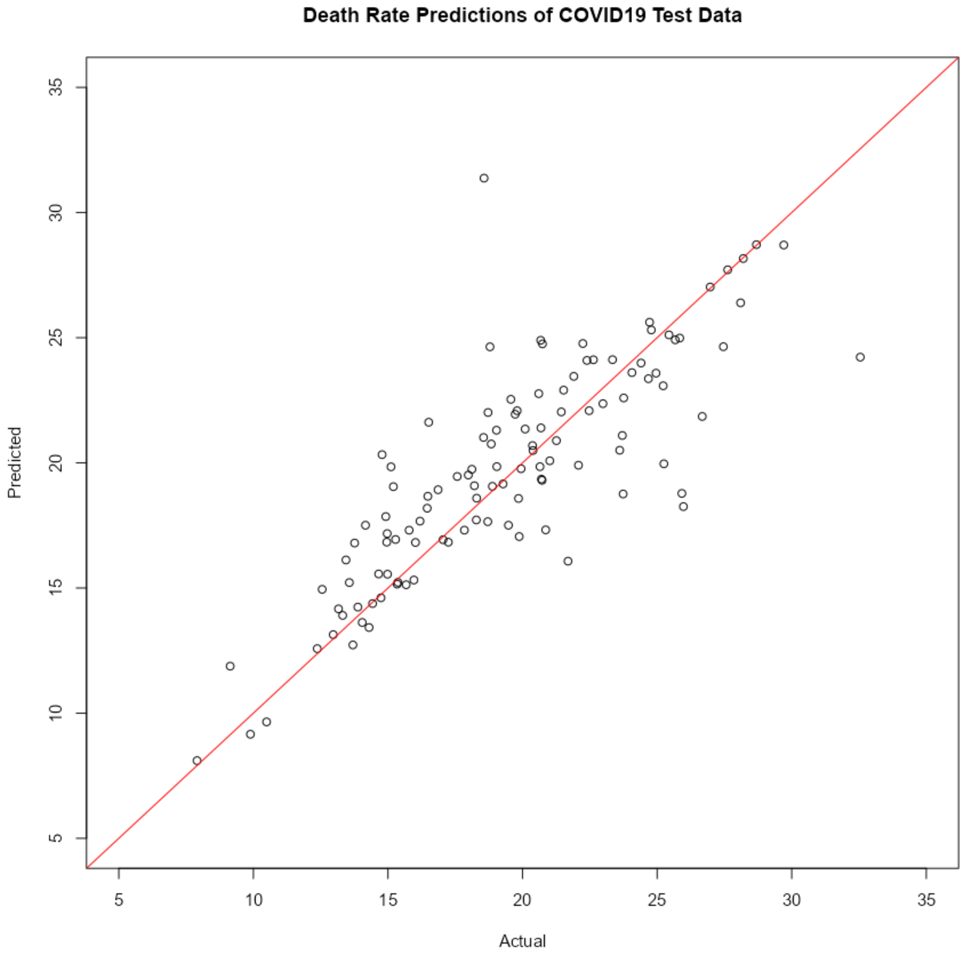
$$||\beta||_2=\sqrt{\sum_{j=1}^{p}\beta_j^2}$$

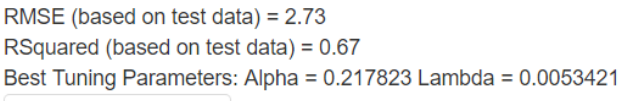
l1 = Lasso, l2 = Ridge

CARET allowed the repeated sampling of the train data to find the best parameters which were Alpha: 0.218 and Lambda: 0.005. GLMNET solves Elastic Net Regression which combines the residual sum of squares plus a penalty factor to all coefficients. The greater the lambda there is, the more penalty is given. This is essentially the combination of ridge and lasso techniques. The penalty is applied to the squares of the coefficients (ridge regression) and the absolute values of all the coefficients (lasso regression). Alpha always takes a value between 1 and 0 and this is how the penalty is split up between ridge and lasso.

Train control was also implemented in this method, this specifies how the repeated cross validation will take place by allowing a select number of partitions and repetitions to be chosen through a 'random' search method. The selection function used was 'best' which by default looks for the lowest RMSE value. Once these features have been used to train the model, it can then be used to generate the prediction graph.

Results:

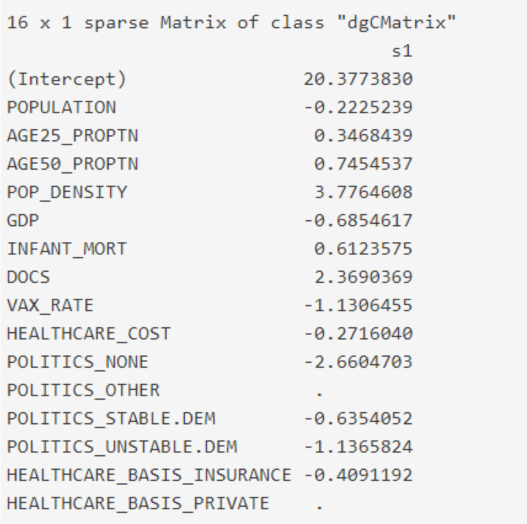




```
RMSE (based on test data) = 2.73
RSquared (based on test data) = 0.67
Best Tuning Parameters: Alpha = 0.217823 Lambda = 0.0053421
```

For this assignment, the best alpha was closer to ridge than lasso as it was closer to 0, this means that it shrinks the coefficients so that they're less sensitive to changes. Viewing the Death Rate

Predictions of COVID19 Test Data graph, most of the values are very close to the line and with the low RMSE of 2.73, the elastic net model did well with the 'unseen' test data.



```
16 x 1 sparse Matrix of class "dgCMatrix"
                                    s1
(Intercept)                 20.3773830
POPULATION                  -0.2225239
AGE25_PROPTN                 0.3468439
AGE50_PROPTN                 0.7454537
POP_DENSITY                  3.7764608
GDP                         -0.6854617
INFANT_MORT                  0.6123575
DOCS                         2.3690369
VAX_RATE                    -1.1306455
HEALTHCARE_COST             -0.2716040
POLITICS_NONE               -2.6604703
POLITICS_OTHER               .
POLITICS_STABLE.DEM         -0.6354052
POLITICS_UNSTABLE.DEM       -1.1365824
HEALTHCARE_BASIS_INSURANCE -0.4091192
HEALTHCARE_BASIS_PRIVATE     .
```

The dots present in the coefficients list means that lasso pushed them out, they didn't provide enough new information as they're more highly correlated to another variable. POP_DENSITY displays a high coefficient – the DEATH_RATE is higher where the POP_DENSITY is higher, and the many variables with a negative value are where DEATH_RATE will lower as those values get higher.

Issued faced were involved with the pairs graph in the cleaned data as the level of missingness threw errors that were related to the lack of data for comparison at certain thresholds. A domain expert would also help confirm the type of missingness this data has present and if POP_DENSITY is an accurate predictor for missingness.