

Covid-19 data

1. Data Dictionary

Covid-19 data: All measurements are as at 2019.

The data relates to characteristics of countries and states that relate to their projected Covid-19 death rate after 10 years if no vaccine is available.

Variables

COUNTRY - Anonymised country or state

GOVERNMENT - Type of government "LEFT", "RIGHT", "UNSTABLE", "AUTOCRATIC"

POPULATION - Total population

AGE25PROP - the proportion of the population that is at or below 25

AGEMEDIAN - the median age of the population

AGE55PROP - the proportion of the population that is at or above 55

POPDENSITY - the population density

GDP2019 - the Gross National Product

INFANTMORT - The infant mortality rate

DOC10 - the number of doctors per 10,000

VAXRATE - the mean vaccination rate for measles, mumps & rubella

HEALTHCARE_BASIS - type of healthcare system "INSURANCE", "PRIVATE", "FREE"

HEALTHCARE_COST - healthcare costs per person where applicable

DEATHRATE - the projected death rate (across ten years)

The outcome variable is **DEATHRATE**

2. Data inspection and initial cleaning

First thing was done was to load up the raw data into R and have a “**raw data table**” and a “**raw data summary**”. From the first look at the **raw data table** there seems to be a few missing values highlighted in yellow and -99 which is a way of representing missing data as well.

[Image 1](#)

COUNTRY	GOVERNMENT	POPULATION	AGE25PROP	AGEMEDIAN	AGE55PROP	POPDENSITY	GDP2019	INFANTMORT	DOC10	VAXRATE
Country1	LEFT	45.3522723	13.59335802	34.64537129	33.45689986	496.080991	22.57267926	16.26391773	22.57577054	39.49819006
Country2	LEFT	45.08909852	17.85335842		36.45560462	502.04048	22.68891847	14.23525026	24.0583818	35.2070913
Country3	LEFT	80.19653304	11.27725748	34.64418495	34.43306411	493.8851338	57.99140875	17.7389792	24.65601864	61.57626386
Country4	RIGHT	7.800894313	16.95537408		32.46770247	505.5077521	56.41513595			
Country5	UNSTABLE	37.51580783	15.08856411		33.05274446	492.3767252	57.37041639	15.2105873	26.60366207	60.57948222
Country6	UNSTABLE	34.06739147	16.23955126	34.77339422	30.26198947	511.0807802	58.05784005	14.0065935	21.88978349	-99
Country7	LEFT	185.947299	16.97791268		38.26336624	493.0798147	60.60280586	14.98587549	22.88021285	62.32694639
Country8	RIGHT	35.14859574	15.42629706	33.91057888	36.6075974	487.6794051	54.58988056	12.34910952	21.2748233	63.91424053
Country9	LEFT	11.85917484	9.577711968		-99		22.21835681	16.34334043	-99	-99
Country10	LEFT	19.7393459	14.00549486		34.73480191	495.8241112	48.92616803	14.74475939	24.07279746	62.55289065

For a summary view of the data we look at the raw data summary, first thing we observe is the dimension of our data. It has 190 rows and 14 variables.

assignment2rawdata

Dimensions: 190 x 14



Duplicates: 0

From the raw data summary, we can see we have 11 numeric variables, **POPULATION** - Total population, **AGE25PROP** - the proportion of the population that is at or below 25, **AGE25PROP** - the proportion of the population that is at or below 25, **AGEMEDIAN** - the median age of the population, **AGE55PROP** - the proportion of the population that is at or above 55, **POPDENSITY** - the population density, **GDP2019** - the Gross National Product, **INFANTMORT** - The infant mortality rate, **DOC10** - the number of doctors per 10,000, **VAXRATE** - the mean vaccination rate for measles, mumps & rubella and the outcome variable **DEATHRATE** - the projected death rate (across ten years).

Also, we have 3 factor variables, **COUNTRY** - Anonymised country or state, **GOVERNMENT** - Type of government "LEFT", "RIGHT", "UNSTABLE", "AUTOCRATIC" and **HEALTHCARE_BASIS** - type of healthcare system "INSURANCE", "PRIVATE", "FREE".

Interesting feature about variable Country is that it is unique for all 190 observations. This may be our **ID role** for our modelling to predict the death rate as it is 100 % unique. Also, the Government variable has 10 observations with “-” and variables age25prop, age55prop, doc10, vaxrate have some observation values with “-99” these seem to be missing data. In R, we want to get to the NA format since NA is R’s way of saying “not available” - this is done as part of data tidying/cleaning.



[Image 2](#)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	COUNTRY [factor]	1. Country1 2. Country10 3. Country100 4. Country101 5. Country102 6. Country103 7. Country104 8. Country105 9. Country106 10. Country107 [180 others]	1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 180 (94.7%)		190 (100%)	0 (0%)
2	GOVERNMENT [factor]	1. -- 2. AUTOCRATIC 3. LEFT 4. RIGHT 5. UNSTABLE	10 (5.3%) 12 (6.3%) 59 (31.1%) 58 (30.5%) 51 (26.8%)		190 (100%)	0 (0%)

AGE55PROP [numeric]	Mean (sd) : 22.8 (38.5) min < med < max: -99 < 34.3 < 43.4 IQR (CV) : 3.9 (1.7)	AGE25PROP [numeric]	Mean (sd) : 10.5 (22.9) min < med < max: -99 < 15.1 < 21.6 IQR (CV) : 3.2 (2.2)
DOC10 [numeric]	Mean (sd) : 11.6 (35.9) min < med < max: -99 < 22.5 < 32 IQR (CV) : 3.5 (3.1)		
VAXRATE [numeric]	Mean (sd) : 37 (52.4) min < med < max: -99 < 58.9 < 67.2 IQR (CV) : 9.6 (1.4)		

Another interesting feature is variable healthcare_basis has level free for 41 observations and variable healthcare_cost has 43 missing observations. This seems to be a case of some of “not applicable values” in the 43 missing observations for healthcare_cost, as free health care basis would mean zero cost or not applicable.

[Image 3](#)

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
HEALTHCARE_BASIS [factor]	1. FREE 2. INSURANCE 3. PRIVATE	41 (22.0%) 111 (59.7%) 34 (18.3%)		186 (97.89%)	4 (2.11%)
HEALTHCARE_COST [numeric]	Mean (sd) : 7316.9 (4265.6) min < med < max: 4475.3 < 5064.5 < 15965.1 IQR (CV) : 448.6 (0.6)	147 distinct values		147 (77.37%)	43 (22.63%)

We deal with this by passing NA to deal with “-99” and “ - - ” missing values and for the observations where healthcare_basis is free assigning missing healthcare_cost to zero as cost would be not applicable to countries or states where healthcare basis is free. After passing the variable we have data table and summary which should just have “NA” for missing values. We can inspect some of the changes made, for example in the data table now we don’t have -99 at in variable Ageprop25, ageprop55 doc10 and vax rate. Health care cost has now 0 for values for which health care basis is free.

[Image 4](#)

Variable	Stats / Values
----------	----------------


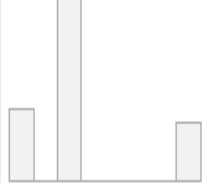
AGE25PROP [numeric]	Mean (sd) : 15.2 (2.3) min < med < max: 9.6 < 15.4 < 21.6 IQR (CV) : 3.4 (0.1)
------------------------	---

AGE55PROP [numeric]	Mean (sd) : 34.8 (2.7) min < med < max: 29.3 < 34.7 < 43.4 IQR (CV) : 3.7 (0.1)
------------------------	--

DOC10 [numeric]	Mean (sd) : 23.1 (2.5) min < med < max: 15.7 < 22.7 < 32 IQR (CV) : 3.2 (0.1)
--------------------	--

VAXRATE [numeric]	Mean (sd) : 56.3 (11.2) min < med < max: 17 < 59.8 < 67.2 IQR (CV) : 5.6 (0.2)
----------------------	---

Image 5

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
GOVERNMENT [factor]	1. -- 2. AUTOCRATIC 3. LEFT 4. RIGHT 5. UNSTABLE	0 (0.0%) 12 (6.7%) 59 (32.8%) 58 (32.2%) 51 (28.3%)		180 (94.74%)	10 (5.26%)
HEALTHCARE_COST [numeric]	Mean (sd) : 5721.2 (4835.7) min < med < max: 0 < 4981.4 < 15965.1 IQR (CV) : 522.1 (0.8)	148 distinct values		188 (98.95%)	2 (1.05%)

3. Missing values

Now we look at missing values. From image 6 we can see overall, we have 85% values present and 15 % missing values. Variables country (ID) and death rate (outcome variable) have no missing observations, this shouldn't surprise us. Variable age median stands out being 59.47 % missing. When observations are clustered by similarity, they form 4 groups in image 7.

Image 6

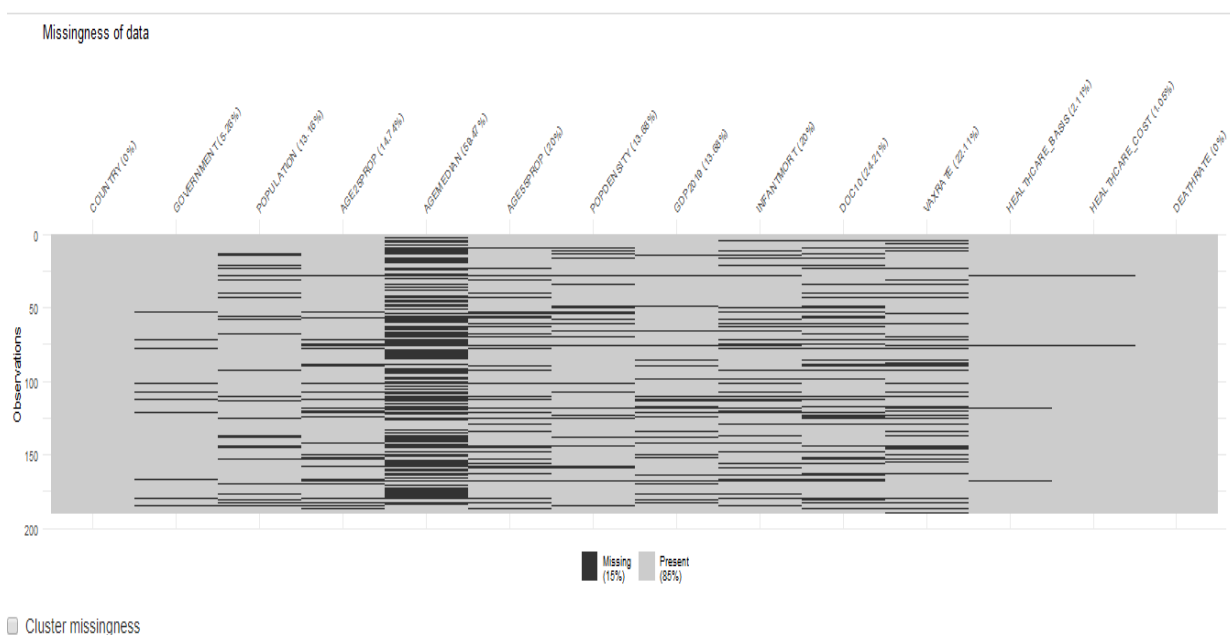
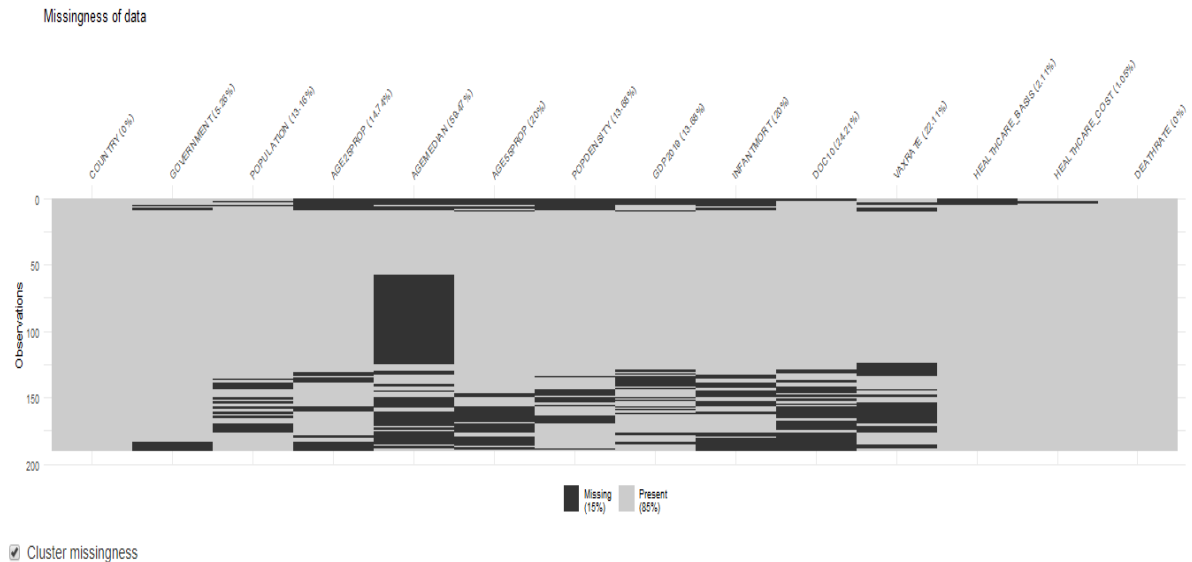
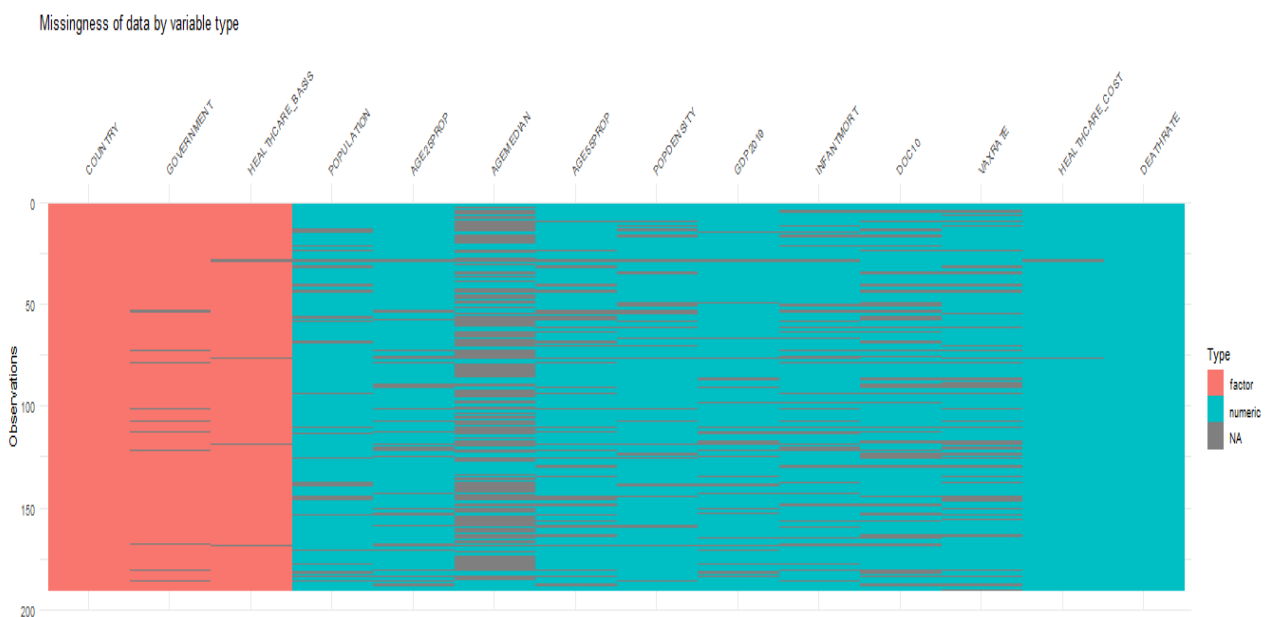


Image 7



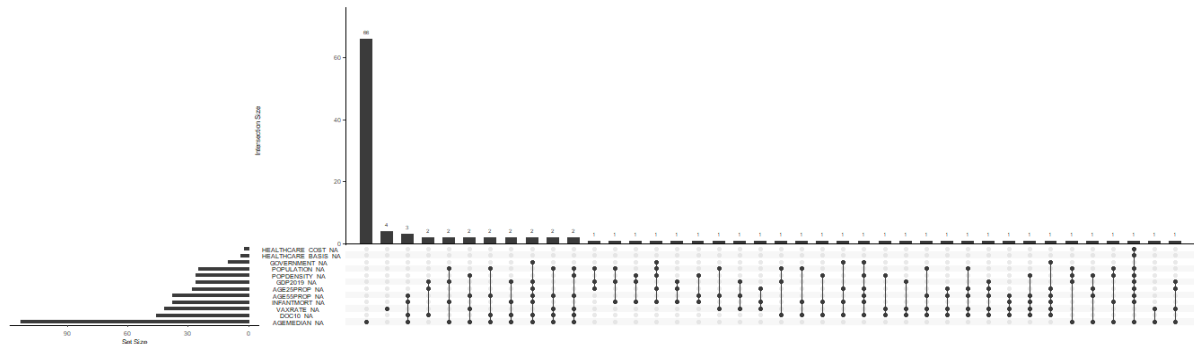
Plot below shows missingness by variable type. We have blue for numeric, pink for factor and NA (missing values in grey). From this visualization we can find out some more insight the two factor variables government and healthcare basis and one numeric variable health care cost have low missingness.

Image 8



In the plot in image 9 we can visualize if there is a pattern between variables in missingness. From the plot we can see variables are missing in pattern, but occasions of the variables missing together 2 more variables missing at once is maximum three. So not overly alarming.

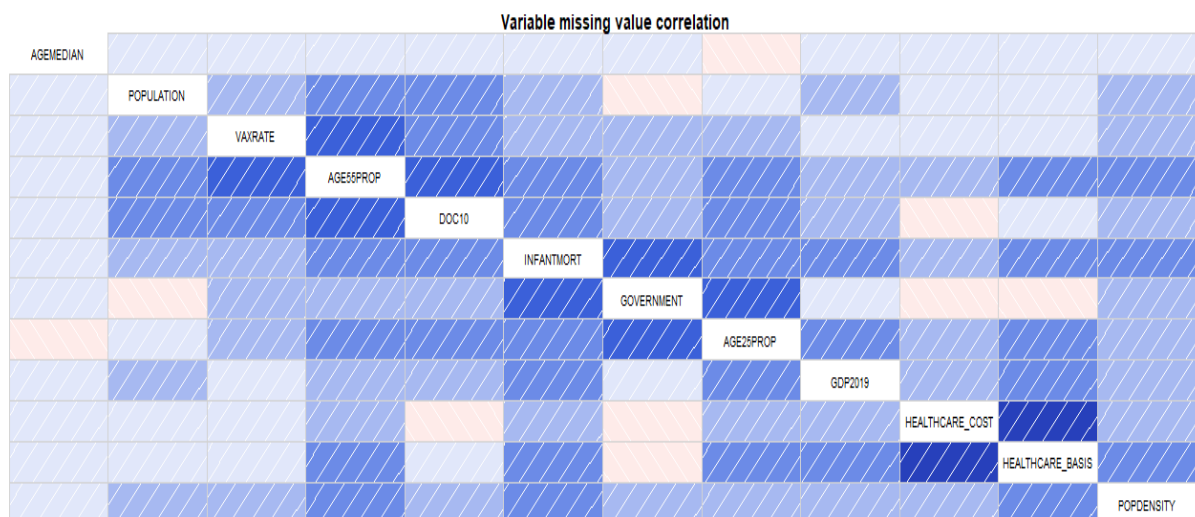
Image 9



The missing data has some discernible pattern for Covid19 data so it cannot be MCAR (missing completely at random). We have no evidence that the values are missing because of their value (that would require insider knowledge which we lack) and would need a domain expert. We shall say the data is MAR and possibly MNAR. We can visualise the correlation structure of the missingness. To assist our comprehension, we have removed any variables that either have no missing values.

From image 10 we can see there is strong spearman correlation in variable missing values between healthcare cost and health care basis which form a set of 2, also three variables Vaxrate, age55prop, and doc10 look to have moderate to strong correlation and form a set of 3. Infant mort, government and age25prop look to have moderate to strong correlation as a set of 3 as well.

Image 10



Correlation method

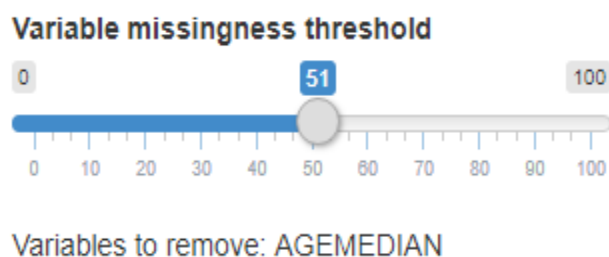
Notice whether variables are missing in sets

Before we model to predict our outcome variable, we need to have a strategy to deal with excessive missingness. The strategy we follow is

- First: remove excessively missing variables
- Second: remove excessively missing observations
- Third: impute missing values.

We saw in the missing value charts `agemedian` had 59.47 % of missing values. It may be prudent to remove variables where the ratio of missing values > threshold (say 50%). A variable that is heavily missing like `agemedian` is a good candidate to drop as a predictor. A reason that might stop us removing the `agemedian` as a predictor that it might be an important variable even though it is largely missing. We can check the variable importance using random forest.

[Image 11](#)

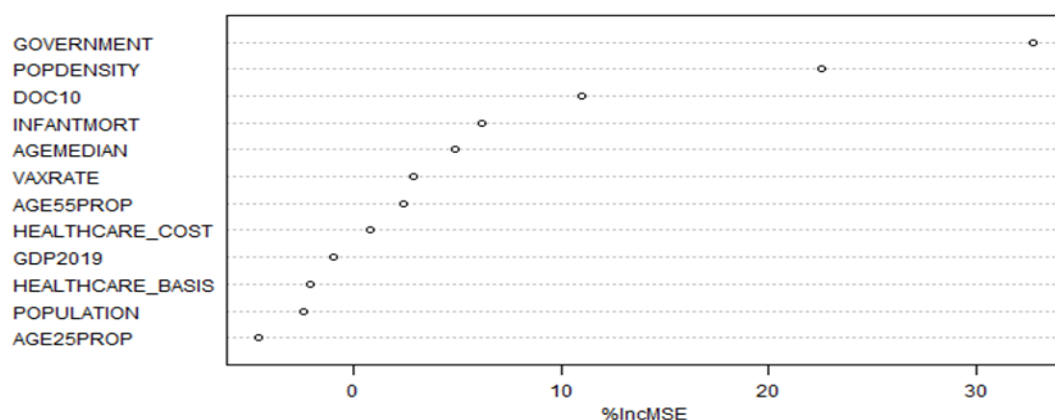


Variable importance using random forest method. Percentage increase in mean square error is analogous to accuracy-based importance and is calculated by shuffling the values of the out-of-bag samples (image 12).

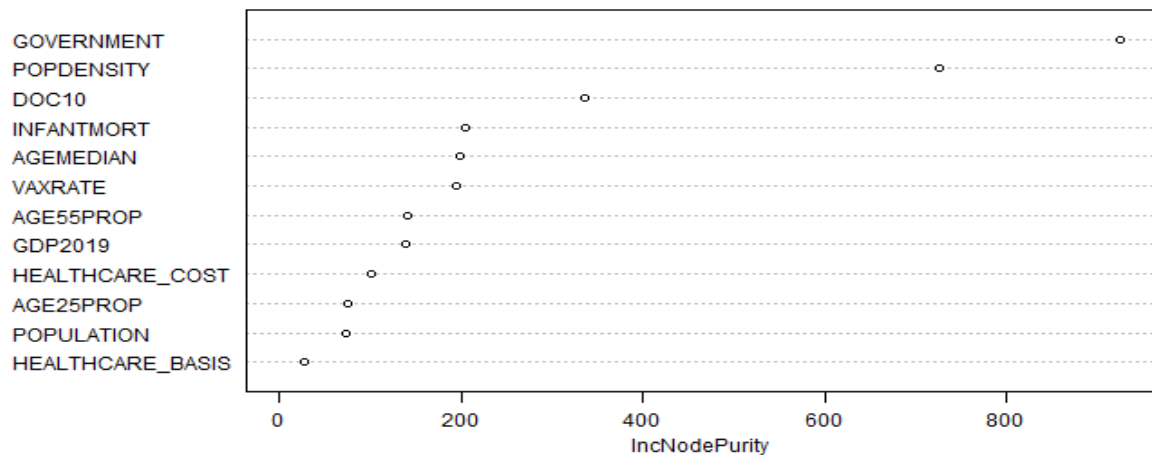
Increase in node purity is analogous to gini-based importance and is calculated based on the reduction in sum of squared errors whenever a variable is chosen to split (image 13).

Neither measure is perfect but viewing both together allows a comparison of the importance ranking of all predictor variables. We can see that age median is not as important a predictor variable in both image 12 and image 13 when compared to government, population and we may remove it.

[Image 12 Random forest variable importance MSE](#)



[Image 13 Random forest variable importance node purity](#)



After removing age median, we can see that now we don't have any variables that are missing more than 24.21 percent which is Doc 10 and imputation should be able to handle this level of missingness well.

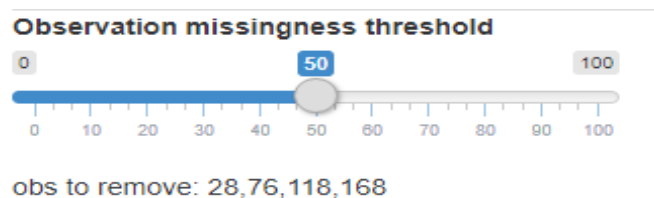
[Image 14](#)



Variables to remove:

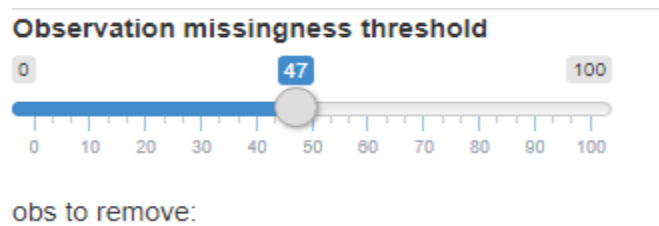
Now we look at if we have any excessively missing observations. Observations that are excessively missing are good candidates to remove, imputing these is mostly an exercise in observation creation. We can see in the image below that observations 28,76,118 and 168 are missing 50% of the time at least. If we leave them in and impute the missing values, we will end up with a largely synthetic observation. That is not a good thing when you have ample observations available.

[Image 15](#)



After removing the above observations from our data set, we have no observations missing more than 47 % so we can handle that via imputation.

Image 16



After steps 1 and 2 removing excessively missing variable and observations, our data dimension is 186 observations for 13 variables.

Image 17

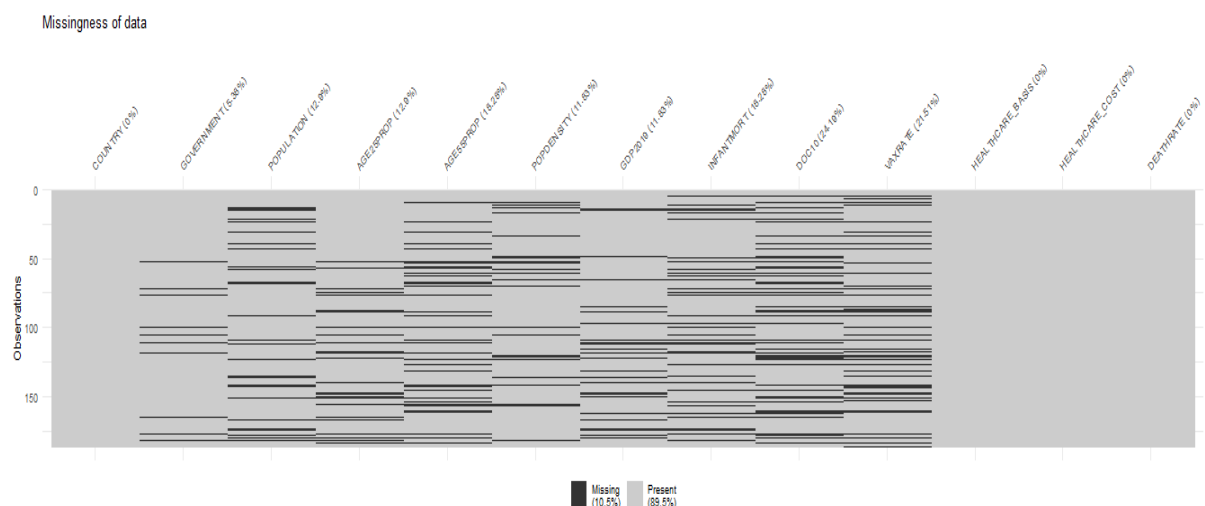
Data Frame Summary

assignment2a

Dimensions: 186 x 13

It's interesting to look at the missingness chart now. Overall missingness is now 10.5 %, down from 15% and we have 89.5 % data present. Also, we have now 4 variables country, healthcare_basis, healthcare_cost and death rate with no missing values, up from two variables country and death rate earlier.

Image 18

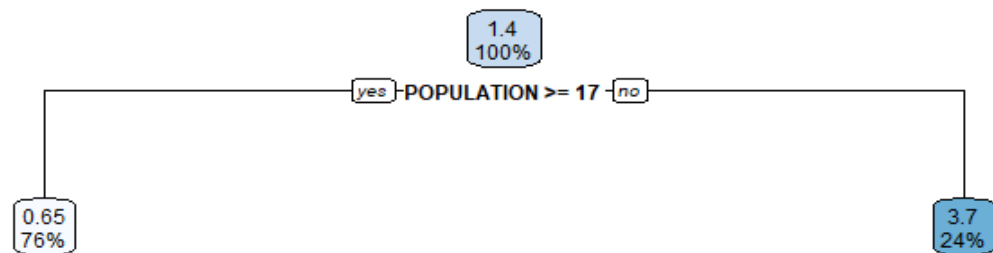


Predicting the missingness can be achieved as well to give us insight. We shall try to predict the number of missing values per observation. We shall model using rpart, not because it is a great method, but because it generates a highly explainable non-linear model. We are primarily interested to see if there is a relationship between missingness and any of the variables. We have included the id (country) and outcome variable (death rate) in this prediction problem. We can see from the image below that there is relationship between population and missingness. Countries with lower than or equal to 17 million population have higher missingness for observations compared to countries with population above 17 million. It might be smaller

countries with lower population might not have robust data collection compared to higher population countries.

[Image 18](#)

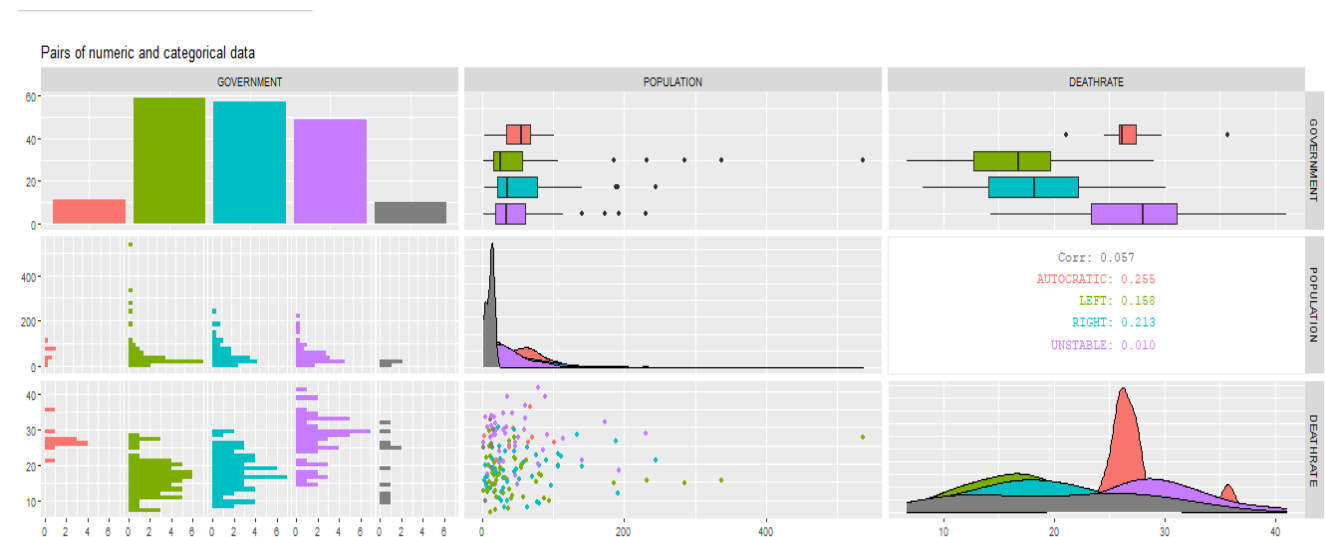
TUNED: Predicting the number of missing variables in an observation



4. Exploratory data analysis

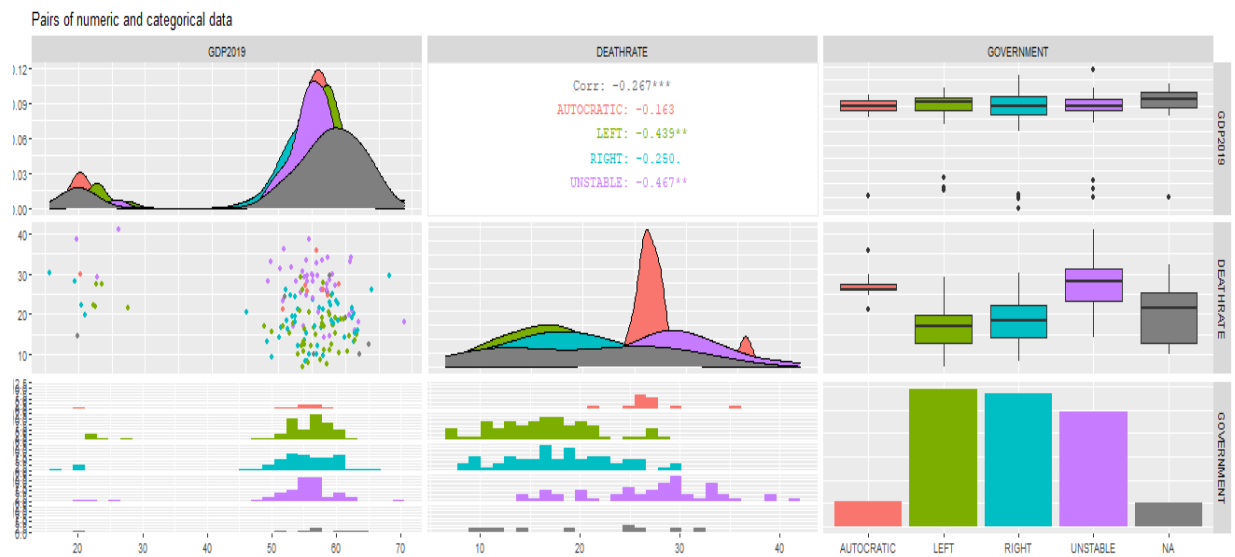
From Pair plot we can see the unstable government purple seems to have high death rate from box plot and wide range as well. Autocratic government looks to have the second-high death rate low range and couple of outlier points, followed by right and then the left government types. Autocratic government has a very small range of values as well as squashed.

[Image 19 – Pairs plot](#)



Interesting feature also is the GDP, GDP and death rate have weak negative correlation for autocratic and right government and moderate negative correlation for the left and unstable government types.

Image 20 Pair plot



From the box plot from image 21 to 28, we can see 8 of our numeric predictor variables have outliers when IQR is set to 1.5. We shall observe them and their points, surprisingly, that this does not necessarily mean that we have a data problem or a modelling problem. Also, to note our outcome variable death rate image 29 doesn't have any outliers when IQR set to 1.5. Also, to note death rate seems to be slightly positively skewed.

Image 21 -Box plot Population

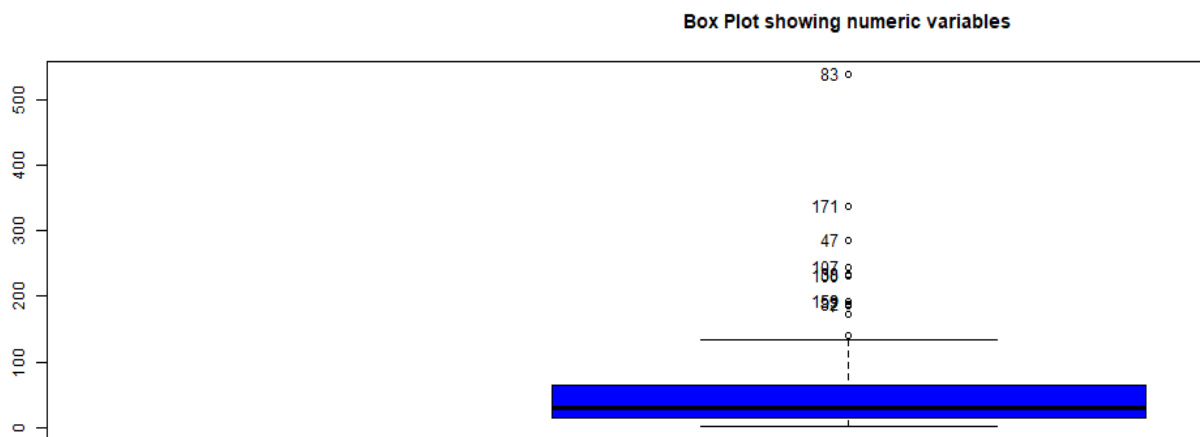


Image 22 – Box plot median age 55 or above

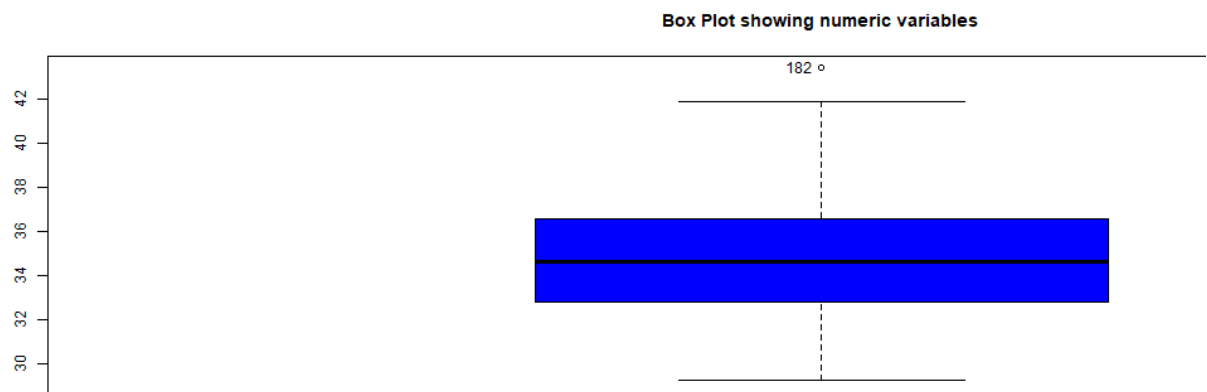


Image 23 - Box plot population density

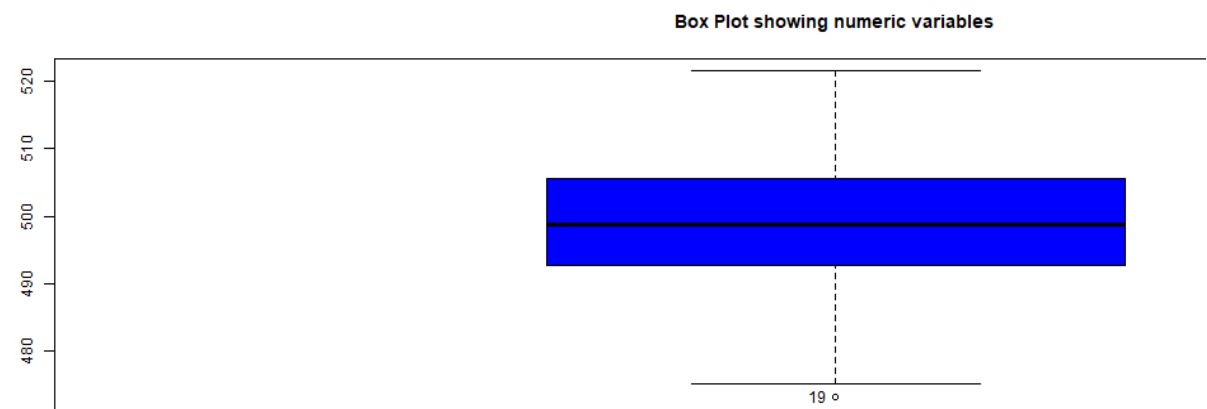


Image 24 - Box plot GDP 2019

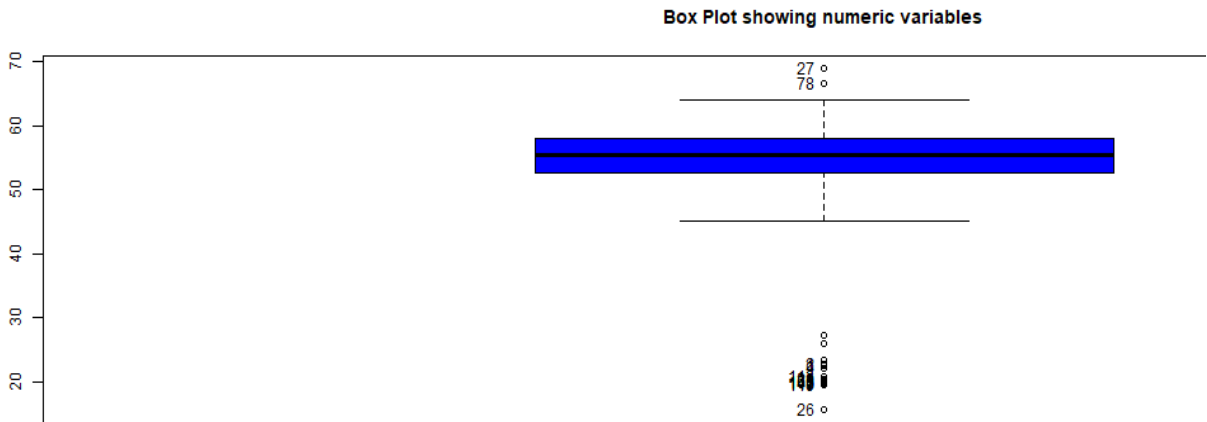


Image 25 - Box plot infant mortality

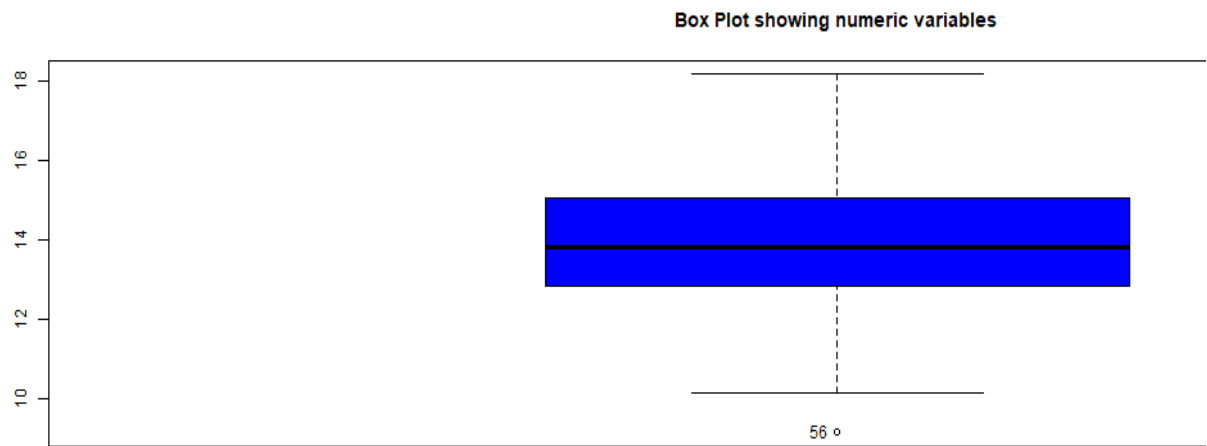


Image 26 – Box plot DOC10 - the number of doctors per 10,000

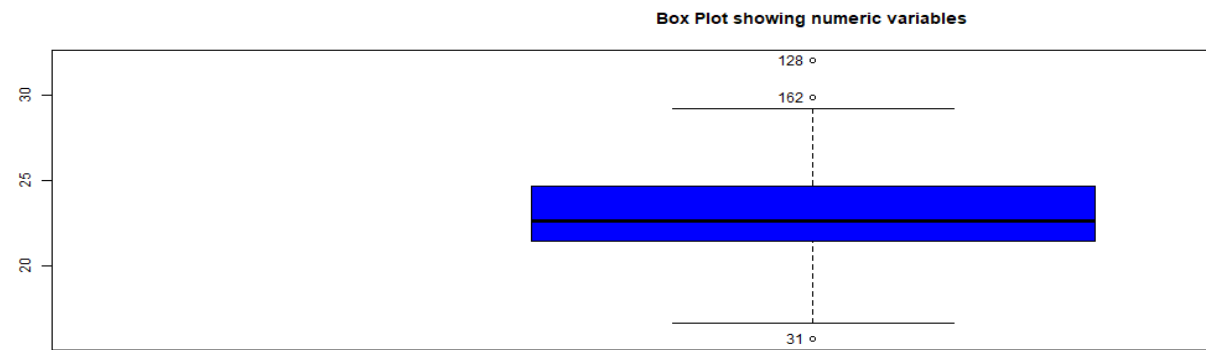
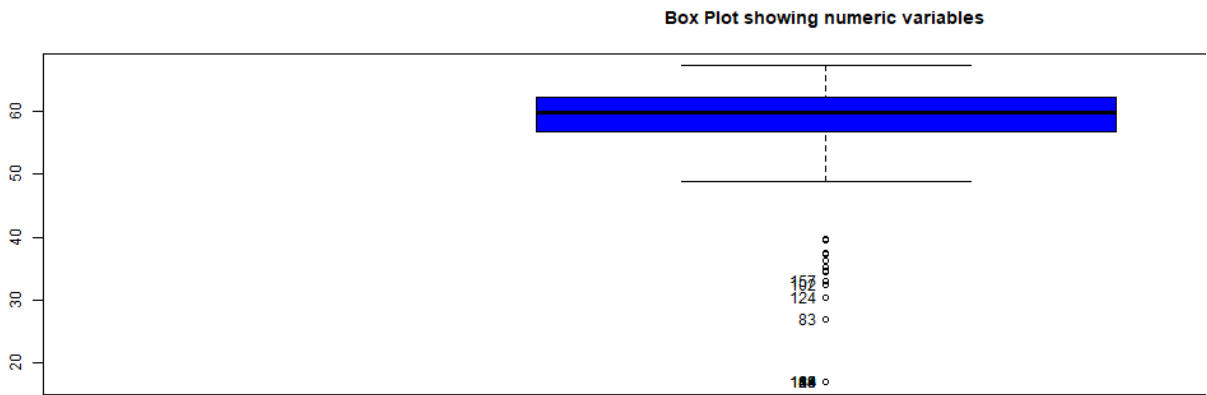
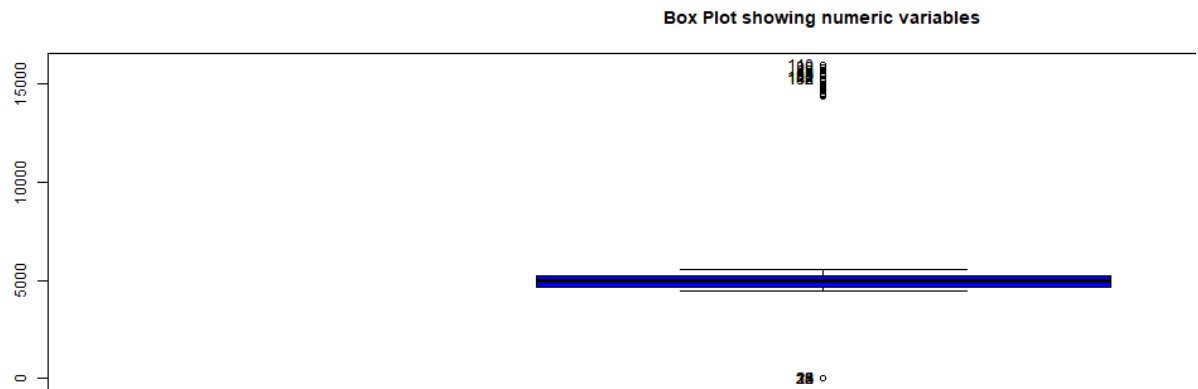


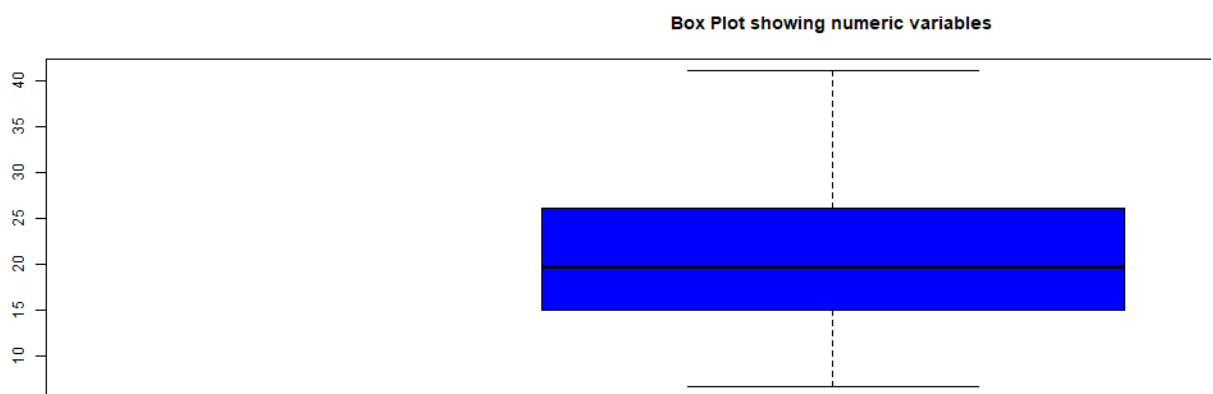
Image 27 – Box plot Vaxrate



[Image 28 – Health care cost](#)



[Image 29 Box plot death rate](#)



5. Train and test split

We create train and test split of our data before passing through recipe-based processing pipeline. We partition data to 0.7 for training the model and 0.3 for test data, so our model can be tested. We can see the training data table in the app has dimension 132 * 13 (132 rows and 13 variables) rest of the 30 percent of observations are in our test data table which has dimension 54*13 (54 rows and 13 variables).

6. Model building, prediction and assessment of the model

Before training our model. We update our recipe-based processing pipeline. Where we assign country variable the ID role and it is not used as a predictor. Impute missing values using KNN impute setting neighbours = 5. Also, outcome role i.e death rate is removed from imputing in the pipeline.

Then we feed out training data to model using the method glmnet. Glmnet fits lasso and elastic-net model paths for regression, logistic and multinomial regression using coordinate descent. The algorithm is extremely fast, and exploits sparsity in the input x matrix where it exists. A variety of predictions can be made from the fitted models. It gets rid of variables without good predictive powers for our outcome variable.

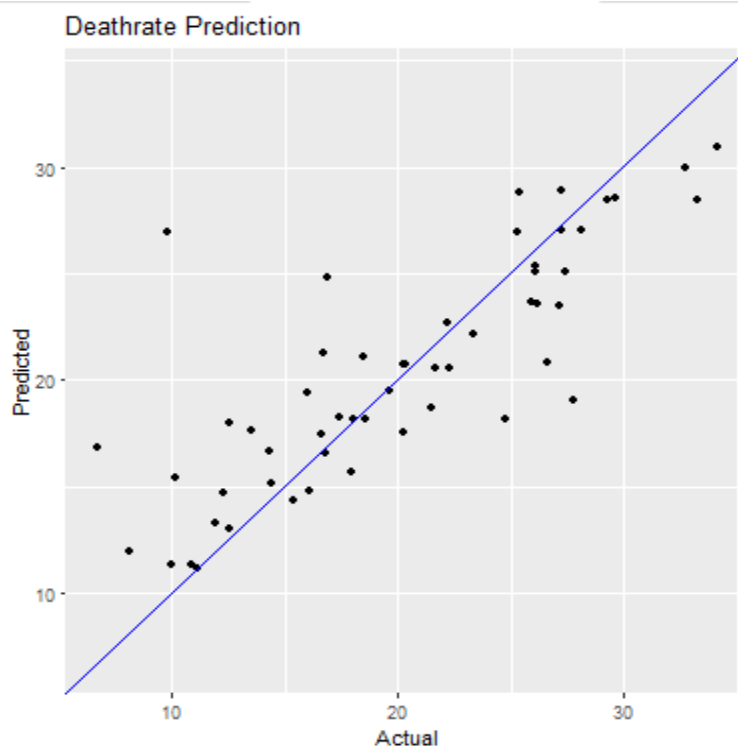
Image 30 is the prediction value for the 54 test cases and test rmse we get is around 4, which is good prediction accuracy for our model. In the predicted vs actual death rate visualization in image 31 we can see most values lie around the regression line, we do have couple of values on left side of the regression line quiet away from the regression line.

[Image 30 Prediction values and RMSE](#)

```
[1] 20.59141 26.94539 18.03710 18.19639 19.41358 30.97358 19.01662 18.14391 22.72844 20.59200
[11] 14.36789 28.80803 18.13303 11.98327 29.97786 24.80552 21.26225 28.54580 14.73360 16.57225
[21] 26.92225 20.77365 12.99323 25.12798 25.04217 16.68794 28.86631 17.62557 14.84467 11.38489
[31] 27.07132 22.17261 17.53629 20.76142 11.31522 15.17628 13.28842 18.25889 16.85669 28.40852
[41] 28.42562 15.71755 17.46850 23.65694 23.47158 11.17024 23.59988 25.38757 21.14230 20.84588
[51] 18.68670 15.40564 19.46743 27.02327
```

```
[1] 4.043038
```

[Image 31 visualization of actual vs predicted](#)



Even from the residual plot we can see that it is slightly left skewed and there are few points which are outlier points 76,71, 105, 153,65 on the left side and 23 on the right side, which we want to make a note of and inspect.

Image 32 visualization of residuals

