Spring 2021 - Final Examination

Maaz Shaikh

Instructions

Your goal for this final exam is to conduct the necessary analyses of vaccination rates in California school districts and then write up a technical report for a scientifically knowledgeable staff member in a California state legislator's office. You should provide sufficient numeric and graphical detail that the staff member can create a comprehensive briefing for a legislator (see question 10 for specific points of interest). You can assume that the staff member understands the concept of statistical significance and other basic concepts like mean, standard deviation, and correlation.

For this exam, the report writing is very important: Your responses will be graded on the basis of clarity; conciseness; inclusion and explanation of specific and appropriate statistical values; inclusion of both frequentist and Bayesian inferential evidence (i.e., it is not sufficient to just examine the data); explanation of any included tabular material and the appropriate use of graphical displays when/if necessary. It is also important to conduct a thorough analysis, including both data exploration and cleaning and appropriate diagnostics. Bonus points will be awarded for work that goes above expectations.

In your answer for each question, make sure you write a narrative with complete sentences that answers the substantive question. You can choose to put important statistical values into a table for readability, or you can include the statistics within your narrative. Be sure that you not only report what a test result was, but also what that result means substantively. Make sure to include enough statistical information so that another analytics professional could review your work. Your report can include graphics created by R, keeping in mind that if you do include a graphic, you will have to provide some accompanying narrative text to explain what it is doing in your report. Finally, be sure to proofread your final knitted submission to ensure that everything is included and readable.

You may not receive assistance, help, coaching, guidance, or support from any human except your instructor at any point during this exam. Your instructor will be available by email throughout the report writing period if you have questions, but don't wait until the last minute!

Data

You have an RData file available on Blackboard area that contains two data sets that pertain to vaccinations for the U.S. as a whole and for Californian school districts. The U.S. vaccine data is a time series and the California data is a sample of end-of-year vaccination reports from n=700 school districts. Here is a description of the datasets:

usVaccines – Time series data from the World Health Organization reporting vaccination rates in the U.S. for five common vaccines

```
Time-Series [1:38, 1:5] from 1980 to 2017:
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:5] "DTP1" "HepB_BD" "Pol3" "Hib3" “MCV1”...
```

(Note: DTP1 = First dose of Diphtheria/Pertussis/Tetanus vaccine (i.e., DTP); HepB_BD = Hepatitis B, Birth Dose (HepB); Pol3 = Polio third dose (Polio); Hib3 – Influenza third dose; MCV1 = Measles first dose (included in MMR))

districts – A sample of California public school districts from the 2017 data collection, along with specific numbers and percentages for each district:

```
'data.frame':   700 obs. of  14 variables:
 $ DistrictName    : Name of the district
 $ WithDTP         : Percentage of students in the district with the DTP vaccine
 $ WithPolio       : Percentage of students in the district with the Polio vaccine
 $ WithMMR         : Percentage of students in the district with the MMR vaccine
 $ WithHepB        : Percentage of students in the district with Hepatitis B vaccine
 $ PctUpToDate     : Percentage of students with completely up-to-date vaccines
 $ DistrictComplete: Boolean showing whether or not district’s reporting was complete
 $ PctBeliefExempt : Percentage of all enrolled students with belief exceptions
 $ PctMedicalExempt: Percentage of all enrolled students with medical exceptions
 $ PctChildPoverty : Percentage of children in district living below the poverty line
 $ PctFamilyPoverty: Percentage of families in district living below the poverty line
 $ PctFreeMeal     : Percentage of students in the district receiving free or reduced cost meals
 $ Enrolled        : Total number of enrolled students in the district
 $ TotalSchools    : Total number of different schools in the district
```

As might be expected, the data are quite skewed: districts range from 1 to 582 schools enrolling from 10 to more than 50,000 students. Further, while most districts have low rates of missing vaccinations, a handful are quite high. Be sure to note problems the data cause for the analysis and address any problems you can.

```
load("C:/Users/shaik/Downloads/datasets15.RData")
```

Descriptive Reporting

1. *Basic Introductory Paragraph*

In your own words, write about three sentences of introduction addressing the staff member in the state legislator's office. Frame the problem/topic that your report addresses.

We have two datasets one is time series which shows how the rate of vaccination has been varied over time. We have the data from the year 1980 to 2017. What are the challenges faced and how to tackle those uusing statistical analysis.

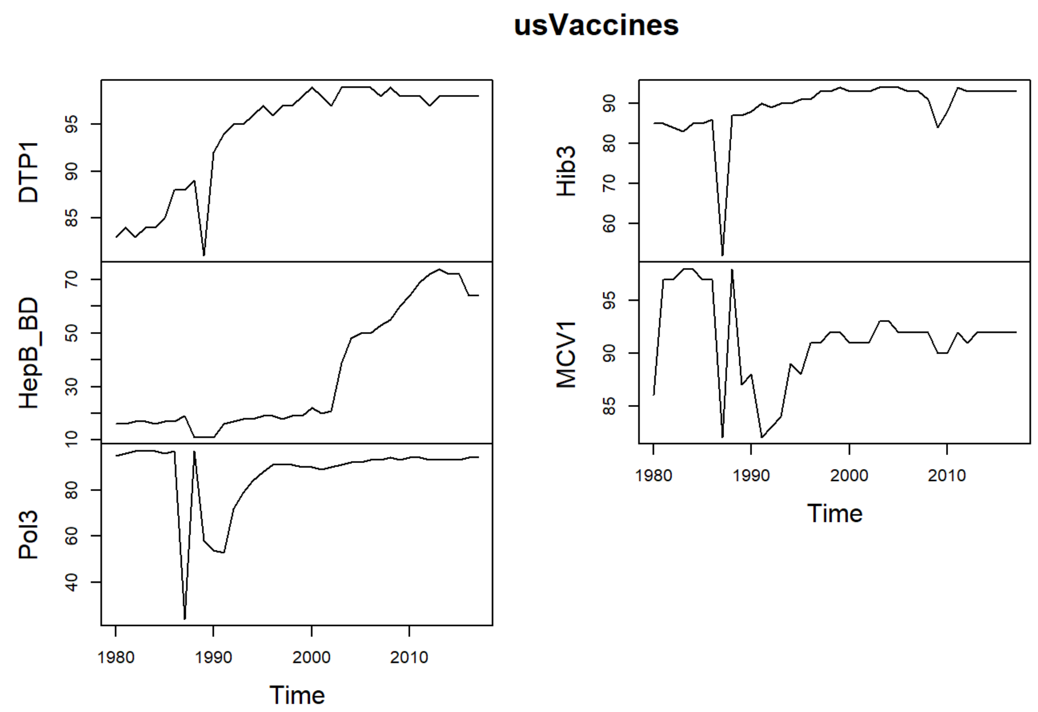
The another datasets which is districts shows how many children are up_to_date with the vaccinatio and how many are not. The challenge here is why others are not vaccinated (up_to_date), what are the challenges faced by them and what are the possible solutions to it ## 2. *Descriptive Overview of U.S. Vaccinations*

You have U.S. vaccination data going back 38 years, but the staff member is only interested in recent vaccination rates as a basis of comparison with California schools.

a. *How have U.S. vaccination rates varied over time?*

Visualizing and undertanding the data

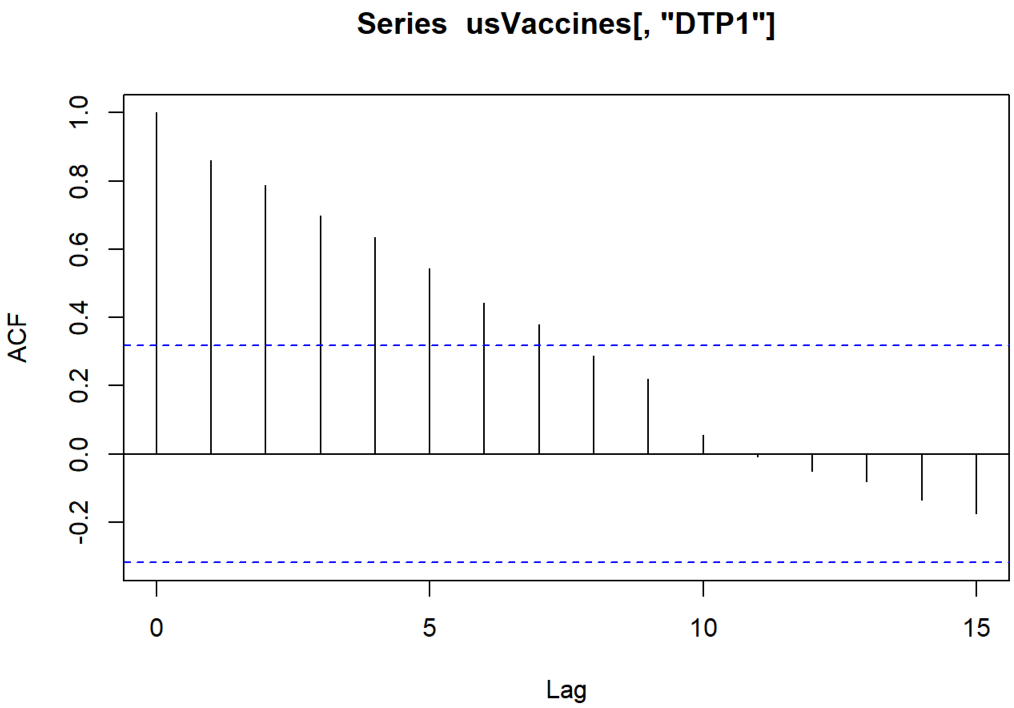
```
plot.ts(usVaccines)
```



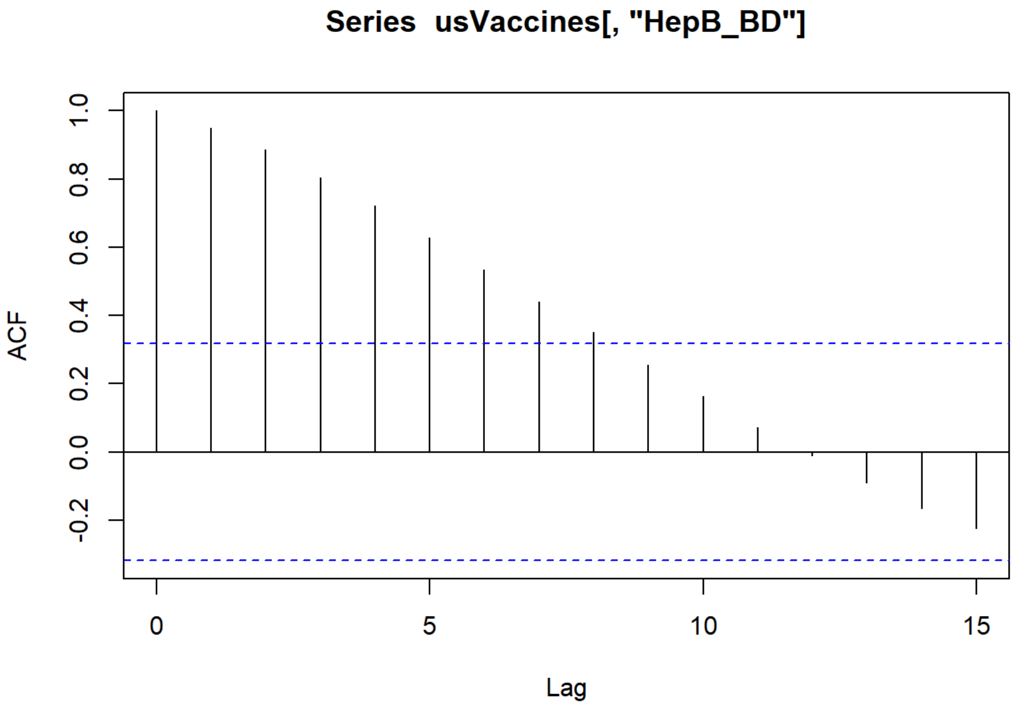
As per the visualization of the graph the important aspect is to note that all of them had a huge drop in rates in the late 1980's and only third dose of polio and first dose of measles shows a drop in rate in early 90's else it shows an overall increase in the rates till the year 2000 and then it shows a constant trend with a negligible increase in all of the vaccines except that of Hepatitis B, Hepatitits B shows an increase till 2010 and a boost in rate in at the initial years after 2010

b. *Are there notable trends or cyclical variation in U.S. vaccination rates?*

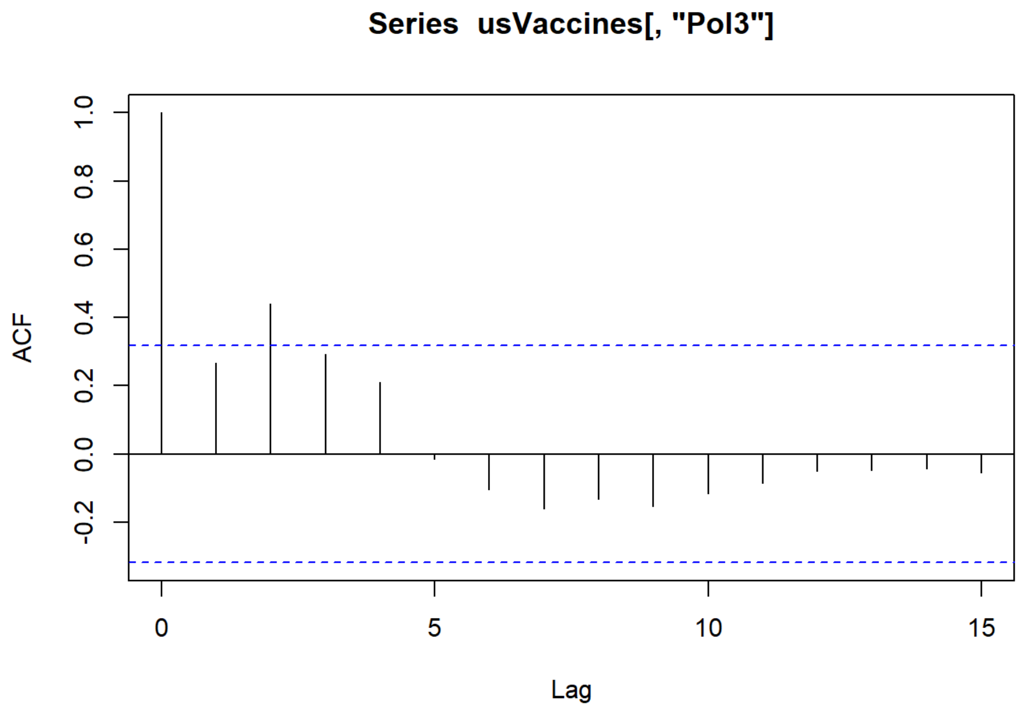
```
acf(usVaccines[,"DTP1"])
```



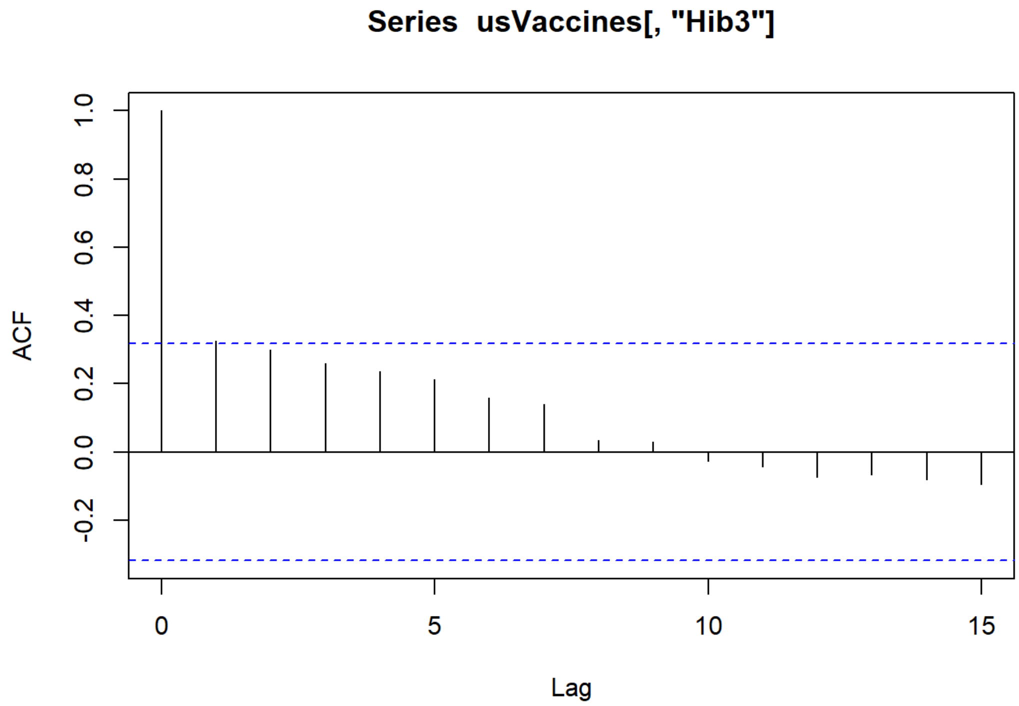

```r
acf(usVaccines[,"HepB_BD"])
```



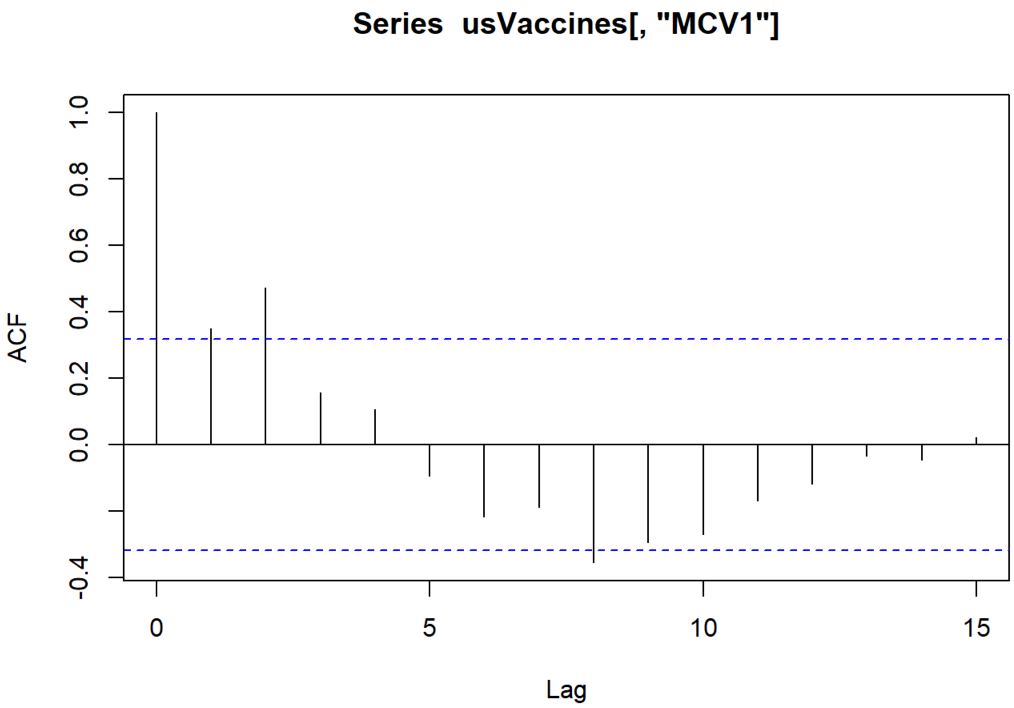

```r
acf(usVaccines[,"Pol3"])
```




```
acf(usVaccines[,"Hib3"])
```




```
acf(usVaccines[,"MCV1"])
```



There can be a trend and cyclicality in First dose of Diphtheria/Pertussis/Tetanus vaccine and Hepatitis B, Birth Dose because the interpreting lines are crossing the dotted lines in the graph, However Polio third dose ,Influenza third dose, Measles first dose (included in MMR does not contain that much of trend and cyclicality because of their frequency

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
adf.test(usVaccines[,"DTP1"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "DTP1"]
## Dickey-Fuller = -0.87963, Lag order = 3, p-value = 0.943
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"HepB_BD"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "HepB_BD"]
## Dickey-Fuller = -1.9729, Lag order = 3, p-value = 0.5839
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"Pol3"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "Pol3"]
## Dickey-Fuller = -2.3918, Lag order = 3, p-value = 0.4202
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"Hib3"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "Hib3"]
## Dickey-Fuller = -2.3377, Lag order = 3, p-value = 0.4414
## alternative hypothesis: stationary
```

```
adf.test(usVaccines[,"MCV1"])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines[, "MCV1"]
## Dickey-Fuller = -2.5324, Lag order = 3, p-value = 0.3652
## alternative hypothesis: stationary
```

The alternative hypothesis suggests all of the vaccines are stationary over time but the p-value to favour the alternate hypothesis is not significant and hence we cannot go with the test model as we can clearly see the trend and cyclicality with respect to time in the visualizations and it is not stationary at all. all of the m shows cyclicality and trends at some point.

c. *What are the mean U.S. vaccination rates when including only recent years in the calculation of the mean (examine your answers to the previous question to decide what a reasonable recent period is, i.e., a period during which the rates are relatively constant)?*

```
Recent <- window(usVaccines, start = 2016, end = 2017)
Recent
```

```
## Time Series:
## Start = 2016
## End = 2017
## Frequency = 1
##      DTP1 HepB_BD Pol3 Hib3 MCV1
## 2016   98      64   94   93   92
## 2017   98      64   94   93   92
```

```
Vaccine <- data.frame(Recent)
Vaccine
```

```
mean(Vaccine$DTP1)
```

```
## [1] 98
```

```
mean(Vaccine$HepB_BD)
```

```
## [1] 64
```

```
mean(Vaccine$Pol3)
```

```
## [1] 94
```

```
mean(Vaccine$Hib3)
```

```
## [1] 93
```

```
mean(Vaccine$MCV1)
```

```
## [1] 92
```

```
oldest <- window(usVaccines, start = 1980, end = 1981)
oldest
```

```
## Time Series:
## Start = 1980
## End = 1981
## Frequency = 1
##      DTP1 HepB_BD Pol3 Hib3 MCV1
## 1980   83      16   95   85   86
## 1981   84      16   96   85   97
```

```
summary(oldest)
```

```
##       DTP1          HepB_BD         Pol3            Hib3          MCV1      
##  Min.   :83.00   Min.   :16   Min.   :95.00   Min.   :85   Min.   :86.00  
##  1st Qu.:83.25   1st Qu.:16   1st Qu.:95.25   1st Qu.:85   1st Qu.:88.75  
##  Median :83.50   Median :16   Median :95.50   Median :85   Median :91.50  
##  Mean   :83.50   Mean   :16   Mean   :95.50   Mean   :85   Mean   :91.50  
##  3rd Qu.:83.75   3rd Qu.:16   3rd Qu.:95.75   3rd Qu.:85   3rd Qu.:94.25  
##  Max.   :84.00   Max.   :16   Max.   :96.00   Max.   :85   Max.   :97.00
```

Mean_DTP1 :- 83.50 Mean_HepB_Bd :- 16 Mean_Pol3 :- 95.5 Mean_Hib3 :- 85 Mean_MCV1 :- 91.50

In overall analysis and comaring the initial variance and visual representations we cann say that the vaccination had higher and constant rates in the year 2016 and 2017

3. *Descriptive Overview of California Vaccinations*

Your districts dataset contains four variables that capture the individual vaccination rates by district: WithDTP, WithPolio, WithMMR, and WithHepB.

a. *What are the mean levels of these variables across districts?*

```
Districts_whole <- subset(districts,select=c(WithDTP,WithPolio,WithMMR,WithHepB,PctUpToDate,PctBeliefExempt,
PctChildPoverty, PctFamilyPoverty,PctFreeMeal,Enrolled,TotalSchools))
summary(Districts_whole)
```

```
##     WithDTP         WithPolio         WithMMR          WithHepB     
##  Min.   : 23.00   Min.   : 23.00   Min.   : 23.00   Min.   : 23.00  
##  1st Qu.: 86.00   1st Qu.: 87.00   1st Qu.: 86.00   1st Qu.: 90.00  
##  Median : 93.00   Median : 94.00   Median : 94.00   Median : 96.00  
##  Mean   : 89.85   Mean   : 90.21   Mean   : 89.81   Mean   : 92.23  
##  3rd Qu.: 97.00   3rd Qu.: 97.00   3rd Qu.: 97.00   3rd Qu.: 98.00  
##  Max.   :100.00   Max.   :100.00   Max.   :100.00   Max.   :100.00  
##   PctUpToDate     PctBeliefExempt   PctChildPoverty PctFamilyPoverty
##  Min.   : 23.00   Min.   :  0.000   Min.   : 3.00   Min.   : 0.00   
##  1st Qu.: 84.00   1st Qu.:  1.000   1st Qu.:13.00   1st Qu.: 5.00   
##  Median : 92.00   Median :  2.000   Median :20.00   Median : 9.00   
##  Mean   : 87.94   Mean   :  6.206   Mean   :22.18   Mean   :11.32   
##  3rd Qu.: 96.00   3rd Qu.:  7.000   3rd Qu.:29.00   3rd Qu.:15.25   
##  Max.   :100.00   Max.   :110.000   Max.   :72.00   Max.   :47.00   
##   PctFreeMeal        Enrolled        TotalSchools    
##  Min.   :  0.00   Min.   :   10.0   Min.   :  1.000  
##  1st Qu.: 30.00   1st Qu.:   49.0   1st Qu.:  1.000  
##  Median : 50.00   Median :  192.5   Median :  3.000  
##  Mean   : 48.42   Mean   :  616.9   Mean   :  7.089  
##  3rd Qu.: 69.00   3rd Qu.:  670.0   3rd Qu.:  8.000  
##  Max.   :100.00   Max.   :54238.0   Max.   :582.000
```

Means are,

WithDTP : 89.85 WithPolio : 90.21 WithMMR : 89.81 WithHepB : 92.23 PctUpToDate : 87.94 PctBeliefExempt : 6.206 PctChildPoverty : 22.18 PctFamilyPoverty: 11.32 PctFreeMeal : 48.42 Enrolled : 616.9 TotalSchools : T7.089

b. *Among districts, how are the vaccination rates for individual vaccines related? In other words, if there are students with one vaccine, are students likely to have*

all of the others?

```
Districts_New1 <- subset(districts,select=c(WithDTP,WithPolio,WithMMR,WithHepB))
cor(Districts_New1)
```

```
##             WithDTP WithPolio   WithMMR  WithHepB
## WithDTP   1.0000000 0.9863267 0.9775216 0.8991932
## WithPolio 0.9863267 1.0000000 0.9709209 0.9067080
## WithMMR   0.9775216 0.9709209 1.0000000 0.8975075
## WithHepB  0.8991932 0.9067080 0.8975075 1.0000000
```

The rates of Polio and WithDTP are highly correlated The rates of MMR, with Polio are highly correlated The rates of With MMR, WithHepB are highly correlated

c. *How do these Californian vaccination levels compare to U.S. vaccination levels (recent years only)? Note any patterns you notice.*

```
#These are the mean US vaccination rates
mean(Vaccine$DTP1)
```

```
## [1] 98
```

```
mean(Vaccine$HepB_BD)
```

```
## [1] 64
```

```
mean(Vaccine$Pol3)
```

```
## [1] 94
```

```
mean(Vaccine$Hib3)
```

```
## [1] 93
```

```
mean(Vaccine$MCV1)
```

```
## [1] 92
```

```
#The below are the California mean vaccination rates

# WithDTP         : 89.85
 #WithPolio       : 90.21
 #WithMMR         : 89.81
 #WithHepB        : 92.23
```

As per the comparision of the means we can say that the DTP vaccine and the MMR vaccine are cheap in California as compared to the Polio and Hepatitis B vaccines

4. *Conclusion Paragraph for Vaccination Rates*

Provide one or two sentences of your professional judgment about where California school districts stand with respect to vaccination rates and in the larger context of the U.S. As per the comparision of the means we can say that the DTP vaccine and the MMR vaccine are cheap in California as compared to the Polio and Hepatitis B vaccines To overcome the challenge

We can notice from the above analysis that there is a variance in the vaccination rates of state and countries. To make it more cost effective the centralized rate or we can say one rate for entire nation is a very good solution as in vaccination the state and central both can contribute the funds making all of the vaccination more ost efective and making it more constant through out the period in future.

Inferential Reporting

For every item below except 7, use PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors. Explore the data *and transform variables as necessary to improve prediction and/or interpretability. Be sure to include appropriate diagnostics and modify your analyses as appropriate.*

5. *Which of the four predictor variables predicts the percentage of all enrolled students with belief exceptions?*

```
library(psych)
library(dlookr)
```

```
##
## Attaching package: 'dlookr'
```

```
## The following object is masked from 'package:psych':
##
##     describe
```

```
## The following object is masked from 'package:base':
##
##     transform
```

```
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x tidyr::extract() masks dlookr::extract()
## x dplyr::filter()  masks mice::filter(), stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
describe(districts)
```

```
summary(districts)
```

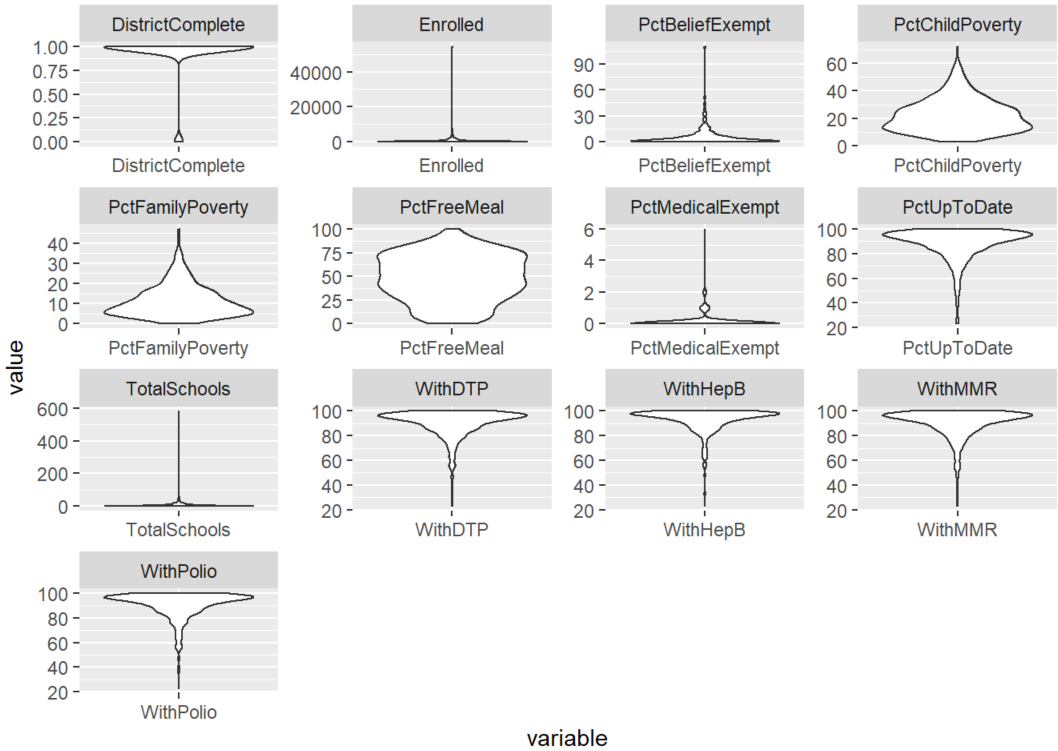
```
##                             DistrictName     WithDTP          WithPolio
##  ABC Unified                          :  1   Min.   : 23.00   Min.   : 23.00
##  Acton-Agua Dulce Unified             :  1   1st Qu.: 86.00   1st Qu.: 87.00
##  Adelanto Elementary                  :  1   Median : 93.00   Median : 94.00
##  Alameda Unified                      :  1   Mean   : 89.85   Mean   : 90.21
##  Albany City Unified                  :  1   3rd Qu.: 97.00   3rd Qu.: 97.00
##  Alexander Valley Union Elementary:  1   Max.   :100.00   Max.   :100.00
##  (Other)                              :694
##     WithMMR          WithHepB        PctUpToDate     DistrictComplete
##  Min.   : 23.00   Min.   : 23.00   Min.   : 23.00   Mode :logical
##  1st Qu.: 86.00   1st Qu.: 90.00   1st Qu.: 84.00   FALSE:37
##  Median : 94.00   Median : 96.00   Median : 92.00   TRUE :663
##  Mean   : 89.81   Mean   : 92.23   Mean   : 87.94
##  3rd Qu.: 97.00   3rd Qu.: 98.00   3rd Qu.: 96.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
##
##  PctBeliefExempt   PctMedicalExempt PctChildPoverty PctFamilyPoverty
##  Min.   :  0.000   Min.   :0.0000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:  1.000   1st Qu.:0.0000   1st Qu.:13.00   1st Qu.: 5.00
##  Median :  2.000   Median :0.0000   Median :20.00   Median : 9.00
##  Mean   :  6.206   Mean   :0.1532   Mean   :22.18   Mean   :11.32
##  3rd Qu.:  7.000   3rd Qu.:0.0000   3rd Qu.:29.00   3rd Qu.:15.25
##  Max.   :110.000   Max.   :6.0000   Max.   :72.00   Max.   :47.00
##                    NA's   :243
##   PctFreeMeal       Enrolled          TotalSchools
##  Min.   :  0.00   Min.   :   10.0   Min.   :  1.000
##  1st Qu.: 30.00   1st Qu.:   49.0   1st Qu.:  1.000
##  Median : 50.00   Median :  192.5   Median :  3.000
##  Mean   : 48.42   Mean   :  616.9   Mean   :  7.089
##  3rd Qu.: 69.00   3rd Qu.:  670.0   3rd Qu.:  8.000
##  Max.   :100.00   Max.   :54238.0   Max.   :582.000
##
```

```
diagnose(districts)
```

```
md.pattern(districts, plot=FALSE)
```

```
##     DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 457            1       1         1       1        1           1
## 243            1       1         1       1        1           1
##                0       0         0       0        0           0
##     DistrictComplete PctBeliefExempt PctChildPoverty PctFamilyPoverty
## 457                1               1               1                1
## 243                1               1               1                1
##                    0               0               0                0
##     PctFreeMeal Enrolled TotalSchools PctMedicalExempt
## 457           1        1            1                1   0
## 243           1        1            1                0   1
##               0        0            0              243 243
```

```
districts %>% pivot_longer(cols=-DistrictName, names_to="variable",
                           values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) + geom_violin() + facet_wrap( ~ variable, scales="free")
```

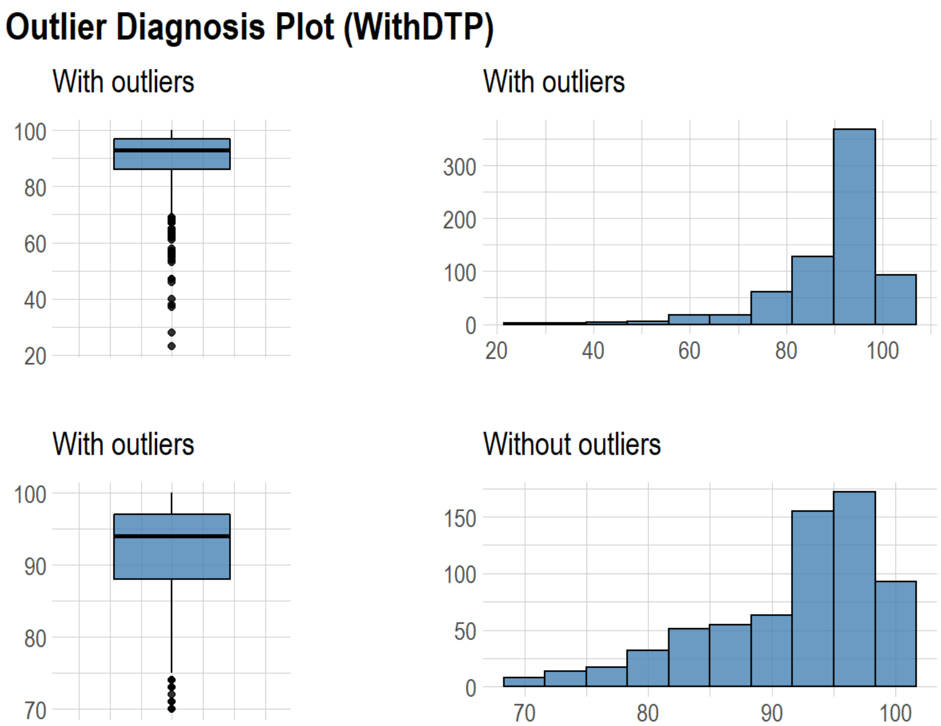


In this via violen plot we are looking whether the data is skewed or not. We are using four variables PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools.We can notice that the distribution of total schools and Enrolled students are highly skewed. To remove the skewness we can either take sqrt function or use log, the reciprocal of the value and so on.

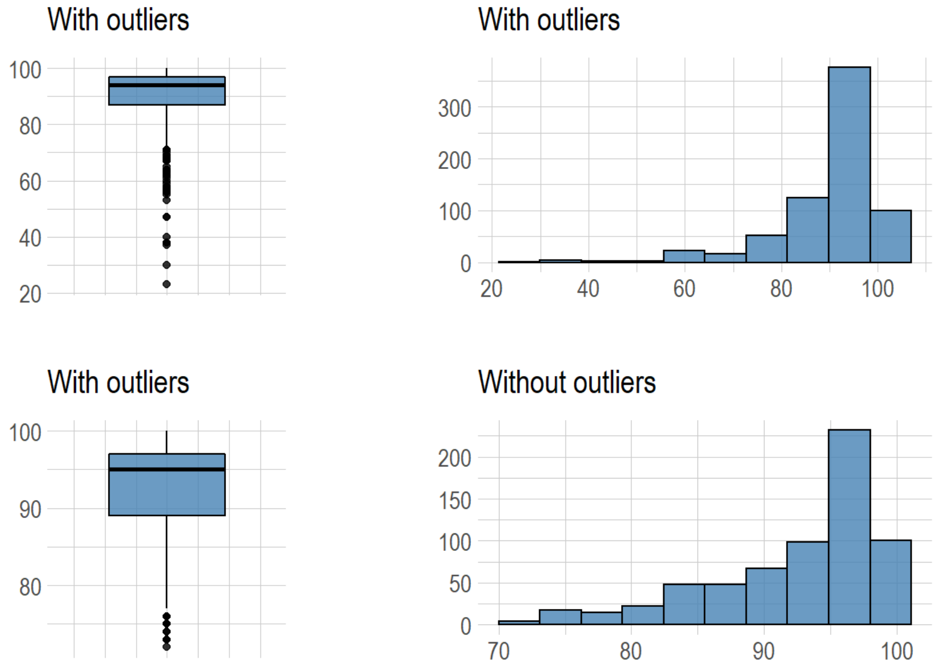
Taking a look at the outliers and and underatanding them while noicing if they are genuine extreme values or an error.

```
library(dlookr)
diagnose_outlier(districts)
```

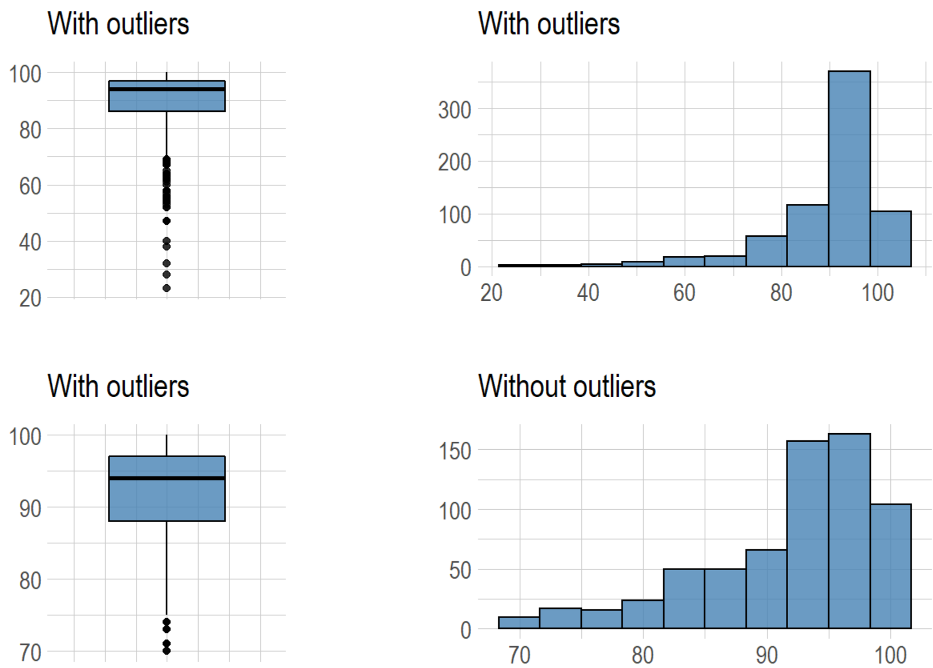
```
plot_outlier(districts)
```



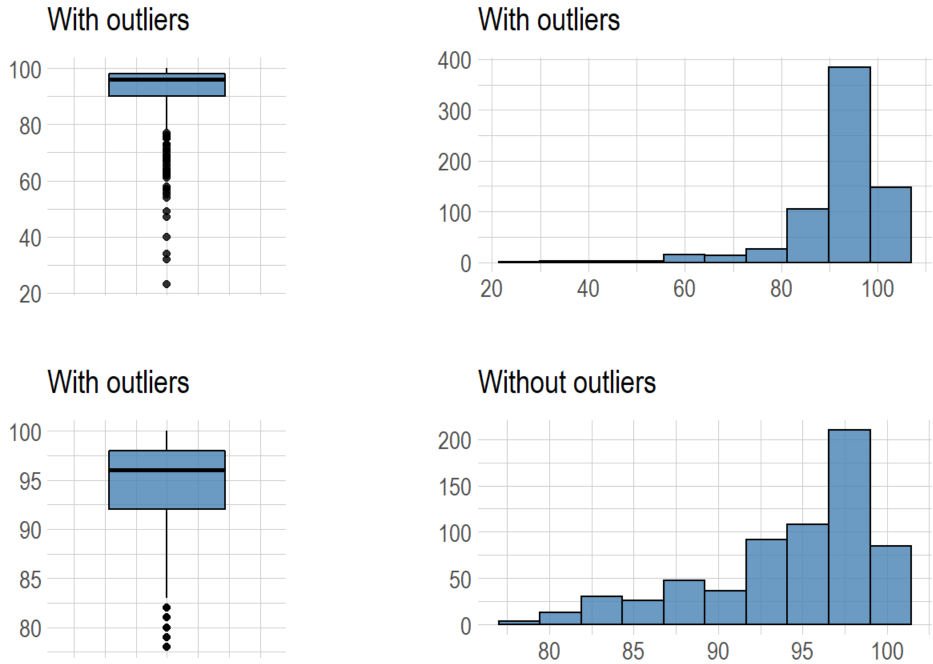
Outlier Diagnosis Plot (WithPolio)



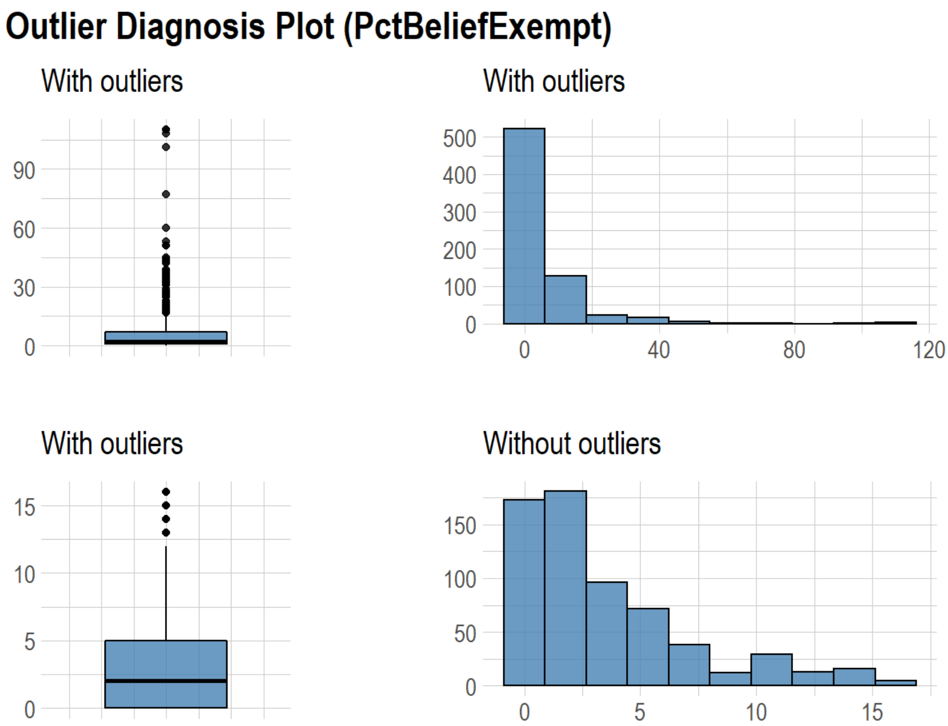
Outlier Diagnosis Plot (WithMMR)

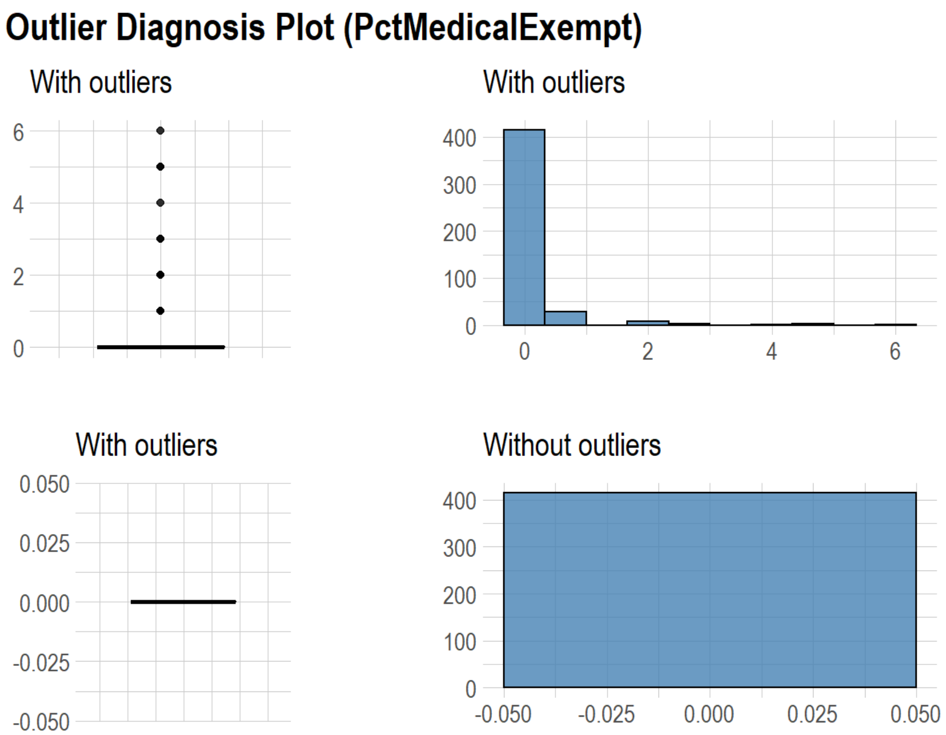


Outlier Diagnosis Plot (WithHepB)

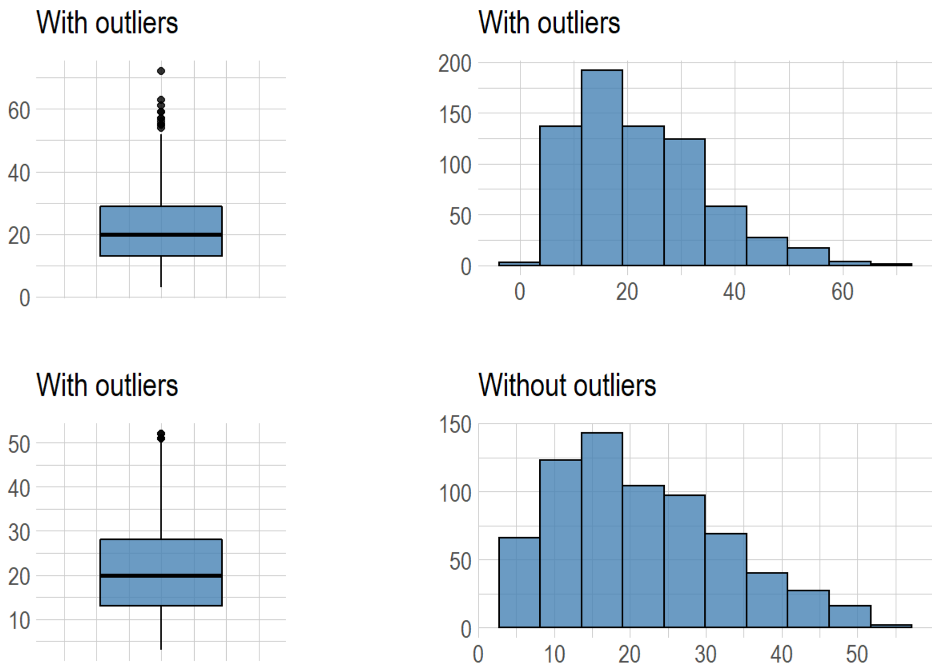


Outlier Diagnosis Plot (PctUpToDate)

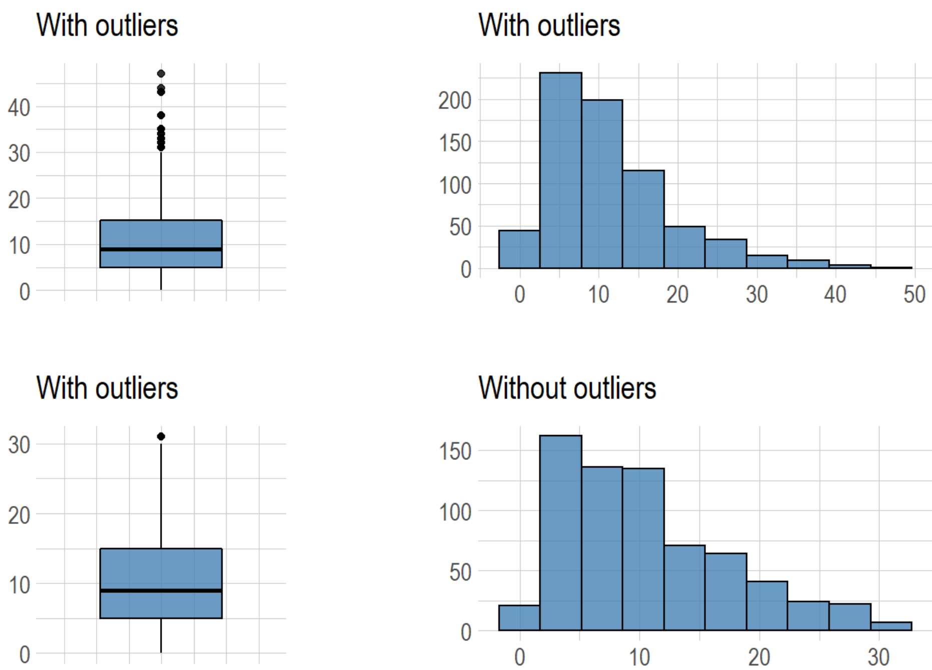

Outlier Diagnosis Plot (PctBeliefExempt)
With outliers
With outliers
90
60
30
0
500
400
300
200
100
0
0
40
80
120
With outliers
Without outliers
15
10
5
0
150
100
50
0
0
5
10
15


Outlier Diagnosis Plot (PctMedicalExempt)
With outliers
With outliers
6
4
2
0
400
300
200
100
0
0
2
4
6
With outliers
Without outliers
0.050
0.025
0.000
-0.025
-0.050
400
300
200
100
0
-0.050
-0.025
0.000
0.025
0.050

Outlier Diagnosis Plot (PctChildPoverty)

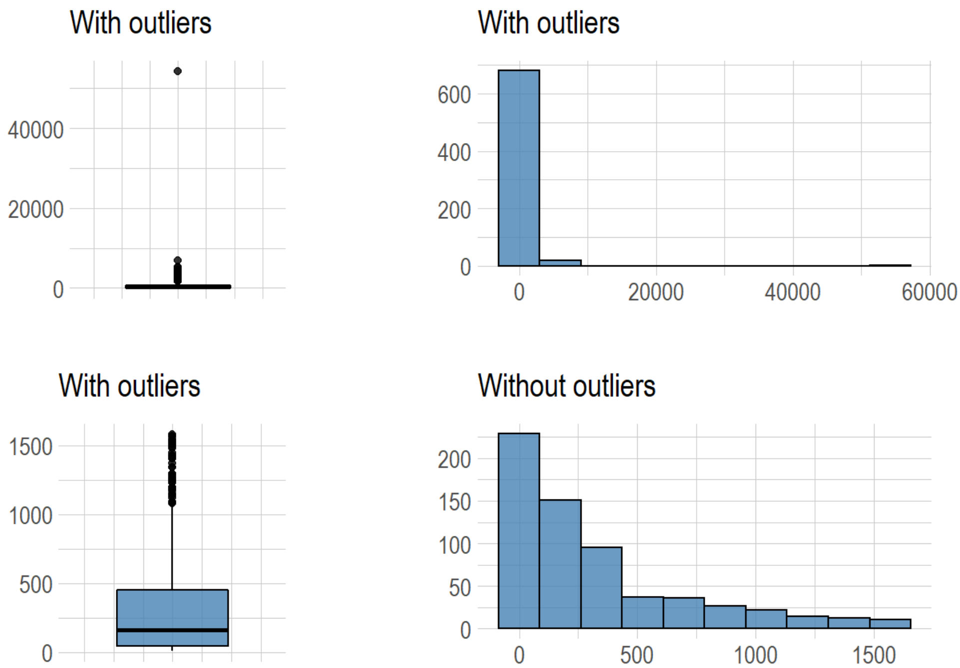


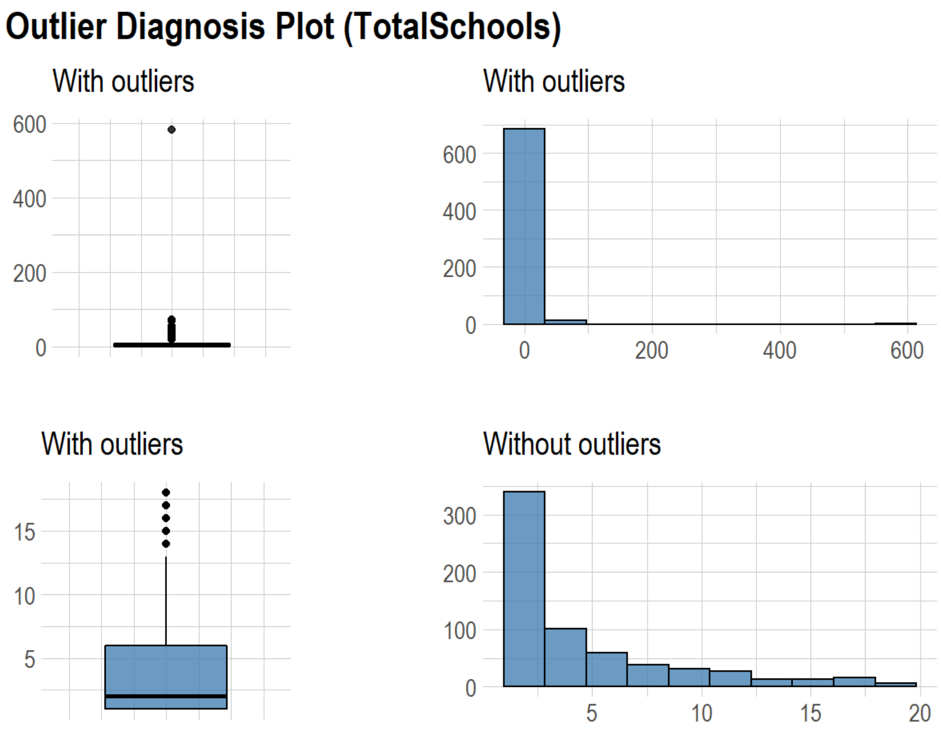
Outlier Diagnosis Plot (PctFamilyPoverty)



Outlier Diagnosis Plot (PctFreeMeal)

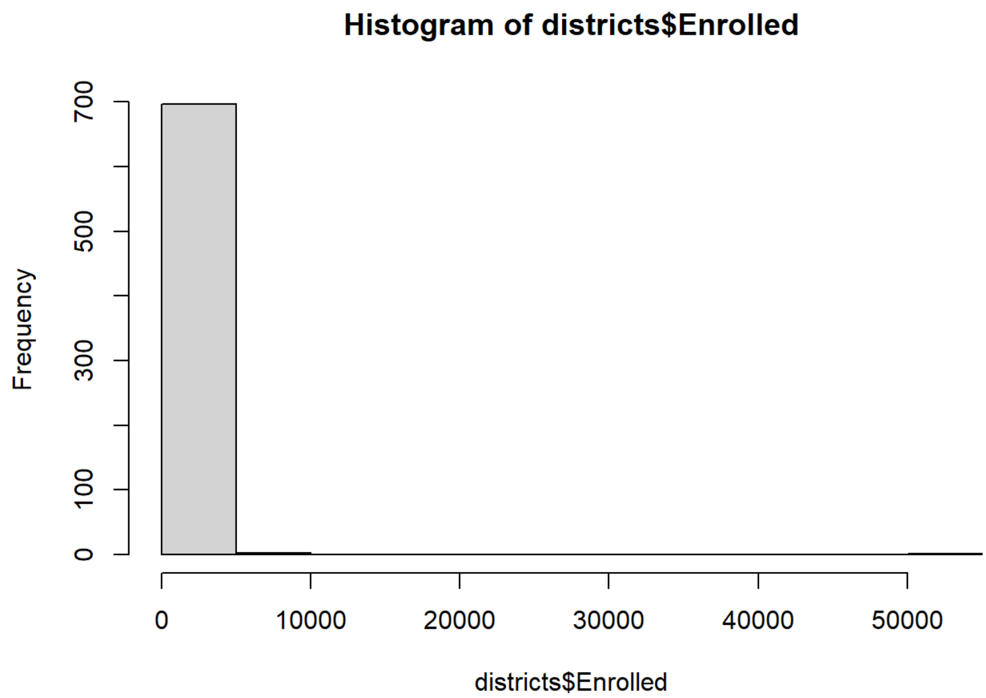
Outlier Diagnosis Plot (Enrolled)



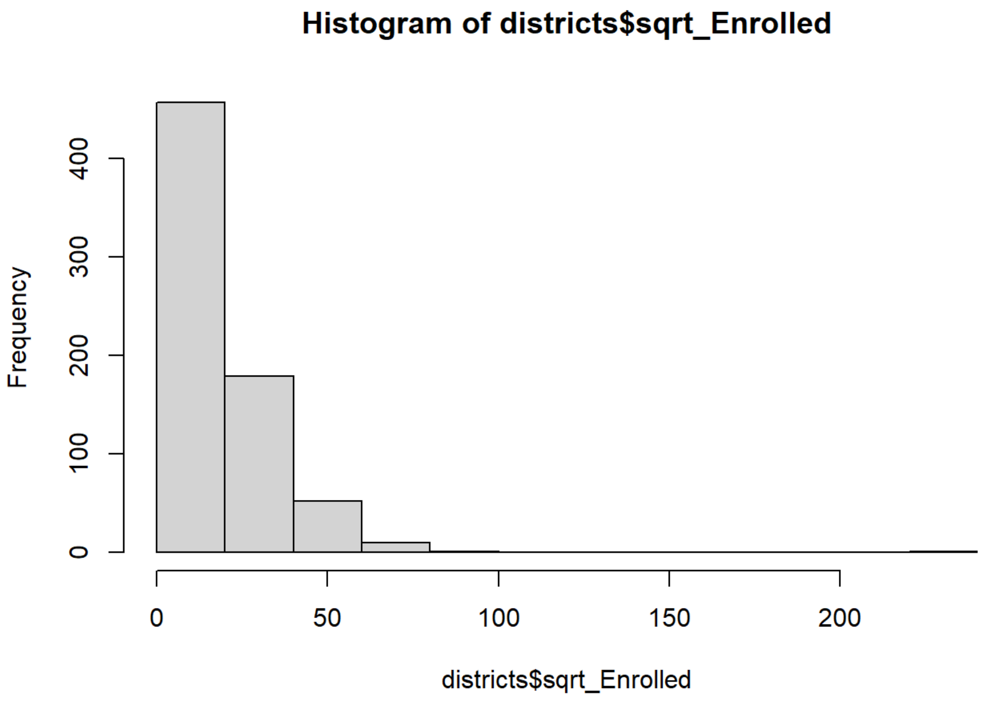


To improve the skewness using squareroot of the value

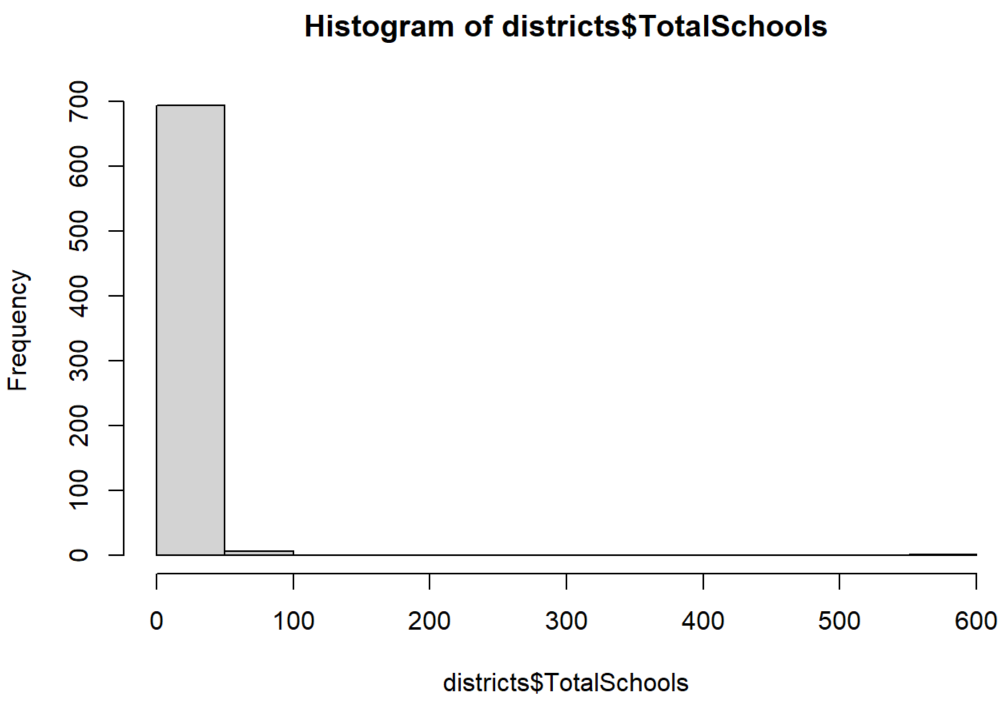
```
districts$sqrt_TotalSchools <- sqrt(districts$TotalSchools)
districts$sqrt_Enrolled <- sqrt(districts$Enrolled)
hist(districts$Enrolled)
```



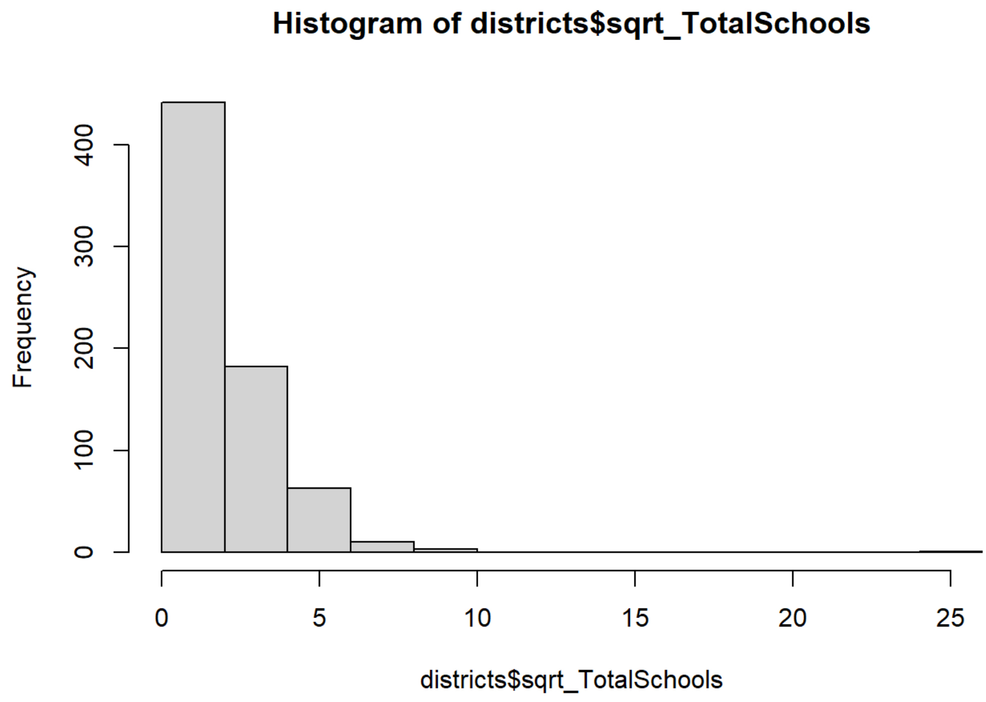

```
hist(districts$sqrt_Enrolled)
```




```
hist(districts$TotalSchools)
```



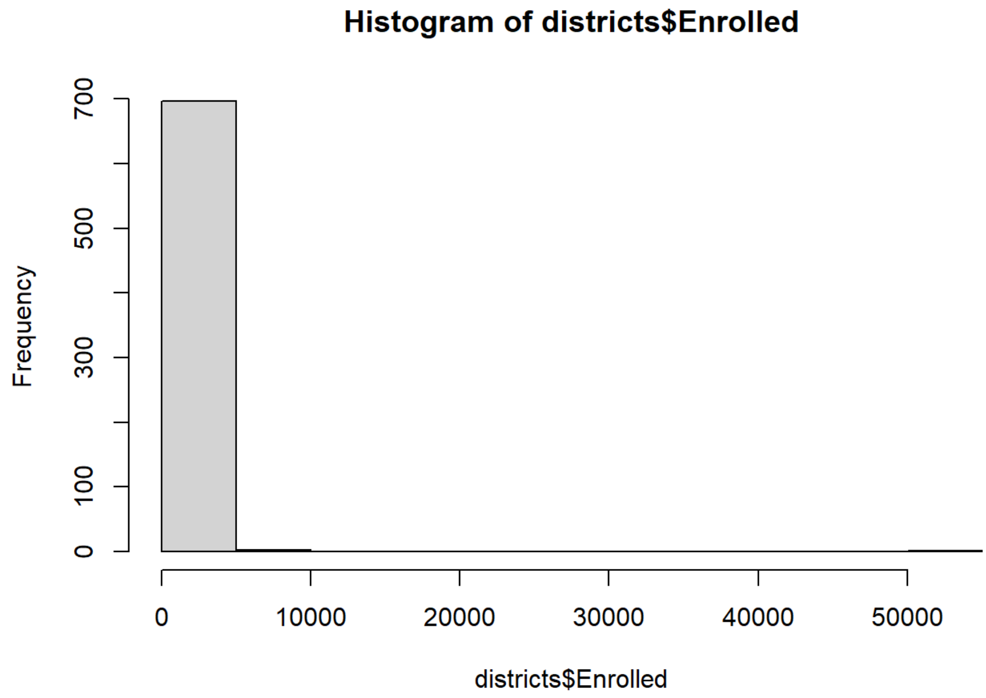

```
hist(districts$sqrt_TotalSchools)
```



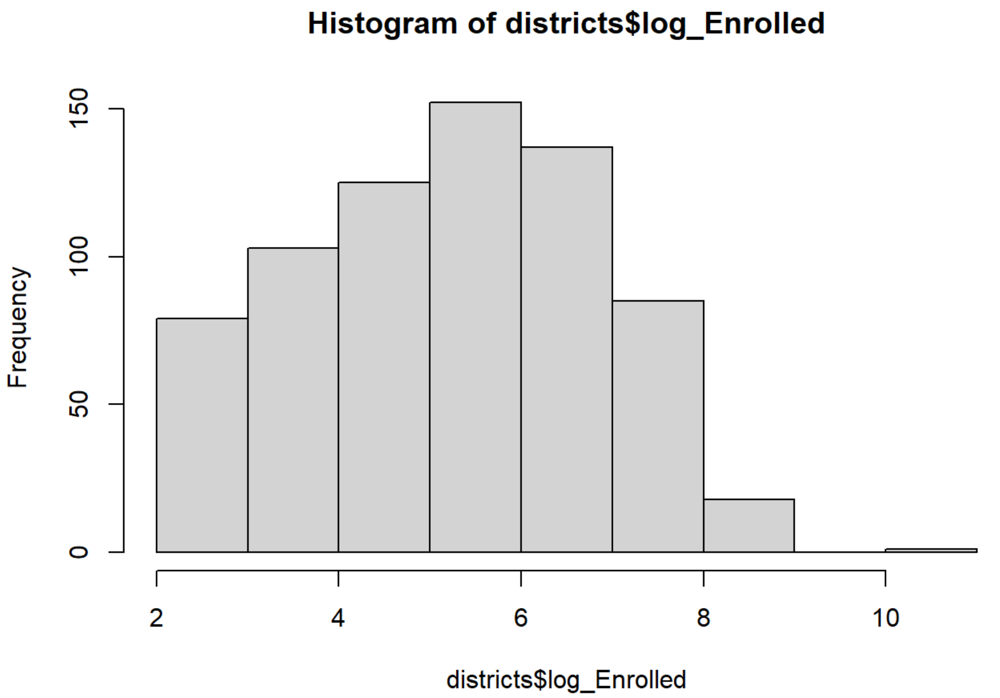
By operating and applying the sqrt we can see a noticable change in the graph but it is not a considerable because the after the operation the resultant shows the skewness as well..

Trying the log to remove skewness.

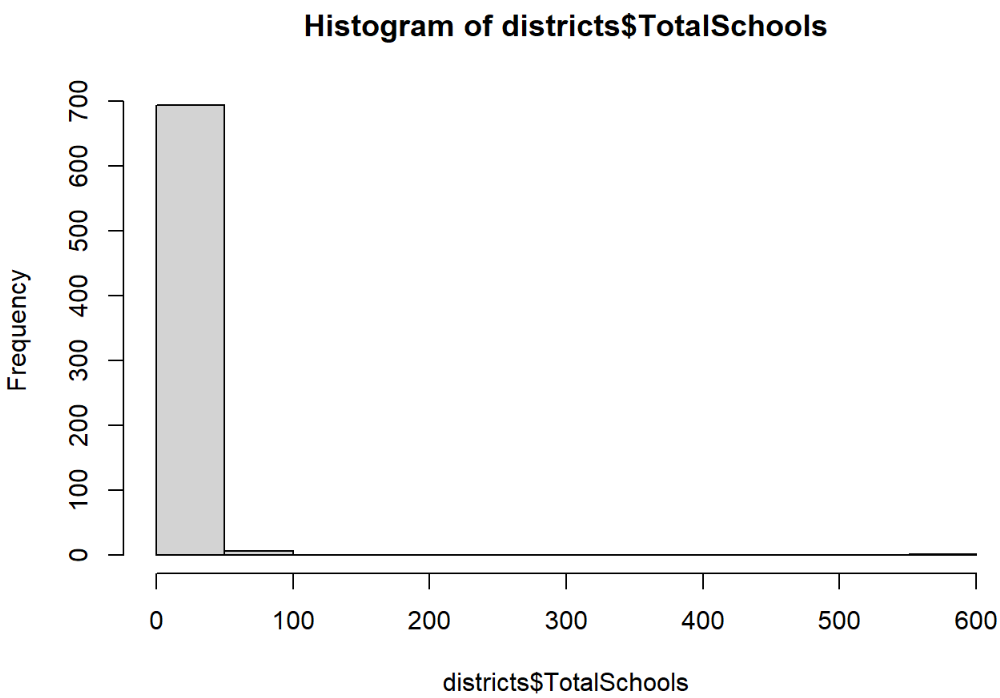
```
districts$log_TotalSchools <- log(districts$TotalSchools)
districts$log_Enrolled <- log(districts$Enrolled)
hist(districts$Enrolled)
```



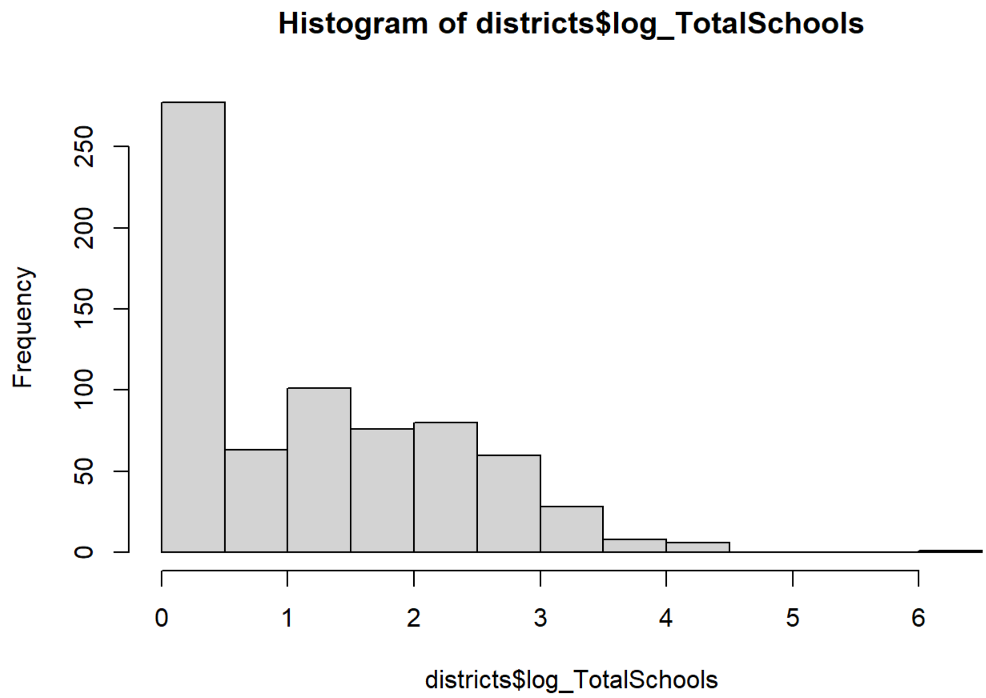

```
hist(districts$log_Enrolled)
```




```
hist(districts$TotalSchools)
```




```
hist(districts$log_TotalSchools)
```



In comparing the sqrt and log

we can see that the skewness is removed more using log than that of sqrt.

Checking how bad the outlier is by taking into consideration the numeric skewness.

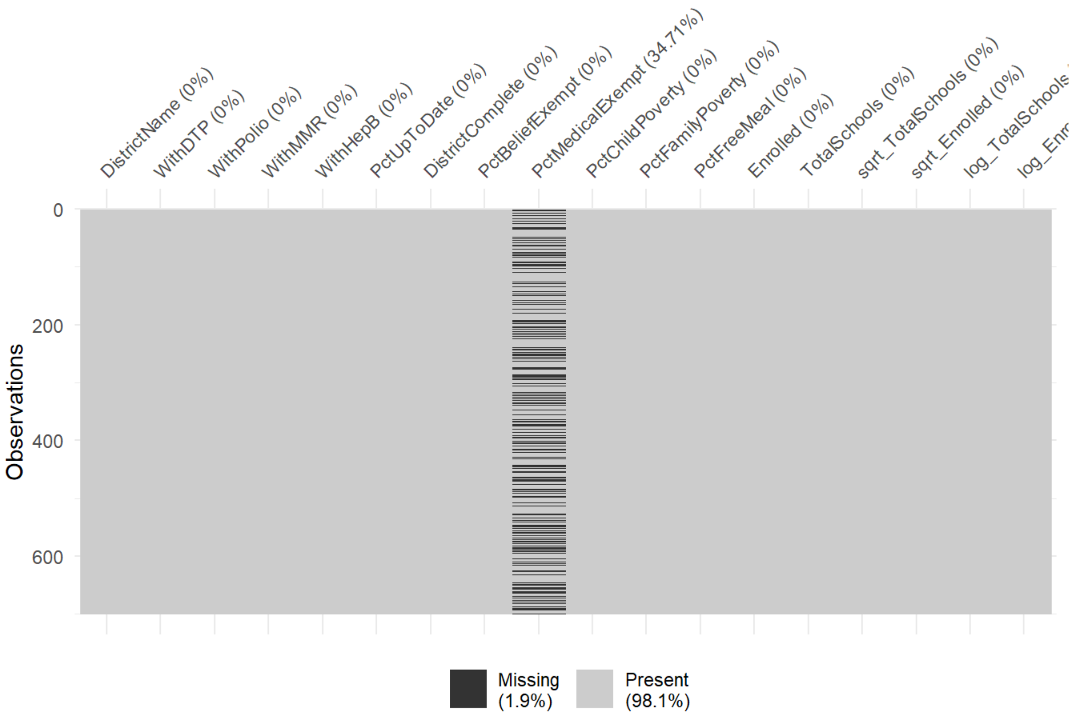
```
with(districts, apply(cbind(PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools), 2, skewness))
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##         0.832989         1.243198        21.069566        20.883276
```

```
with(districts, apply(cbind(PctChildPoverty, PctFamilyPoverty, districts$log_Enrolled, districts$log_TotalSc
hools ), 2, skewness))
```

```
##  PctChildPoverty PctFamilyPoverty
##      0.832988957      1.243198499      0.002892155      0.647966127
```

```
library(visdat)
vis_miss(districts)
```

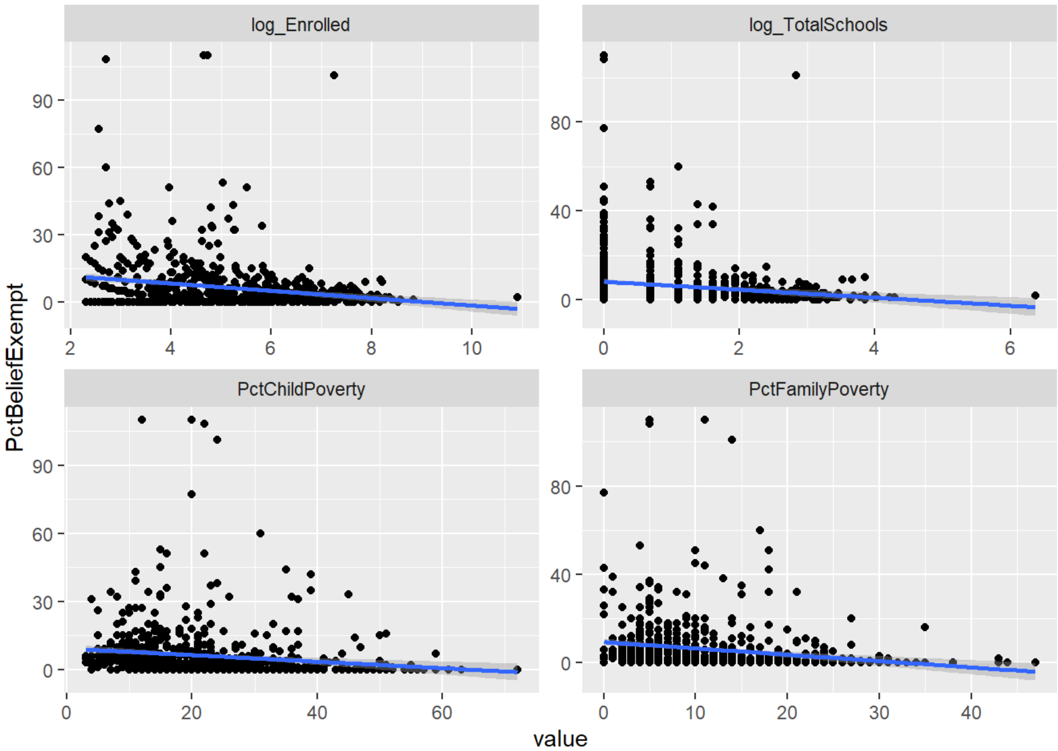


Here Percentage in Medical Exempt shows that there are 34.71 percentage of missing values. For our analysis since aprroximately 35% of the data is missing. I will be excluding th column.

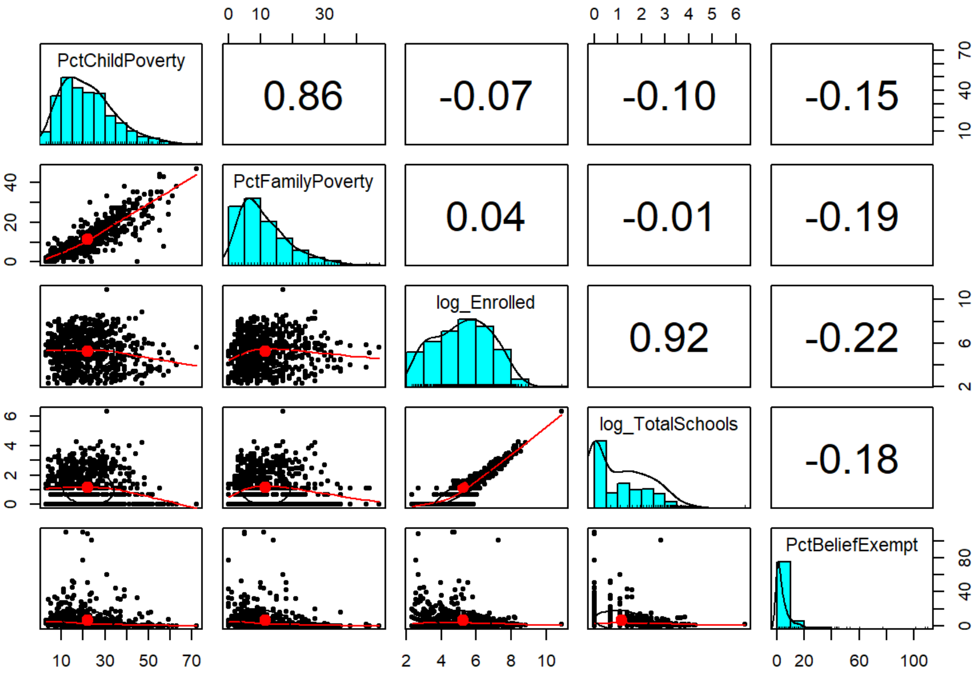
```
Districts <- subset(districts,select=c(PctChildPoverty, PctFamilyPoverty, log_Enrolled,log_TotalSchools, Pct
BeliefExempt ))
```

```
require(tidyverse)
Districts %>% pivot_longer(-PctBeliefExempt, names_to="variable", values_to="value", values_drop_na = TRUE)
%>%
            ggplot(aes(x=value, y=PctBeliefExempt)) + geom_point() +
                geom_smooth(method = "lm") + facet_wrap( ~ variable, scales="free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```r
library(psych)
pairs.panels(Districts)
```



From the above graphical interpretation and correlation values we can say that the schools and enrolled values are highly correlated.

```r
Districts <- subset(districts,select=c(PctChildPoverty, PctFamilyPoverty, log_Enrolled,log_TotalSchools, Pct
BeliefExempt ))

x <- lm(PctBeliefExempt ~ log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty, data=Distric
ts)
summary(x)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ log_Enrolled + PctChildPoverty +
##     log_TotalSchools + PctFamilyPoverty, data = Districts)
##
## Residuals:
##     Min       1Q  Median      3Q     Max
## -14.954  -4.495  -1.903   1.065 103.882
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      20.61366    2.75642   7.478 2.27e-13 ***
## log_Enrolled     -2.42380    0.67351  -3.599 0.000342 ***
## PctChildPoverty  -0.01974    0.07011  -0.281 0.778419
## log_TotalSchools  1.25026    0.92636   1.350 0.177566
## PctFamilyPoverty -0.23829    0.10585  -2.251 0.024688 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 695 degrees of freedom
## Multiple R-squared:  0.08699,    Adjusted R-squared:  0.08174
## F-statistic: 16.56 on 4 and 695 DF,  p-value: 5.745e-13
```

```r
library(BayesFactor)
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey@gma
il.com).
##
## Type BFManual() to open the manual.
## ************
```

```
y <- lmBF( PctBeliefExempt ~ log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty, data=Dist
ricts, posterior=TRUE, iterations=10000)
summary(y)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                       Mean      SD  Naive SE Time-series SE
## mu                 6.21199 0.42054 0.0042054      0.0042054
## log_Enrolled      -2.35017 0.66398 0.0066398      0.0066398
## PctChildPoverty   -0.01905 0.06884 0.0006884      0.0007004
## log_TotalSchools   1.21809 0.90807 0.0090807      0.0092277
## PctFamilyPoverty  -0.23078 0.10450 0.0010450      0.0010719
## sig2             125.31261 6.71267 0.0671267      0.0683035
## g                  0.07210 0.09232 0.0009232      0.0009232
##
## 2. Quantiles for each variable:
##
##                       2.5%       25%       50%       75%     97.5%
## mu                 5.37759   5.92951   6.21603   6.49295   7.03783
## log_Enrolled      -3.65476  -2.78751  -2.35452  -1.90456  -1.01999
## PctChildPoverty   -0.15378  -0.06564  -0.01879   0.02778   0.11502
## log_TotalSchools  -0.57857   0.60511   1.22177   1.82493   3.00868
## PctFamilyPoverty  -0.43470  -0.30125  -0.23060  -0.16043  -0.02761
## sig2             113.03839 120.70970 124.98860 129.64713 139.22842
## g                  0.01627   0.03274   0.04990   0.08163   0.26026
```

```
library(BayesFactor)
z <- lmBF( PctBeliefExempt ~ log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty, data=Dist
ricts)
z
```

```
## Bayes factor analysis
## --------------
## [1] log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty : 6.118e+09 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

Interpretation:-

I performed the linear regression to predict the percentage of the belief exempt from log_enrolled students, log_total schools, percentage of child poverty and percentage of family poverty.

Before performing the regression the violin plot showed the skewness in the variables. Total schools and enrolled where highly skewed which I confirmed using the histogram, outliers diagnosis plot and the numeric representation of the skewness.

To improve the skewness I used sqrt but it did not affect the variables. The resultant was also skewed. Therefore to deal with the skewness the log function is performed on enrolled and total schools. This helped me improving not only the skewness but also the non-linearity.

A linear regression found strong support for the relationship ($F(4, 695)=16.56$, $p<0.001$, adjusted $R2 = 0.08174$). PctFamilyPoverty and log_Enrolled are statistically significant and the only variable which we can consider because of statistical significance. Rest of the variables are not statistically significant on basis of their p-value and hence we cannot consider them in interpretation of the Belief Exempt

A Bayesian regression also found overwhelming evidence in support of a model with percentage of family poverty and log_Enrolled.The sampled coefficients had similar values, a mean of -2.34740 for log_Enrolled with an HDI of -0.59572 (lower bound) to -1.05459(upper bound),The mean of -0.02553 for log_ Family Poverty with an HDI of -0.43820 (lower bound) to 2.97586(upper bound).

The bayes factor gives us the odds ratio of 6.118e+09 : 1 which gives us the very strong evidence in the favour of alternative hypothesis that means log_Enrolled and Percentage of family poverty will predict the percentage of belief exemptions in the population data and it is rejecting the intercept only model.

6. *Which of the four predictor variables predicts the percentage of all enrolled students with completely up-to-date vaccines?*

Taking a proper visualization of the dataset and understanding it.

```
library(psych)
library(dlookr)
library(mice)
library(tidyverse)
describe(districts)
```

```
summary(districts)
```

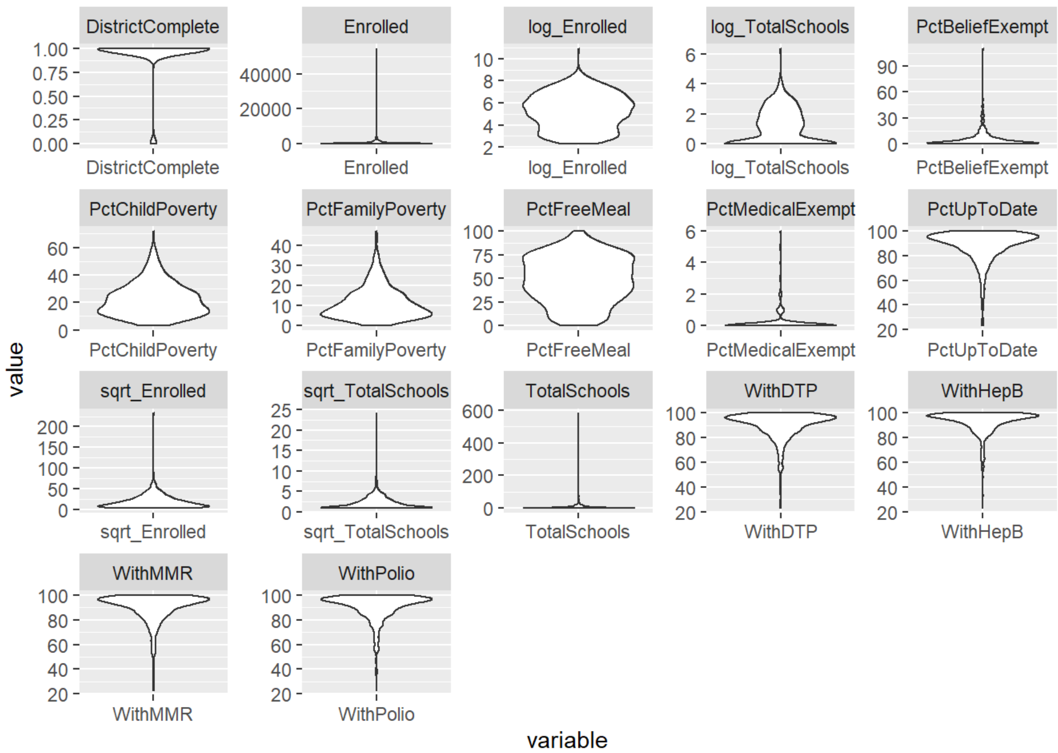
```
##                            DistrictName    WithDTP          WithPolio
##  ABC Unified                     :  1   Min.   : 23.00   Min.   : 23.00
##  Acton-Agua Dulce Unified        :  1   1st Qu.: 86.00   1st Qu.: 87.00
##  Adelanto Elementary             :  1   Median : 93.00   Median : 94.00
##  Alameda Unified                 :  1   Mean   : 89.85   Mean   : 90.21
##  Albany City Unified             :  1   3rd Qu.: 97.00   3rd Qu.: 97.00
##  Alexander Valley Union Elementary:  1   Max.   :100.00   Max.   :100.00
##  (Other)                         :694
##     WithMMR          WithHepB        PctUpToDate     DistrictComplete
##  Min.   : 23.00   Min.   : 23.00   Min.   : 23.00   Mode :logical
##  1st Qu.: 86.00   1st Qu.: 90.00   1st Qu.: 84.00   FALSE:37
##  Median : 94.00   Median : 96.00   Median : 92.00   TRUE :663
##  Mean   : 89.81   Mean   : 92.23   Mean   : 87.94
##  3rd Qu.: 97.00   3rd Qu.: 98.00   3rd Qu.: 96.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
##
##  PctBeliefExempt   PctMedicalExempt PctChildPoverty PctFamilyPoverty
##  Min.   :  0.000   Min.   :0.0000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:  1.000   1st Qu.:0.0000   1st Qu.:13.00   1st Qu.: 5.00
##  Median :  2.000   Median :0.0000   Median :20.00   Median : 9.00
##  Mean   :  6.206   Mean   :0.1532   Mean   :22.18   Mean   :11.32
##  3rd Qu.:  7.000   3rd Qu.:0.0000   3rd Qu.:29.00   3rd Qu.:15.25
##  Max.   :110.000   Max.   :6.0000   Max.   :72.00   Max.   :47.00
##                    NA's   :243
##   PctFreeMeal        Enrolled         TotalSchools     sqrt_TotalSchools
##  Min.   :  0.00   Min.   :   10.0   Min.   :  1.000   Min.   : 1.000
##  1st Qu.: 30.00   1st Qu.:   49.0   1st Qu.:  1.000   1st Qu.: 1.000
##  Median : 50.00   Median :  192.5   Median :  3.000   Median : 1.732
##  Mean   : 48.42   Mean   :  616.9   Mean   :  7.089   Mean   : 2.130
##  3rd Qu.: 69.00   3rd Qu.:  670.0   3rd Qu.:  8.000   3rd Qu.: 2.828
##  Max.   :100.00   Max.   :54238.0   Max.   :582.000   Max.   :24.125
##
##  sqrt_Enrolled     log_TotalSchools  log_Enrolled
##  Min.   :  3.162   Min.   :0.000    Min.   : 2.303
##  1st Qu.:  7.000   1st Qu.:0.000    1st Qu.: 3.892
##  Median : 13.874   Median :1.099    Median : 5.260
##  Mean   : 18.706   Mean   :1.143    Mean   : 5.240
##  3rd Qu.: 25.884   3rd Qu.:2.079    3rd Qu.: 6.507
##  Max.   :232.891   Max.   :6.366    Max.   :10.901
##
```

```
diagnose(districts)
```

```
md.pattern(districts, plot=FALSE)
```

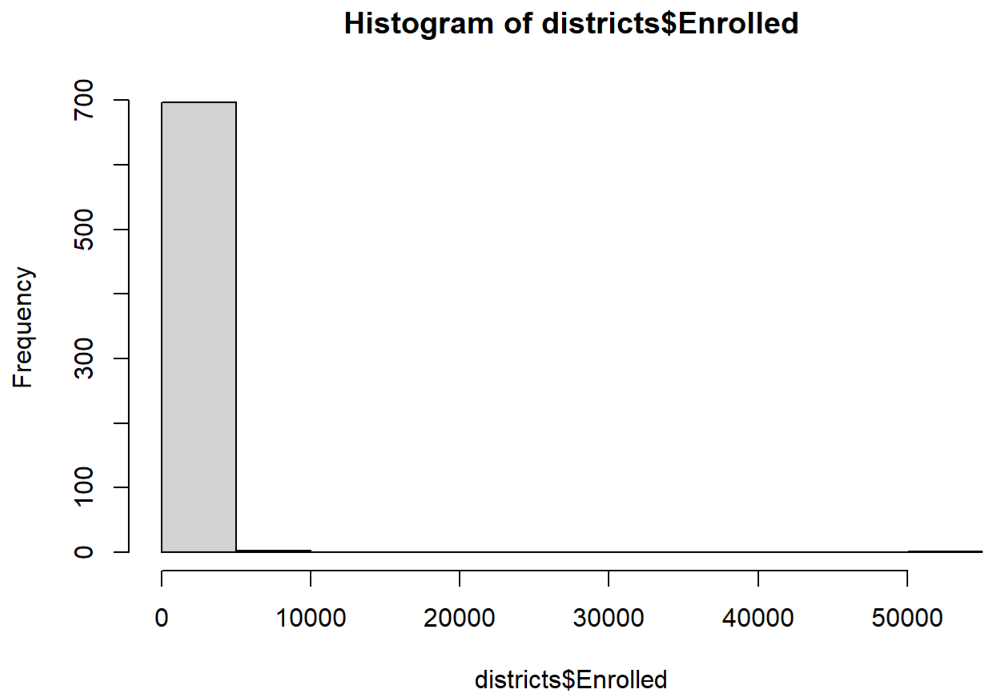
```
##     DistrictName WithDTP WithPolio WithMMR WithHepB PctUpToDate
## 457            1       1         1       1        1           1
## 243            1       1         1       1        1           1
##                0       0         0       0        0           0
##     DistrictComplete PctBeliefExempt PctChildPoverty PctFamilyPoverty
## 457                1               1               1                1
## 243                1               1               1                1
##                    0               0               0                0
##     PctFreeMeal Enrolled TotalSchools sqrt_TotalSchools sqrt_Enrolled
## 457           1        1            1                 1             1
## 243           1        1            1                 1             1
##               0        0            0                 0             0
##     log_TotalSchools log_Enrolled PctMedicalExempt
## 457                1            1                1   0
## 243                1            1                0   1
##                    0            0              243 243
```

```
districts %>% pivot_longer(cols=-DistrictName, names_to="variable",
                                values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) + geom_violin() + facet_wrap( ~ variable, scales="free")
```

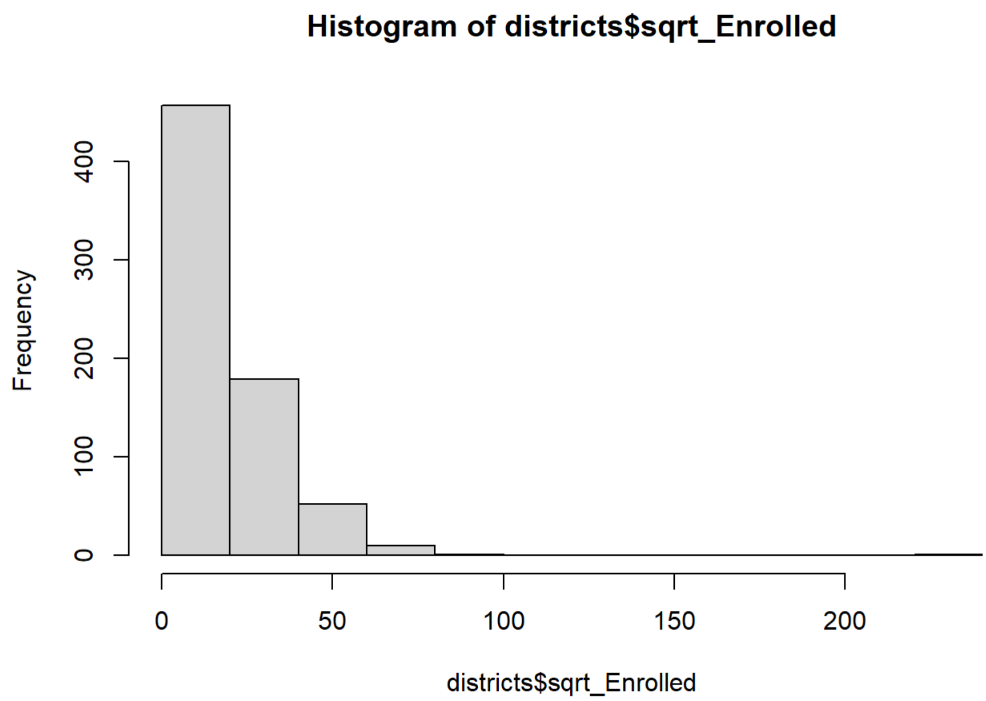


To improve the skewness using squareroot of the value

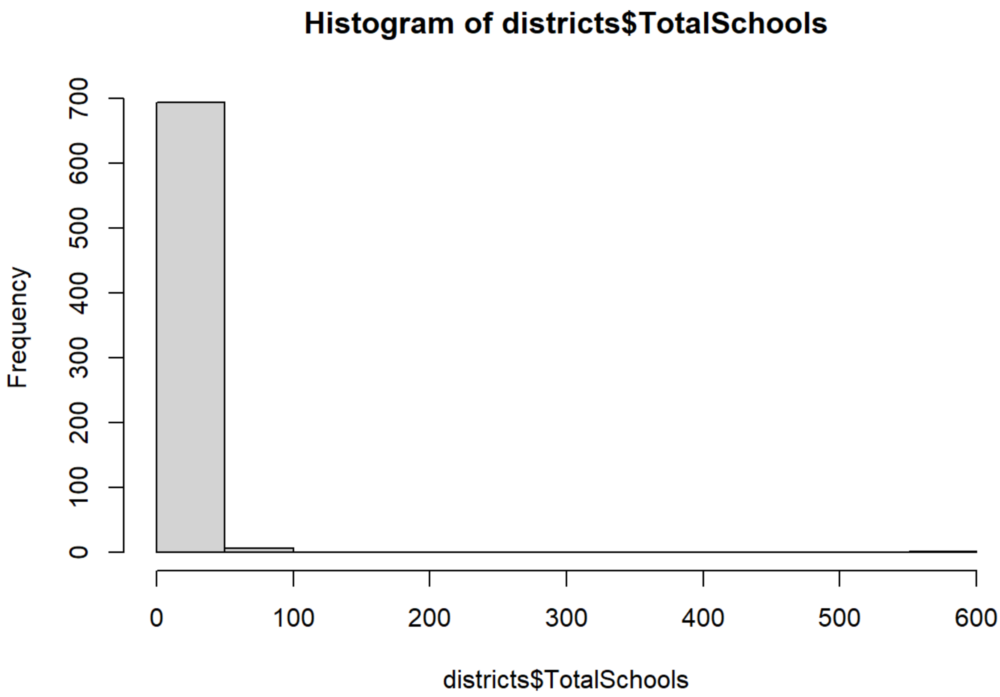
```
districts$sqrt_TotalSchools <- sqrt(districts$TotalSchools)
districts$sqrt_Enrolled <- sqrt(districts$Enrolled)
hist(districts$Enrolled)
```



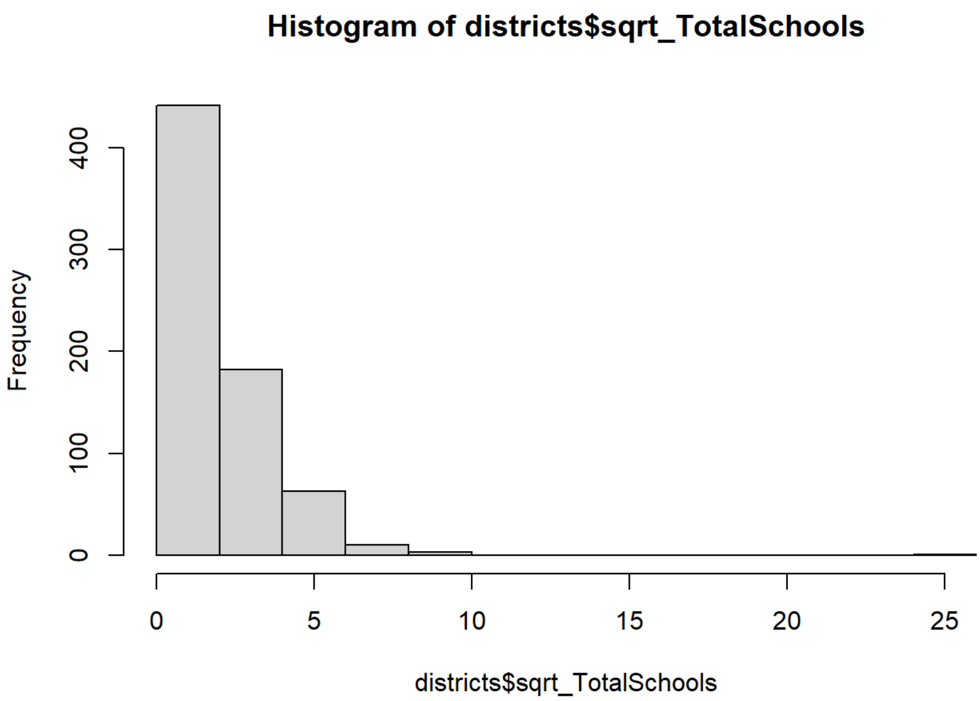

```
hist(districts$sqrt_Enrolled)
```




```
hist(districts$TotalSchools)
```



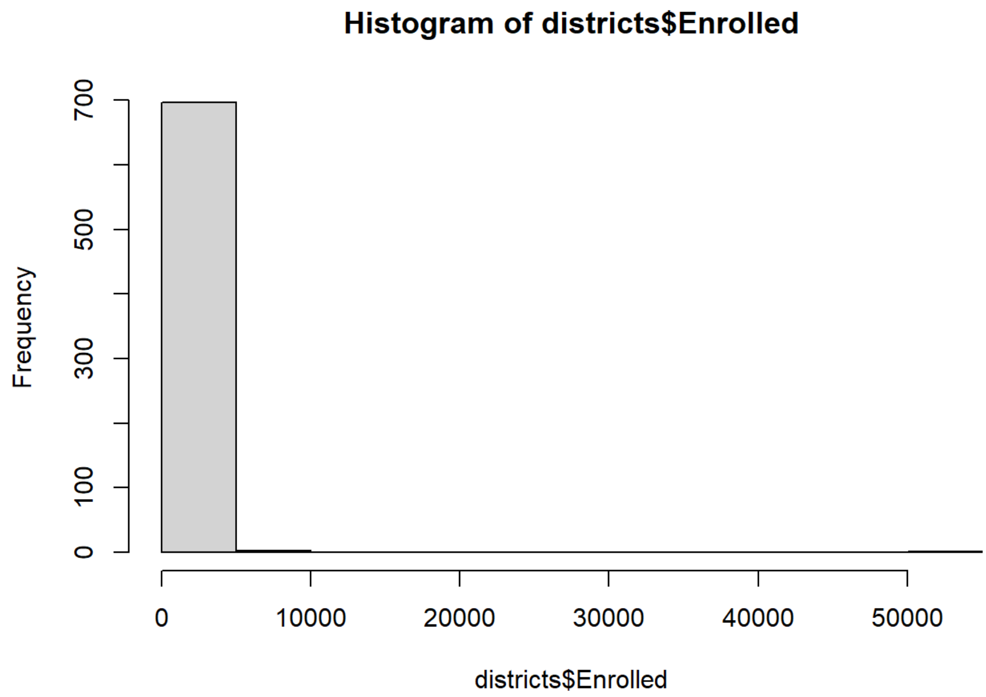

```
hist(districts$sqrt_TotalSchools)
```



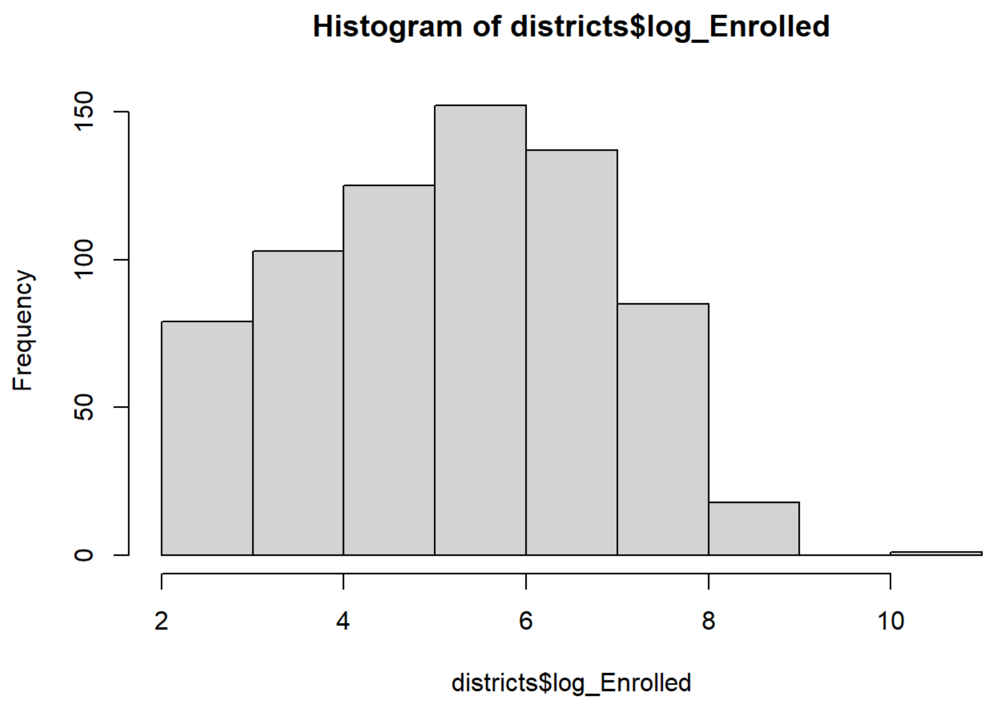
By operating and applying the sqrt we can see a noticable change in the graph but it is not a considerable because the after the operation the resultant shows the skewness as well..

Trying the log to remove skewness.

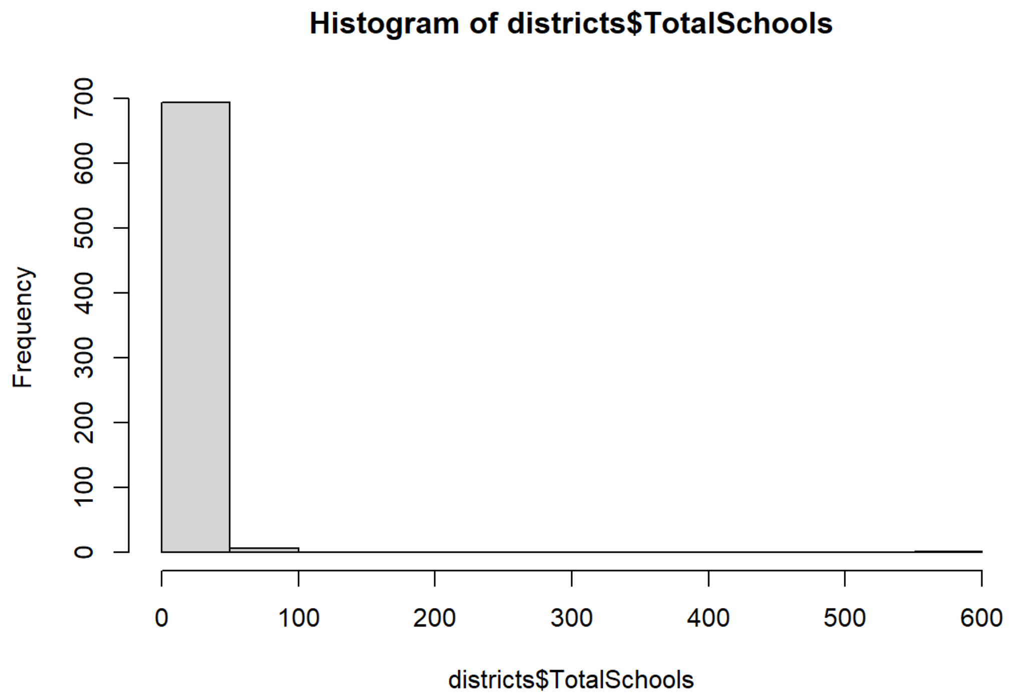
```
districts$log_TotalSchools <- log(districts$TotalSchools)
districts$log_Enrolled <- log(districts$Enrolled)
hist(districts$Enrolled)
```



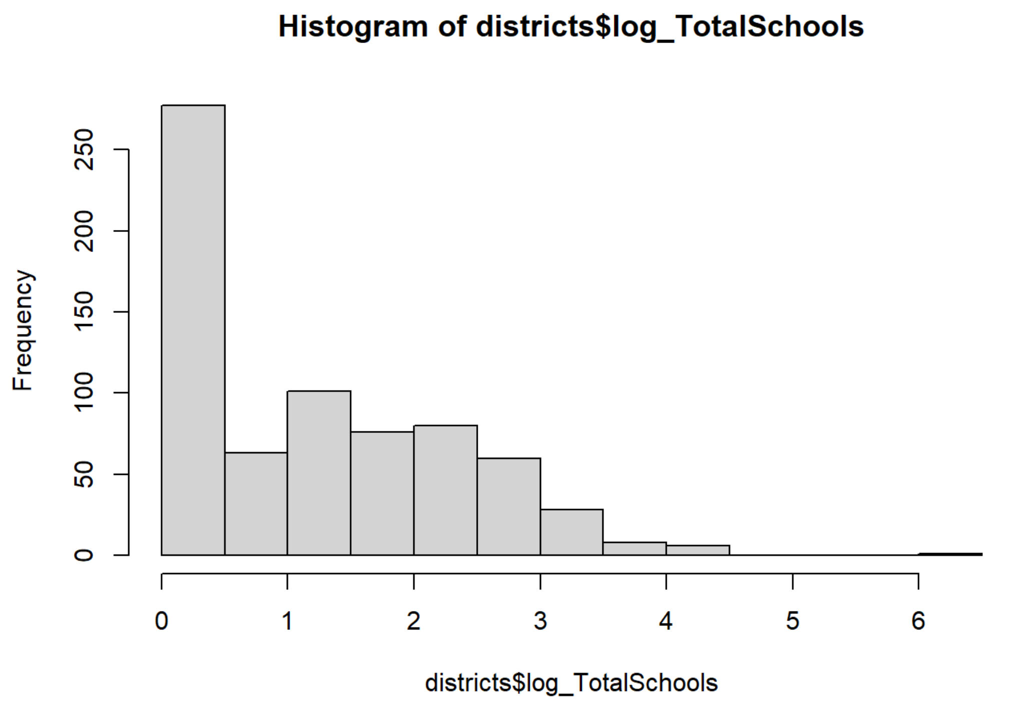

```
hist(districts$log_Enrolled)
```




```
hist(districts$TotalSchools)
```




```
hist(districts$log_TotalSchools)
```



In comparing the sqrt and log we can see that the skewness is removed more using log than that of sqrt. Checking how bad the outlier is by taking into consideration the numeric skewness.

```
library(e1071)
```

```
##
## Attaching package: 'e1071'
```

```
## The following objects are masked from 'package:dlookr':
##
##     kurtosis, skewness
```

```
with(districts, apply(cbind(PctChildPoverty, PctFamilyPoverty, Enrolled, TotalSchools), 2, skewness))
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##        0.8312046        1.2405355       21.0244326       20.8385423
```

```
with(districts, apply(cbind(PctChildPoverty, PctFamilyPoverty, districts$log_Enrolled, districts$log_TotalSc
hools ), 2, skewness))
```

```
##  PctChildPoverty PctFamilyPoverty
##       0.83120462       1.24053545       0.00288596       0.64657812
```

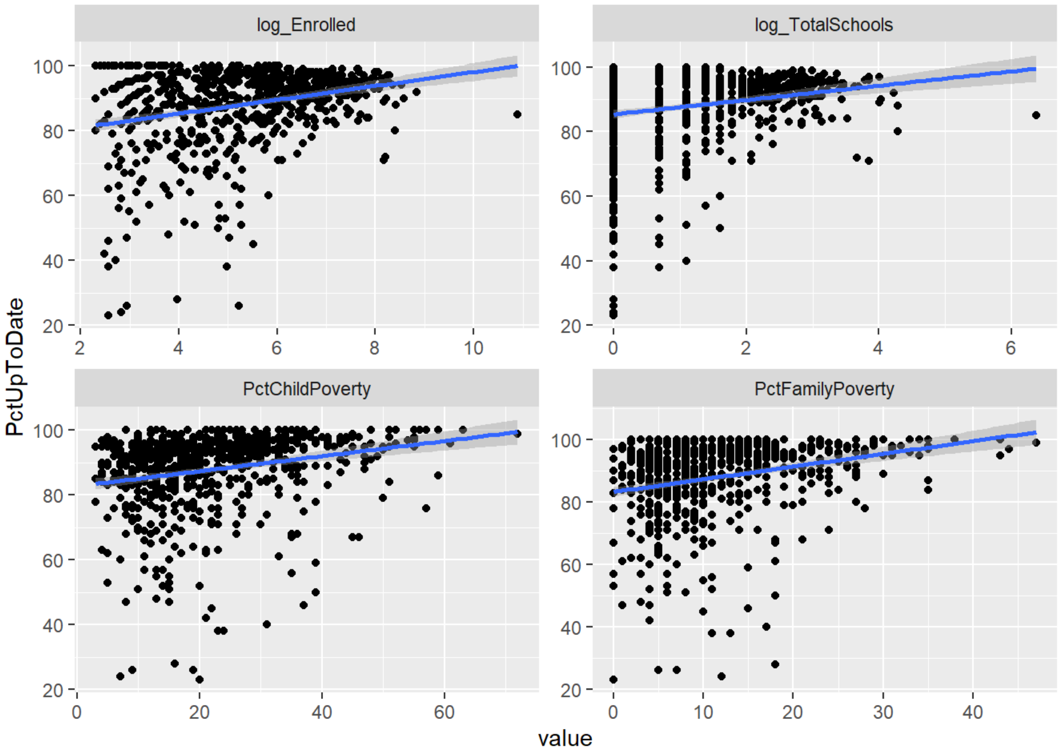
```
library(visdat)
vis_miss(districts)
```

Here Percentage in Medical Exempt shows that there are 34.71 percentage of missing values. For our analysis since aprroximately 35% of the data is missing. I will be excluding th column.

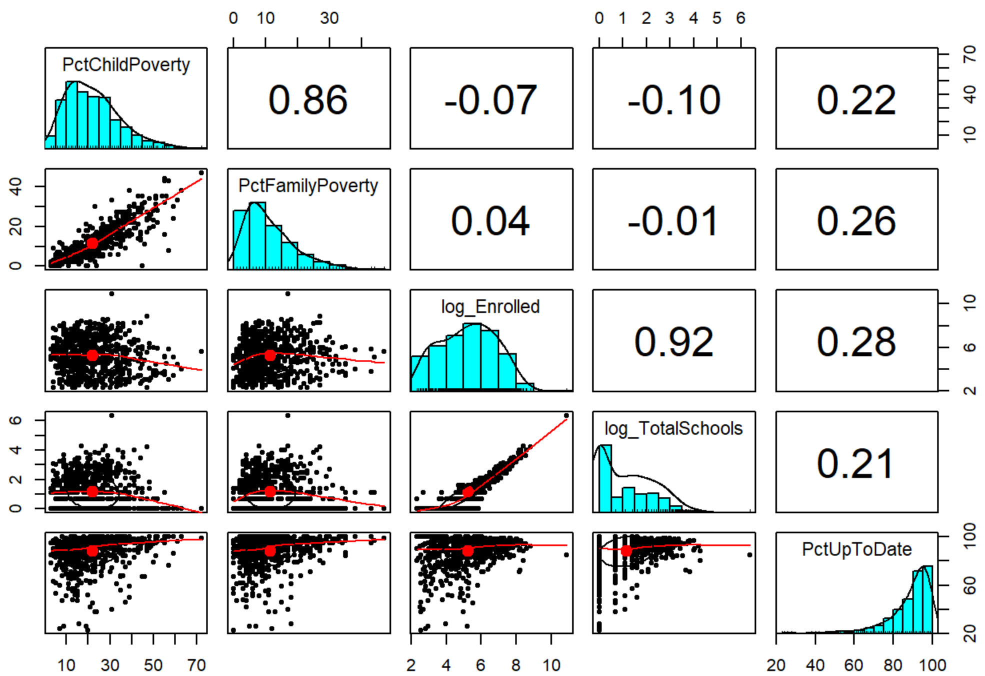
```
Districts1 <- subset(districts,select=c(PctChildPoverty, PctFamilyPoverty, log_Enrolled,log_TotalSchools, Pc
tUpToDate ))
View(Districts1)
```

```
require(tidyverse)
Districts1 %>% pivot_longer(-PctUpToDate, names_to="variable", values_to="value", values_drop_na = TRUE) %>%

             ggplot(aes(x=value, y=PctUpToDate)) + geom_point() +
                  geom_smooth(method = "lm") + facet_wrap( ~ variable, scales="free")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
library(psych)
pairs.panels(Districts1)
```



From the above graphical interpretation and correlation values we can say that the schools and enrolled values are highly correlated.

```r
View(Districts)
x <- lm(PctUpToDate ~ log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty, data=Districts1)
summary(x)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ log_Enrolled + PctChildPoverty + log_TotalSchools +
##     PctFamilyPoverty, data = Districts1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.642  -4.140   2.094   6.439  23.680
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      65.78290    2.82544  23.282  < 2e-16 ***
## log_Enrolled      3.77261    0.69038   5.465 6.47e-08 ***
## PctChildPoverty   0.12784    0.07187   1.779   0.0757 .
## log_TotalSchools -2.43935    0.94955  -2.569   0.0104 *
## PctFamilyPoverty  0.20701    0.10850   1.908   0.0568 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.47 on 695 degrees of freedom
## Multiple R-squared:  0.1494, Adjusted R-squared:  0.1445
## F-statistic: 30.51 on 4 and 695 DF,  p-value: < 2.2e-16
```

```
library(BayesFactor)
y <- lmBF( PctUpToDate ~ log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty, data= Distric
ts1, posterior=TRUE, iterations=10000)
summary(y)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                       Mean     SD Naive SE Time-series SE
## mu                87.95086 0.4315 0.004315       0.004315
## log_Enrolled       3.68272 0.6749 0.006749       0.006749
## PctChildPoverty    0.12478 0.0712 0.000712       0.000712
## log_TotalSchools  -2.37618 0.9303 0.009303       0.009303
## PctFamilyPoverty   0.20200 0.1072 0.001072       0.001155
## sig2             131.70481 7.1991 0.071991       0.071991
## g                  0.09877 0.1067 0.001067       0.001067
##
## 2. Quantiles for each variable:
##
##                        2.5%       25%       50%      75%    97.5%
## mu                87.107451  87.66372  87.95523  88.2416  88.7960
## log_Enrolled       2.352687   3.23271   3.68638   4.1321   5.0171
## PctChildPoverty   -0.014215   0.07727   0.12491   0.1724   0.2640
## log_TotalSchools  -4.173741  -2.99103  -2.37552  -1.7537  -0.5470
## PctFamilyPoverty  -0.008362   0.12808   0.20119   0.2751   0.4106
## sig2             118.379389 126.83030 131.53026 136.4212 146.1962
## g                  0.022770   0.04462   0.06872   0.1120   0.3659
```

```
library(BayesFactor)
z <- lmBF( PctUpToDate ~ log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty, data=District
s1)
summary(z)
```

```
## Bayes factor analysis
## --------------
## [1] log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty : 1.584279e+20 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```
z
```


```
## Bayes factor analysis
## --------------
## [1] log_Enrolled + PctChildPoverty + log_TotalSchools + PctFamilyPoverty : 1.584279e+20 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```


Interpretation:-

I performed the linear regression to predict the percentage of the up_to_date from total enrolled students, total schools, percentage of child poverty and percentage of family poverty.

Before performing the regression the violin plot showed the skewness in the variables. Total schools and enrolled where highly skewed which I confirmed using the histogram, outliers diagnosis plot and the numeric representation of the skewness.

To improve the skewness I used sqrt but it did not affect the variables. The resultant was also skewed. Therefore to deal with the skewness the log function is performed on enrolled and total schools. This helped me improving not only the skewness but also the non-linearity.

A linear regression found strong support for the relationship ($F(4, 695)=30.51$, $p<0.001$, adjusted $R2 = 0.1445$).log_Enrolled, log_Total Schools and percentage of family poverty are the only variables which are statistically significant and the only variables which we can consider because of statistical significance. Percentage of the child poverty is not statistically significant on basis of its p-value and hence we cannot consider it for the interpretation of the Percentage of up to date vaccine

A Bayesian regression also found overwhelming evidence in support of a model with log_Enrolled, log_Total Schools and percentage off family poverty. The sampled coefficients had similar values, a mean of 3.68849 for log_Enrolled with an HDI of 2.34731 (lower bound) to 5.0279(upper bound),The mean of -2.38878 for log_ TotalSchools with an HDI of -4.20420(lower bound) to 0.4185(upper bound), The mean of 0.20346 for percentage of family poverty with an HDI of -0.00563(lower bound) to -0.5639(upper bound)

The odds ratio is 1.584279e+20 : 1 which is strongly in the favour of alternate of hypothesis that means the log_Total school, log_Enrolled, and percentage od family poverty will predict the up_to_date vaccine and it rejects the null hypothesis or the intercept only model.

7. *Using any set of predictors that you want to use, what’s the best R-squared you can achieve in predicting the percentage of all enrolled students with completely up-to-date vaccines while still having an acceptable regression?*

We can use the step-wise regression to see what predictors are giving the best results, another approach is to use the correlation matrix and check which predictors are highly correlated and use one of them. We have to predict the up_to_date vaccines.

```
Districts_whole <- subset(districts,select=c(WithDTP,WithPolio,WithMMR,WithHepB,PctUpToDate,PctBeliefExempt,
PctChildPoverty, PctFamilyPoverty,PctFreeMeal,Enrolled,TotalSchools))
cor(Districts_whole)
```

```
##                    WithDTP   WithPolio     WithMMR    WithHepB  PctUpToDate
## WithDTP          1.00000000  0.98632671  0.97752159  0.89919324  0.95932288
## WithPolio        0.98632671  1.00000000  0.97092094  0.90670803  0.94997085
## WithMMR          0.97752159  0.97092094  1.00000000  0.89750751  0.96880917
## WithHepB         0.89919324  0.90670803  0.89750751  1.00000000  0.85552830
## PctUpToDate      0.95932288  0.94997085  0.96880917  0.85552830  1.00000000
## PctBeliefExempt -0.59406414 -0.60679766 -0.58291317 -0.68760688 -0.53146917
## PctChildPoverty  0.21339233  0.20686287  0.21313036  0.21865882  0.22403712
## PctFamilyPoverty 0.25612734  0.25023209  0.25545838  0.26852729  0.26085175
## PctFreeMeal      0.27107949  0.27728238  0.27514857  0.32212151  0.27222896
## Enrolled         0.07057304  0.07467829  0.07639021  0.08320725  0.06226121
## TotalSchools     0.05668883  0.06093742  0.06335096  0.07046193  0.04861402
##                 PctBeliefExempt PctChildPoverty PctFamilyPoverty PctFreeMeal
## WithDTP              -0.59406414      0.21339233       0.25612734  0.27107949
## WithPolio            -0.60679766      0.20686287       0.25023209  0.27728238
## WithMMR              -0.58291317      0.21313036       0.25545838  0.27514857
## WithHepB             -0.68760688      0.21865882       0.26852729  0.32212151
## PctUpToDate          -0.53146917      0.22403712       0.26085175  0.27222896
## PctBeliefExempt       1.00000000     -0.14726824      -0.19464030 -0.23314931
## PctChildPoverty      -0.14726824      1.00000000       0.85597722  0.73848178
## PctFamilyPoverty     -0.19464030      0.85597722       1.00000000  0.71042038
## PctFreeMeal          -0.23314931      0.73848178       0.71042038  1.00000000
## Enrolled             -0.07315774      0.02627437       0.04740043  0.06614970
## TotalSchools         -0.06584042      0.02188522       0.04093328  0.06101766
##                     Enrolled TotalSchools
## WithDTP          0.07057304   0.05668883
## WithPolio        0.07467829   0.06093742
## WithMMR          0.07639021   0.06335096
## WithHepB         0.08320725   0.07046193
## PctUpToDate      0.06226121   0.04861402
## PctBeliefExempt -0.07315774  -0.06584042
## PctChildPoverty  0.02627437   0.02188522
## PctFamilyPoverty 0.04740043   0.04093328
## PctFreeMeal      0.06614970   0.06101766
## Enrolled         1.00000000   0.99421966
## TotalSchools     0.99421966   1.00000000
```

```
View(Districts_whole)
```

DTP is highly correlated to Polio

Polio is highly correlated to MMR

MMR is highly correlated to PctUp_to_date

Pct uptodate is highly correlated to With Pct_Belief Exempt

PctBelief Exempt is highly correlated with PctFreeMeal

PctChild Poverty is highly correlated to PctFamilypoverty

Pct Family poverty is highly correlated to pct free meal

Pct free meal is strongly correlated to enrolled students

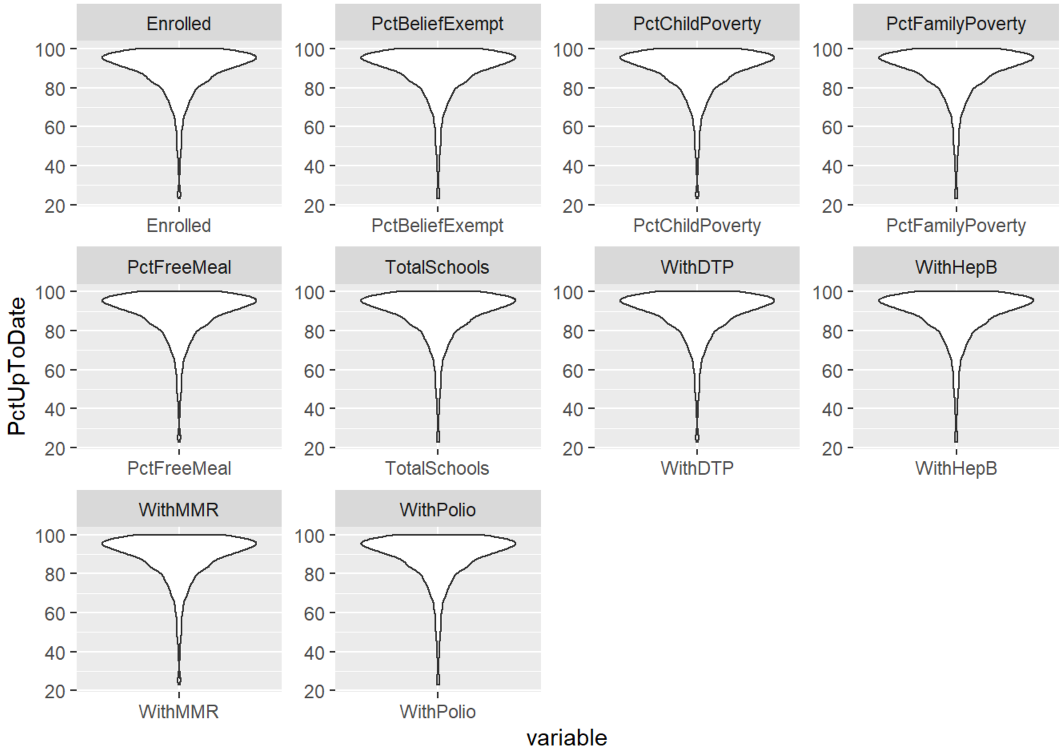
Enrolled students is highly correlated to Total schools

```
library(psych)
library(dlookr)
library(mice)
library(tidyverse)

md.pattern(Districts_whole, plot=FALSE)
```

```
##   /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```

```
##     WithDTP WithPolio WithMMR WithHepB PctUpToDate PctBeliefExempt
## 700       1         1       1        1           1               1
##           0         0       0        0           0               0
##     PctChildPoverty PctFamilyPoverty PctFreeMeal Enrolled TotalSchools
## 700               1                1           1        1            1 0
##                   0                0           0        0            0 0
```

```
Districts_whole %>% pivot_longer( cols = -PctUpToDate  , names_to="variable",
                        values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=PctUpToDate)) + geom_violin() + facet_wrap( ~ variable, scales="free")
```



As we can see that all the predictors are skewed but not so highly skewed that we need to improve the skewness of each predictor by using the operations. We will perform the analysis on the same predictors as it is.

```
lm.outwhole <- lm(PctUpToDate  ~ WithDTP + WithPolio+ Enrolled + WithMMR + WithHepB + PctBeliefExempt + PctC
hildPoverty + PctFamilyPoverty + PctFreeMeal + TotalSchools, data = Districts_whole)
summary(lm.outwhole)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithDTP + WithPolio + Enrolled + WithMMR +
##     WithHepB + PctBeliefExempt + PctChildPoverty + PctFamilyPoverty +
##     PctFreeMeal + TotalSchools, data = Districts_whole)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.005  -0.451   0.521   1.244  12.150
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.020e+01  1.524e+00  -6.691 4.58e-11 ***
## WithDTP           3.280e-01  7.092e-02   4.625 4.46e-06 ***
## WithPolio         6.126e-02  6.438e-02   0.952 0.341676
## Enrolled          2.993e-04  4.720e-04   0.634 0.526182
## WithMMR           7.971e-01  4.797e-02  16.618  < 2e-16 ***
## WithHepB         -9.940e-02  3.008e-02  -3.305 0.001000 **
## PctBeliefExempt   4.791e-02  1.309e-02   3.658 0.000273 ***
## PctChildPoverty   1.939e-02  1.923e-02   1.008 0.313658
## PctFamilyPoverty  2.245e-03  2.774e-02   0.081 0.935515
## PctFreeMeal       2.114e-04  7.063e-03   0.030 0.976128
## TotalSchools     -3.240e-02  4.382e-02  -0.739 0.459884
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 689 degrees of freedom
## Multiple R-squared:  0.9458, Adjusted R-squared:  0.945
## F-statistic:  1202 on 10 and 689 DF,  p-value: < 2.2e-16
```

```
lm.out <- lm( PctUpToDate  ~ WithDTP + WithMMR + WithHepB +  PctBeliefExempt , data = Districts_whole)
summary(lm.out)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithDTP + WithMMR + WithHepB + PctBeliefExempt,
##     data = Districts_whole)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.271  -0.436   0.517   1.230  11.970
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -10.31376    1.50396  -6.858 1.55e-11 ***
## WithDTP           0.38054    0.04906   7.756 3.12e-14 ***
## WithMMR           0.80288    0.04740  16.937  < 2e-16 ***
## WithHepB         -0.09040    0.02909  -3.108 0.001959 **
## PctBeliefExempt   0.04764    0.01307   3.645 0.000288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.911 on 695 degrees of freedom
## Multiple R-squared:  0.9452, Adjusted R-squared:  0.9449
## F-statistic:  2998 on 4 and 695 DF,  p-value: < 2.2e-16
```

```
lm.out1 <- lm(PctUpToDate ~ WithDTP + WithMMR + WithHepB +  PctBeliefExempt + PctChildPoverty ,  data = Dist
ricts_whole)
summary(lm.out1)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ WithDTP + WithMMR + WithHepB + PctBeliefExempt +
##     PctChildPoverty, data = Districts_whole)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.036  -0.430   0.524   1.230  11.863
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -10.272800   1.500019  -6.848 1.65e-11 ***
## WithDTP           0.379486   0.048933   7.755 3.15e-14 ***
## WithMMR           0.801890   0.047277  16.962  < 2e-16 ***
## WithHepB         -0.093802   0.029049  -3.229 0.001300 **
## PctBeliefExempt   0.047620   0.013034   3.653 0.000278 ***
## PctChildPoverty   0.020575   0.009421   2.184 0.029301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.903 on 694 degrees of freedom
## Multiple R-squared:  0.9456, Adjusted R-squared:  0.9452
## F-statistic:  2412 on 5 and 694 DF,  p-value: < 2.2e-16
```

We can see that the following predictors are not producing significant results on basis of their p-value and hence we dont need them for our predictions. They are as follows

WithPolio,PctFreeMeals,Enrolled,TotalSchools, PctChildPoverty, PctFamilyPoverty

Removing all of this from our model

So our first model generates the adjusted Rsquared of 0.949 with an inclusion of all the predictors some of them are giving the significant .

The second model generates the adjusted Rsquared of 94.49 with all of the significant predictors.

The third model generates the adjusted Rsquared of 0.9452 with all of the significant predictors which is predicting the percentage up_to_date and they are WithDTP, WithMMR, WithHepB, PctBeliefExempt and Pct Child Poverty.

So on basis of the results we can say that the best predictors which are predicting the Percentage of students with completely up-to-date vaccines are WithDTP, WithMMR, WithHepB, PctBeliefExempt,, PctChildPoverty, with F-statistic of 2412 on 5 and 694 Degrees of Freedom and a significant p-value of 2.2e-16. The Adjusted R squared is of 0.9452 i.e the model can predict upto the accuracy of 94.52%

```
library(BayesFactor)
lmBF.outwhole <- lmBF(PctUpToDate  ~ WithDTP + WithPolio+ Enrolled + WithMMR + WithHepB + PctBeliefExempt +
PctChildPoverty + PctFamilyPoverty + PctFreeMeal + TotalSchools, data = Districts_whole, posterior=TRUE, ite
rations=10000)
summary(lmBF.outwhole)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                         Mean       SD  Naive SE Time-series SE
## mu                87.9431407 0.113665 1.137e-03      1.105e-03
## WithDTP            0.3271516 0.079705 7.971e-04      7.971e-04
## WithPolio          0.0609679 0.076973 7.697e-04      7.808e-04
## Enrolled           0.0002939 0.000505 5.050e-06      4.887e-06
## WithMMR            0.7973142 0.048734 4.873e-04      4.937e-04
## WithHepB          -0.0994773 0.029968 2.997e-04      2.997e-04
## PctBeliefExempt    0.0478294 0.013081 1.308e-04      1.308e-04
## PctChildPoverty    0.0191859 0.019936 1.994e-04      2.026e-04
## PctFamilyPoverty   0.0021634 0.027724 2.772e-04      2.772e-04
## PctFreeMeal        0.0002664 0.007385 7.385e-05      7.385e-05
## TotalSchools      -0.0318570 0.047150 4.715e-04      4.567e-04
## sig2               8.5026307 1.133907 1.134e-02      1.157e-02
## g                  1.9137584 1.006123 1.006e-02      1.006e-02
##
## 2. Quantiles for each variable:
##
##                        2.5%        25%        50%        75%     97.5%
## mu                87.730862  8.787e+01 87.9439173 88.0170335 88.155036
## WithDTP            0.187913  2.786e-01  0.3269136  0.3735581  0.467439
## WithPolio         -0.066184  1.883e-02  0.0613307  0.1044499  0.188765
## Enrolled          -0.000623 -1.124e-05  0.0002903  0.0006144  0.001209
## WithMMR            0.703014  7.654e-01  0.7973416  0.8296883  0.892303
## WithHepB          -0.156973 -1.197e-01 -0.0995808 -0.0792510 -0.041232
## PctBeliefExempt    0.022243  3.914e-02  0.0478762  0.0565573  0.073125
## PctChildPoverty   -0.018469  5.944e-03  0.0192065  0.0322795  0.056577
## PctFamilyPoverty -0.052596 -1.633e-02  0.0022572  0.0203942  0.056778
## PctFreeMeal       -0.013558 -4.554e-03  0.0002528  0.0050215  0.014199
## TotalSchools      -0.116648 -6.172e-02 -0.0315303 -0.0036417  0.053262
## sig2               7.625039  8.180e+00  8.4814292  8.7852629  9.436731
## g                  0.779697  1.261e+00  1.6738744  2.2620982  4.518078
```

```
lmBF.out <- lmBF( PctUpToDate  ~ WithDTP + WithMMR + WithHepB +  PctBeliefExempt , data = Districts_whole, p
osterior=TRUE, iterations=10000)
summary(lmBF.out)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean      SD  Naive SE Time-series SE
## mu               87.94473 0.11372 0.0011372      0.0011149
## WithDTP           0.37915 0.05177 0.0005177      0.0005177
## WithMMR           0.80344 0.05544 0.0005544      0.0005544
## WithHepB         -0.08993 0.03270 0.0003270      0.0003332
## PctBeliefExempt   0.04762 0.01364 0.0001364      0.0001364
## sig2              8.52641 1.12358 0.0112358      0.0114137
## g                 5.63140 6.60825 0.0660825      0.0660825
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%      50%      75%    97.5%
## mu               87.72455 87.87028 87.94540 88.01972 88.16092
## WithDTP           0.28282  0.34569  0.37894  0.41253  0.47543
## WithMMR           0.70983  0.77167  0.80344  0.83623  0.89618
## WithHepB         -0.14726 -0.10979 -0.09000 -0.07037 -0.03308
## PctBeliefExempt   0.02187  0.03869  0.04755  0.05642  0.07368
## sig2              7.65732  8.18886  8.50529  8.82098  9.47630
## g                 1.31949  2.59463  3.98659  6.46827 19.64889
```

```r
lmBF.out1 <- lmBF(PctUpToDate ~ WithDTP + WithMMR + WithHepB +  PctBeliefExempt + PctChildPoverty ,  data =
Districts_whole, posterior=TRUE, iterations=10000)
summary(lmBF.out1)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean       SD  Naive SE Time-series SE
## mu               87.94437 0.142529 1.425e-03      0.0014253
## WithDTP           0.37837 0.053941 5.394e-04      0.0005555
## WithMMR           0.80211 0.055221 5.522e-04      0.0005715
## WithHepB         -0.09344 0.031321 3.132e-04      0.0003132
## PctBeliefExempt   0.04761 0.013744 1.374e-04      0.0001374
## PctChildPoverty   0.02054 0.009739 9.739e-05      0.0001011
## sig2              8.48330 1.508313 1.508e-02      0.0150831
## g                 4.27927 4.063999 4.064e-02      0.0406400
##
## 2. Quantiles for each variable:
##
##                       2.5%      25%      50%      75%    97.5%
## mu               87.728450 87.86786 87.94360 88.01846 88.15863
## WithDTP           0.283028  0.34446  0.37814  0.41235  0.47568
## WithMMR           0.706149  0.76955  0.80207  0.83421  0.89472
## WithHepB         -0.149917 -0.11270 -0.09326 -0.07435 -0.03529
## PctBeliefExempt   0.021624  0.03885  0.04758  0.05644  0.07404
## PctChildPoverty   0.002425  0.01414  0.02059  0.02696  0.03911
## sig2              7.615339  8.14760  8.45458  8.77122  9.41756
## g                 1.172289  2.18158  3.21123  4.95857 13.70147
```

```r
library(BayesFactor)

Bayesian_Result1 <- lmBF(PctUpToDate  ~ WithDTP + WithPolio+ Enrolled + WithMMR + WithHepB + PctBeliefExempt
+ PctChildPoverty + PctFamilyPoverty + PctFreeMeal + TotalSchools, data = Districts_whole)
Bayesian_Result2 <- lmBF( PctUpToDate  ~ WithDTP + WithMMR + WithHepB +  PctBeliefExempt , data =  Districts
_whole )
Bayesian_Result3 <- lmBF(PctUpToDate ~ WithDTP + WithMMR + WithHepB +  PctBeliefExempt + PctChildPoverty ,
data = Districts_whole)

Bayesian_Result1
```

```
## Bayes factor analysis
## --------------
## [1] WithDTP + WithPolio + Enrolled + WithMMR + WithHepB + PctBeliefExempt + PctChildPoverty + PctFamilyPo
verty + PctFreeMeal + TotalSchools : 7.673795e+423 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```r
Bayesian_Result2
```

```
## Bayes factor analysis
## --------------
## [1] WithDTP + WithMMR + WithHepB + PctBeliefExempt : 1.244568e+432 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

```r
Bayesian_Result3
```

```
## Bayes factor analysis
## --------------
## [1] WithDTP + WithMMR + WithHepB + PctBeliefExempt + PctChildPoverty : 2.587821e+431 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

#Bayesian Interpretation

A Bayesian regression also found overwhelming evidence in support of a model 3 which provides the accuracy upto 94.52 percentage or the adjusted Rsquared is 0.9452 in the linear regression model, Percentage of students with completely up-to-date vaccines, Percentage of students in the district with the MMR vaccine, Percentage of students in the district with Hepatitis B vaccine, Percentage of students in the district with Hepatitis B vaccine, Percentage of all enrolled students with medical exceptions, Percentage of children in district living below the poverty line are the excellent significant predictors. The sampled coefficients had similar values, a mean of 0.37935 for WithDTP with an HDI of 0.283431 (lower bound) to 0.47659(upper bound),The mean of 0.80186 for WithMMR with an HDI of 0.80186(lower bound) to 0.89568(upper bound), The mean of -0.09436 for WithHepB with an HDI of -0.151667(lower bound) to -0.03890(upper bound), The mean of 0.04754 forpercentage of belief exempt with an HDI of 0.021999(lower bound) to 0.07240(upper bound), The mean of 0.02085 forpercentage of child poverty with an HDI of 0.001827(lower bound) to 0.03901(upper bound)

The odds ratio is 2.587821e+431 ± 0% which is strongly in the favour of alternate of hypothesis that means Percentage of students in the district with the MMR vaccine, Percentage of students in the district with Hepatitis B vaccine, Percentage of students in the district with Hepatitis B vaccine, Percentage of all enrolled students with medical exceptions, Percentage of children in district living below the poverty line are perfectly predicting the perrcentage of enrolled students with up_to_date vaccine rejecting the null hypothesis or the intercept only model.

8. *In predicting the percentage of all enrolled students with completely up-to-date vaccines, is there an interaction between PctChildPoverty and Enrolled?*

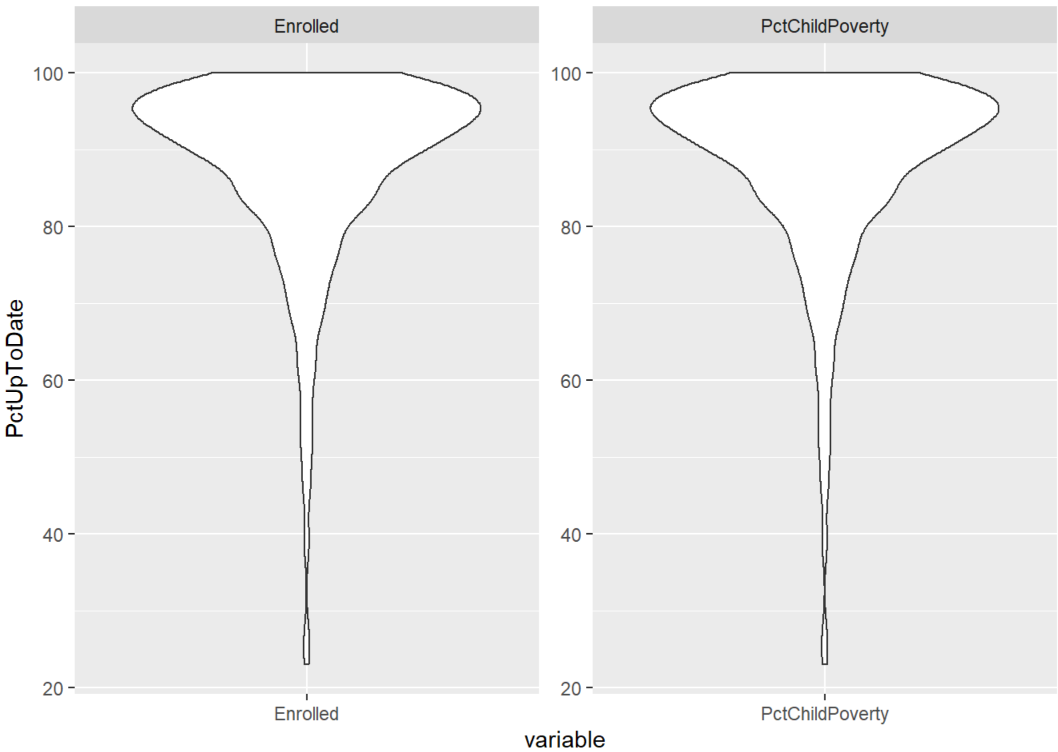
```
Interaction_Whole <- subset(districts,select=c(PctUpToDate,PctChildPoverty,Enrolled))
View(Interaction_Whole)
```

```
library(psych)
library(dlookr)
library(mice)
library(tidyverse)

md.pattern(Interaction_Whole, plot=FALSE)
```

```
##   /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```

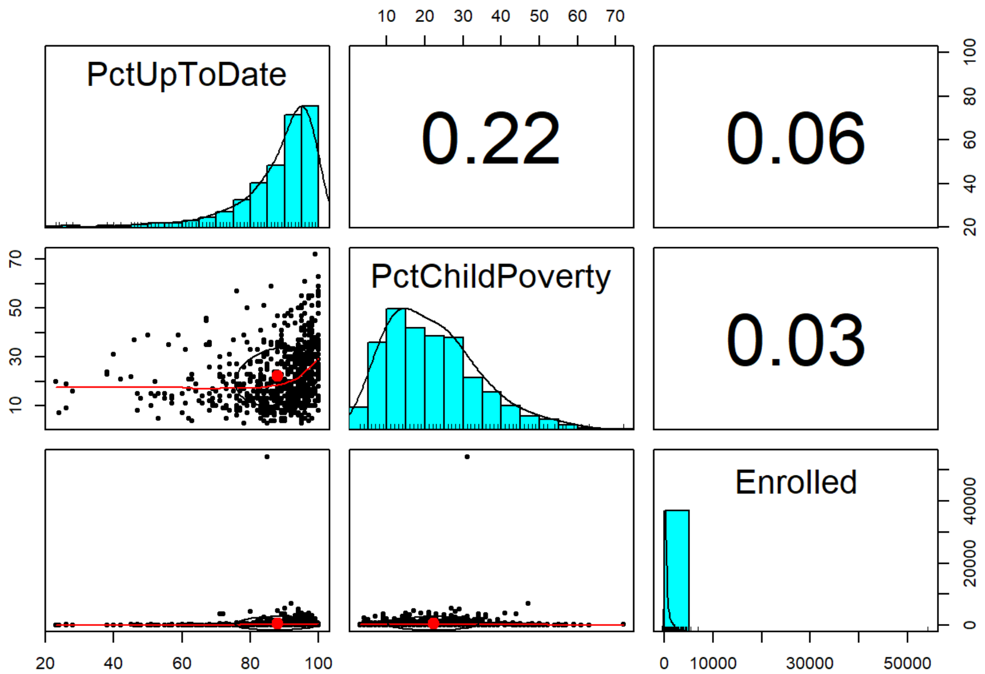
```
##     PctUpToDate PctChildPoverty Enrolled
## 700           1               1        1 0
##               0               0        0 0
```

```
Interaction_Whole %>% pivot_longer( cols = -PctUpToDate  , names_to="variable",
                        values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=PctUpToDate)) + geom_violin() + facet_wrap( ~ variable, scales="free")
```

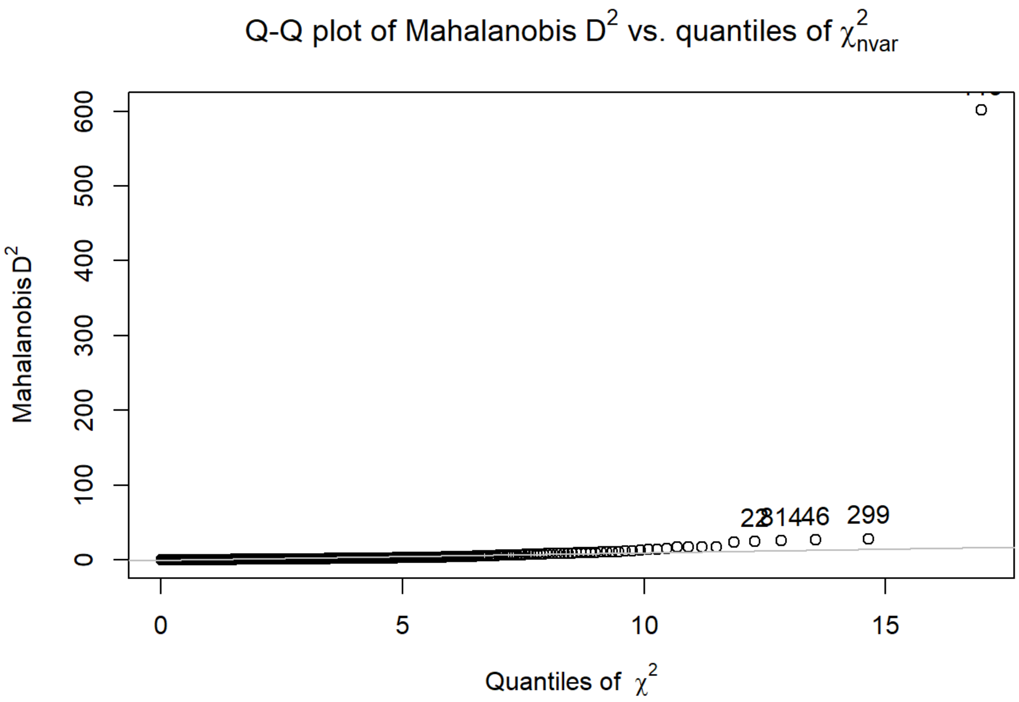



```
Centred_PctUpToDate <- scale(Interaction_Whole$PctUpToDate, center=T, scale= F)

Centred_PctChildPoverty <- scale(Interaction_Whole$PctChildPoverty, center=T, scale= F)

Centred_Enrolled <- scale(Interaction_Whole$Enrolled, center=T, scale= F)
```

```
library(psych)
pairs.panels(Interaction_Whole)
```




```
library(dlookr)
outlier(Interaction_Whole)
```




```
##          114          200          269          682          210           74
##   0.46186509   0.42338313   1.67438832   1.07850722   0.26756116   1.89964175
##          120          406          653          234          596          784
##   0.62005158   2.26751750   1.04729450   1.82532337   4.39478925   2.01413017
##          598          475           89           97          132           81
##   0.82696236   0.82206428   3.07109706   0.23759344   0.28108147   0.65267787
##          338          126          290           93          808           17
##   3.34304181   0.58052123   0.31375708   0.74209054  13.97234851   6.37551666
##          802          337          623          144          321          183
##   1.72151926   1.19220305   0.57534814   0.34107073   1.13193017   0.04772471
##          493          640          732          824          481          223
##   1.69005211   0.24938077   0.88914312   0.30562977   0.08181699   2.01074866
##          396           11          102          565          615          717
##   0.53192133   0.65794459   0.38274044   2.54801469   6.27561107   1.00211262
```

```
##          211          489          261          634          564          806 
##   9.00668922   0.21582102   1.12660336   0.93815793   1.48536762   0.33158592 
##          358          668          719          669          750          203 
##   5.04361391   1.25528609   0.31602483   8.40591361   1.15496858  12.59962872 
##          783          557          454          498          495          662 
##   0.52900554   1.84689524   1.19647868   2.41514236   1.06149887   0.19587482 
##          492          757          388          752          547          352 
##   3.09284525   1.44433940   1.90835495   0.26377642   2.59814054   0.81711595 
##          382          523          641          368          486          370 
##   0.19319773   0.64098915   0.61339100   0.15880853   0.72006758   3.91335327 
##          785          314          294          429           77          258 
##   0.60171024   3.43513592   7.64790843   0.94735563   1.18586741   1.88725665 
##          458          240          748          444          148          594 
##   1.71082680   0.72230452   0.70148806   0.37718197   0.27592079   0.49404118 
##          552          177          271          655          514           91 
##   5.10055939   4.63547294   1.40135744   0.91472850   1.06258465   0.34758301 
##           52          702          723          661          273          778 
##   0.85660020   1.40991749   0.34680995   0.60180651   1.10322785   0.65496608 
##          544          767          817           12          179          551 
##   1.25296801   1.89721758   9.39002444   0.39295970   1.07864287   5.36622701 
##          673          360           51          504          160          395 
##  10.91894949   1.27395216   0.09748337   0.51261757   1.57312235   0.28416301 
##          685          518          384          287          648          763 
##   1.02686511   1.17164485   0.71062481   3.21859928   0.29681587   0.08733569 
##          535          766          735          422          607          264 
##   2.04927489   0.48184547   0.33609821   2.52028640   1.50817379   0.56088772 
##          229          100          280            6          118          830 
##   8.62359313   2.24347330   2.05959544   1.82262125   1.07019786   0.93039272 
##           43          185          629          220          751           98 
##   0.51669549   0.21290974   0.20641036   1.02040526   0.58037516   1.65484553 
##          818          573          571          707          482          467 
##   0.77479700   1.21162477   2.27436187   3.75294218   0.52435477   1.23780619 
##          745          389          375          805          632           10 
##  10.32901481   3.12258768   8.63818054   0.69247675   0.25970045   0.19270661 
##          832          829           26          608          459          299 
##   1.63615738   0.46752236   3.21905260   2.85729912   2.25336579  28.45442726 
##          411          686          116           72          721          456 
##   3.64593721   2.40213134   2.63096508   2.30132418   0.33259803   1.62039016 
##          367          521           54          260          227          154 
##   2.60339438   1.02468358   1.58732060   1.72988943   3.77950751   0.25225675 
##          507          543          448          128          410          324 
##   0.14963482   0.62612532   2.17108107   0.05127674   1.35852015   0.86097961 
##          452          364          651          381          642          674 
##   1.67253420   4.81222135   0.35435626   2.69937619   1.02495423   0.41663293 
##          716           28          180           29          616           85 
##   0.73009625   0.89517210   0.27399640   0.78287922   1.67492374   1.61012432 
##          419          140           36          150          143           73 
##   0.77696926   2.46366492   1.34306621   1.22855834   0.01905916   0.91597366 
##          190          442          503          188          434          581 
##   0.62090111   0.18075507   0.98316928   0.22858166   0.47593856   0.92463913 
##          725          765          605          241          263          213 
##  15.64496491  23.69858804   1.17145790   1.37913188   0.62298683   0.82823560 
##          831          516          726           92          416          463 
##   0.35031872   0.02848003   7.06724137   3.01337553   1.09339526   1.15911778 
##          638          113          205           19          468          636 
##   1.09242034   1.20533892   2.84455689   2.54609752   2.59559822   3.12287958 
##          687          109          121          427          278          340 
##   2.59966296   0.72445765   0.25435822   1.88691930   3.14346727   1.10031701 
##          125           31          304          377          488          156 
##   0.44915842   1.27668595   1.16098802   6.65079005   1.29448301   1.93265550 
##          821          239          277          363          282          130 
##   0.57519463   3.45050378   1.95985179   0.45546467   1.56059996   1.21270117 
##          744           86          530          566          418          106 
##   1.03321876   0.33643203   3.76910729   1.44010097   0.51536836   0.81195552 
##          754          756          133          122          545          746 
##   1.19933031   0.14030486   0.77292622   0.20255643   1.47799202   0.45863990 
##          580          798          293          609          797          355 
##   0.46714401   0.29615539   5.16655225   2.61844557   0.37402120   3.12480009 
##          825          666          151          563          281          532 
##   2.05164757   0.95073702   0.85556716   0.36206488   1.33884449   2.79819246 
##          378          469          149           24          804          536 
##   0.94346639   0.82459329   0.62924111   1.06885953   1.26857544   0.98164334 
##          820          284          618          801          494          274 
```

```
##  0.78510506  0.96996337  0.83132803  4.19562070  1.64911694  1.46400067
##         538          78         414         470         709          22
##  0.35980728  1.21450027  2.71211322  0.13777171  0.21080416 24.95838895
##         195         209         311         568         349         325
##  0.68274110  0.16131316  3.21217662  0.72158012  0.35590650  0.84451059
##         537         249         323         244          30         619
##  4.77369682  0.93652501  2.55372343  3.82535094  0.74444428  2.92893825
##         131          49         759         737         208         527
##  1.39877188  0.87140447  2.82976827  0.86641195  1.09474495  4.00950656
##         346           8         412         289         379         192
##  0.57659747  1.47858033  2.94650857  0.64054013  6.52648944  1.58173503
##         621         473         809         542         276         155
##  5.26959702  0.45148290  9.51574347  2.26289815  1.95570812  1.22361028
##          50         110          42         230         137         471
##  4.71551391  0.54349860  0.33367821  1.02777009  0.31666442  1.19730877
##         313         198         292         639         814         541
##  2.02458225  0.35276577  3.55194483  0.70150632 25.71150146  0.80959033
##          39         592         334         430         597          60
##  0.72829321  1.35184074  0.89226605  0.91811028  0.11522867  1.75439279
##         595         296         303         351         511         186
##  0.25964437  2.47899623  3.75513670  0.04440256  3.07992079  8.39755525
##         718         466         664         297         112         464
##  0.23151432  1.27736968  0.97753675  1.13618379  0.18554949  3.53811544
##         555          87          18         660         512         451
##  2.10066988  0.22840063  0.66977193  0.17493841  0.48220786  3.03436605
##         484         620          53         626         630         222
##  2.32185734  0.75938770  4.54281893  0.65884842  2.71252669  1.70775491
##         404         694         217         779         515         589
##  1.61696199  7.10398167  1.59679129  1.21247237  0.49387369  0.22586515
##         168         335         614         729         417         684
##  0.63960778  1.18317451  1.66403429  2.85379472  5.69364715  1.92881899
##         654         359         780         235         129         317
##  2.65493784  7.63218747  0.24797817  1.92052105  0.47717749  0.71511074
##          16         266         776         194          20         437
##  0.06792457  1.56408406 11.29395488  0.47405547  2.16165046  0.88872253
##         480         799         714         720         813         705
##  2.32224230  0.08195327  0.52322834  0.44605580  1.67977241  1.01359395
##         828         699         369         243         816         409
##  1.05315299  0.19457402  4.95237266  5.01822048  1.90835495  0.24956739
##         420         792         658         760         546         214
##  1.06241852  0.31749017  1.63676436 17.65413450  2.32913659  0.29847836
##         520         525         166         790         298         585
##  0.61487678  0.79247517  1.46476755  0.72379842  9.66151176  2.21167121
##           5         665         602         519         483          35
##  0.68410330  1.53526117  0.91426819  2.64301638  2.15026947  0.58931121
##         800         601         380         251         710         622
##  4.50245668  1.92706132  1.78759184  4.57042333  4.54237842  0.41944978
##         531         391         487          79         262         201
##  1.46627679  3.56638990  1.39933641  0.04413435  0.80638484  2.20766777
##         145         407         490         231         424         172
##  0.33463666  1.03908545  1.45547017  0.70340583  1.08963516  0.39525019
##         365         549         402         643         646         134
##  0.19529519  0.43154185  0.76272194  0.53626988  1.24345691  1.03843820
##         142           1         218         786         476         450
##  1.80368306  0.63878188  1.51047477  1.49936364  2.99349708  0.63374633
##         329         236         117         307         733          25
##  0.69339101  0.59101092  0.37796203 10.65548316 17.20872624  2.63513241
##         228         688         678         627          56          33
##  1.15769278  1.07974593  0.47129780  1.27828539  2.19873412  8.87791131
##         184         257         446         681         161         162
##  0.84235040  0.50608052  0.27542604  0.33158647  1.38777498  1.30536212
##         305         366         570          88         432         197
##  0.93080379  2.19995788  0.27594414  0.58248694  6.26298784  0.66546470
##         252         439         577         617         739         328
##  1.09555029  0.19554800  0.87152799  1.49388647  0.37520393  7.65871215
##         479         104         212          44         270         465
##  1.83456065  1.85045503  6.77616274  0.43047044  1.42610042  3.49628214
##         255         399         309         225           9         675
##  1.15977626  1.00679818 11.83142415  1.00442750  1.35109520  0.83073367
##           4         300         461         722         425         246
##  2.92001891  4.20049783  1.65216229  0.82197993  1.68732155  0.46296428
##         683         219         272         712         250          32
```

```
##   2.20435511   0.47011815   8.80763081   0.80870476   0.60533116   1.95339036
##          224          823          187          372           34           68
##   2.07245535   1.62809927   7.94188008   3.60773263   3.33844356   4.43290993
##          645          819          330          575          576          667
##   4.04109973   0.51518248   0.58366307   0.84203457   0.29584013   1.09278581
##            7          362          221           75          567          777
##   2.57452654   2.47247471   1.27409245   2.01528968   0.68226524   2.31708008
##          259           62          500          428           66          174
##   5.44213821   1.41138526   0.46831288   3.11175162   1.56324004   0.58333815
##          603          275          762           64          457          354
##   2.91152550   1.07205558   0.30613077   0.58806107   1.83729921   5.91893417
##          606          232          123          173          327          795
##   1.66234945   0.77748608   1.47865659   0.44790536   1.15767174   0.63637988
##          136          812          306          539          226          522
##   0.90932462   2.66284112   5.51764288   0.62993475   0.13233777   3.45087201
##          528          663          517          163          138          371
##   0.87360149   1.16584777   0.63443400   0.35766423   0.42802271   2.08665090
##          583          554          772          268          390          127
##   0.51268768   3.07830071   8.63797537   5.32282145   1.17163601   0.90306160
##          671           27          286          582          393          103
##   1.23151215   2.12774421   0.14648784   1.79385010   3.21492673   1.23160771
##          165          215          308          698          826          189
##   1.70182083   1.07071074   1.23077313   0.42432323   1.99063241   1.34685682
##          604          693          453          447          247          526
##   0.63814889   0.48007727   2.24345994   3.53033126   0.78100943   0.55960527
##          343          443          233          193          659          392
##   2.37362655   0.22408455   3.22818036   2.52268514   0.21265357   1.52954245
##          460          237           70          204           47          730
##   2.41463490   1.76492866   0.44156948   0.73238193   0.65114889   0.35649640
##           55          559          591          600          650          558
##   1.09151755   0.45863990  10.37865924   1.27148646   3.32476373   0.95946197
##          383          631          206          788          176          505
##   0.54373384   1.47793473   1.19103987   1.01694222   0.27109116   0.39124812
##           71          347          119          807           96           65
##   0.33309116   4.43203813 601.16723087   7.09826794   1.39606270   1.78474543
##          413          169          561          386          408          341
##   0.47032469   0.20484248   1.60262996  12.21888167  11.11252508   1.29772323
##          677          700          449          533          644          322
##   2.20501761   1.20125963   2.57406826   3.23000114   1.83706611   1.23196378
##          747          387          572          599          242          135
##   0.08071211   0.55757510   0.63262546   1.47843098   0.84835992   0.65551501
##          455           95          811           37          181          344
##   3.41812661   0.60889361  14.28853575   4.33583434   1.21092153   6.13505209
##          586          319          167          633          462          191
##   1.74393475   3.24694203   0.15388502   0.85683853   0.08225428   2.33906448
##          440           48          436          353          385          711
##   3.40625151   1.65427269   2.01963210   4.70041191   2.25160432   0.66310895
##           84          637          556          153          727          312
##   2.01028284   1.73995294   1.91186320   0.65101764   0.79377604   7.21987785
##          394          357          423          279          692          574
##   2.16831320   1.66191044   1.56994479   3.45599411   0.97074244   0.44323457
##          295          579          764          769          796          401
##  17.43821723   1.56314519   0.98917575   0.22842677   5.08977724   0.12703644
##          350           99          794          115          202          612
##   6.00331210   0.27845091   4.94861125   0.34041940   1.17771755   1.72541625
##          477          147          318           38          441          703
##   2.35303755   1.49325964   1.90173742   0.66008582   2.12282704   0.41215677
##          342          152           41          170          791          758
##   0.92927488   0.45526145   0.81580807   1.21158286   2.16444606   1.37608050
##          361          101          708          238          421          171
##   7.79963485   0.49005337   0.81525331   0.86961032   1.86036699   1.35558024
##          496          672          291           69          743            3
##   1.86868316   0.27283793   8.23480286   1.01070382   1.06772332   0.35356263
##          679          584          713          175          426          696
##   0.61970424   1.66682121   6.25396087   1.46474539   0.31569937   0.82409171
##           46          288          506          740          753          588
##  26.59834221   2.16397176   2.35864572   0.78442514   2.73354499   0.85811669
##           83          139          670          111          431          415
##   0.63363779   1.88297945   1.08460239   1.58901165   1.93972761   2.07887097
##          649           82          691          283          773           59
##   2.52523523   1.51484582   0.15555473   0.13215806   1.59152700   0.55184622
##          562          108          770          706          611          736
##  17.38798620   1.57122629   0.65402371   1.37037109   1.16088215   1.38950214
```

```
##          497          593          445           67
##   1.11329312   0.79519756   0.76814476   0.70678508
```

```
lm.Interaction_Whole <- lm(formula = Centred_PctUpToDate ~ Centred_Enrolled * Centred_PctChildPoverty , data
= Interaction_Whole)
summary(lm.Interaction_Whole)
```

```
##
## Call:
## lm(formula = Centred_PctUpToDate ~ Centred_Enrolled * Centred_PctChildPoverty,
##     data = Interaction_Whole)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.285  -3.392   3.462   7.397  17.441
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               1.307e-01  4.519e-01   0.289    0.772
## Centred_Enrolled                          1.845e-03  4.043e-04   4.564 5.94e-06
## Centred_PctChildPoverty                   1.947e-01  3.865e-02   5.037 6.03e-07
## Centred_Enrolled:Centred_PctChildPoverty -1.900e-04  4.333e-05  -4.385 1.34e-05
##
## (Intercept)
## Centred_Enrolled                         ***
## Centred_PctChildPoverty                  ***
## Centred_Enrolled:Centred_PctChildPoverty ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.93 on 696 degrees of freedom
## Multiple R-squared:  0.07882,    Adjusted R-squared:  0.07485
## F-statistic: 19.85 on 3 and 696 DF,  p-value: 2.345e-12
```

```
library(DHARMa)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

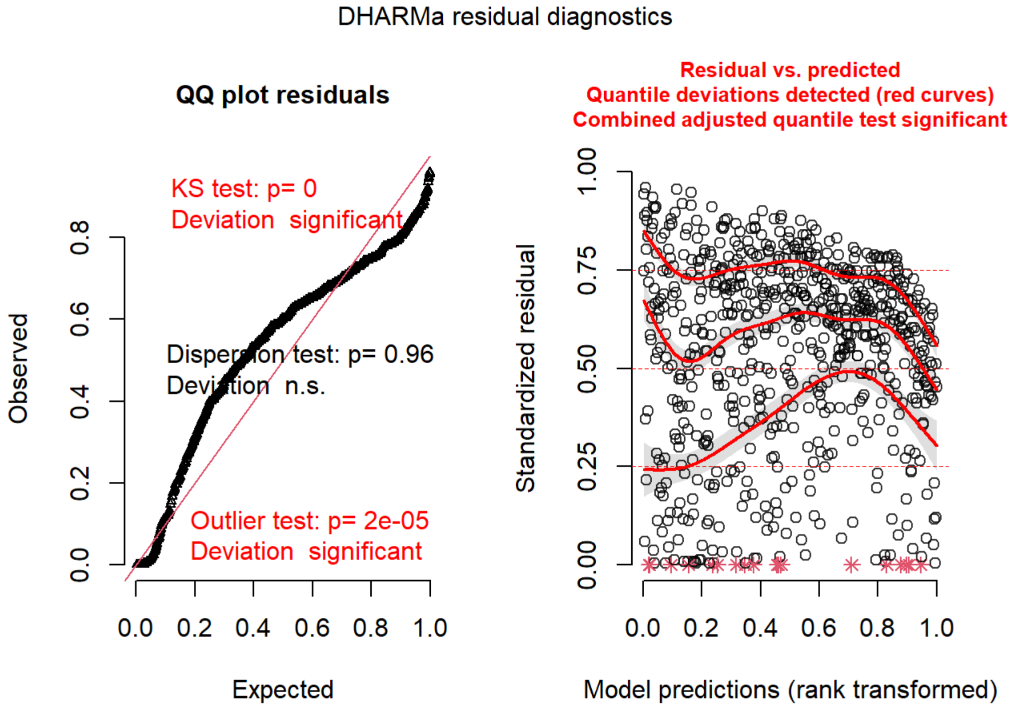
```
## This is DHARMa 0.4.1. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa') Note
: Syntax of plotResiduals has changed in 0.3.0, see ?plotResiduals for details
```

```
Residuals1 <- simulateResiduals(fittedModel = lm.Interaction_Whole, n=250)
```

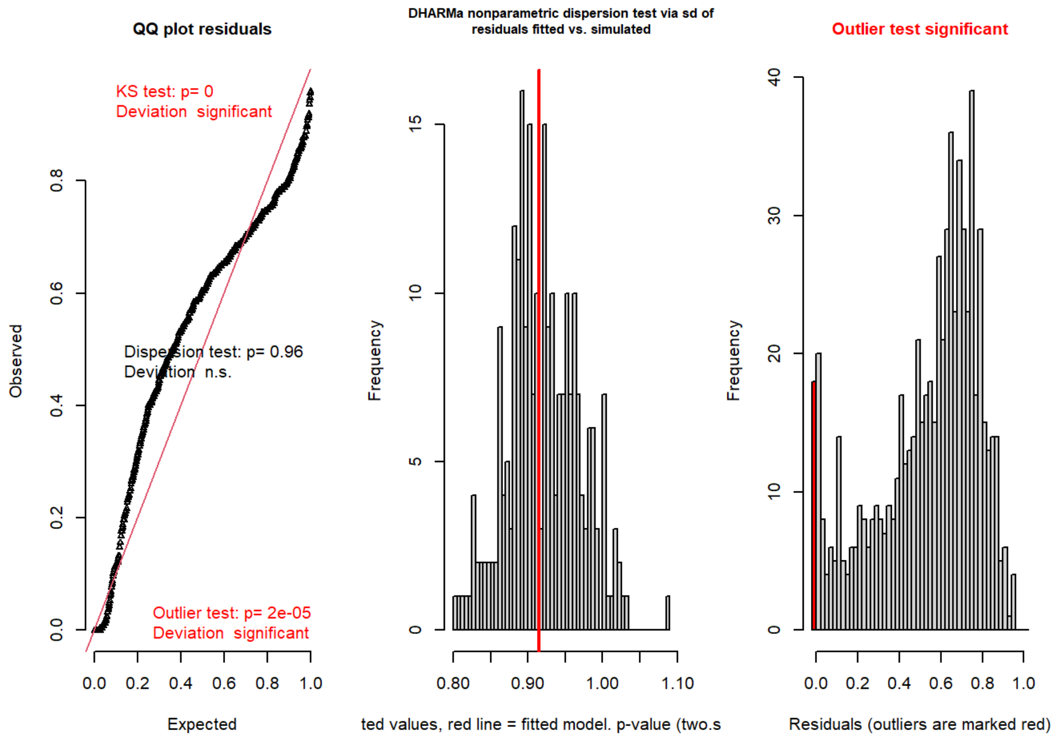
```
## Warning in securityAssertion("nKcase - wrong dimensions of response"): Message from DHARMa: During the ex
ecution of a DHARMa function, some unexpected conditions occurred. Even if you didn't get an error, your res
ults may not be reliable. Please check with the help if you use the functions as intended. If you think that
the error is not on your side, I would be grateful if you could report the problem at https://github.com/flo
rianhartig/DHARMa/issues
##
##  Context: nKcase - wrong dimensions of response
```

```
## Warning in securityAssertion("nKcase - wrong family"): Message from DHARMa: During the execution of a DHA
RMa function, some unexpected conditions occurred. Even if you didn't get an error, your results may not be
reliable. Please check with the help if you use the functions as intended. If you think that the error is no
t on your side, I would be grateful if you could report the problem at https://github.com/florianhartig/DHAR
Ma/issues
##
##  Context: nKcase - wrong family
```

```
plot(Residuals1)
```




```
testResiduals(Residuals1)
```



```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.152, p-value = 1.787e-14
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.99295, p-value = 0.96
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 18, observations = 700, p-value =
## 2.037e-05
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.01530952 0.04033605
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.02571429
```


```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.152, p-value = 1.787e-14
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.99295, p-value = 0.96
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 18, observations = 700, p-value =
## 2.037e-05
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.01530952 0.04033605
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                              0.02571429
```


#Conducting Bayesian Analysis

```r
library(BayesFactor)
bayes.out <- lmBF( formula = PctUpToDate ~ Enrolled * PctChildPoverty , data = Interaction_Whole, posterior=
TRUE, iterations=10000)
summary(bayes.out)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                                Mean        SD  Naive SE Time-series SE
## mu                        8.795e+01 4.502e-01 4.502e-03      4.502e-03
## Enrolled                  5.918e-03 1.299e-03 1.299e-05      1.299e-05
## PctChildPoverty           3.035e-01 4.161e-02 4.161e-04      4.234e-04
## Enrolled.&.PctChildPoverty -1.857e-04 4.254e-05 4.254e-07      4.254e-07
## sig2                      1.425e+02 7.701e+00 7.701e-02      7.701e-02
## g                         1.071e-01 2.056e-01 2.056e-03      2.056e-03
##
## 2. Quantiles for each variable:
##
##                                2.5%        25%        50%        75%
## mu                        8.706e+01  8.764e+01  8.795e+01  8.824e+01
## Enrolled                  3.331e-03  5.058e-03  5.913e-03  6.802e-03
## PctChildPoverty           2.214e-01  2.757e-01  3.033e-01  3.316e-01
## Enrolled.&.PctChildPoverty -2.693e-04 -2.145e-04 -1.855e-04 -1.577e-04
## sig2                      1.284e+02  1.372e+02  1.422e+02  1.475e+02
## g                         1.859e-02  3.851e-02  6.324e-02  1.114e-01
##                               97.5%
## mu                        8.884e+01
## Enrolled                  8.498e-03
## PctChildPoverty           3.847e-01
## Enrolled.&.PctChildPoverty -1.015e-04
## sig2                      1.586e+02
## g                         4.479e-01
```

```r
library(BayesFactor)
Bayes.output <- lmBF( formula = PctUpToDate ~ Enrolled * PctChildPoverty , data = Interaction_Whole)
Bayes.output
```

```
## Bayes factor analysis
## --------------
## [1] Enrolled * PctChildPoverty : 1980915230 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

#Interpretation :-

The Enrolled students and percentage of child poverty perfectly interacts with each other for prediction of percentage of students with up_to_date vaccine as all of them are statistically significant. The interaction with a p-value less than that of standard alpha value which is 1.34e-05 ** and the Percentage of child poverty with the p-value of 6.03e-07 *** and the percentage of enrolled students with a p-value of5.94e-06 ***. The model give us the F statistics of 19.85 on 3 and 695 degrees of freedom and the Adjusted R squared r the accuracy which we can gate from this model is upto0.07485 and 7.485 respectively. The model favors the alternate hypothesis and rejects the null hypothesis which says its an intercept only model.

Bayesian Representation

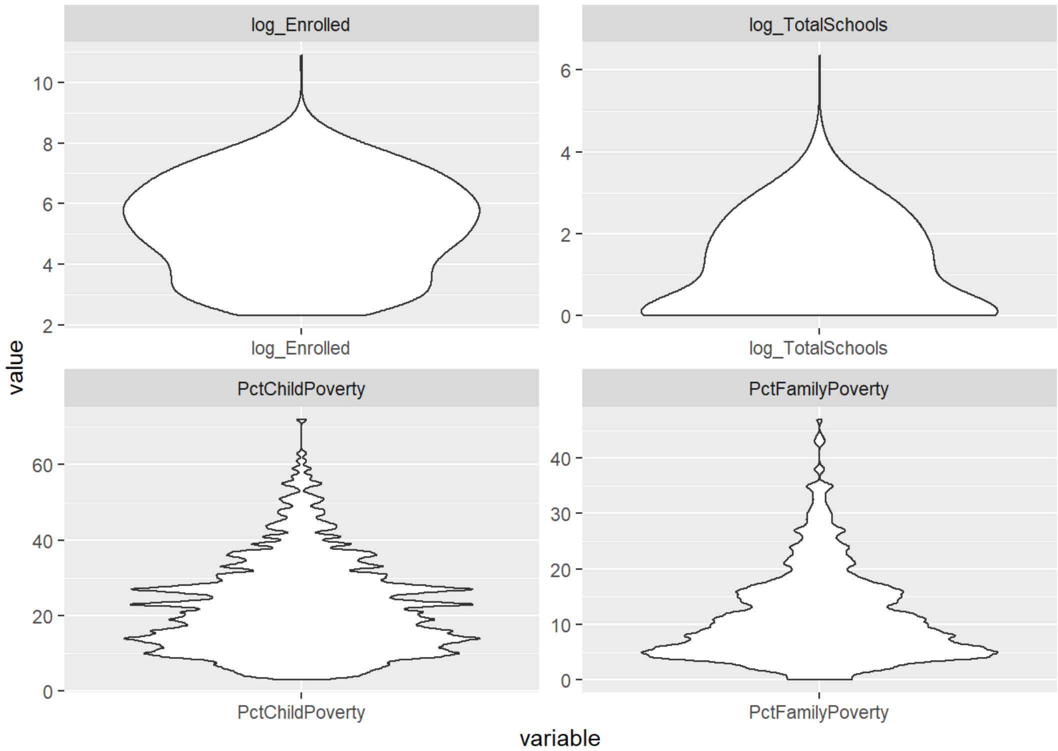
A Bayesian regression also found overwhelming evidence in support of a model which provides the accuracy upto 7.485 percentage or the adjusted Rsquared is 0.07485 in the linear regression modelThe sampled coefficients had similar values, a mean of 0.37935 for WithDTP with an HDI of 0.283431 (lower bound) to 0.47659(upper bound),The mean of 0.80186 for WithMMR with an HDI of 0.80186(lower bound) to 0.89568(upper bound), The mean of -0.09436 for WithHepB with an HDI of -0.151667(lower bound) to -0.03890(upper bound), The mean of 0.04754 forpercentage of belief exempt with an HDI of 0.021999(lower bound) to 0.07240(upper bound), The mean of 0.02085 forpercentage of child poverty with an HDI of 0.001827(lower bound) to 0.03901(upper bound)

The odds ratio is 1980915230 ±0% which is strongly in the favour of alternate of hypothesis that means The Enrolled students and percentage of child poverty perfectly interacts with each other for prediction of percentage of students with up_to_date vaccine rejecting the null hypothesis or the intercept only model.

The sampled coefficients had similar values, a mean of 0.37935 for WithDTP with an HDI of 0.283431 (lower bound) to 0.47659(upper bound),The mean of 5.894e-03 for Enrolled students with an HDI of 3.332e-03(lower bound) to 8.465e-03(upper bound), The mean of 3.038e-01 for Percentage of child poverty with an HDI of 3.038e-01(lower bound) to 8.465e-03(upper bound), The mean of -1.849e-04 for the interaction between the enrolled and the child poverty with an HDI of -2.676e-04 (lower bound) to -1.005e-04(upper bound)

9. *Which, if any, of the four predictor variables predict whether or not a district's reporting was complete?*

```
Districts_New <- subset(districts,select=c(PctChildPoverty, PctFamilyPoverty, log_Enrolled,log_TotalSchools,
DistrictComplete ))
library(tidyverse)
Districts_New %>% pivot_longer(cols=-c(DistrictComplete), names_to="variable",
                        values_to="value", values_drop_na = TRUE) %>%
ggplot(aes(x=variable, y=value)) + geom_violin(bw=.5) + facet_wrap( ~ variable, scales="free")
```



On basis of the observation of violin plot we can say that the log_Enrolled and log_TotalSchool's skewness is improved after we performed the log on enrolled and total schools.

```
summary(Districts_New)
```

```
##  PctChildPoverty PctFamilyPoverty  log_Enrolled    log_TotalSchools
##  Min.   : 3.00   Min.   : 0.00    Min.   : 2.303   Min.   :0.000
##  1st Qu.:13.00   1st Qu.: 5.00    1st Qu.: 3.892   1st Qu.:0.000
##  Median :20.00   Median : 9.00    Median : 5.260   Median :1.099
##  Mean   :22.18   Mean   :11.32    Mean   : 5.240   Mean   :1.143
##  3rd Qu.:29.00   3rd Qu.:15.25    3rd Qu.: 6.507   3rd Qu.:2.079
##  Max.   :72.00   Max.   :47.00    Max.   :10.901   Max.   :6.366
##  DistrictComplete
##  Mode :logical
##  FALSE:37
##  TRUE :663
##
##
##
```

```
cor(Districts_New)
```

```
##                    PctChildPoverty PctFamilyPoverty log_Enrolled log_TotalSchools
## PctChildPoverty        1.00000000      0.855977219  -0.07029536     -0.095150144
## PctFamilyPoverty       0.85597722      1.000000000   0.04361788     -0.007324958
## log_Enrolled          -0.07029536      0.043617881   1.00000000      0.916661848
## log_TotalSchools      -0.09515014     -0.007324958   0.91666185      1.000000000
## DistrictComplete      -0.08670704     -0.112000876  -0.13850355     -0.224278205
##                    DistrictComplete
## PctChildPoverty         -0.08670704
## PctFamilyPoverty        -0.11200088
## log_Enrolled            -0.13850355
## log_TotalSchools        -0.22427820
## DistrictComplete         1.00000000
```

From the results we can see that the highly correlated data are :-

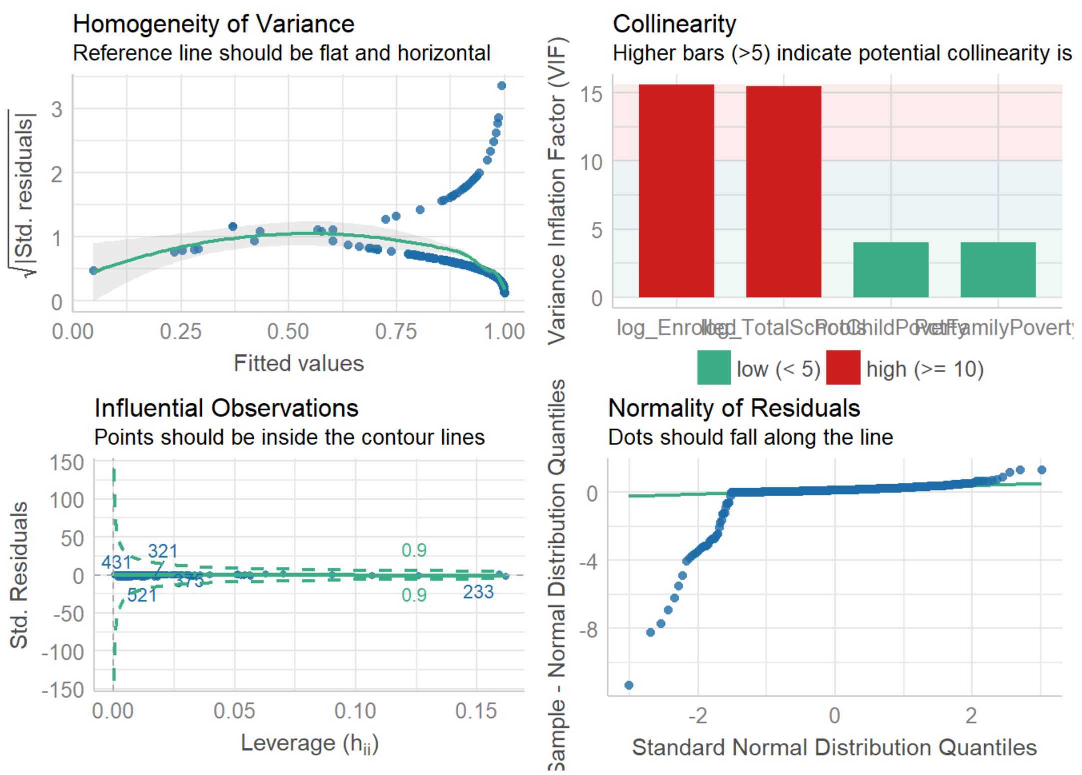
Percentage of child poverty and percentage of family poverty are strongly correlated log_Enrolled is highly correlated to log_TotalSchools Percentage of family poverty is correlated to log_enrolled

We are excluding the District Complete because it is not a numerical value.

#Running a logistic regression model and understanding the visual representations

```
library(performance)
library(see)

DistrictsNew.glm <- glm(formula = DistrictComplete ~ PctChildPoverty + PctFamilyPoverty + log_Enrolled + log
_TotalSchools, family = binomial(link="logit"), data = Districts_New)
check_model(DistrictsNew.glm)
```

```
## Loading required namespace: qqplotr
```




```
summary(DistrictsNew.glm )
```

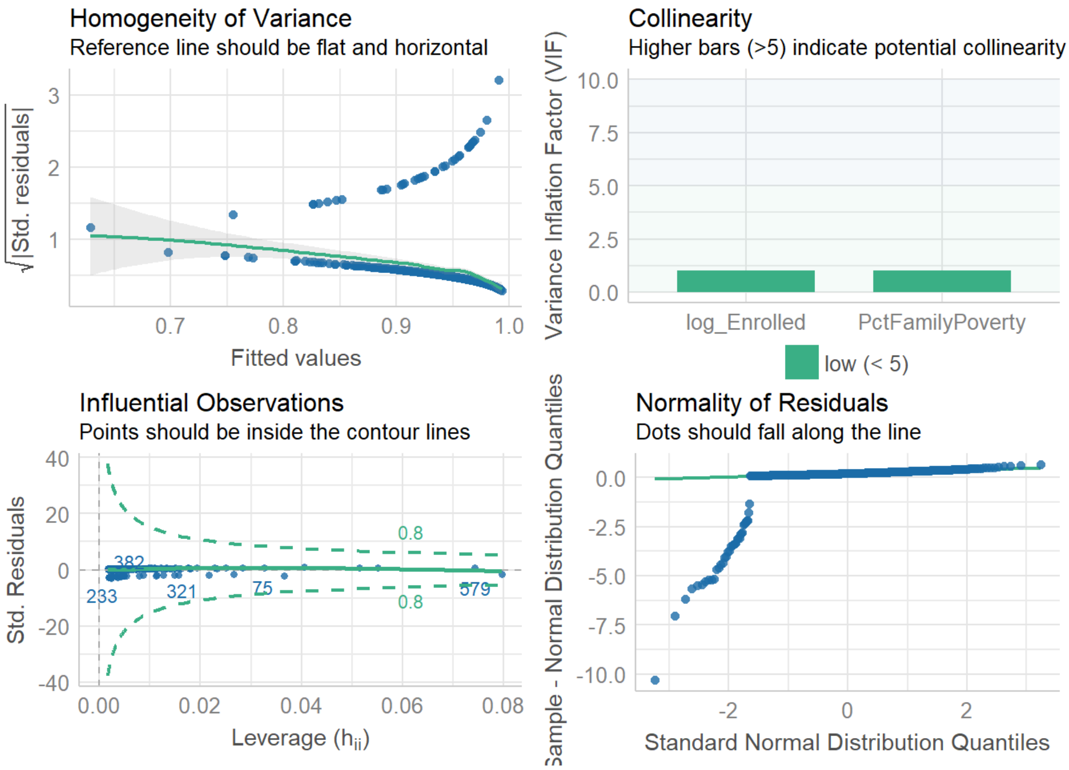
```
##
## Call:
## glm(formula = DistrictComplete ~ PctChildPoverty + PctFamilyPoverty +
##     log_Enrolled + log_TotalSchools, family = binomial(link = "logit"),
##     data = Districts_New)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1187   0.1062   0.2071   0.3195   1.4126
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.587491   1.299050  -1.222   0.2217
## PctChildPoverty   0.008808   0.033015   0.267   0.7896
## PctFamilyPoverty -0.080851   0.045306  -1.785   0.0743 .
## log_Enrolled      1.890938   0.373025   5.069 4.00e-07 ***
## log_TotalSchools -3.290770   0.554701  -5.933 2.98e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 289.58  on 699  degrees of freedom
## Residual deviance: 219.57  on 695  degrees of freedom
## AIC: 229.57
##
## Number of Fisher Scoring iterations: 7
```

As per the observation we can see that for homogeneity of variance the reference line is flat and horizontal, For influential Observations points are inside the contour lines, For normality of rsiduals Dots are falling along the line. The only issue is with collinearity it is showing us that multiple variables are collinear and the model needs an improvement in removing the multi-collinearity issue.

Now here we can use our results of correlation matrix. We can see that the log_Enrolled and log_TotalSchool, percentage of child poverty and percentage of family poverty were highly correlated so we will consider only on among each of them.

As from summary we can see that the p-value of child poverty is not statistically significant so we can remove that from the model to deal with multi-collinearity issue and along with it we can remove log_totalschools as the standard error of log_totalschools is more of it than that of the log_Enrolled

```
DistrictsNew1.glm <- glm(formula = DistrictComplete ~ PctFamilyPoverty + log_Enrolled , family = binomial(li
nk="logit"), data = Districts_New)
check_model(DistrictsNew1.glm)
```



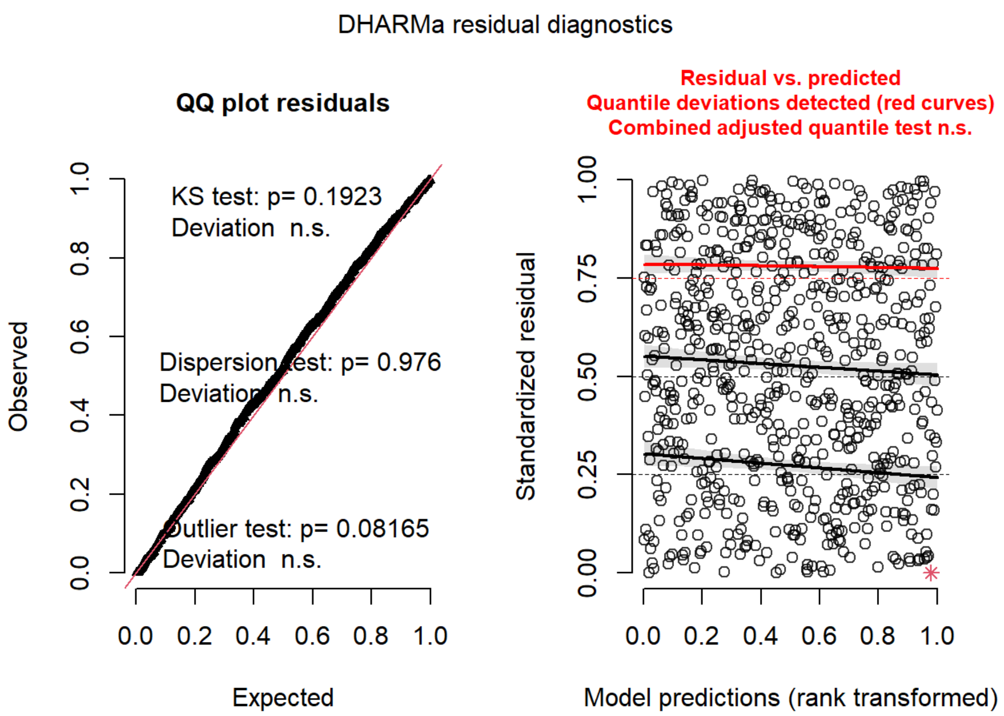
```
summary(DistrictsNew1.glm )
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctFamilyPoverty + log_Enrolled,
##     family = binomial(link = "logit"), data = Districts_New)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0572   0.1999   0.2764   0.3695   0.8485
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       5.93706    0.79438   7.474 7.79e-14 ***
## PctFamilyPoverty -0.05421    0.01875  -2.892 0.003829 **
## log_Enrolled     -0.41168    0.11549  -3.565 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 289.58  on 699  degrees of freedom
## Residual deviance: 268.25  on 697  degrees of freedom
## AIC: 274.25
##
## Number of Fisher Scoring iterations: 6
```

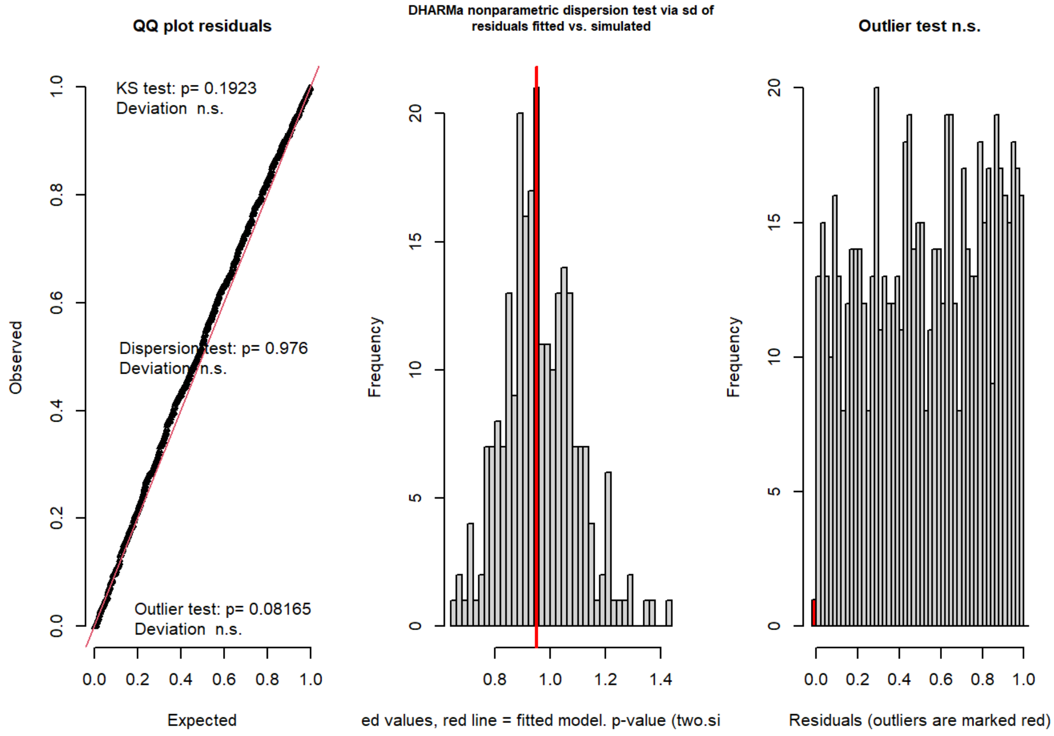
We have now cleared the multi-collinearity issue from the model.

To create interpret the residuals and test them Dharma residual diagnostics is the best simulation based approach

```
library(DHARMa)
Residuals <- simulateResiduals(fittedModel = DistrictsNew1.glm, n=250)
plot(Residuals)
```




```
testResiduals(Residuals)
```




```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.040891, p-value = 0.1923
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.98701, p-value = 0.976
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 10, observations = 700, p-value = 0.08165
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.006871248 0.026114596
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                            0.01428571
```

```
## $uniformity
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  simulationOutput$scaledResiduals
## D = 0.040891, p-value = 0.1923
## alternative hypothesis: two-sided
##
##
## $dispersion
##
##  DHARMa nonparametric dispersion test via sd of residuals fitted vs.
##  simulated
##
## data:  simulationOutput
## dispersion = 0.98701, p-value = 0.976
## alternative hypothesis: two.sided
##
##
## $outliers
##
##  DHARMa outlier test based on exact binomial test with approximate
##  expectations
##
## data:  simulationOutput
## outliers at both margin(s) = 10, observations = 700, p-value = 0.08165
## alternative hypothesis: true probability of success is not equal to 0.007968127
## 95 percent confidence interval:
##  0.006871248 0.026114596
## sample estimates:
## frequency of outliers (expected: 0.00796812749003984 )
##                                             0.01428571
```

Here in a qq plot the expected plots are coinciding with the resultants. The model predictions shows us the residual vs predicted plots the resultants around first and second quantile is good to go but the third quantile is not coinciding In the Residuals plot the outliers is around 0 marked in red which is not affecting the model that effectiely.

It looks like a normal distribution

In order to check wether the models are statistically significant or not and further description we will summarised the glm model.

```
summary(DistrictsNew1.glm)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctFamilyPoverty + log_Enrolled,
##     family = binomial(link = "logit"), data = Districts_New)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0572   0.1999   0.2764   0.3695   0.8485
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       5.93706    0.79438   7.474 7.79e-14 ***
## PctFamilyPoverty -0.05421    0.01875  -2.892 0.003829 **
## log_Enrolled     -0.41168    0.11549  -3.565 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 289.58  on 699  degrees of freedom
## Residual deviance: 268.25  on 697  degrees of freedom
## AIC: 274.25
##
## Number of Fisher Scoring iterations: 6
```

To obtain the result it performed 6 iterations. The predictors which are percentage of family poverty and percentage of log_enrolled are statistically significant with a p-value less than that of the 0.05.

Here the AIC is 274.25 which is calculating the stress on the model, the lower the aic the better the model is. The Null deviance represents the null hypothesis and residual deviance represents the alternate hypothesis. The residual deviance should be smaller than the null deviance in order to obtain the results in favor of predictors.

The intercept is using one degree of freedom with n = 700 therefore for Null deviance the degrees of freedom is n-1 = 700 - 1 =699 with a null deviance of 289.58

For the Residual deviance the Percentage of family poverty and log_enrolled are using one degree of freedom each and one degree of freedom is used by intercept which makes it 3. Therefore the degrees of freedom is 699 - 2 = 697 for residual deivance of 268.25

we converted the regular odds into log odds for prediction, converting it back into the regular odds using exponential function

```
exp(coef(DistrictsNew1.glm))
```

```
##      (Intercept) PctFamilyPoverty     log_Enrolled
##      378.8196952        0.9472318        0.6625383
```

```
exp(confint(DistrictsNew1.glm) )
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %       97.5 %
## (Intercept)      86.9126433 1980.3161360
## PctFamilyPoverty  0.9137179    0.9839753
## log_Enrolled      0.5241409    0.8261604
```

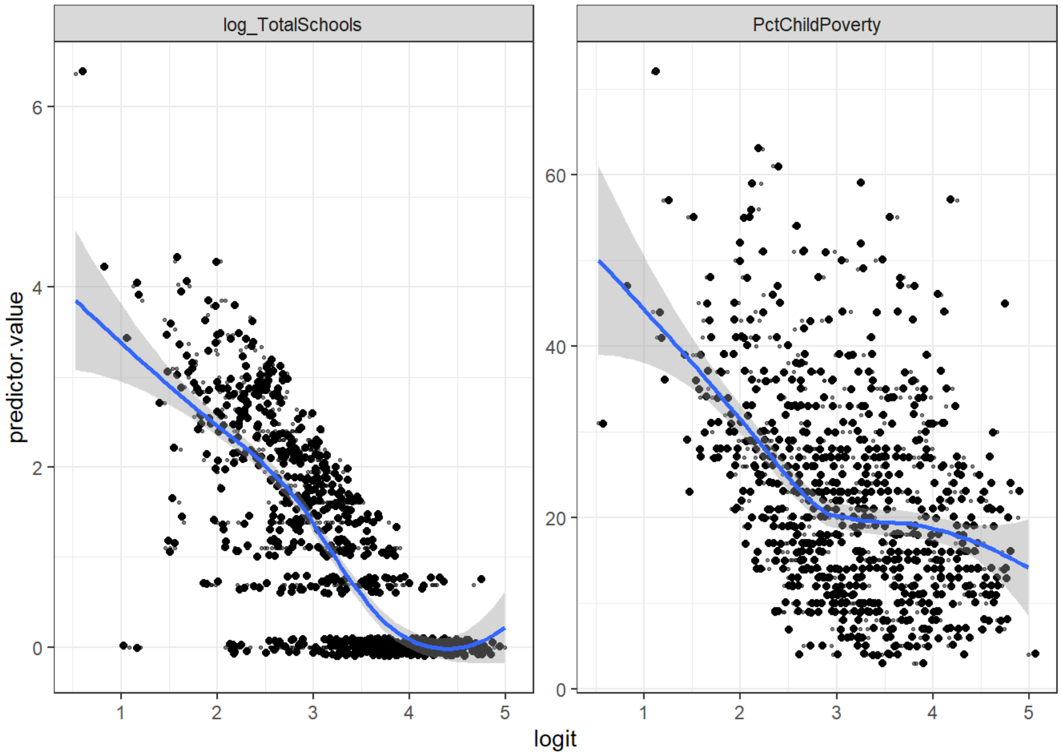
```
probabilities <- predict(DistrictsNew1.glm, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")
head(predicted.classes)
```

```
##   114   200   269   682   210    74
## "pos" "pos" "pos" "pos" "pos" "pos"
```

```
library(DHARMa)
require(dplyr)
library(tidyverse)
# Select only numeric predictors
District_New.n <- Districts_New %>% dplyr::select_if(is.numeric) %>% dplyr::select(-c(PctFamilyPoverty, log_
Enrolled))
predictors <- colnames(District_New.n)
# Bind the logit and tidying the data for plot
District_New.n <- District_New.n %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

```
library(ggplot2)
ggplot(District_New.n, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_jitter(height=.1,width=.1) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



library(performance) library(caret) model_performance(DistrictsNew1.glm) g <- Districts_New g \(DistrictComplete <- as.factor(g\)DistrictComplete) predictedDistrictNew <-round(predict(DistrictsNew1.glm, type="response")) sum(predictedDistrictNew) confusion<-table(predictedDistrictNew, ifelse(Districts_New$DistrictComplete == "TRUE",1,0)) confusion addmargins(confusion) confusionMatrix(confusion, positive="1")

A logistic regression was performed on the data with 700 Districts to predict whether the district's reporting was complete or not. To predict the reporting we used the Percentage of families in district living below the poverty line, Total number of enrolled students in the district as predictors. We can see both of the predictors are statistically significant. We can see that the 95% confidence interval for our Percentage of family poverty and log_Enrolled are the variable—representing our District Complete from a low of 0.9137179:1 up to a high of 0.9839753:1 for percentage of family poverty and a low of0.5241409:1 up to a high of 0.8261604:1 for the log_enrolled.

The model showed the performance of with a Tjur's R2 of 3.9% and accuracy of

#Conducting Bayesian Logistic Regression /analysis

```
library(MCMCpack)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
```

```
## ## Copyright (C) 2003-2021 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
```

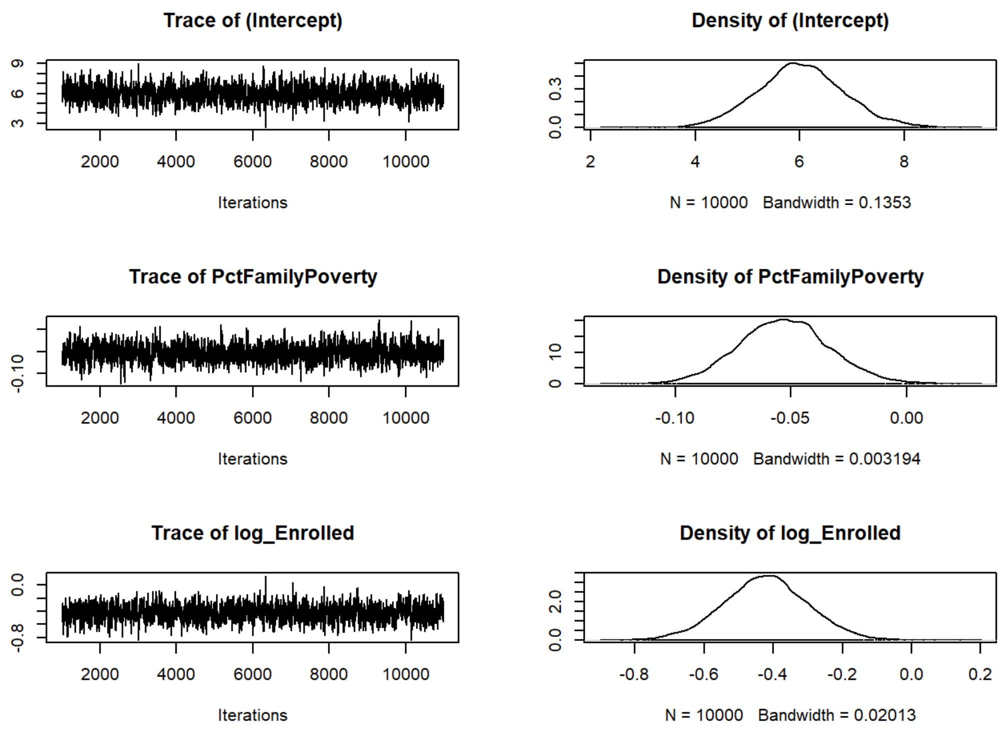
```
## ##
## ## Support provided by the U.S. National Science Foundation
```

```
## ## (Grants SES-0350646 and SES-0350613)
## ##
```

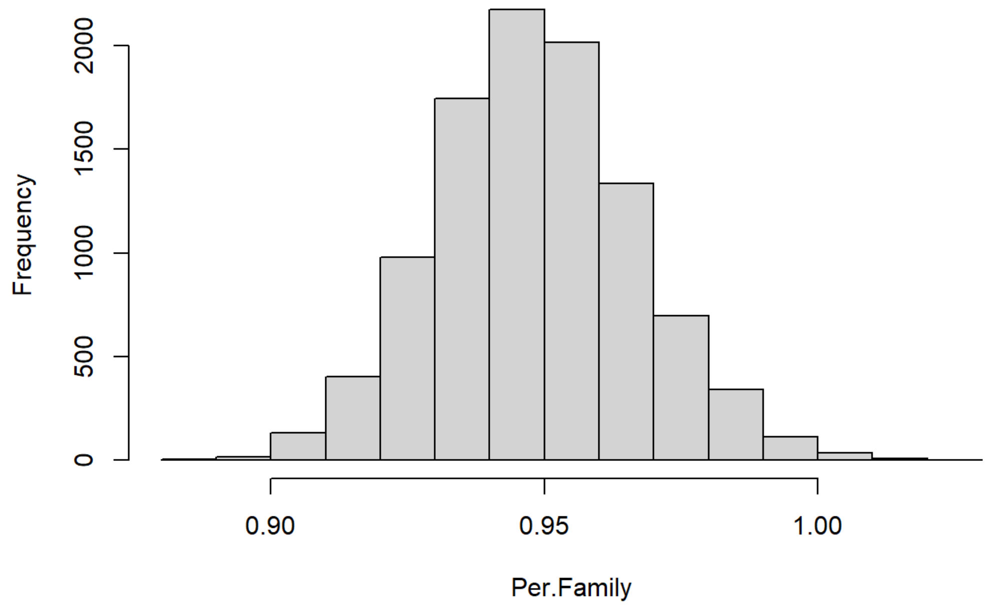
```
MCMC <- MCMClogit(formula = DistrictComplete ~ PctFamilyPoverty + log_Enrolled, data = Districts_New)
summary(MCMC)
```

```
##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean      SD  Naive SE Time-series SE
## (Intercept)       6.01735 0.81969 0.0081969      0.0272248
## PctFamilyPoverty -0.05306 0.01944 0.0001944      0.0006699
## log_Enrolled     -0.42263 0.12099 0.0012099      0.0039630
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%     50%      75%    97.5%
## (Intercept)       4.42942  5.47148  6.0028  6.55045  7.71116
## PctFamilyPoverty -0.09119 -0.06619 -0.0533 -0.04071 -0.01405
## log_Enrolled     -0.66821 -0.50355 -0.4229 -0.34302 -0.18743
```

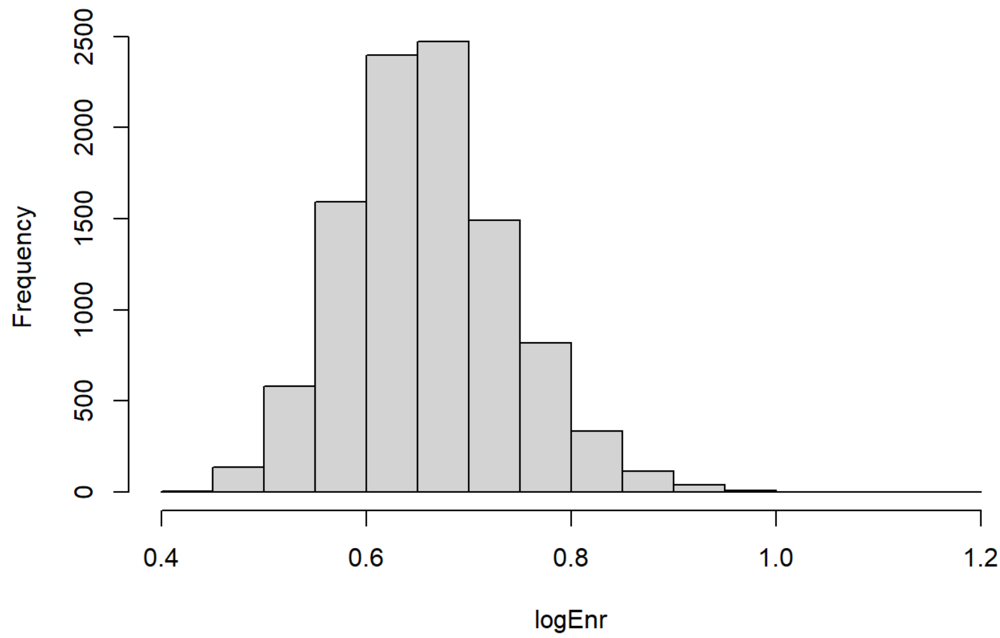
```
plot(MCMC)
```



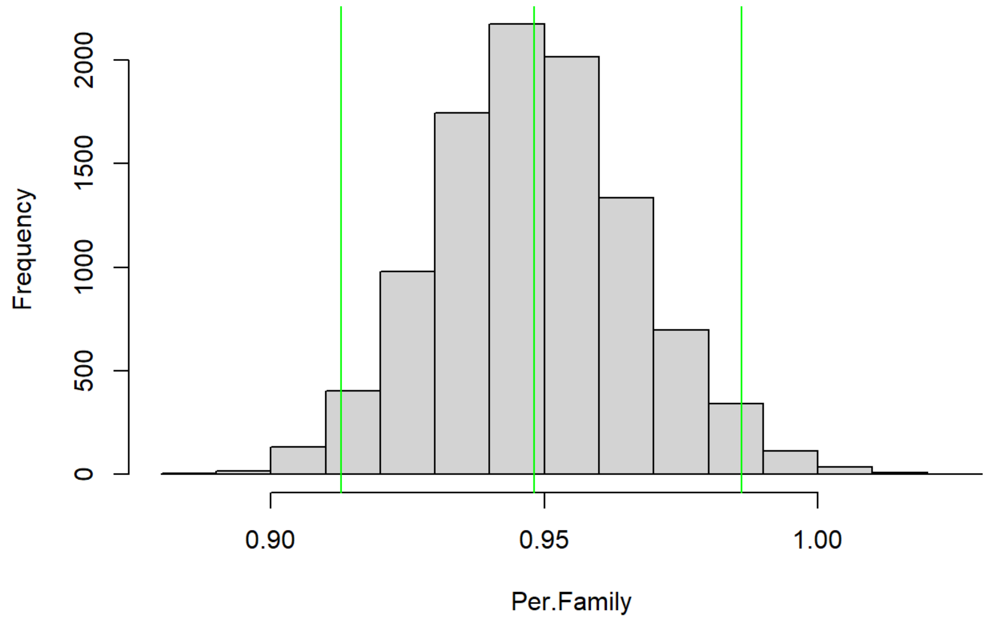

```
Per.Family <- as.matrix(MCMC[,"PctFamilyPoverty"])
logEnr <- as.matrix(MCMC[,"log_Enrolled"] )
Per.Family <- exp(Per.Family)
hist(Per.Family, main=NULL)
```



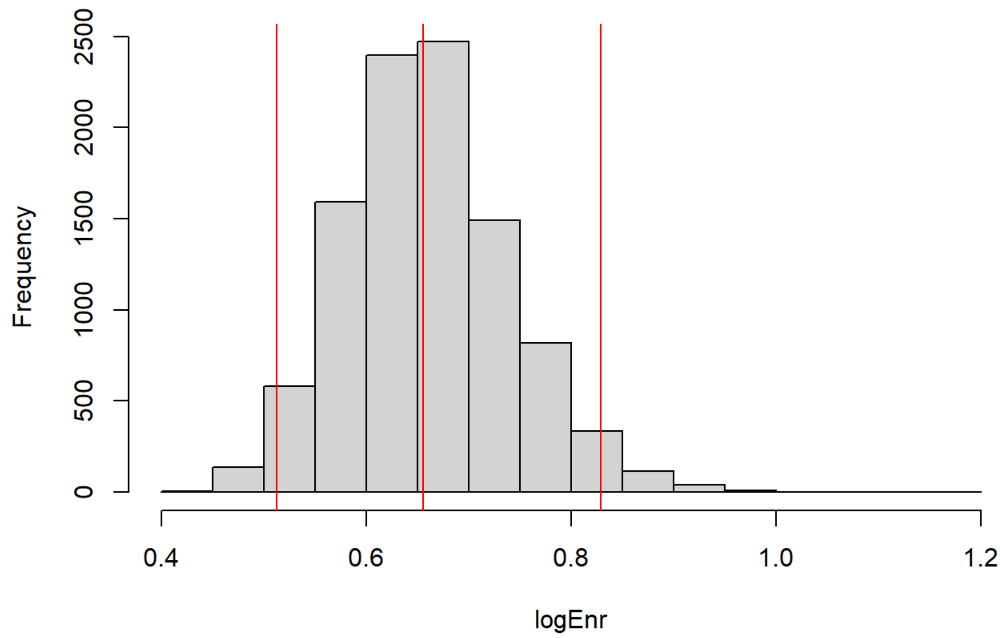

```
logEnr <- exp(logEnr)
hist(logEnr, main=NULL)
```




```
hist(Per.Family, main=NULL)
abline(v=quantile(Per.Family,c(0.025, 0.5, 0.975)),col="Green")
```




```
hist(logEnr, main=NULL)
abline(v=quantile(logEnr,c(0.025, 0.5, 0.975)),col="Red")
```



Iterpretation A logistic regression was performed on the data with 700 Districts to predict whether the district's reporting was complete or not. To predict the reporting we used the Percentage of families in district living below the poverty line, Total number of enrolled students in the district as predictors. We can see both of the predictors are statistically significant. We can see that the 95% confidence interval for our Percentage of family poverty and log_Enrolled are the variable—representing our District Complete from a low of 0.9137179:1 up to a high of 0.9839753:1 for percentage of family poverty and a low of0.5241409:1 up to a high of 0.8261604:1 for the log_enrolled. # The result may vary because of Bayesian analysis as there are 10000 iterations and I didn't set seed.

10. *Concluding Paragraph*

Describe your conclusions, based on all of the foregoing analyses. As well, the staff member in the state legislator's office is interested to know how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance. Make sure you have at least one sentence that makes a recommendation about improving vaccination rates. Make sure you have at least one sentence that makes a recommendation about improving reporting rates. Finally, say what further analyses might be helpful to answer these *questions and any additional data you would like to have.*

Conclusion 1:-

While performing the time series analysis we noticed there is a cyclicity and the trends in rates of the vaccine. To improve the vaccination rate there should be one rate through out the countries and state and a constant rate which we can achieve by allocating the funds with a joint venture of central state legislators.

Relation :-

The analysis shows a deep relation in enrolled students and the total no. of schools. This shows the more no. of schools in a district the more no. ppf students in the school

The other high relation is between the child poverty and family poverty. The analysis suggests that when the child in a district is below poverty line there is a lot of possibility that it is from a family in the district which is below poverty line as well.

With the help pf other analysis we found a strong relation between The no. of children in district living below the poverty line and family who's living in the district below poverty line and the Percentage of students in the district receiving free or reduced cost meals

The more no. of children who are below poverty line are more likely to belong to the families in the district from below poverty line which are from the percentage of ste=udents in a district who recieves ffree or cost_reducin meals

Conclusion2:-

We can conclude that the more no. of schools attracts more no. of students. More no. of student includes all of the students from below poverty line and others.

The people who cannot afford even a meal cannot afford a vaccine.

Building more schools reducing the rate of vaccine and providing free meals and subssidy to families below poverty line will help in improving the overall scenarios.