Fall 2021 - Final Examination

Trishla Jain

Instructions

Your goal for this final exam is to conduct the necessary analyses of vaccination rates in California schools and 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 7 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, so you do not need to define those.

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 and say what you see); 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. Please place the answers in the text (not R comments) after the relevant analysis. 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 for the question you are answering. Please keep your answers concise and focused on the question asked. 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 (e.g., that the code does not run off the edge of the page).

You may not receive assistance, help, coaching, guidance, or support from any human except your instructor at any point during this exam. Obtaining improper assistance will result in a 0 for 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 a personalized 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 me
als
 $ 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 (NB. your sample may be slightly different). 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. Note that the data are about districts, not individual students, so be careful that you do not commit an ecological fallacy by stating conclusions about individuals.

In addition, you will find on Blackboard a CSV file, All Schools.csv, with data about 7,381 individual schools.

```
'data.frame' 7,381 obs. of 18 variables:
 $ SCHOOL CODE              : School ID number
 $ PUBLIC/ PRIVATE          : School status, "PUBLIC" or "PRIVATE" (note the space in the var
iable name: you can access it as `PUBLIC/ PRIVATE`)
 $ Public School District ID: School district ID (only if public)
 $ PUBLIC SCHOOL DISTRICT   : School district name (only if public)
 $ CITY                     : City name
 $ COUNTY                   : Country name
 $ SCHOOL NAME              : School name
 $ ENROLLMENT               : Total number of enrolled students in the school
 $ UP_TO_DATE               : Number of students with completely up-to-date vaccines
 $ CONDITIONAL              : Number of students missing some vaccine without an exemption
 $ PME                      : Number of students with a medical exemption
 $ PBE_BETA                 : Number of students with a personal belief exemption
 $ DTP                      : Number of students in the district with the DTP vaccine
 $ POLIO                    : Number of students in the district with the Polio vaccine
 $ MMR                      : Number of students in the district with the MMR vaccine
 $ HEPB                     : Number of students in the district with Hepatitis B vaccine
 $ VARICELLA                : Number of students in the district with Varicella vaccine
 $ REPORTED                 : Whether the school reported vaccination data (Y or N)
```

Please Note:

1. Data exploration and cleaning has been done at the start of the file (seperately for all 3 data sets provided) to better understand the data and the cleaned data set has been used directly in the questions.
2. There are places where non significant variables haven't been reported to avoid readers fatigue.
3. There are places where base R plots have been used to check linearity and Dharma and check_model plots haven't been used intentionally to avoid repeating and analyzing the same thing again and again.
4. Apologies for any typos (if any!) I tried my best to remove!

Answers:::

Before we begin to answer the questions lets explore the data and do any pre processing or cleaning that is needed so that we can answer the questions with better insights.

Loading the libraries

```
#options(warn=-1)
suppressMessages(library(tidyverse))
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressMessages(library(DHARMa))
suppressMessages(library(dlookr))
suppressMessages(library(e1071))
suppressMessages(library(moments))
suppressMessages(library(GGally))
suppressMessages(library(psych))
suppressMessages(library(pairsD3))
```

```
## Warning: package 'pairsD3' was built under R version 4.1.2
```

```
suppressMessages(library(corrplot))
```

```
## Warning: package 'corrplot' was built under R version 4.1.2
```

```
suppressMessages(library(TSA))
```

```
## Warning: package 'TSA' was built under R version 4.1.2
```

```
suppressMessages(library(tseries))
suppressMessages(library(changepoint))
```

```
## Warning: package 'changepoint' was built under R version 4.1.2
```

```
suppressMessages(library(BEST))
suppressMessages(library(car))
suppressMessages(library(BayesFactor))
suppressMessages(library(lm.beta))
suppressMessages(library(HSAUR))
suppressMessages(library(see))
suppressMessages(library(performance))
suppressMessages(library(MCMCpack))
```

Loading the necessary files

```
load("C:/Users/trish/Desktop/syracuse/Sem 1/IST.772.M001.FALL21.Quant Reasoning Data Science
 17460.1221/final exam/datasets9.RData")
schools <- read.csv("C:/Users/trish/Desktop/syracuse/Sem 1/IST.772.M001.FALL21.Quant Reasonin
g Data Science 17460.1221/final exam/All Schools.csv")
#schools <- read_csv("All Schools.csv")
```

Data Exploration Data Preprocessing and Cleaning For districts data set

1. Checking for NA's in the datasets

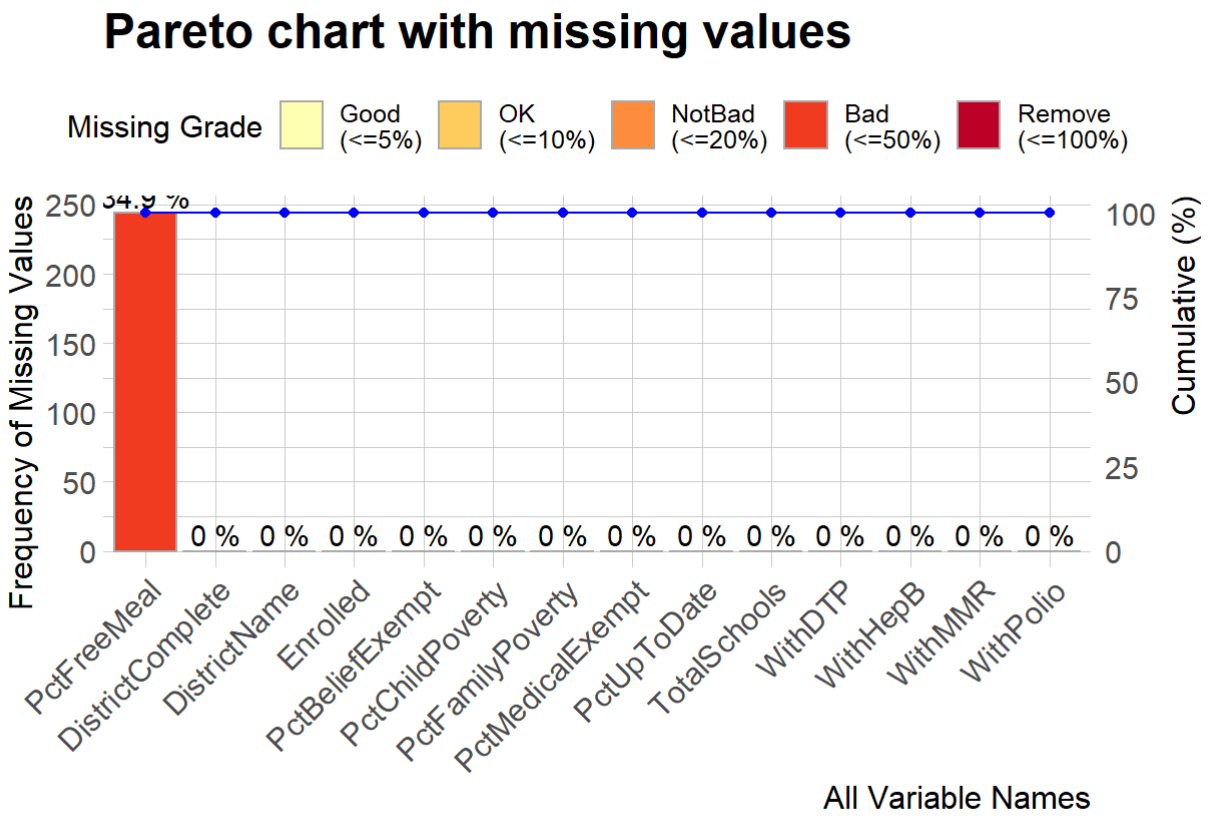
```
summary(districts)
```

```
##                   DistrictName    WithDTP        WithPolio
##  ABC Unified              :  1   Min.   : 23.0   Min.   : 23.0
##  Ackerman Charter         :  1   1st Qu.: 86.0   1st Qu.: 87.0
##  Acton-Agua Dulce Unified:  1   Median : 93.0   Median : 94.0
##  Adelanto Elementary      :  1   Mean   : 89.8   Mean   : 90.2
##  Alameda Unified          :  1   3rd Qu.: 97.0   3rd Qu.: 97.0
##  Albany City Unified      :  1   Max.   :100.0   Max.   :100.0
##  (Other)                 :694
##     WithMMR          WithHepB       PctUpToDate    DistrictComplete
##  Min.   : 23.00   Min.   : 23.00   Min.   : 23.0   Mode :logical
##  1st Qu.: 86.00   1st Qu.: 90.00   1st Qu.: 84.0   FALSE:43
##  Median : 94.00   Median : 96.00   Median : 92.0   TRUE :657
##  Mean   : 89.79   Mean   : 92.26   Mean   : 88.4
##  3rd Qu.: 97.00   3rd Qu.: 98.00   3rd Qu.: 96.0
##  Max.   :100.00   Max.   :100.00   Max.   :200.0
##
##  PctBeliefExempt  PctMedicalExempt PctChildPoverty PctFamilyPoverty
##  Min.   : 0.000   Min.   :0.0000   Min.   : 2.00   Min.   : 0.00
##  1st Qu.: 0.750   1st Qu.:0.0000   1st Qu.:13.00   1st Qu.: 5.00
##  Median : 2.000   Median :0.0000   Median :21.00   Median : 9.00
##  Mean   : 5.623   Mean   :0.1471   Mean   :22.33   Mean   :11.48
##  3rd Qu.: 7.000   3rd Qu.:0.0000   3rd Qu.:29.00   3rd Qu.:16.00
##  Max.   :77.000   Max.   :8.0000   Max.   :72.00   Max.   :47.00
##
##   PctFreeMeal        Enrolled        TotalSchools
##  Min.   :  0.00   Min.   :   10.0   Min.   :  1.000
##  1st Qu.: 28.75   1st Qu.:   50.5   1st Qu.:  1.000
##  Median : 48.00   Median :  201.5   Median :  3.000
##  Mean   : 47.65   Mean   :  630.8   Mean   :  7.253
##  3rd Qu.: 69.00   3rd Qu.:  684.2   3rd Qu.:  8.000
##  Max.   :100.00   Max.   :54238.0   Max.   :582.000
##  NA's   :244
```

```
sum(is.na(districts))
```

```
## [1] 244
```

```
districts %>% plot_na_pareto( col = "blue")
```



We can see that there are 244 NA's in PctFreeMeal column in the districts dataset. Looking at the parento plot, we can see that it would be a good idea to remove these but since removing rows will lead to loss of data with only 456 out of 700 records we will instead remove this column from analysis .

```
districts_noNA <- subset(districts, select = -c(PctFreeMeal) )
#districts %>% drop_na() -> districts_noNA

# checking if Na's are out of the dataset
sum(is.na(districts_noNA))
```

```
## [1] 0
```

Now we have no Na's in districts

2. Checking for Null's in the datasets

```
sapply(districts_noNA,function(x) sum(is.null(x)))
```

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

We have no NULL records in the dataset.

3. Checking for outliers

```
diagnose_outlier(districts_noNA)
```

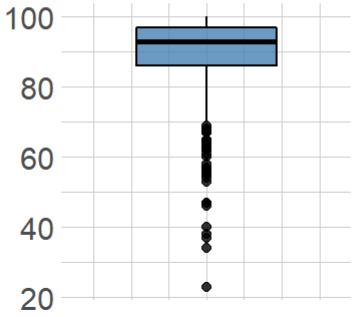
variables <chr>	outliers_cnt <int>	outliers_ratio <dbl>	outliers_mean <dbl>	with_mean <dbl>	without_me <d
WithDTP	42	6.000000	56.857143	89.7957143	91.898
WithPolio	48	6.857143	58.770833	90.2042857	92.518
WithMMR	44	6.285714	56.772727	89.7871429	92.001
WithHepB	46	6.571429	62.434783	92.2628571	94.360
PctUpToDate	48	6.857143	62.583333	88.4014286	90.302
PctBeliefExempt	59	8.428571	29.372881	5.6228571	3.436
PctMedicalExempt	61	8.714286	1.688525	0.1471429	0.000
PctChildPoverty	14	2.000000	58.214286	22.3257143	21.593
PctFamilyPoverty	17	2.428571	37.352941	11.4785714	10.834
Enrolled	63	9.000000	3617.079365	630.7500000	335.398

1-10 of 11 rows Previous **1** 2 Next

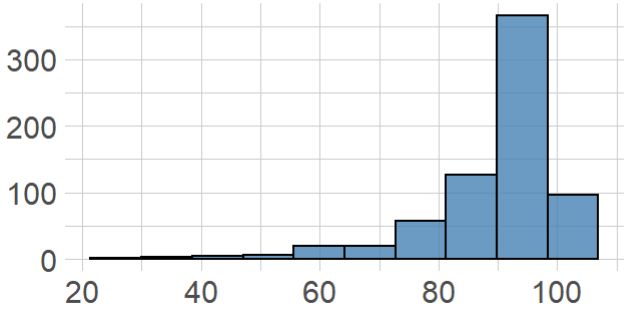
```
plot_outlier(districts)
```

Outlier Diagnosis Plot (WithDTP)

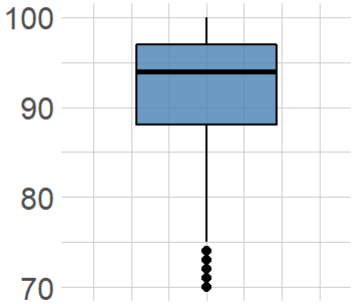
With outliers



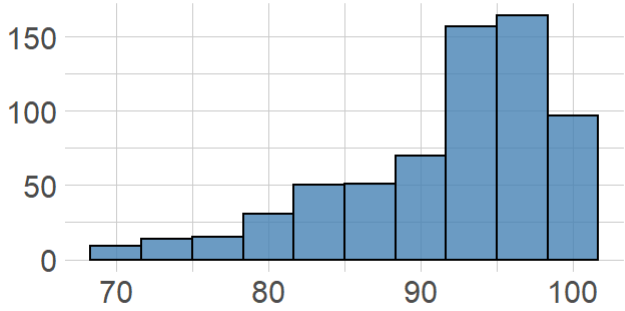
With outliers



Without outliers

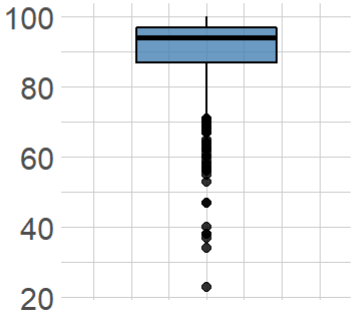


Without outliers

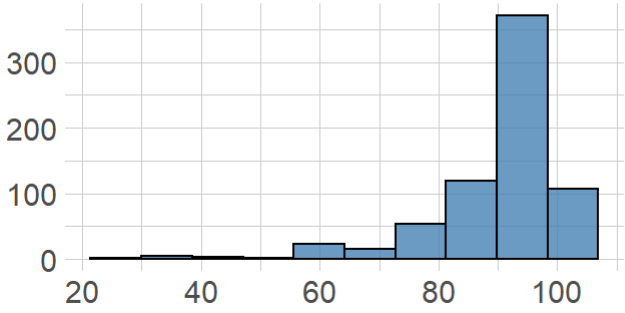


Outlier Diagnosis Plot (WithPolio)

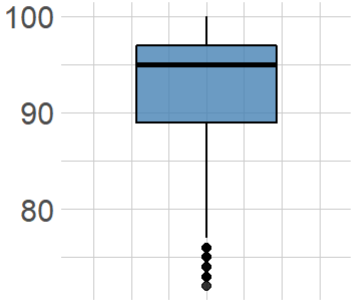
With outliers



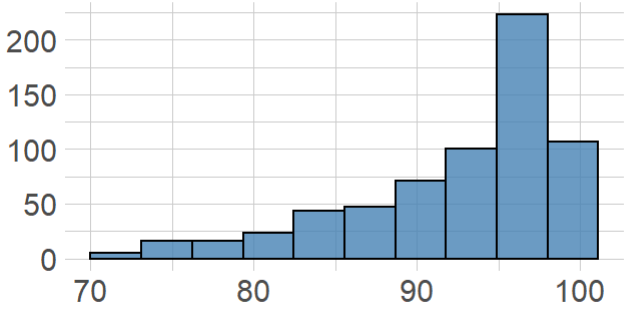
With outliers



Without outliers



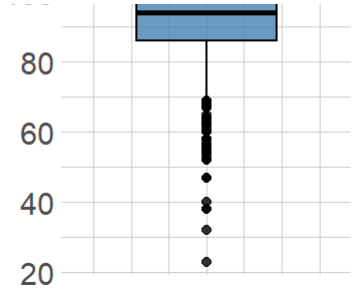
Without outliers

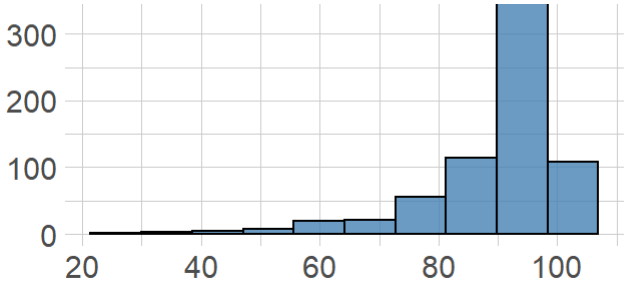


Outlier Diagnosis Plot (WithMMR)

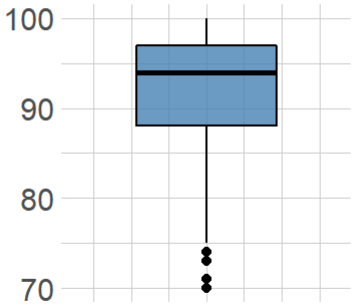
With outliers

With outliers

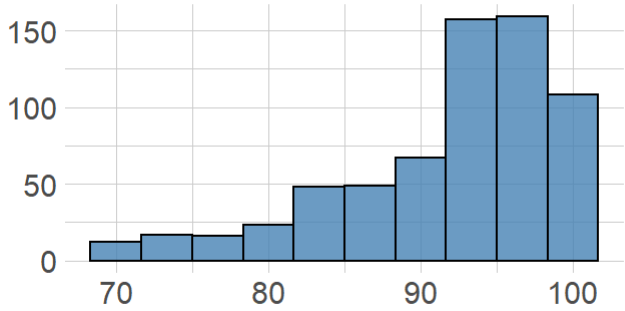




Without outliers

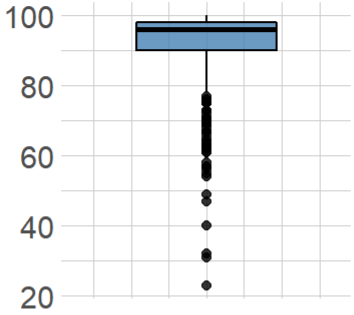


Without outliers

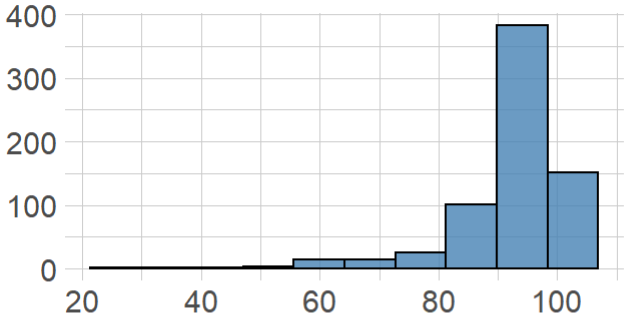


Outlier Diagnosis Plot (WithHepB)

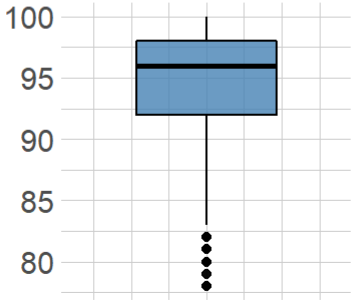
With outliers



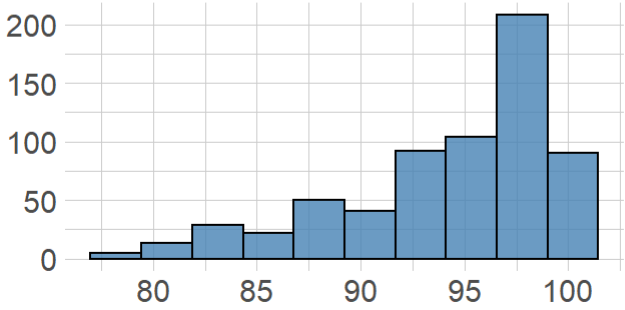
With outliers



Without outliers

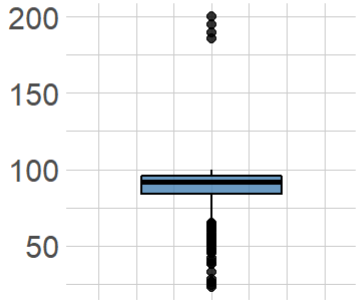


Without outliers

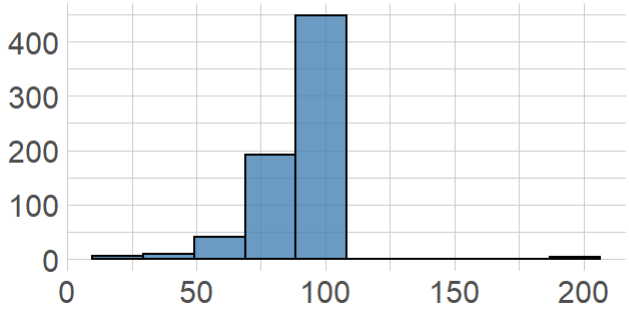


Outlier Diagnosis Plot (PctUpToDate)

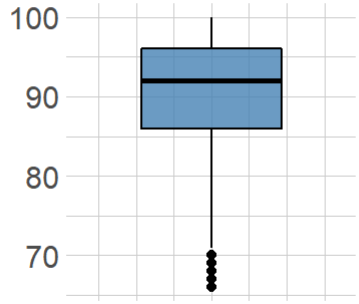
With outliers



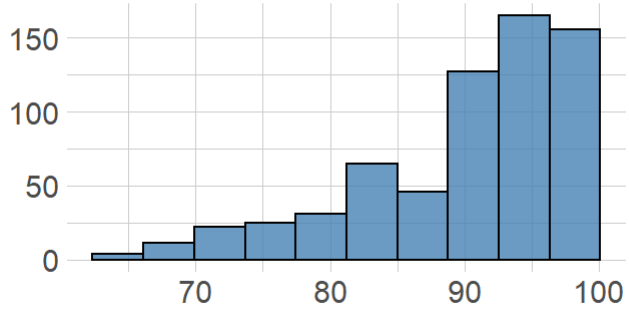
With outliers



Without outliers

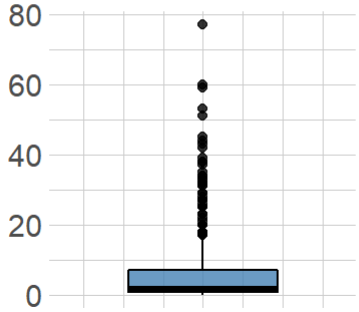


Without outliers

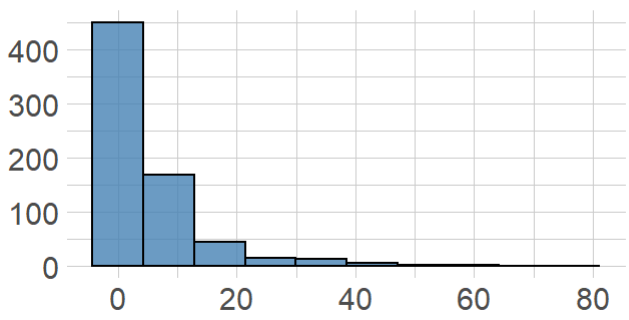


Outlier Diagnosis Plot (PctBeliefExempt)

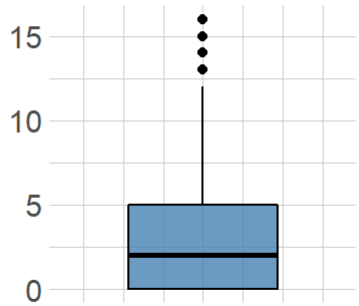
With outliers



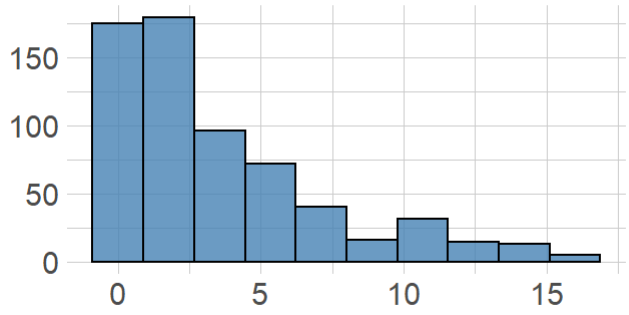
With outliers



Without outliers

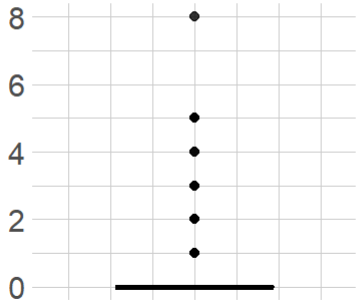


Without outliers

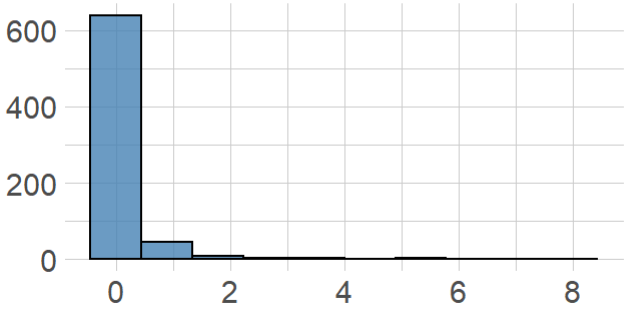


Outlier Diagnosis Plot (PctMedicalExempt)

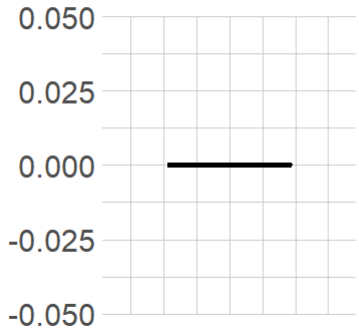
With outliers



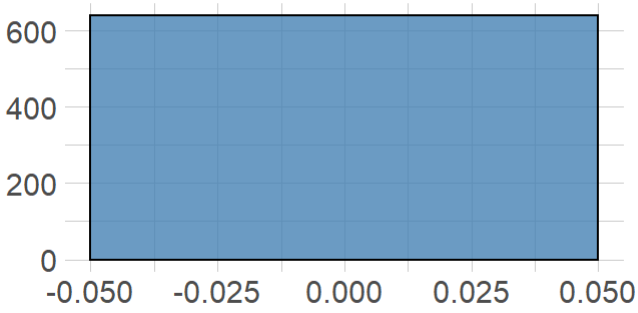
With outliers



Without outliers

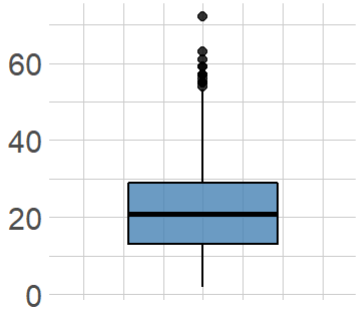


Without outliers

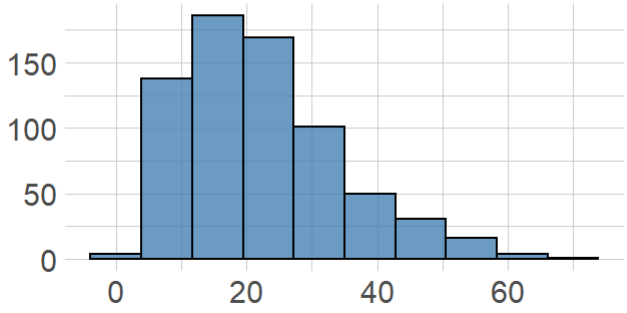


Outlier Diagnosis Plot (PctChildPoverty)

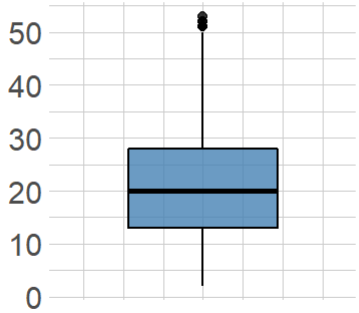
With outliers



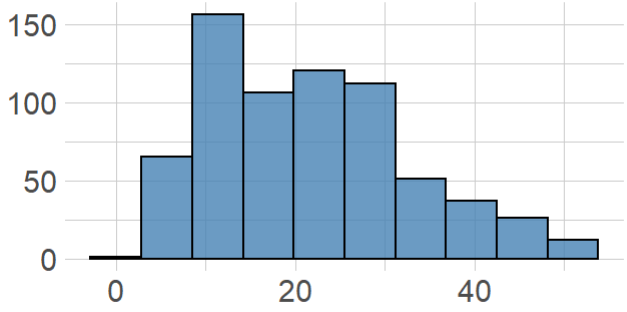
With outliers



Without outliers

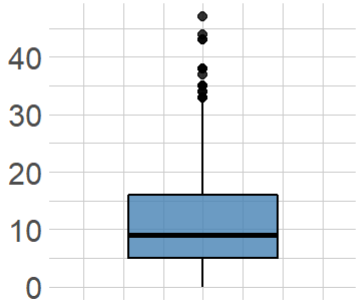


Without outliers

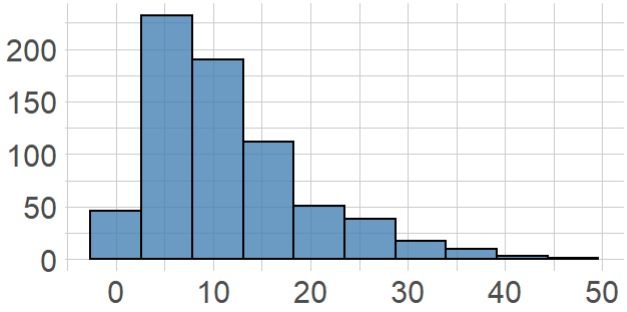


Outlier Diagnosis Plot (PctFamilyPoverty)

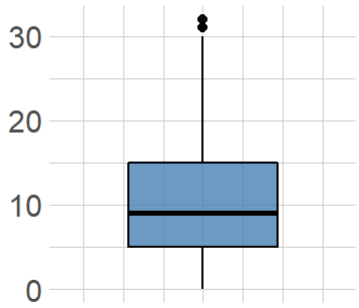
With outliers



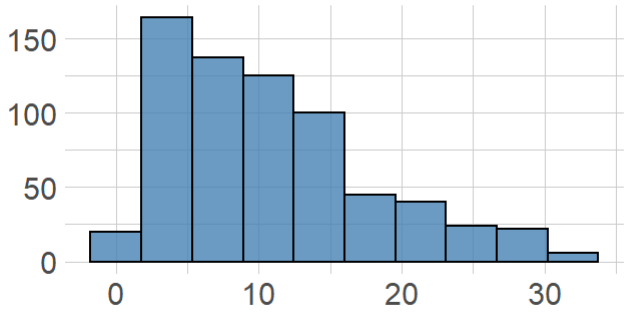
With outliers



Without outliers

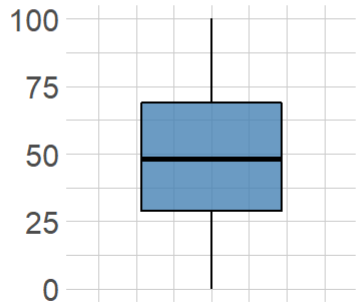


Without outliers

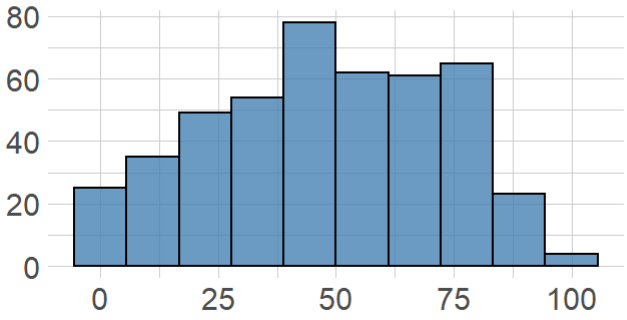


Outlier Diagnosis Plot (PctFreeMeal)

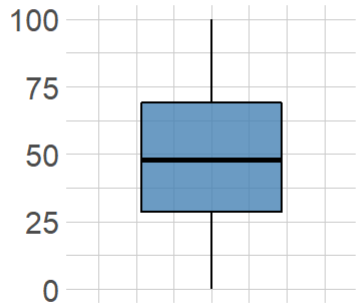
With outliers



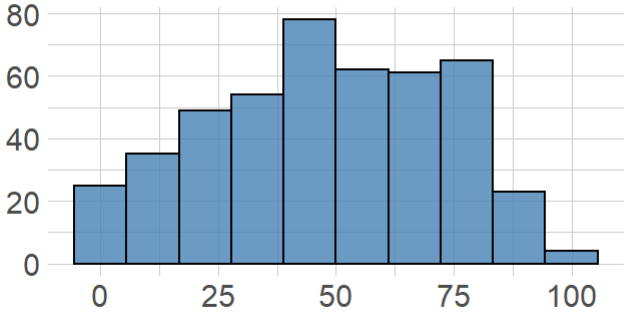
With outliers



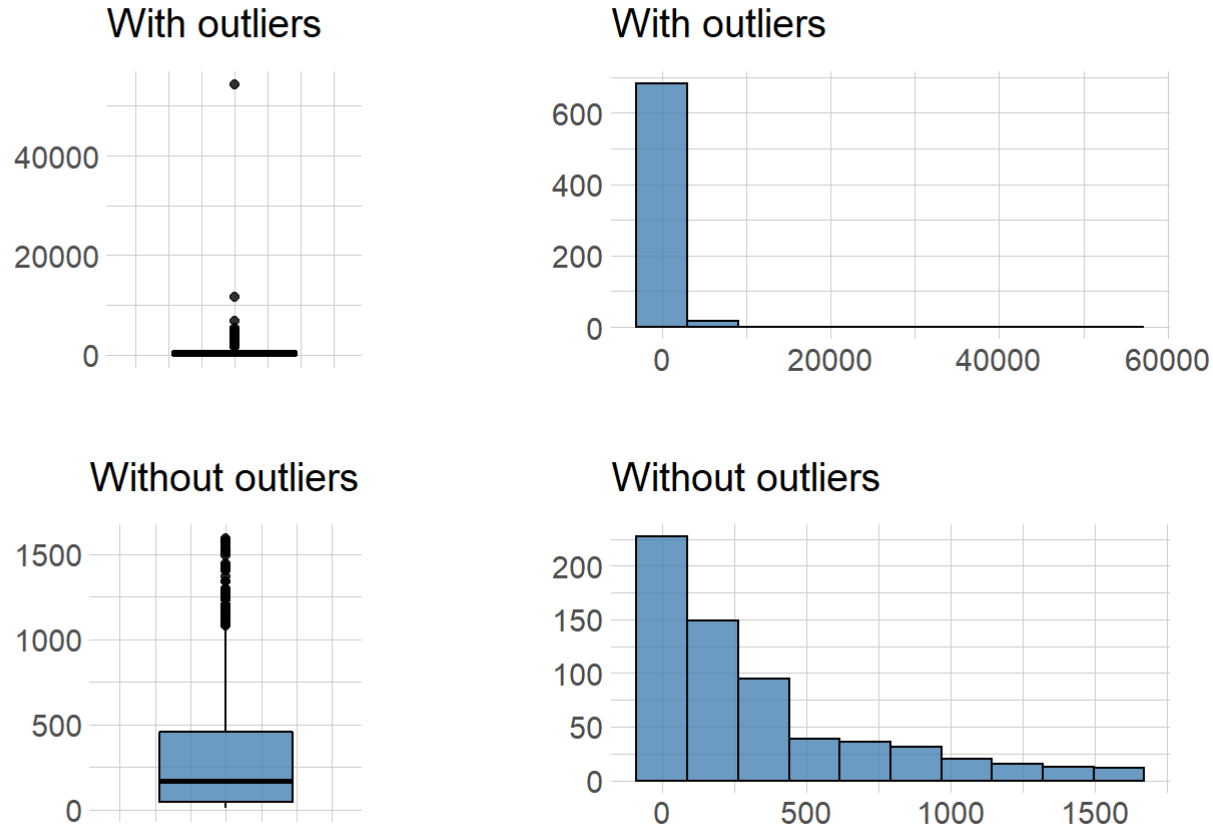
Without outliers



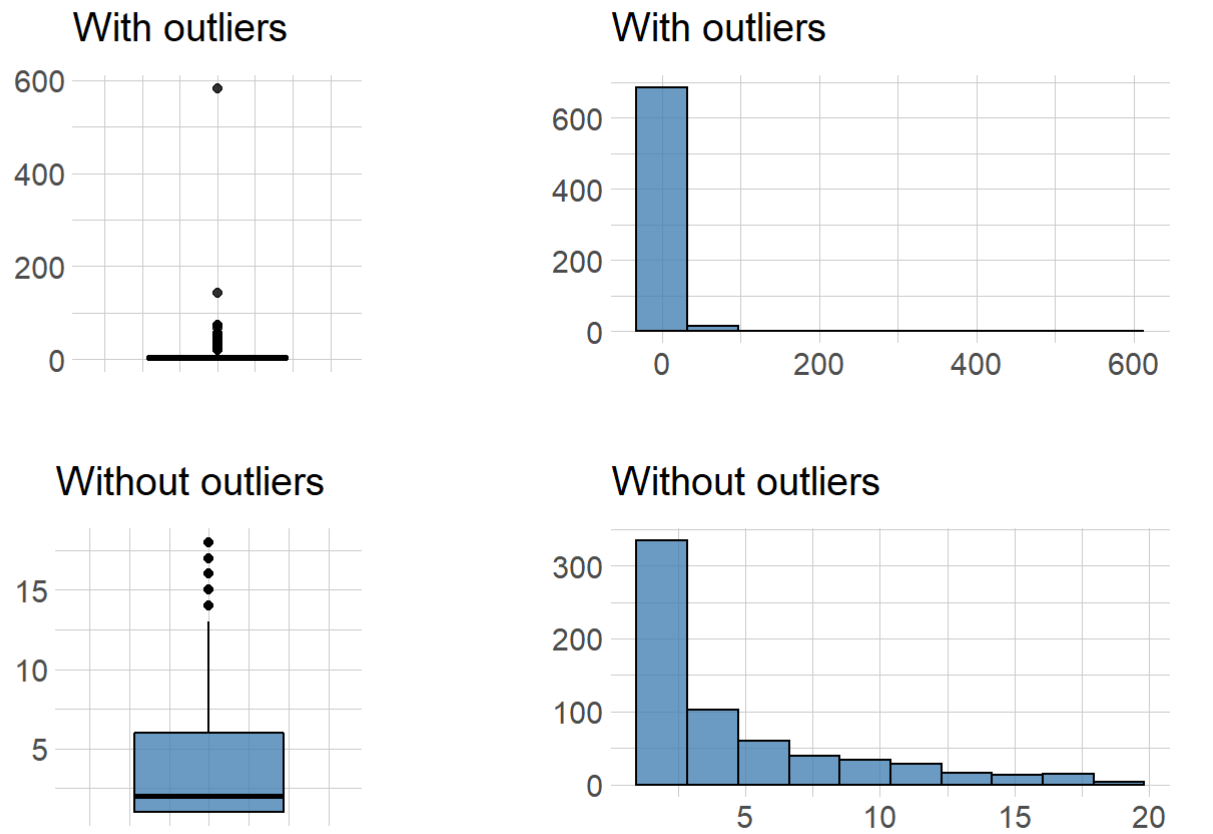
Without outliers



Outlier Diagnosis Plot (Enrolled)



Outlier Diagnosis Plot (TotalSchools)



We can see that maximum ouliers lie in column Enrolled and TotalSchools and rest lie around 45-50 (taking 50 as the average outliers as the diagnose shows us mostly all columns have around 45-50 outliers, we can see that around 7% of the data is comprising of outliers). Looking at the graphs for these columns, we can see that

despite removing outliers, the data is still skewed and only the scale changes or reduces in scale thereby making the bars thicker. Since we don't get a normal distribution post removal of outliers, we can keep the data for now and see if the removal is needed later. Eg: the outlier diagnosis plot of total schools with outliers is with a scale of 0 to 600 and without outliers is 0 to 20, but despite removing the outlier we can see that it is still right skewed.

4. Checking skewness

```
#skewness(districts_noNA %>% select(-DistrictName))
#with(Data_noType, apply(cbind(education, income, women, prestige,census), 2, skewness))
with(districts_noNA, apply(cbind(WithDTP,WithPolio,WithMMR, WithHepB,
                                 PctUpToDate, DistrictComplete,PctBeliefExempt
                                 ,PctMedicalExempt, PctChildPoverty,
                                 PctFamilyPoverty, Enrolled,
                                 TotalSchools), 2, skewness))
```

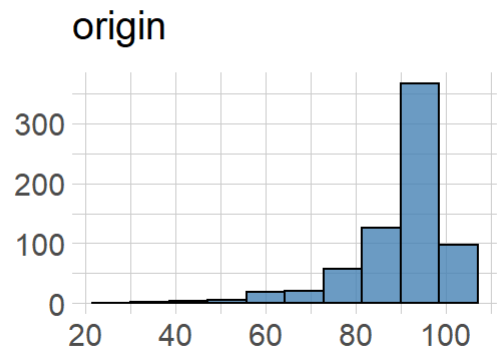
```
##          WithDTP        WithPolio          WithMMR         WithHepB
##       -2.1660928       -2.2652120       -2.1099440       -2.8230543
##      PctUpToDate DistrictComplete  PctBeliefExempt PctMedicalExempt
##        0.5686406       -3.6530150        3.2665435        6.7550566
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##        0.8317925        1.2139424       20.2670495       19.8999038
```

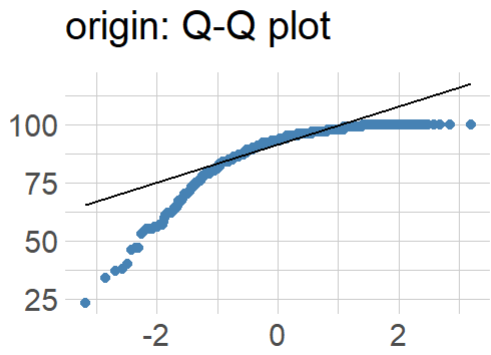
We can see high skewness for Enrolled and TotalSchools.

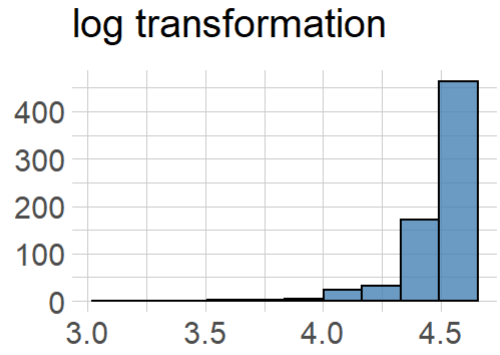
Checking which transformation can help reduce this skweness

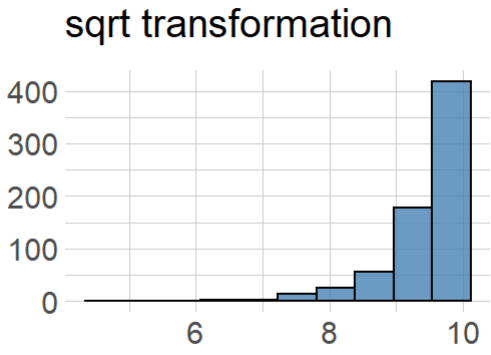
```
plot_normality(districts_noNA)
```

Normality Diagnosis Plot (WithDTP)

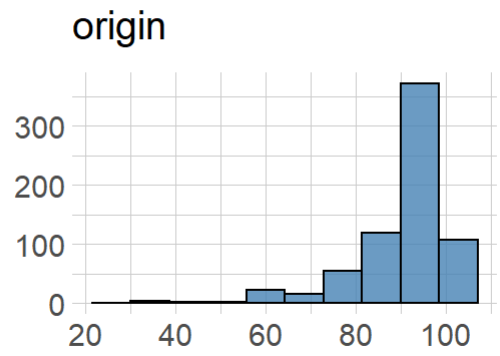


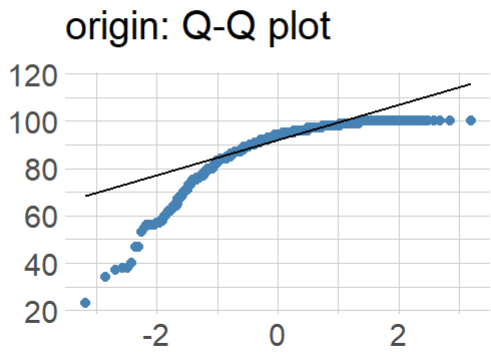


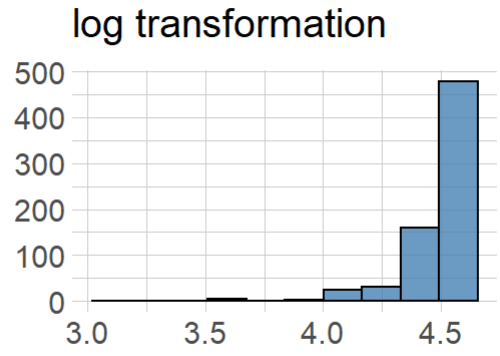


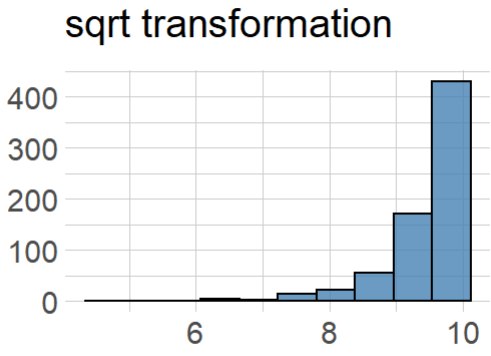


Normality Diagnosis Plot (WithPolio)





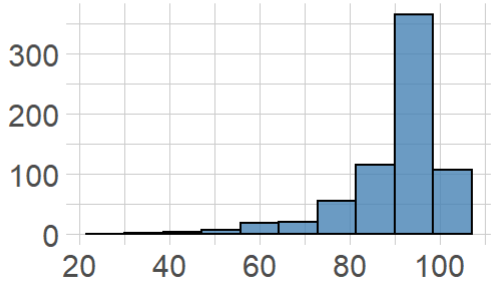


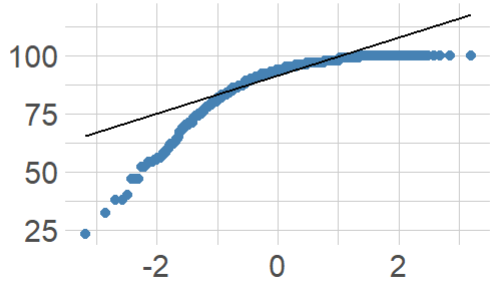


Normality Diagnosis Plot (WithMMR)

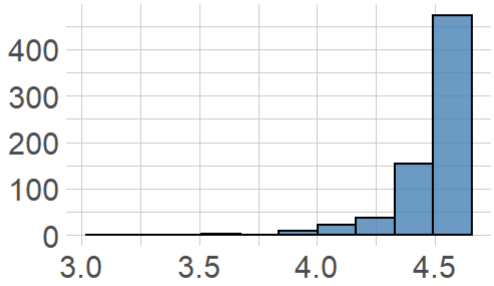
origin

origin: Q-Q plot

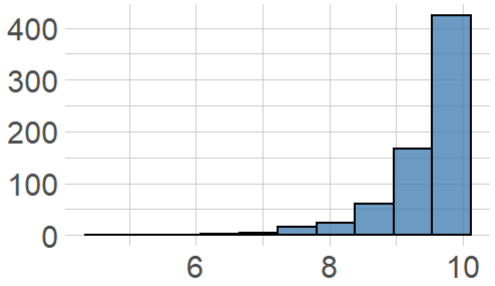




log transformation

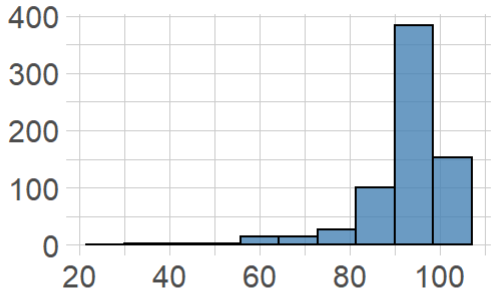


sqrt transformation

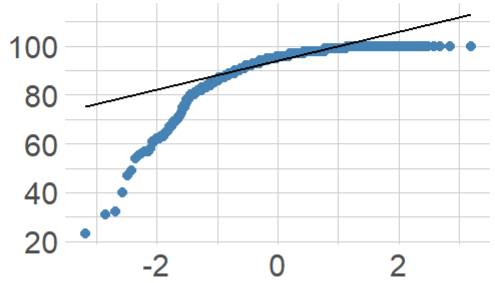


Normality Diagnosis Plot (WithHepB)

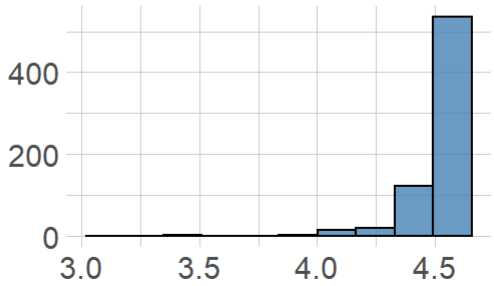
origin



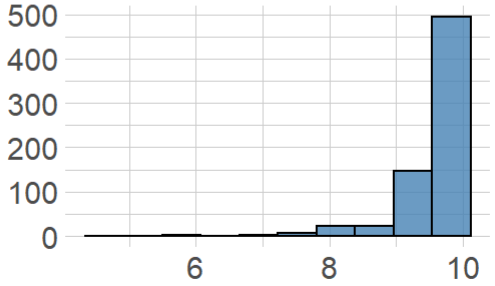
origin: Q-Q plot



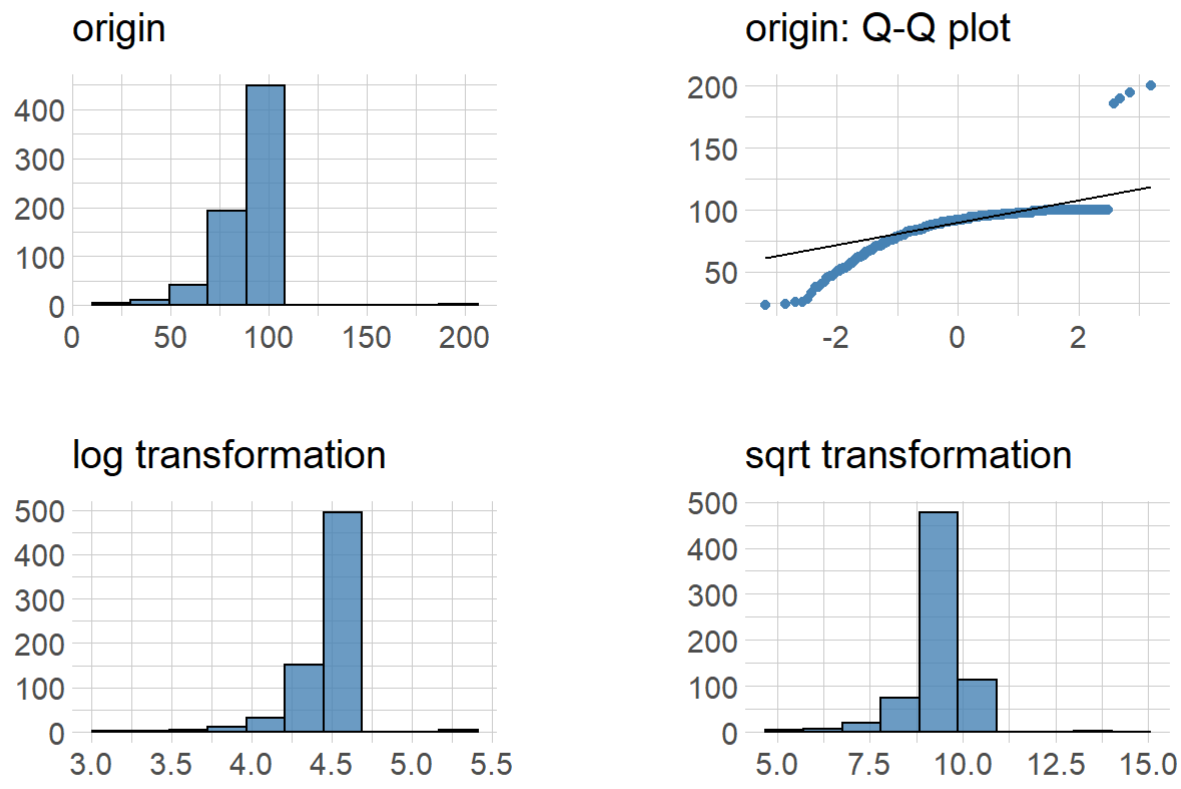
log transformation



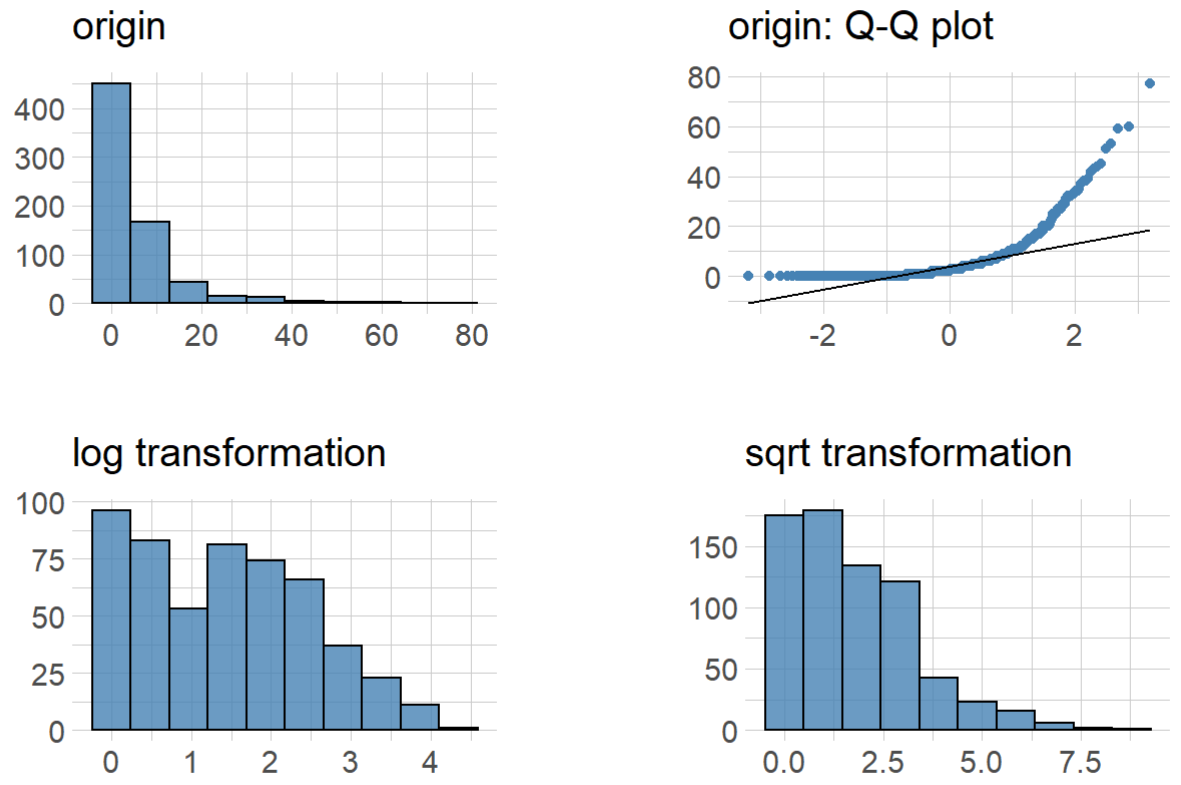
sqrt transformation



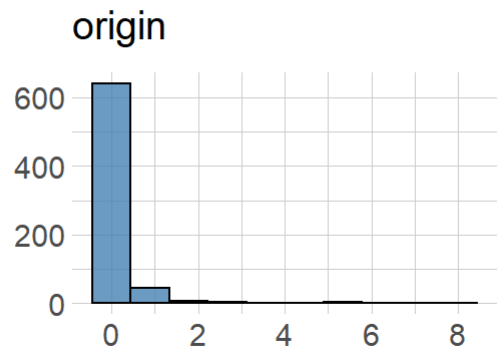
Normality Diagnosis Plot (PctUpToDate)

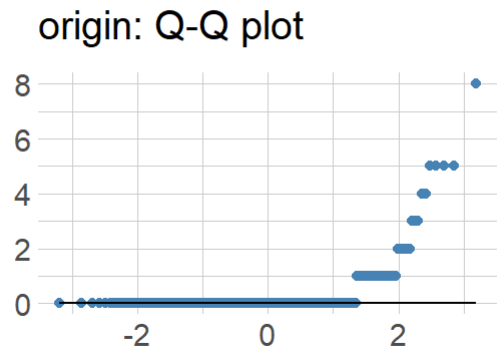


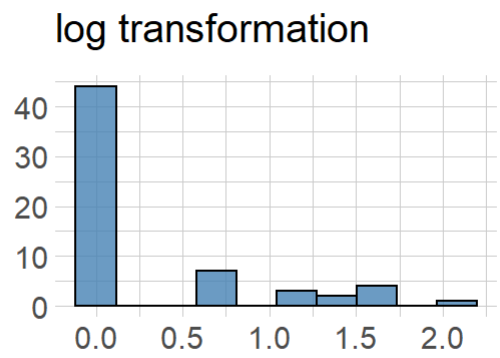
Normality Diagnosis Plot (PctBeliefExempt)

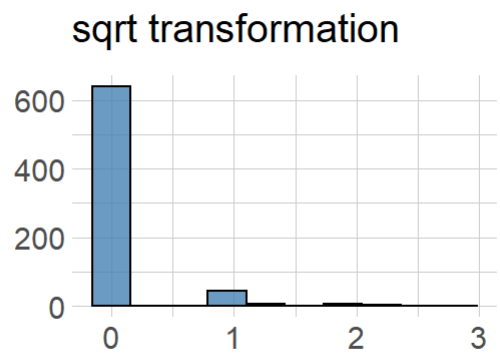


Normality Diagnosis Plot (PctMedicalExempt)

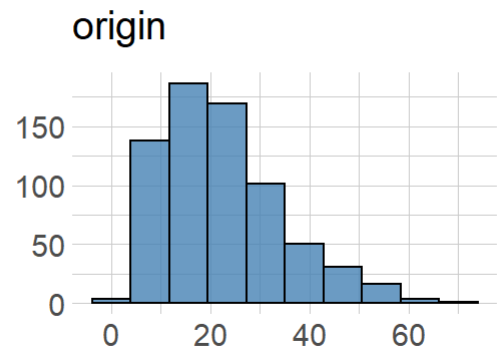


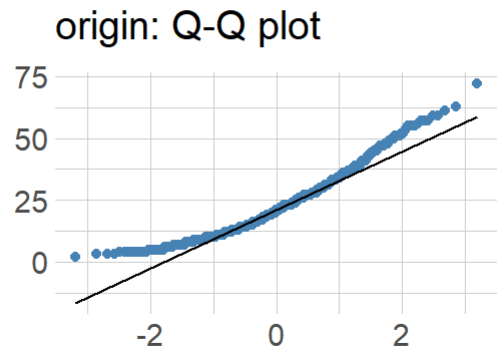


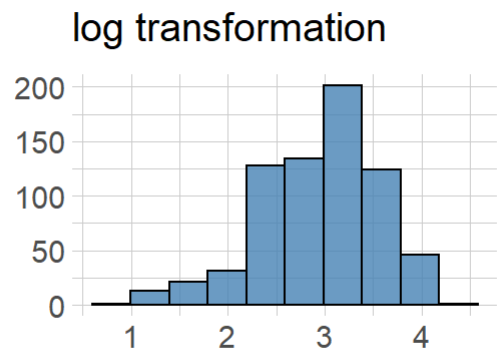


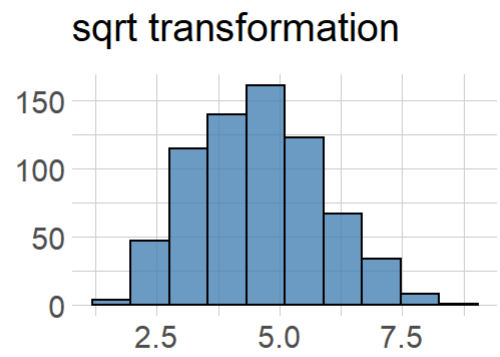


Normality Diagnosis Plot (PctChildPoverty)

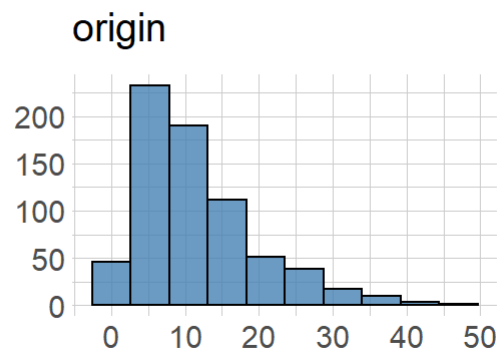


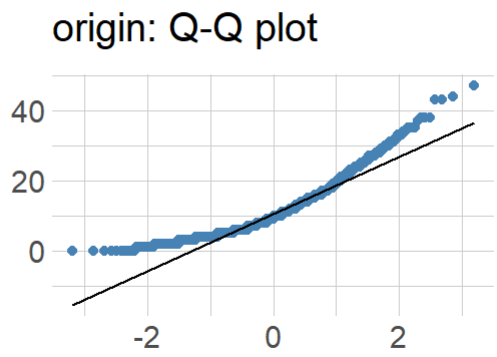


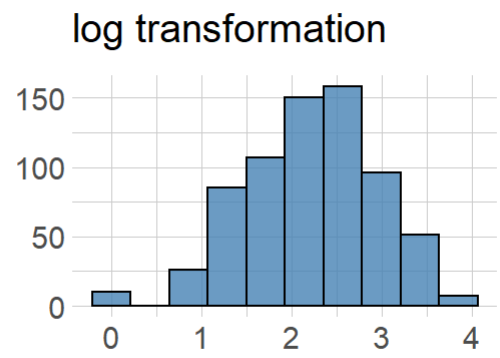


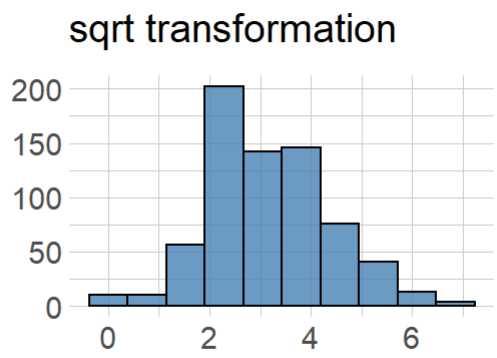


Normality Diagnosis Plot (PctFamilyPoverty)

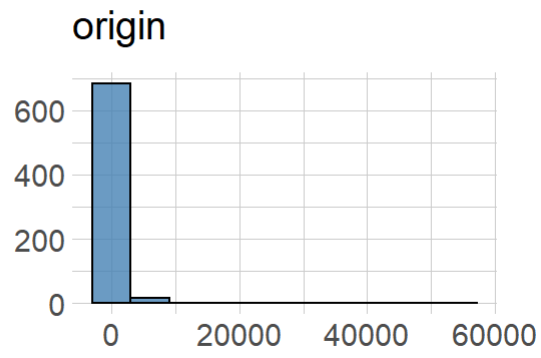


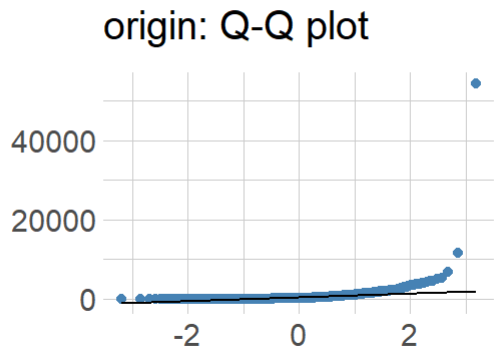


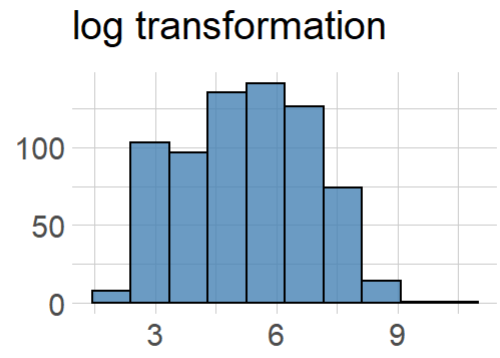


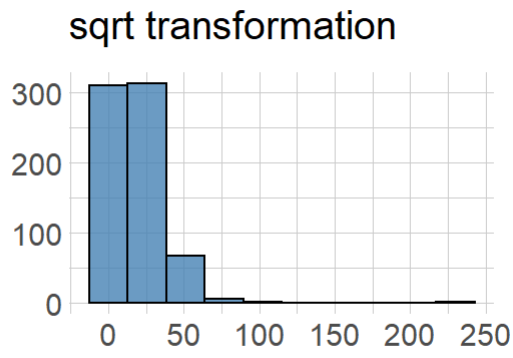


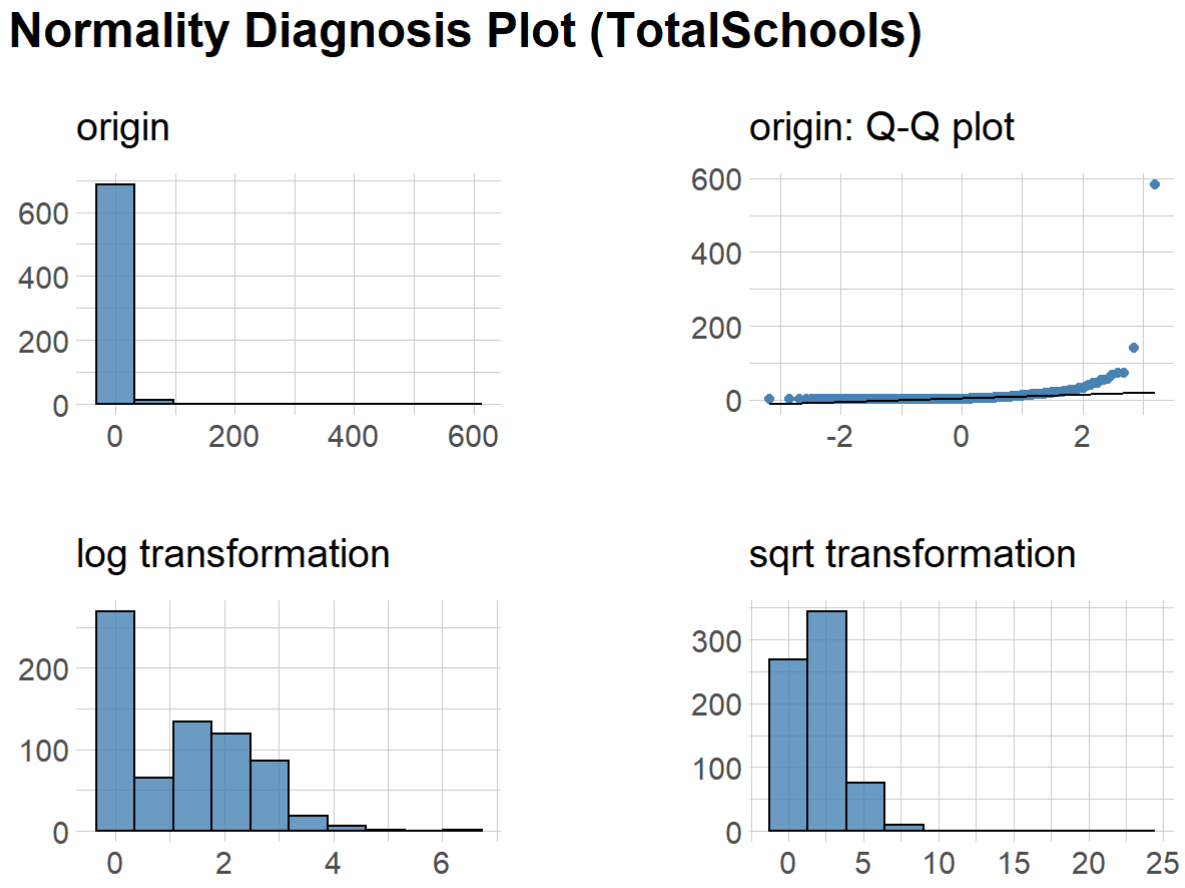
Normality Diagnosis Plot (Enrolled)











To remove this skewness lets create a new dataframe with the columns with log transformation instead and then lets check the skewness. We won't be taking log transformation on the other variables becuase there isn't alot of skewness and we don't want to mess up with the percent values and make it harder to interpret.

```
# taking out type variable so that we can take correlation in the next step
districts_log <- districts_noNA
# taking log transformation on the income column in the newly created dataset
districts_log$Enrolled <- log(districts_log$Enrolled)
districts_log$TotalSchools <- log(districts_log$TotalSchools)
with(districts_log, apply(cbind(WithDTP,WithPolio,WithMMR, WithHepB,
                                PctUpToDate, DistrictComplete,PctBeliefExempt
                                ,PctMedicalExempt, PctChildPoverty,
                                PctFamilyPoverty, Enrolled,
                                TotalSchools), 2, skewness))
```

```
##          WithDTP        WithPolio          WithMMR         WithHepB
##     -2.166092832     -2.265212049     -2.109944012     -2.823054321
##      PctUpToDate DistrictComplete  PctBeliefExempt PctMedicalExempt
##      0.568640615     -3.653015026      3.266543501      6.755056641
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##      0.831792543      1.213942412     -0.002078338      0.636935703
```

```
#skewness(subset(districts_log, select = -c(DistrictName) ))
```

As we can see the skewness reduced alot for the Enrolled and TotalSchools columns.

Let's check if the outliers were taken care of as well due to this for these two columns

```
diagnose_outlier(districts_log)
```

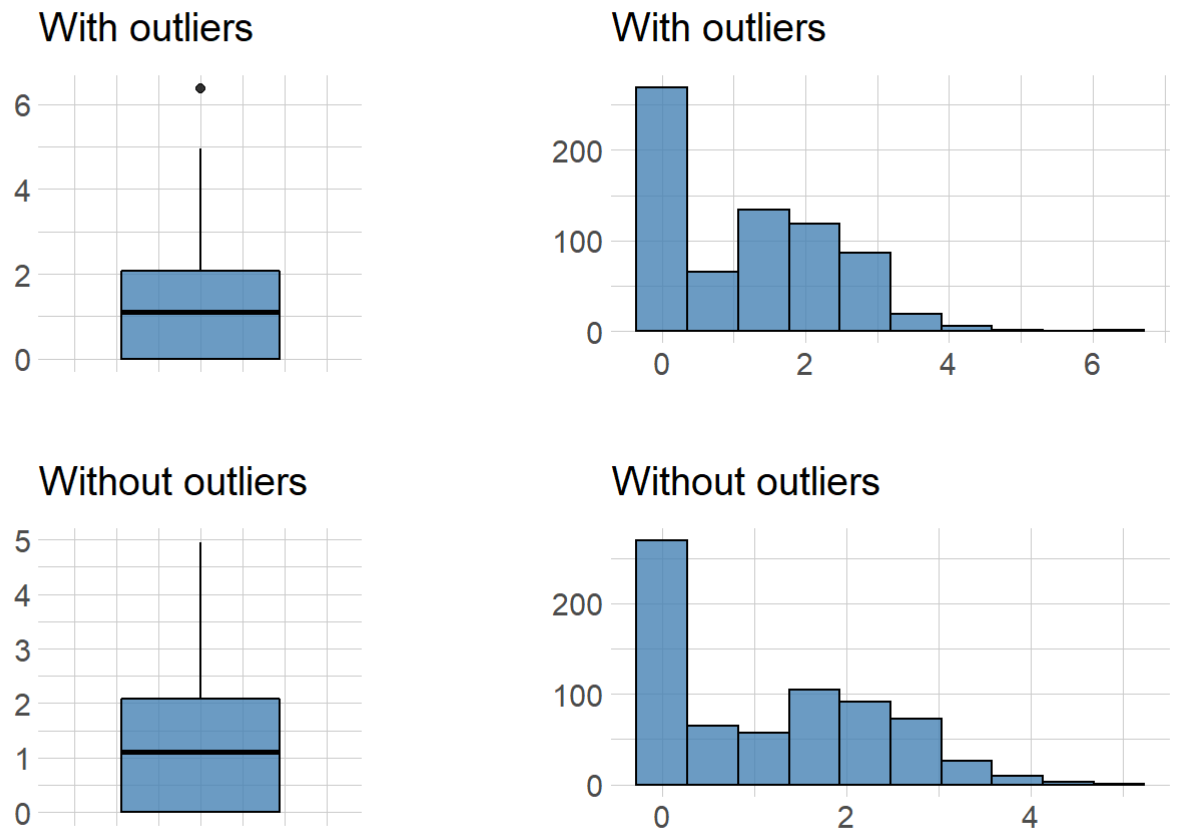
variables <chr>	outliers_cnt <int>	outliers_ratio <dbl>	outliers_mean <dbl>	with_mean <dbl>	without_mea <db
WithDTP	42	6.0000000	56.857143	89.7957143	91.89817
WithPolio	48	6.8571429	58.770833	90.2042857	92.51840
WithMMR	44	6.2857143	56.772727	89.7871429	92.00152
WithHepB	46	6.5714286	62.434783	92.2628571	94.36085
PctUpToDate	48	6.8571429	62.583333	88.4014286	90.30214
PctBeliefExempt	59	8.4285714	29.372881	5.6228571	3.43681
PctMedicalExempt	61	8.7142857	1.688525	0.1471429	0.00000
PctChildPoverty	14	2.0000000	58.214286	22.3257143	21.59329
PctFamilyPoverty	17	2.4285714	37.352941	11.4785714	10.83455
Enrolled	1	0.1428571	10.901137	5.2614283	5.25336

1-10 of 11 rows Previous **1** 2 Next

```
plot_outlier(subset(districts_log, select = c(Enrolled, TotalSchools) ))
```

Outlier Diagnosis Plot (Enrolled)

Outlier Diagnosis Plot (TotalSchools)



As we can see the outliers from 63 for Enrolled and 55 for TotalSchools has reduced to 1 which is great. Looking at the outlier plots for these two columns we see we are way close to normal distribution now due to log transformation and have removed the right skewness that existed in the original dataset.

5. Checking if the data is normal or not

```
normality(districts_log)
```

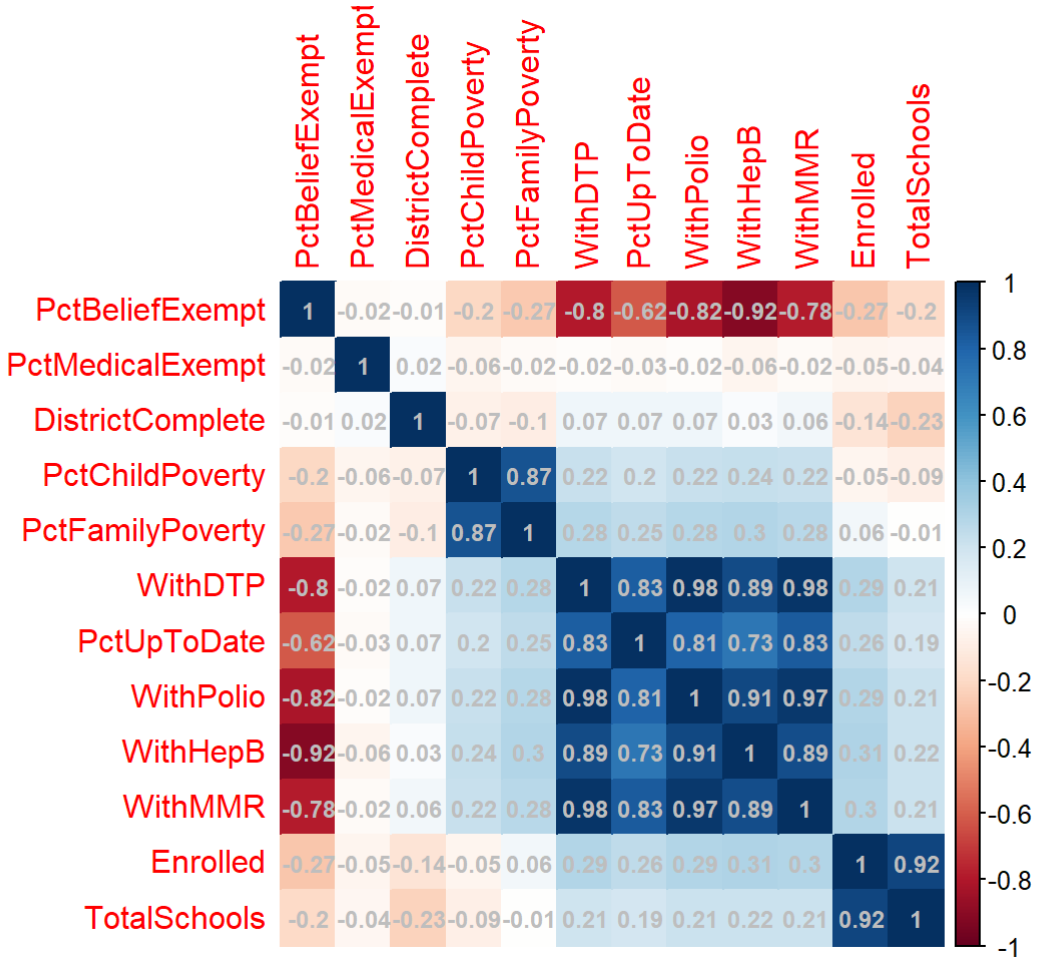
vars <chr>	statistic <dbl>	p_value <dbl>	sample <dbl>
WithDTP	0.7795654	1.186252e-29	700
WithPolio	0.7647191	1.931550e-30	700
WithMMR	0.7781442	9.929743e-30	700
WithHepB	0.7025956	2.301871e-33	700
PctUpToDate	0.7332535	5.459710e-32	700
PctBeliefExempt	0.6314522	3.649174e-36	700
PctMedicalExempt	0.2448039	2.176640e-46	700
PctChildPoverty	0.9494523	9.667296e-15	700
PctFamilyPoverty	0.9037986	1.418027e-20	700
Enrolled	0.9760526	2.732022e-09	700

1-10 of 11 rows Previous **1** 2 Next

The data looks good.

6. Checking correlation

```
#cor(districts_log %>% select(-c(DistrictName)))
#plot_correlate(districts_log %>% select(-c(DistrictName)))
corrplot(cor(districts_log %>% dplyr::select(-c(DistrictName))),
   method = "color",
   addCoef.col="grey",
   order = "AOE",
   number.cex=0.75)
```



We can see that DistrictComplete, PctMedicalExempt, PctFamilyPoverty , Enrolled, TotalSchools are not highly correlated, rest of the columns which are WithMMR, WithHepB, PctUpToDate have high positive correlation which makes sense as the parents getting their children vaccination for one kind of disease isn't an anti vaxer so will end up getting their children all the vaccines and keep the vaccines up to date and PctBeliefExempt has high negative correlation which makes sense as people with medical exemption won't be taking a vaccine or having thier vaccinations up to date.

Data Exploration Data Preprocessing and Cleaning For schools data set

1. Checking for NA's in the datasets

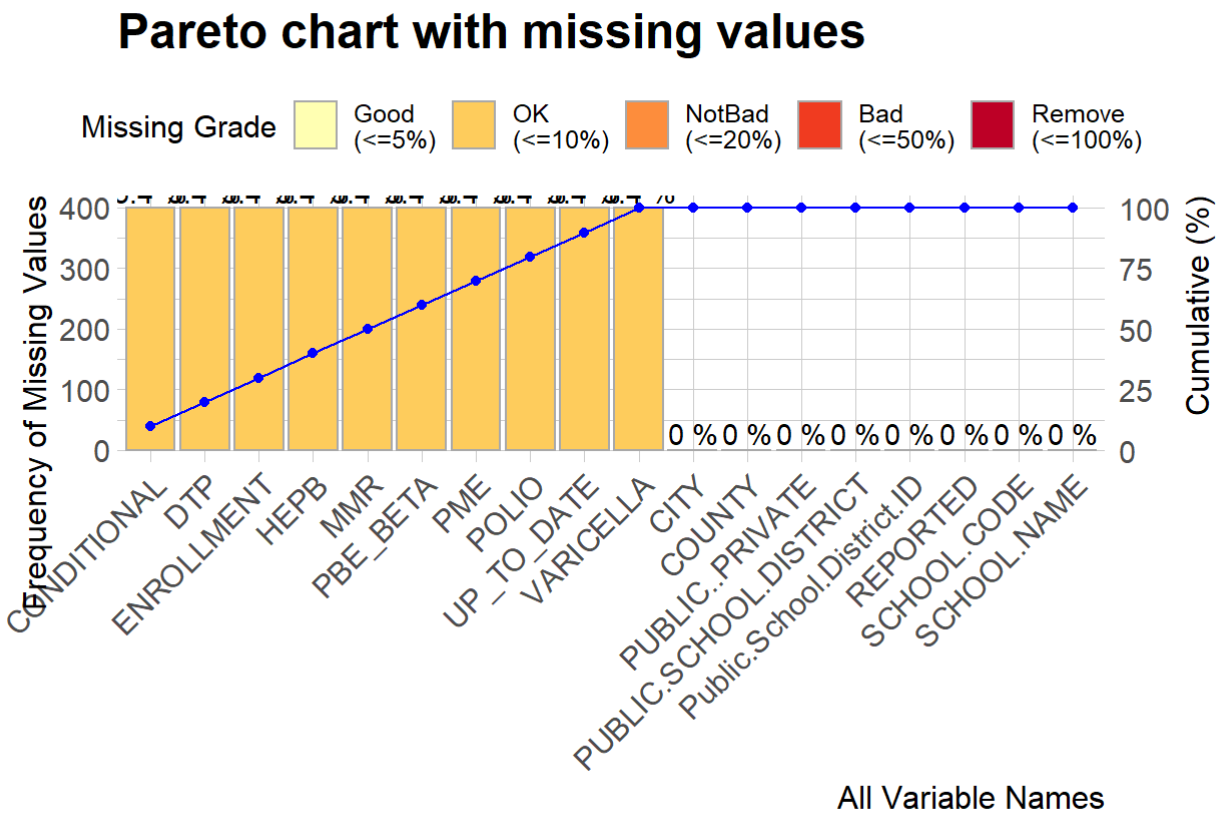
```
summary(schools)
```

```
##   SCHOOL.CODE      PUBLIC..PRIVATE    Public.School.District.ID
##  Min.   : 100016   Length:7381        Length:7381
##  1st Qu.:6016406   Class :character   Class :character
##  Median :6043806   Mode  :character   Mode  :character
##  Mean   :5546463
##  3rd Qu.:6116404
##  Max.   :7105125
##
##  PUBLIC.SCHOOL.DISTRICT     CITY              COUNTY
##  Length:7381            Length:7381        Length:7381
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##  SCHOOL.NAME          ENROLLMENT       UP_TO_DATE      CONDITIONAL
##  Length:7381        Min.   : 10.00   Min.   :  0.00   Min.   :  0.000
##  Class :character   1st Qu.: 41.00   1st Qu.: 34.00   1st Qu.:  0.000
##  Mode  :character   Median : 74.00   Median : 66.00   Median :  2.000
##                     Mean   : 75.99   Mean   : 68.56   Mean   :  4.931
##                     3rd Qu.:104.00   3rd Qu.: 95.00   3rd Qu.:  6.000
##                     Max.   :544.00   Max.   :350.00   Max.   :127.000
##                     NA's   :399      NA's   :399      NA's   :399
##       PME             PBE_BETA            DTP            POLIO
##  Min.   : 0.0000   Min.   :  0.000   Min.   :  0.0   Min.   :  0.00
##  1st Qu.: 0.0000   1st Qu.:  0.000   1st Qu.: 36.0   1st Qu.: 36.00
##  Median : 0.0000   Median :  1.000   Median : 68.0   Median : 68.00
##  Mean   : 0.1408   Mean   :  2.351   Mean   : 70.1   Mean   : 70.44
##  3rd Qu.: 0.0000   3rd Qu.:  3.000   3rd Qu.: 96.0   3rd Qu.: 97.00
##  Max.   :19.0000   Max.   :127.000   Max.   :395.0   Max.   :381.00
##  NA's   :399       NA's   :399       NA's   :399     NA's   :399
##       MMR              HEPB          VARICELLA        REPORTED
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Length:7381
##  1st Qu.: 36.00   1st Qu.: 37.00   1st Qu.: 37.00   Class :character
##  Median : 68.00   Median : 70.00   Median : 71.00   Mode  :character
##  Mean   : 70.21   Mean   : 72.06   Mean   : 72.44
##  3rd Qu.: 97.00   3rd Qu.: 99.00   3rd Qu.: 99.00
##  Max.   :381.00   Max.   :387.00   Max.   :374.00
##  NA's   :399      NA's   :399      NA's   :399
```

```
sum(is.na(schools))
```

```
## [1] 3990
```

```
schools %>% plot_na_pareto( col = "blue")
```



In the school dataset, there are a total of 3990 NA's and looking at the summary we can see that there are 10 columns each having 399 NA's (hence the number 3990) and the columns are ENROLLMENT, UP_TO_DATE, CONDITIONAL, PME, PBE_BETA, DTP, POLIO,MMR, HEPB and VARICELLA. Looking at the parento plot it seems like these aren't that much of an issue.

Let's check if we can find any pattern we can to which null's occur and if there are full rows of NA's as we can't just remove 3990 records and reduce the data set by half in the process and looking at the data manually we can see most of the records have reported as N. Checking if this pattern is true and checking how many such records exists with the full row N

```
View(schools[which(schools$REPORTED == 'N'), ] )
#schools %>% drop_na() -> schools_noNA

# checking how many such records are there
sum(schools$REPORTED == 'N')
```

```
## [1] 400
```

Removing these 400 records but since there are 399 NA's there is one good record in this 400 and we wont be dropping that.

```
schools %>% drop_na() -> schools_noNA
```

Rechecking our NA's

```
sum(is.na(schools_noNA))
```

```
## [1] 0
```

2.Checking for Null's in the datasets

```
sapply(schools,function(x) sum(is.null(x)))
```

```
##               SCHOOL.CODE          PUBLIC..PRIVATE Public.School.District.ID
##                         0                        0                         0
##    PUBLIC.SCHOOL.DISTRICT                     CITY                   COUNTY
##                         0                        0                         0
##               SCHOOL.NAME                ENROLLMENT                UP_TO_DATE
##                         0                        0                         0
##               CONDITIONAL                      PME                  PBE_BETA
##                         0                        0                         0
##                       DTP                    POLIO                       MMR
##                         0                        0                         0
##                      HEPB                VARICELLA                  REPORTED
##                         0                        0                         0
```

We have no NULL records in the dataset.

3. Checking for outliers

```
diagnose_outlier(schools)
```

variables <chr>	outliers_cnt <int>	outliers_ratio <dbl>	outliers_mean <dbl>	with_mean <dbl>	without_mean <dbl>
SCHOOL.CODE	2263	30.6598022	4.399404e+06	5.546463e+06	6.053652e+06
ENROLLMENT	49	0.6638667	2.417551e+02	7.598539e+01	7.481379e+01
UP_TO_DATE	54	0.7316082	2.203704e+02	6.856216e+01	6.737890e+01
CONDITIONAL	550	7.4515648	2.765818e+01	4.931252e+00	2.987873e+00
PME	601	8.1425281	1.635607e+00	1.407906e-01	0.000000e+00
PBE_BETA	498	6.7470532	1.473092e+01	2.351189e+00	1.400370e+00
DTP	58	0.7858014	2.213793e+02	7.009797e+01	6.883073e+01
POLIO	54	0.7316082	2.239074e+02	7.043999e+01	6.924379e+01
MMR	55	0.7451565	2.228727e+02	7.021398e+01	6.900188e+01
HEPB	54	0.7316082	2.280741e+02	7.205643e+01	7.084036e+01

1-10 of 11 rows Previous **1** 2 Next

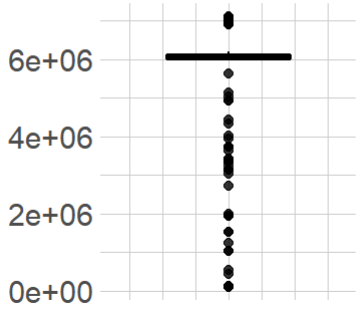
It seems like there are a lot of outliers in school code, conditional, pme and pbe_beta.

Plotting and checking if removing these outliers is a good idea or not

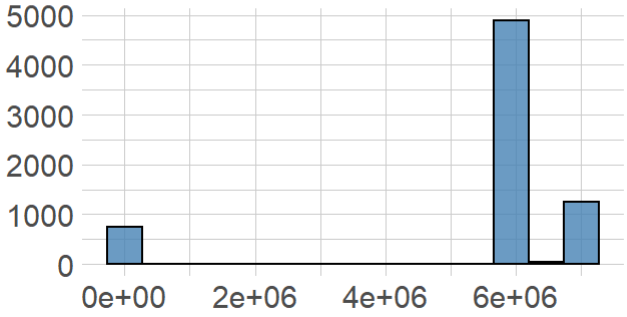
```
plot_outlier(schools_noNA)
```

Outlier Diagnosis Plot (SCHOOL.CODE)

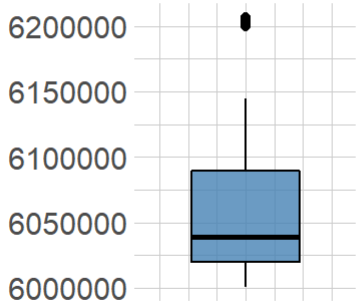
With outliers



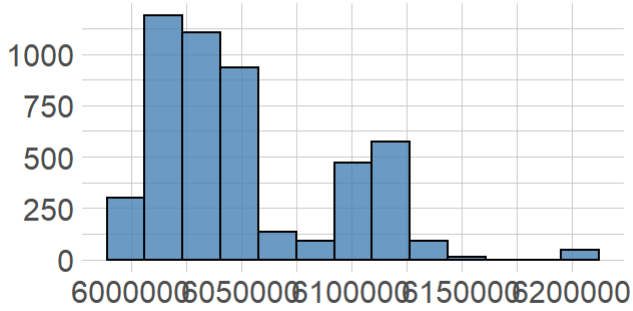
With outliers



Without outliers

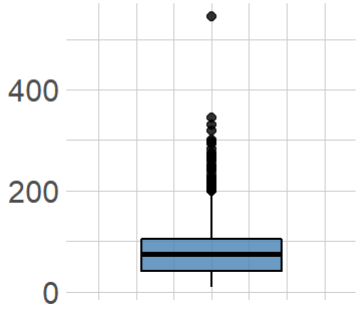


Without outliers

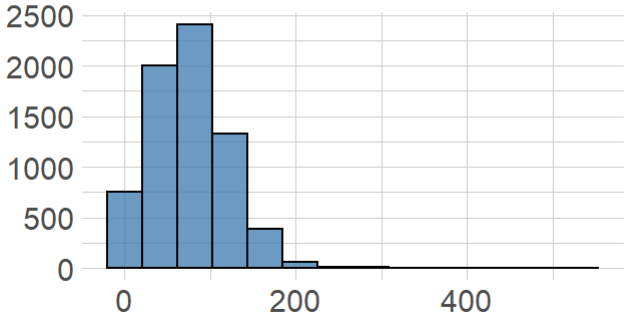


Outlier Diagnosis Plot (ENROLLMENT)

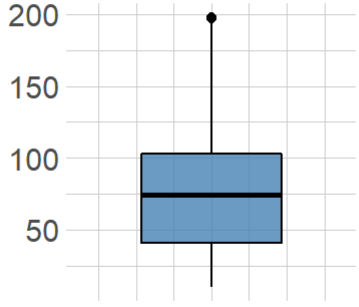
With outliers



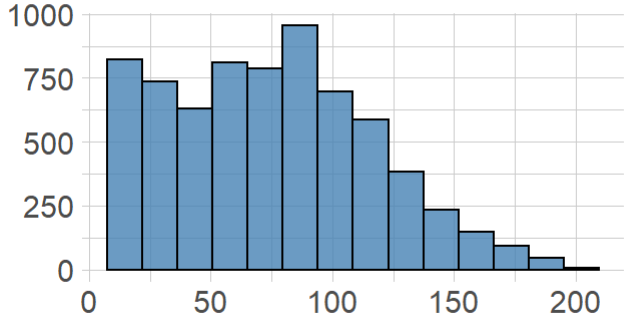
With outliers



Without outliers



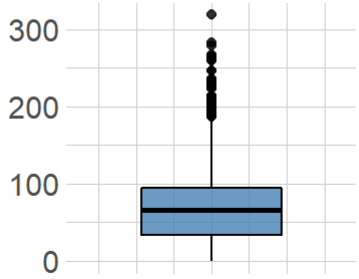
Without outliers

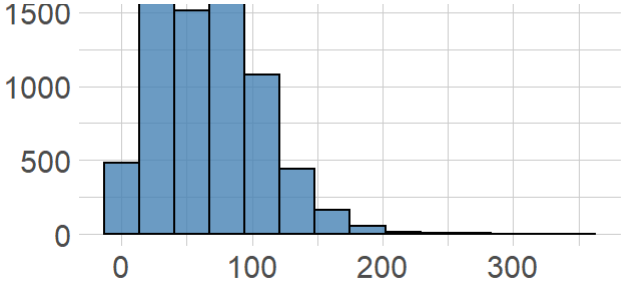


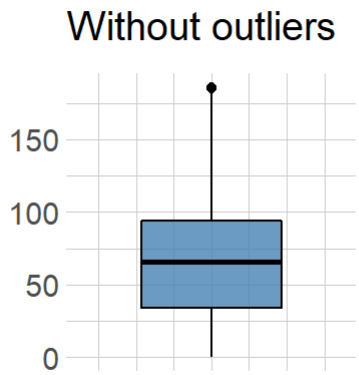
Outlier Diagnosis Plot (UP_TO_DATE)

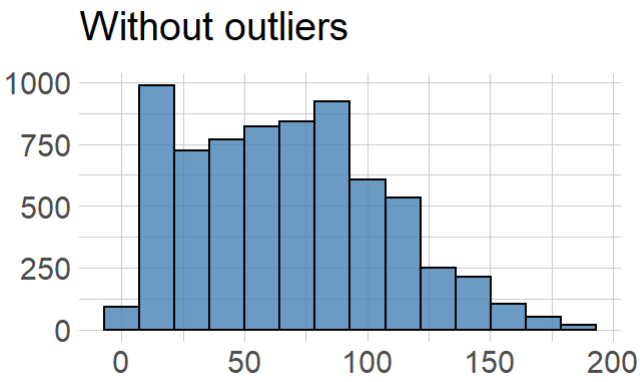
With outliers

With outliers

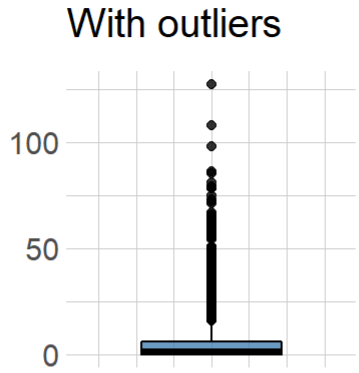


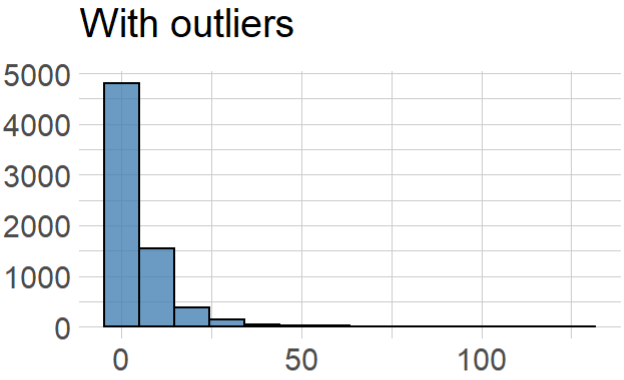


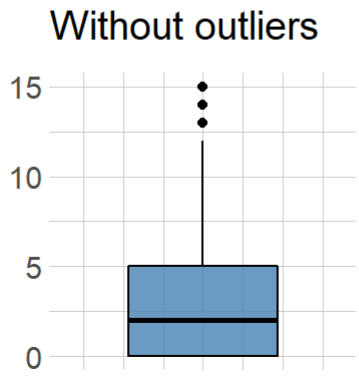


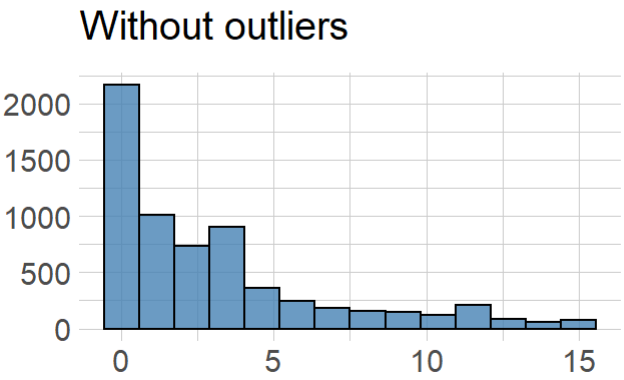


Outlier Diagnosis Plot (CONDITIONAL)

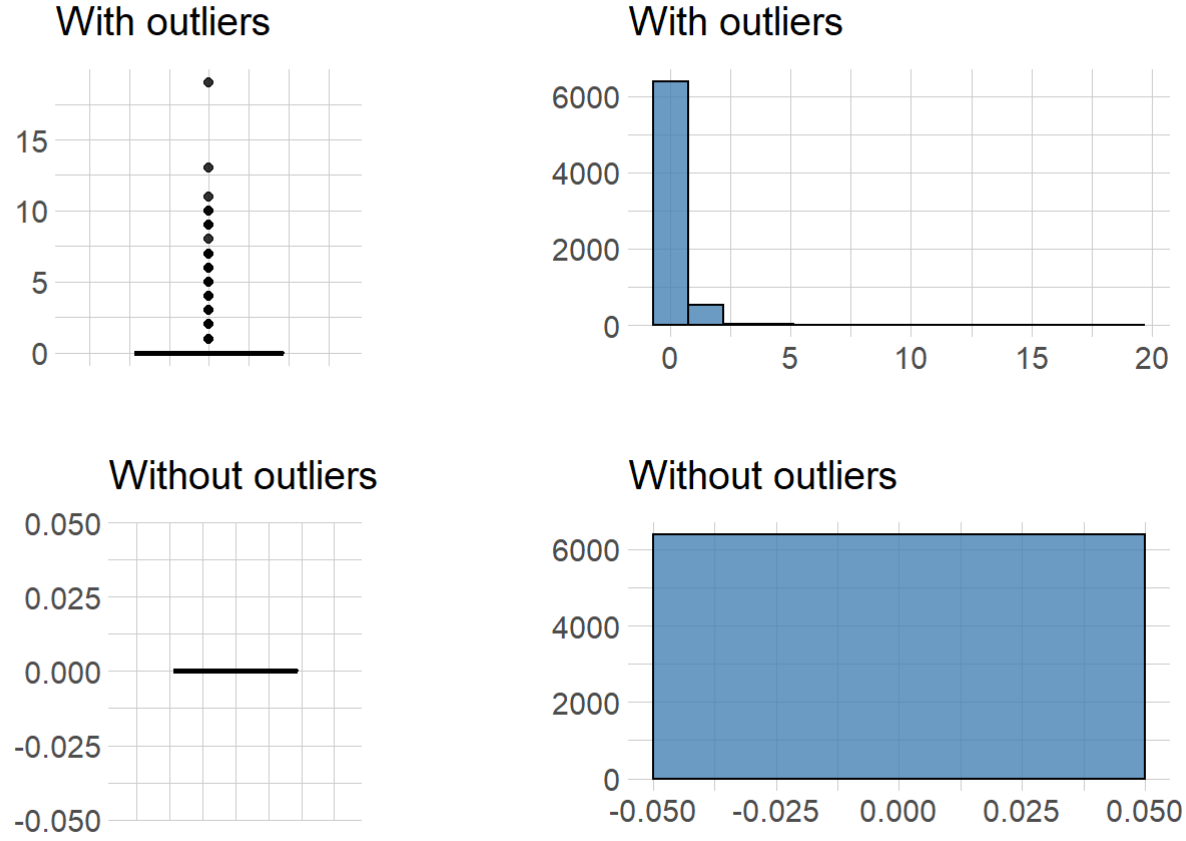




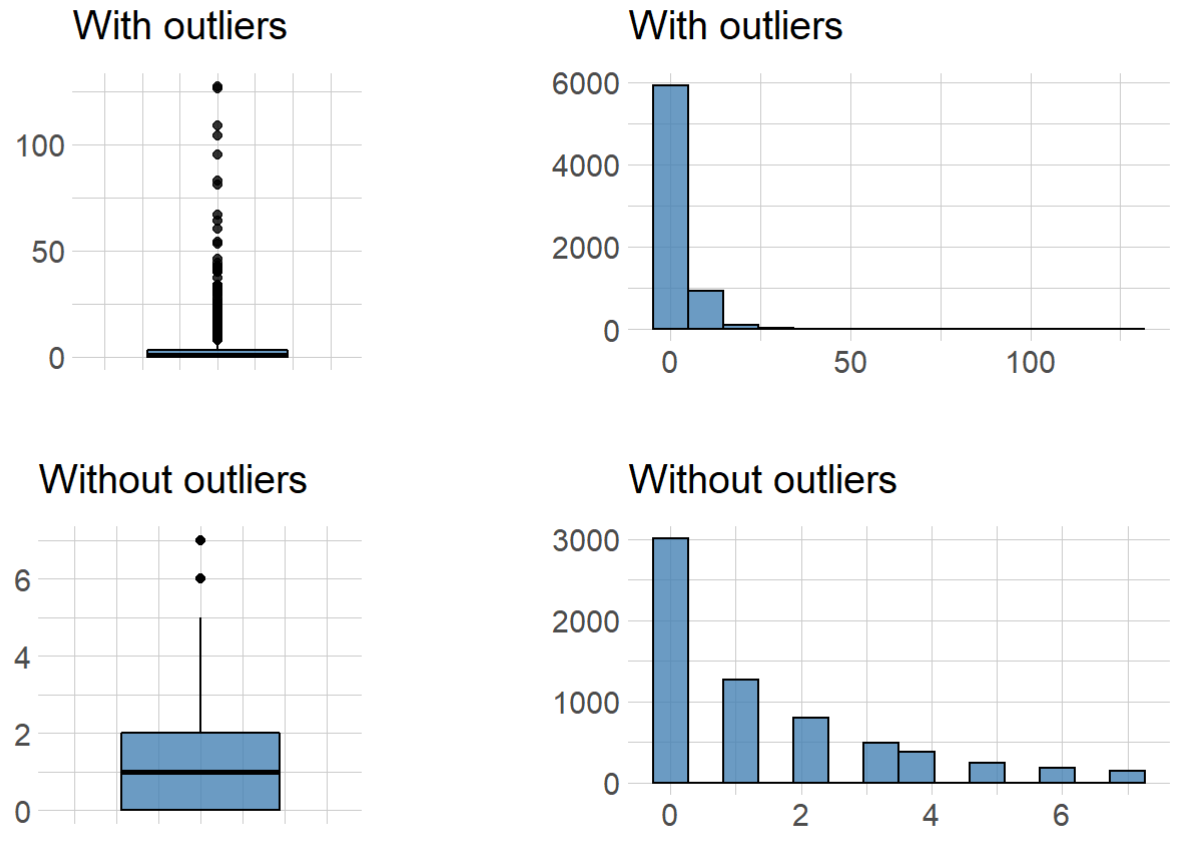




Outlier Diagnosis Plot (PME)

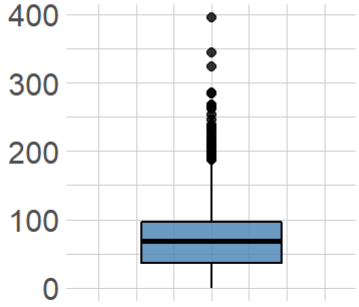


Outlier Diagnosis Plot (PBE_BETA)

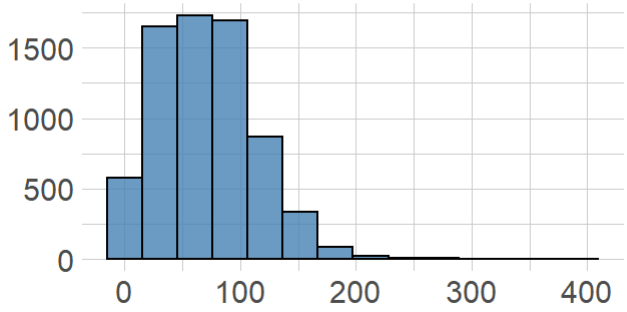


Outlier Diagnosis Plot (DTP)

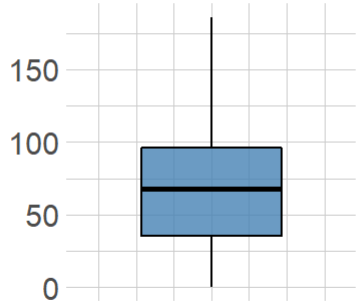
With outliers



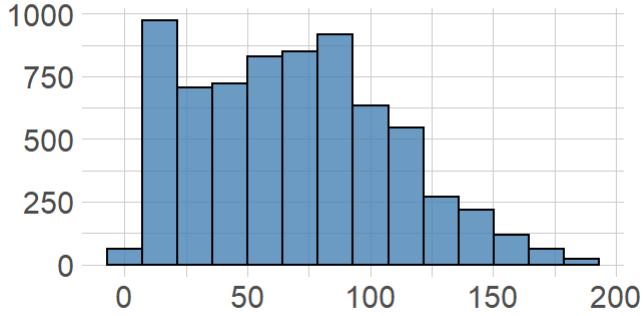
With outliers



Without outliers

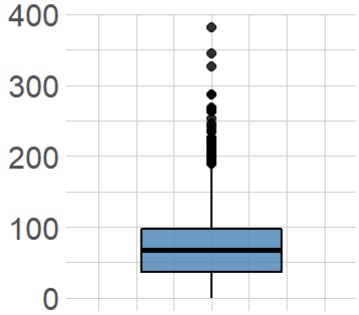


Without outliers

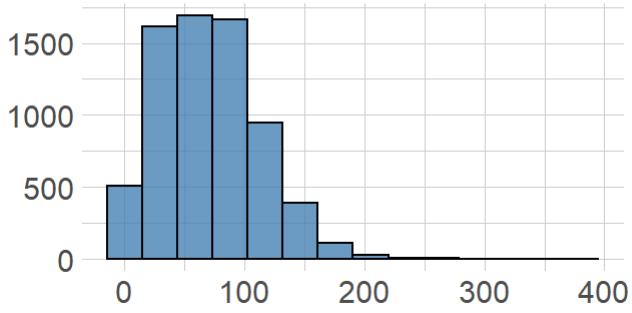


Outlier Diagnosis Plot (POLIO)

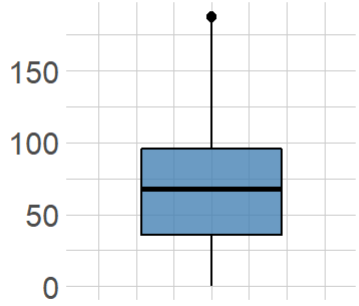
With outliers



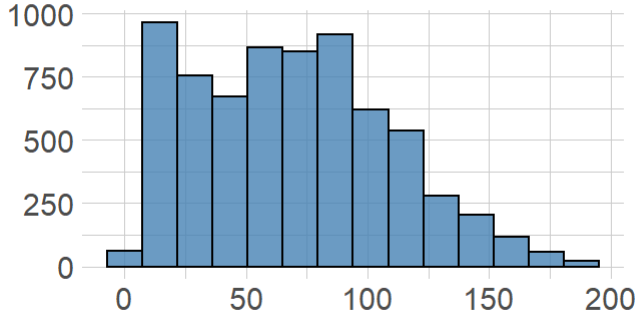
With outliers



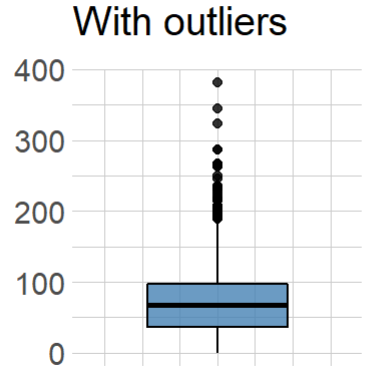
Without outliers

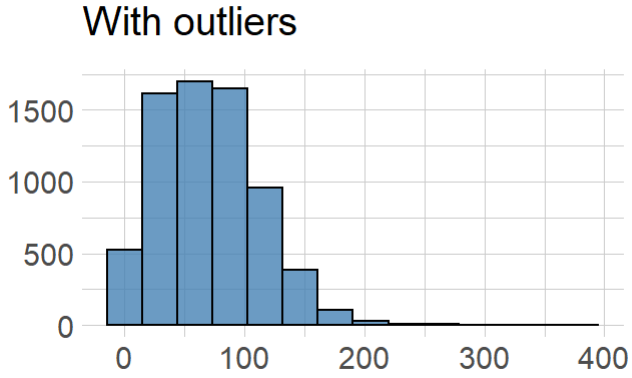


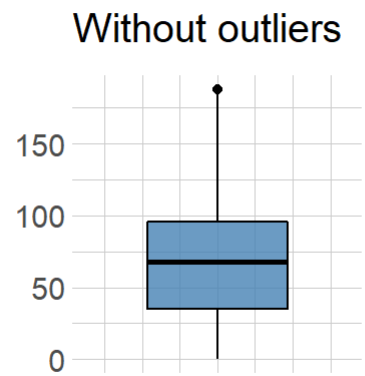
Without outliers

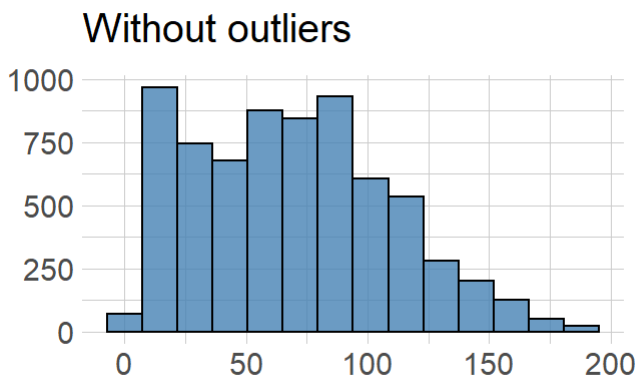


Outlier Diagnosis Plot (MMR)

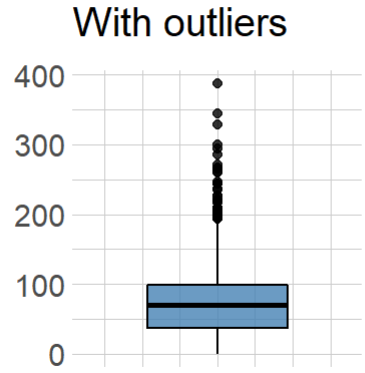


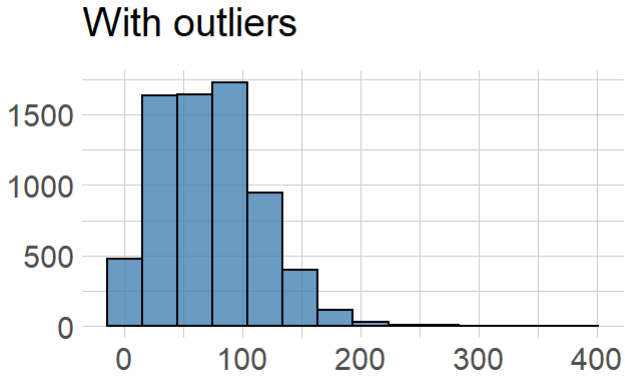


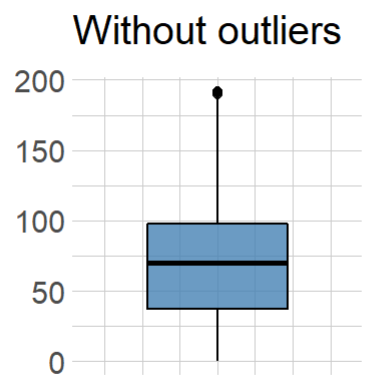


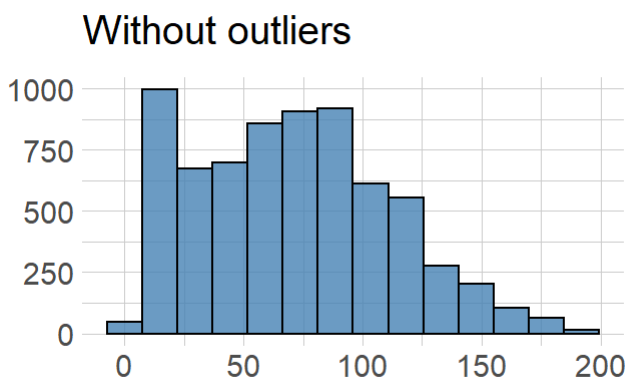


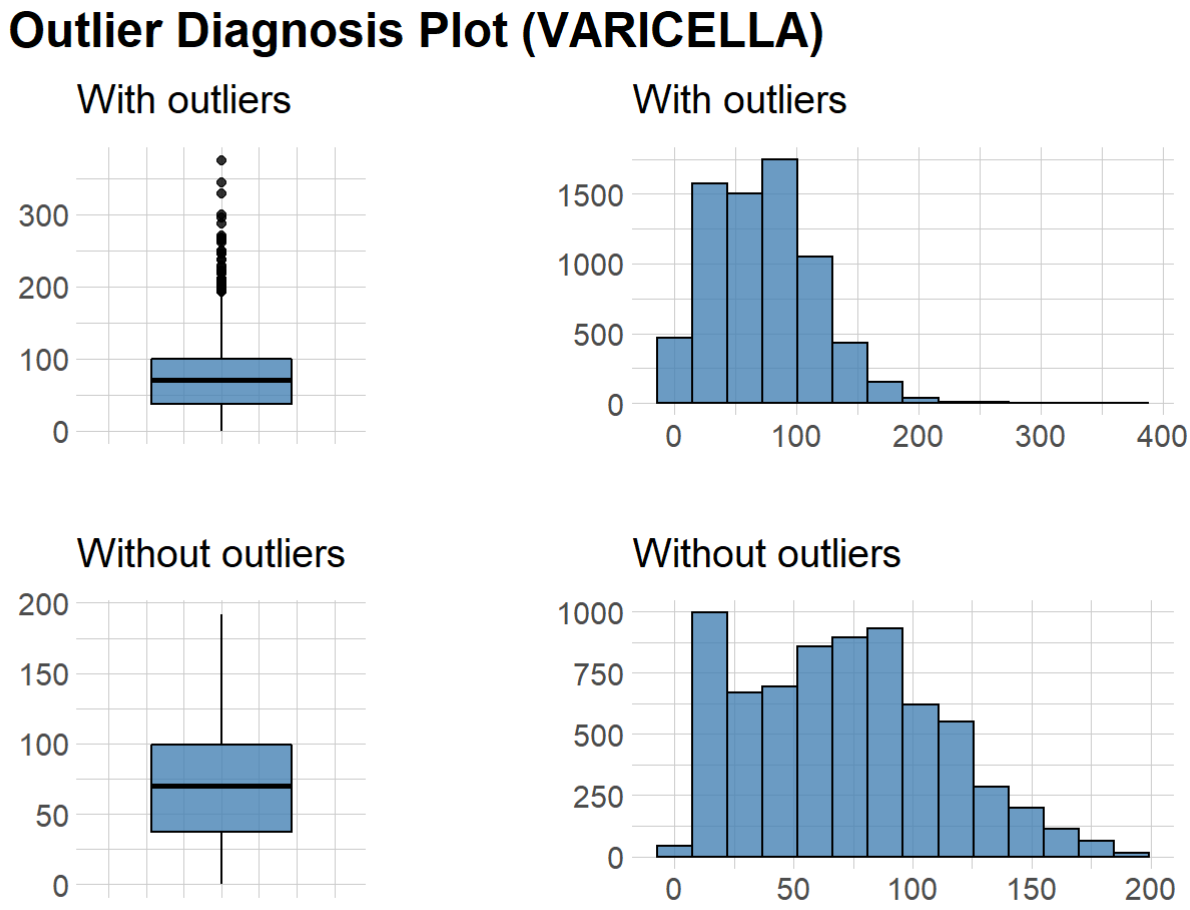
Outlier Diagnosis Plot (HEPB)











We can see that removing the outliers in school code column makes the graph bimodal rather than left skewed, almost normal for enrollment, varicella, HEPB, MMR, Polio, DPT, Up_TO_DATE, conditional remains right skewed, PME becomes uniform, PBE_BETA remains skewed to the right.

4. Checking skewness

```
skewness(select_if(schools_noNA, is.numeric))
```

```
## Warning in mean.default(x): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(x, ...): argument is not numeric or logical: returning
## NA
```

```
## Warning in mean.default((x - mean(x, ...))^2): argument is not numeric or
## logical: returning NA
```

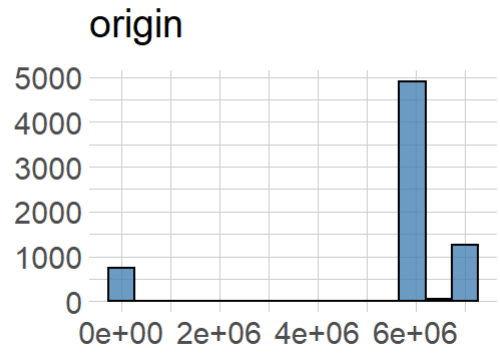
```
## [1] NA
```

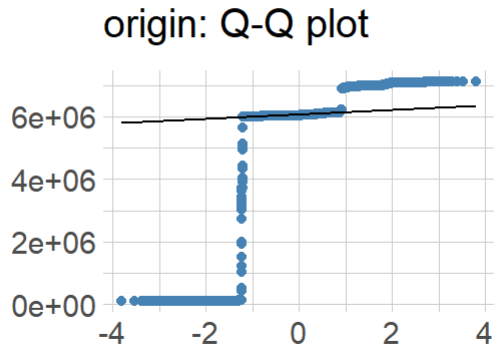
PME and PBE_BETA seem to have high correlation and there is some skewness in school code and conditional. On further inspection, we can see that school code is School ID number which doesn't seem important so we wont be trying to reduce skewness for this.

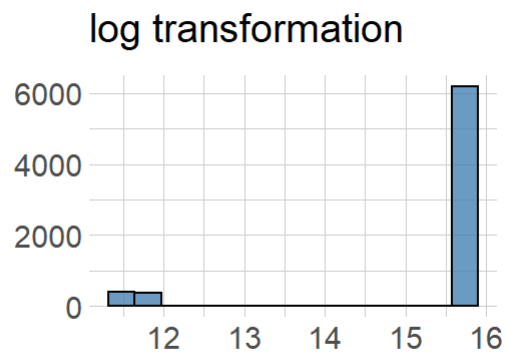
Checking which transformation can lead to normal data

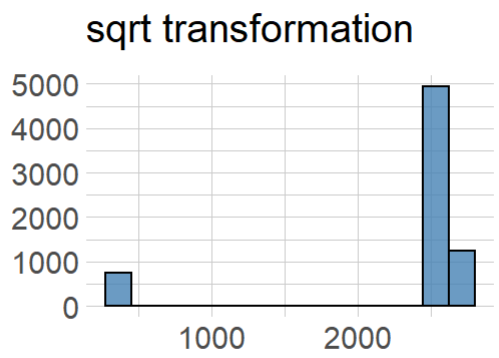
```
plot_normality(schools_noNA)
```

Normality Diagnosis Plot (SCHOOL.CODE)

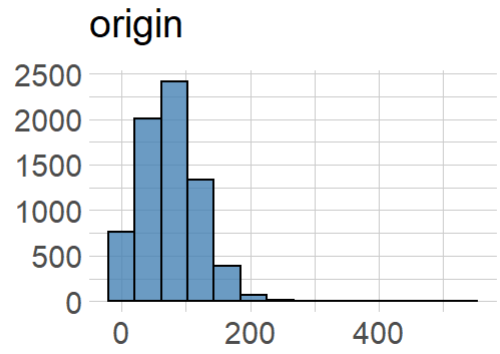


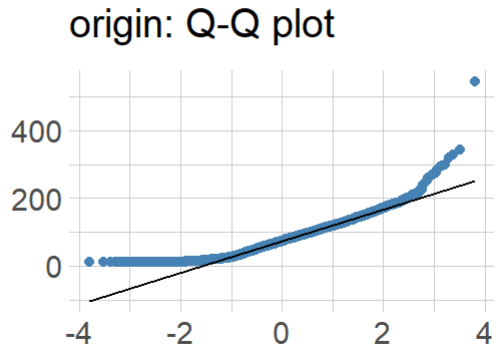


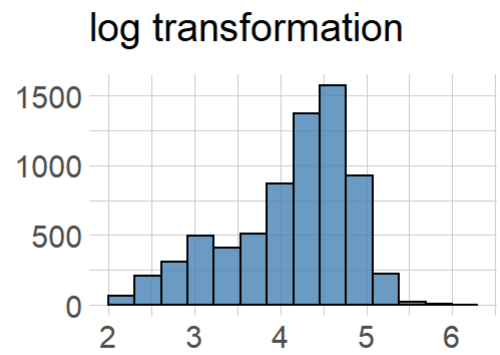


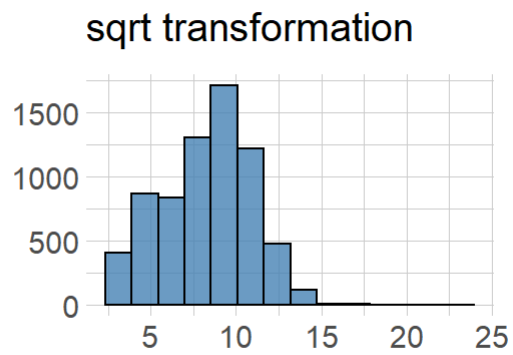


Normality Diagnosis Plot (ENROLLMENT)





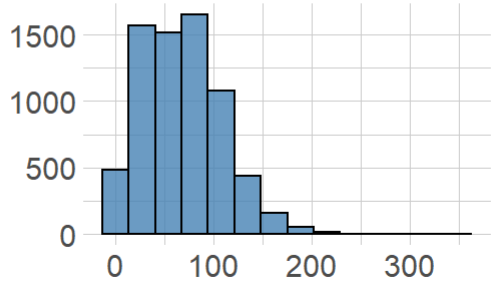


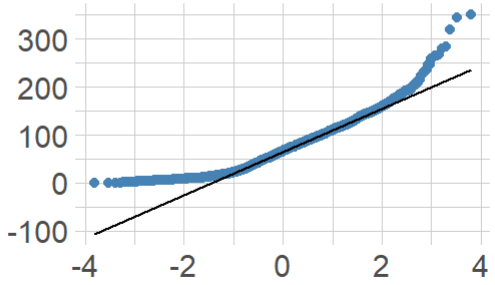


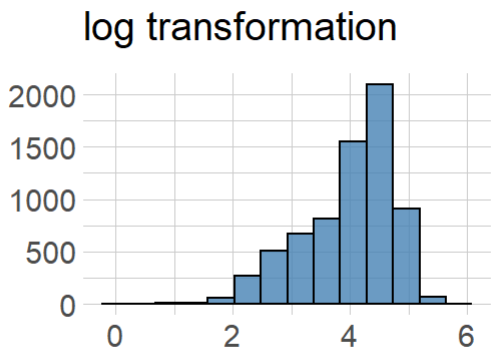
Normality Diagnosis Plot (UP_TO_DATE)

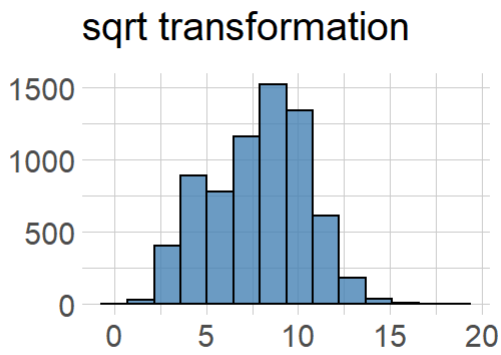
origin

origin: Q-Q plot

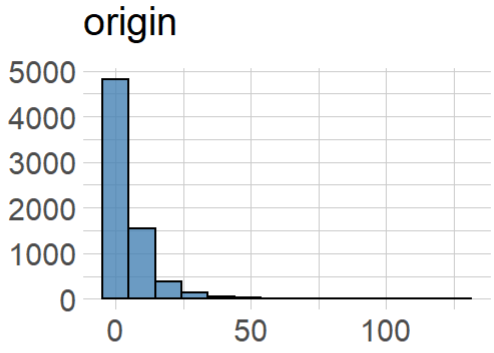


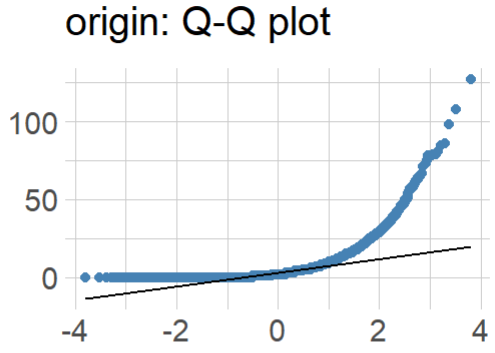


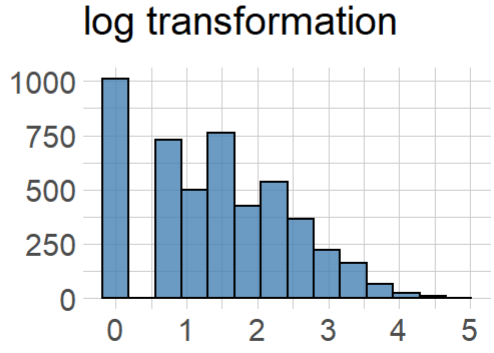


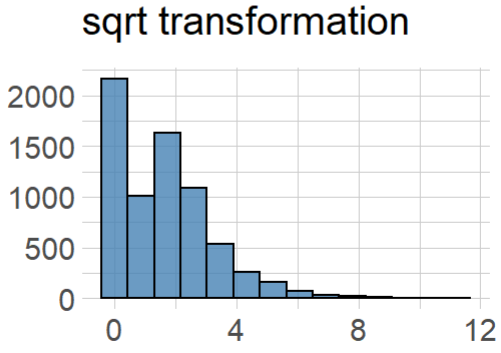


Normality Diagnosis Plot (CONDITIONAL)

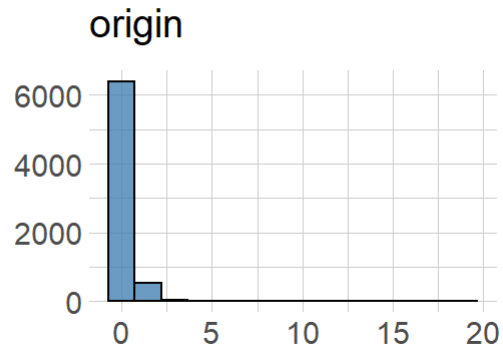


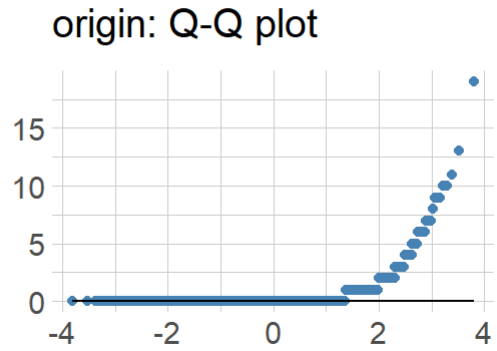


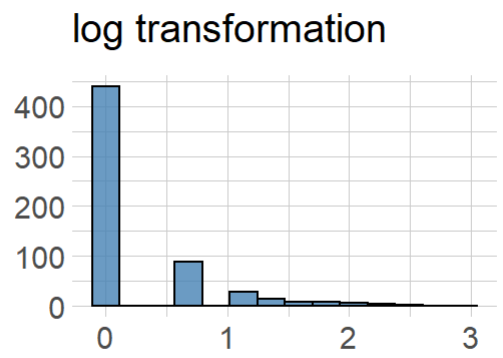


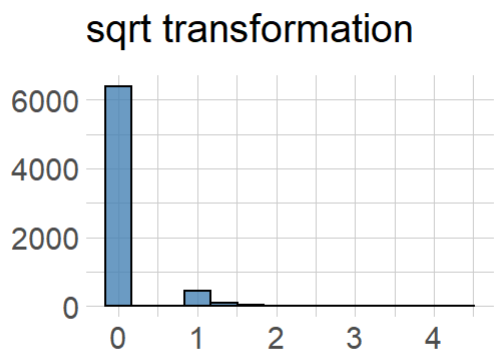


Normality Diagnosis Plot (PME)

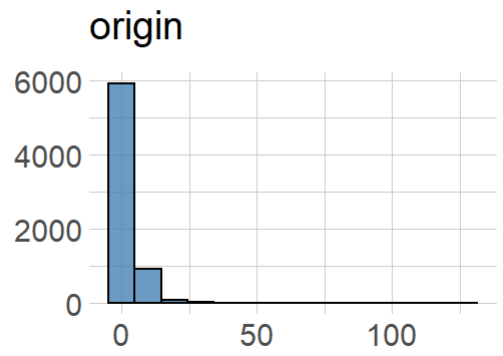


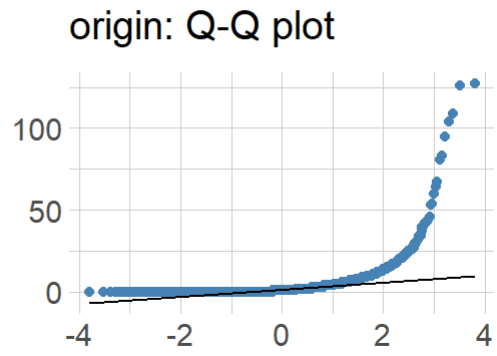


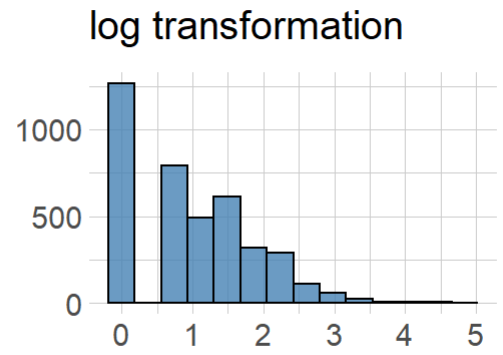


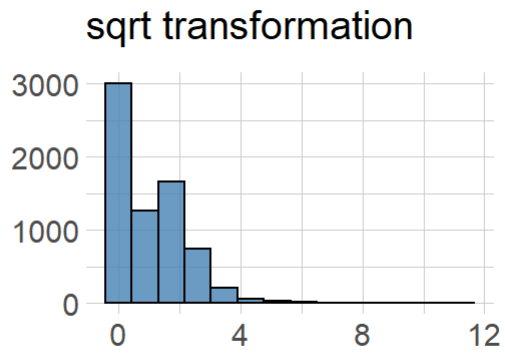


Normality Diagnosis Plot (PBE_BETA)

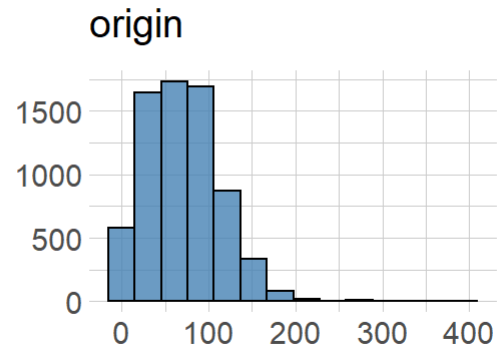


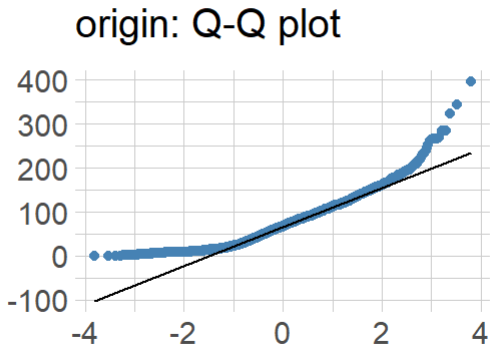


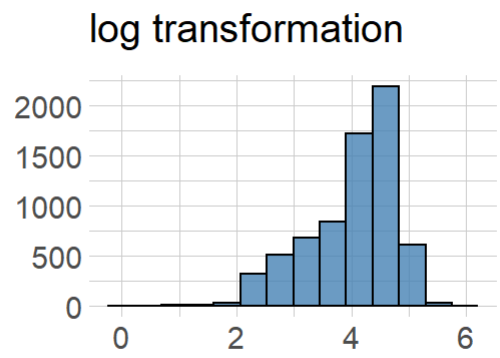


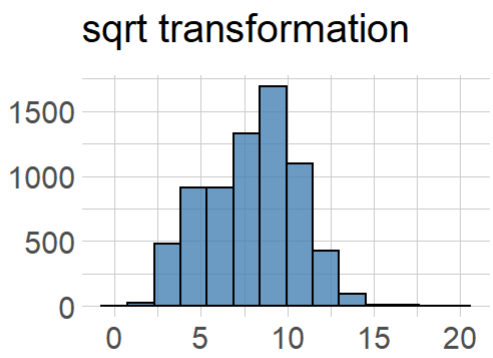


Normality Diagnosis Plot (DTP)

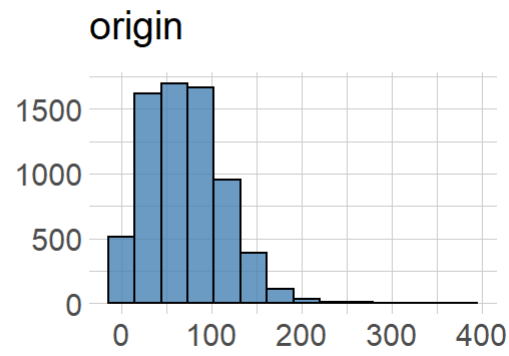


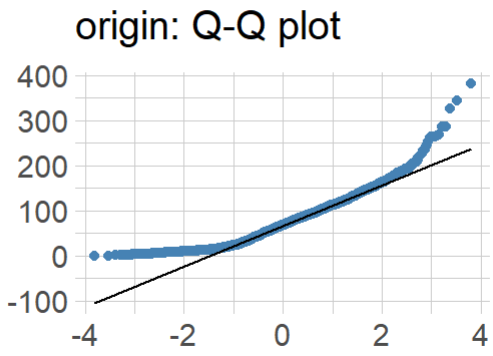


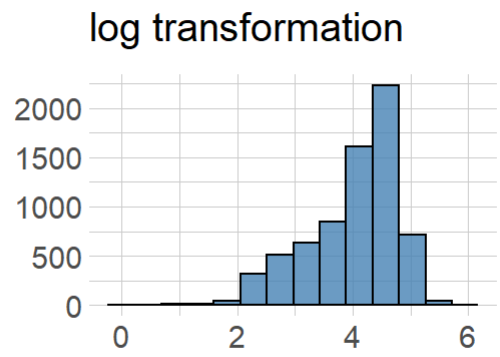


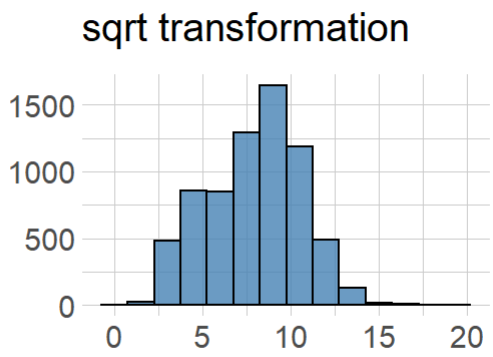


Normality Diagnosis Plot (POLIO)

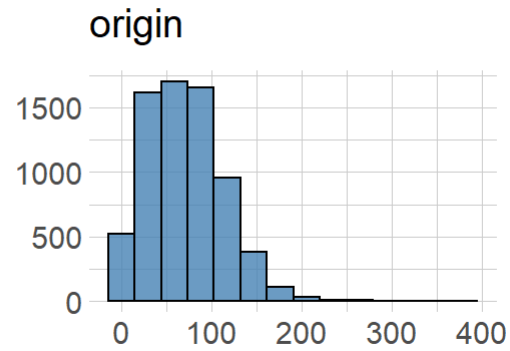


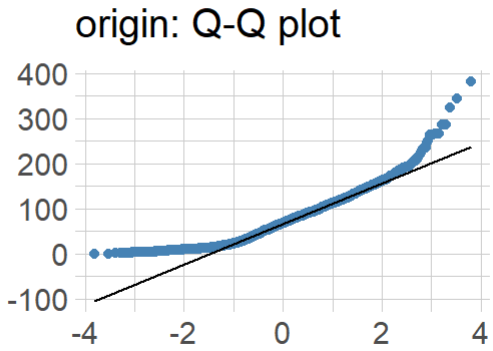


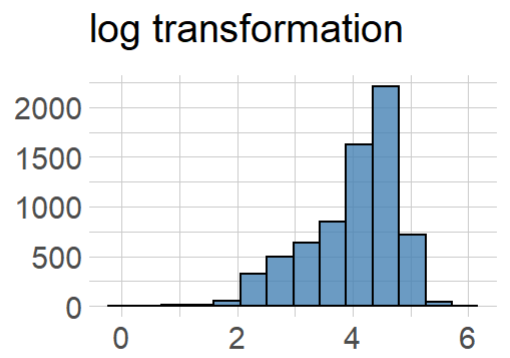


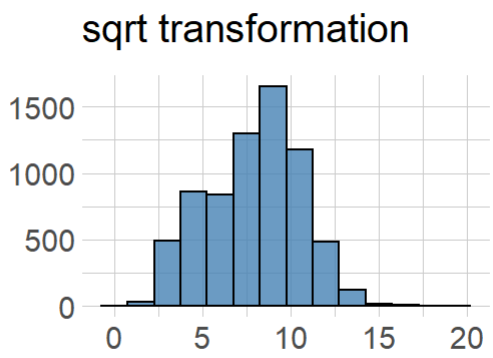


Normality Diagnosis Plot (MMR)

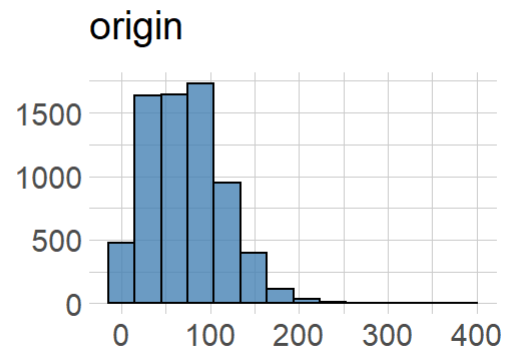


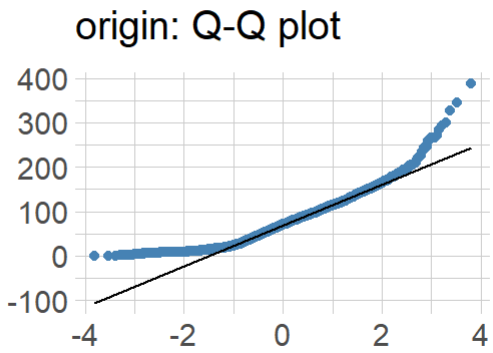


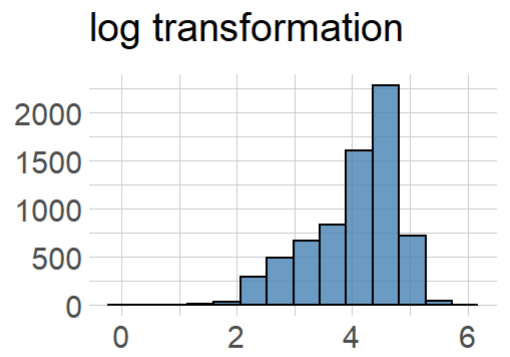


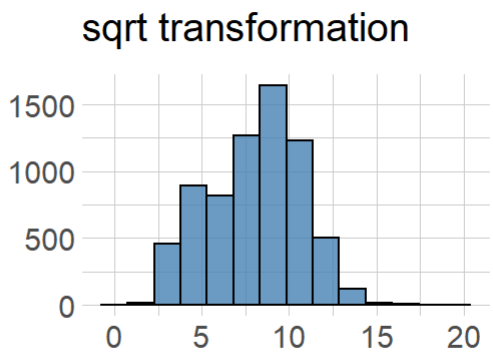


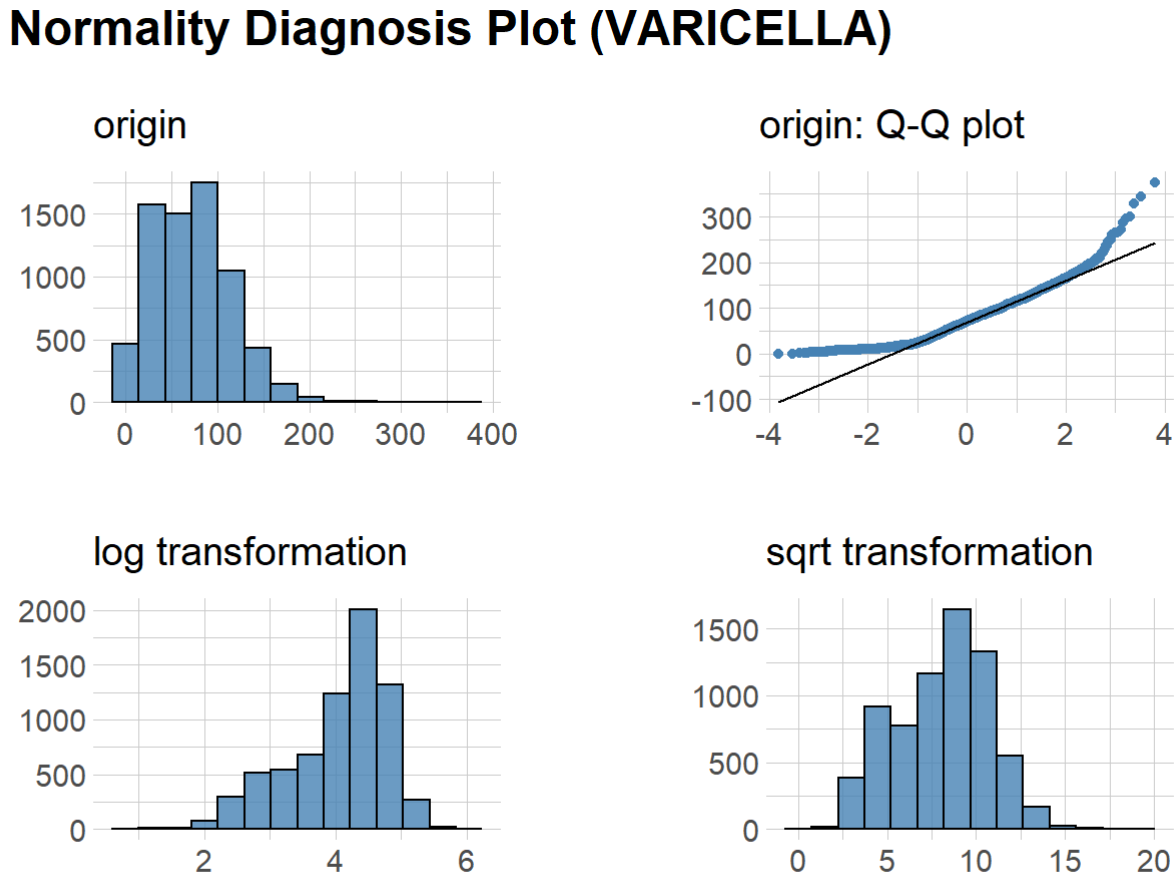
Normality Diagnosis Plot (HEPB)











We will address this skewness using sqrt transformation rather than log transformation as that can induce NaN in the dataset.

```
# taking out type variable so that we can take correlation in the next step
schools_sqrt <- schools_noNA
# taking sqrt transformation on the income column in the newly created dataset
schools_sqrt$PME <- sqrt(schools_sqrt$PME)
schools_sqrt$PBE_BETA <- sqrt(schools_sqrt$PBE_BETA)
schools_sqrt$CONDITIONAL <- sqrt(schools_sqrt$CONDITIONAL)
skewness(select_if(schools_noNA, is.numeric))
```

```
## Warning in mean.default(x): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(x, ...): argument is not numeric or logical: returning
## NA
```

```
## Warning in mean.default((x - mean(x, ...))^2): argument is not numeric or
## logical: returning NA
```

```
## [1] NA
```

```
skewness(select_if(schools_sqrt, is.numeric))
```

```
## Warning in mean.default(x): argument is not numeric or logical: returning NA
```

```
## Warning in mean.default(x, ...): argument is not numeric or logical: returning
## NA
```

```
## Warning in mean.default((x - mean(x, ...))^2): argument is not numeric or
## logical: returning NA
```

```
## [1] NA
```

As we can see above, the skewness has reduced. Now checking if the outliers have reduced as well:

```
diagnose_outlier(schools_sqrt)
```

variables <chr>	outliers_cnt <int>	outliers_ratio <dbl>	outliers_mean <dbl>	with_mean <dbl>	without_mean <dbl>
SCHOOL.CODE	2046	29.3039244	4.374762e+06	5.560174e+06	6.051534e+06
ENROLLMENT	49	0.7018046	2.417551e+02	7.598539e+01	7.481379e+01
UP_TO_DATE	54	0.7734174	2.203704e+02	6.856216e+01	6.737890e+01
CONDITIONAL	85	1.2174162	7.326479e+00	1.599870e+00	1.529294e+00
PME	601	8.6078488	1.204372e+00	1.036705e-01	0.000000e+00
PBE_BETA	82	1.1744486	5.668172e+00	1.024208e+00	9.690189e-01
DTP	58	0.8307075	2.213793e+02	7.009797e+01	6.883073e+01
POLIO	54	0.7734174	2.239074e+02	7.043999e+01	6.924379e+01
MMR	55	0.7877399	2.228727e+02	7.021398e+01	6.900188e+01
HEPB	54	0.7734174	2.280741e+02	7.205643e+01	7.084036e+01

1-10 of 11 rows Previous **1** 2 Next

We can see that the outlier counts for Conditional, PBE_BETA have reduced but PME remained the same.

5. Checking the normality

```
normality(schools_sqrt)
```

vars <chr>	statistic <dbl>	p_value <dbl>	sample <dbl>
SCHOOL.CODE	0.5089732	9.298602e-80	5000
ENROLLMENT	0.9498346	2.606030e-38	5000
UP_TO_DATE	0.9562005	2.474822e-36	5000
CONDITIONAL	0.8804568	2.918659e-52	5000
PME	0.3161375	4.574182e-87	5000
PBE_BETA	0.8061700	5.348593e-61	5000

vars <chr>	statistic <dbl>	p_value <dbl>	sample <dbl>
DTP	0.9552212	1.188670e-36	5000
POLIO	0.9560662	2.236423e-36	5000
MMR	0.9563921	2.861050e-36	5000
HEPB	0.9563199	2.708661e-36	5000

1-10 of 11 rows Previous **1** 2 Next

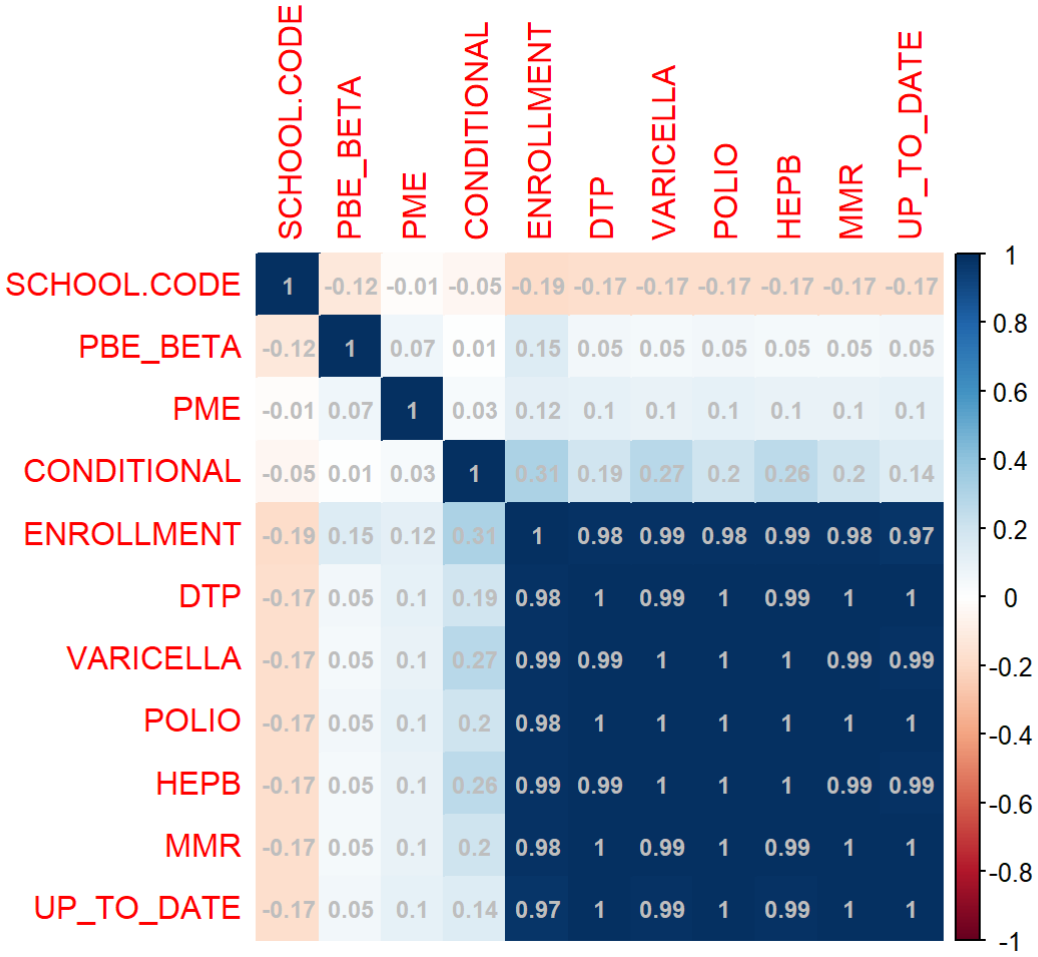
PME like the other columns seems to have normal data as the p - value is less than the threshold of 0.05, hence we won't do any clean up for the outliers.

6. Checking the correlation:

```
cor(select_if(schools_sqrt, is.numeric))
```

```
##             SCHOOL.CODE ENROLLMENT  UP_TO_DATE  CONDITIONAL         PME
## SCHOOL.CODE  1.00000000 -0.1892666 -0.17197653 -0.045902093 -0.01316548
## ENROLLMENT  -0.18926660  1.0000000  0.97260752  0.313085667  0.11802510
## UP_TO_DATE  -0.17197653  0.9726075  1.00000000  0.142918402  0.10015913
## CONDITIONAL -0.04590209  0.3130857  0.14291840  1.000000000  0.03373593
## PME         -0.01316548  0.1180251  0.10015913  0.033735934  1.00000000
## PBE_BETA    -0.12070754  0.1482038  0.05044765  0.009929792  0.06760275
## DTP         -0.17225359  0.9814412  0.99699843  0.194564310  0.10138750
## POLIO       -0.17222073  0.9826588  0.99637880  0.204547760  0.10166968
## MMR         -0.17331383  0.9819507  0.99578978  0.201351713  0.09976863
## HEPB        -0.17161476  0.9886980  0.98927448  0.263837616  0.09983844
## VARICELLA   -0.17170350  0.9891083  0.98788688  0.272572031  0.09966983
##                PBE_BETA         DTP      POLIO         MMR        HEPB
## SCHOOL.CODE -0.120707536 -0.17225359 -0.1722207 -0.17331383 -0.17161476
## ENROLLMENT   0.148203769  0.98144116  0.9826588  0.98195069  0.98869798
## UP_TO_DATE   0.050447653  0.99699843  0.9963788  0.99578978  0.98927448
## CONDITIONAL  0.009929792  0.19456431  0.2045478  0.20135171  0.26383762
## PME          0.067602750  0.10138750  0.1016697  0.09976863  0.09983844
## PBE_BETA     1.000000000  0.05343112  0.0508381  0.04918563  0.04880640
## DTP          0.053431115  1.00000000  0.9992549  0.99815403  0.99478415
## POLIO        0.050838096  0.99925490  1.0000000  0.99835413  0.99584641
## MMR          0.049185629  0.99815403  0.9983541  1.00000000  0.99483096
## HEPB         0.048806400  0.99478415  0.9958464  0.99483096  1.00000000
## VARICELLA    0.049805087  0.99400203  0.9952135  0.99442589  0.99917553
##              VARICELLA
## SCHOOL.CODE -0.17170350
## ENROLLMENT   0.98910828
## UP_TO_DATE   0.98788688
## CONDITIONAL  0.27257203
## PME          0.09966983
## PBE_BETA     0.04980509
## DTP          0.99400203
## POLIO        0.99521351
## MMR          0.99442589
## HEPB         0.99917553
## VARICELLA    1.00000000
```

```
corrplot(cor(select_if(schools_sqrt, is.numeric)),
         method = "color",
         addCoef.col="grey",
         order = "AOE",
         number.cex=0.75)
```



We can see that Enrollment, DTP, varicella, polio, hepb, mmr, up_to_date are highly correlated, rest of the columns dont have a good correlation which makes sense as the parents getting their children vaccination for one kind of disease isn't an anti vaxer so will end up getting their children all the vaccines and keep the vaccines up to date and it is good to see that enrolled students are vaccinated, which could mean that the schools might be mandating vaccines as a pre requisite in the school.

Data Exploration Data Preprocessing and Cleaning For usVaccines time series data

1. Checking basics structure of the time series data

```
summary(usVaccines)
```

```
##       DTP1          HepB_BD           Pol3            Hib3
##  Min.   :81.00   Min.   :11.00   Min.   :24.00   Min.   :52.00
##  1st Qu.:89.75   1st Qu.:17.00   1st Qu.:90.00   1st Qu.:87.00
##  Median :97.00   Median :19.00   Median :93.00   Median :91.00
##  Mean   :94.05   Mean   :34.21   Mean   :87.16   Mean   :89.21
##  3rd Qu.:98.00   3rd Qu.:54.50   3rd Qu.:94.00   3rd Qu.:93.00
##  Max.   :99.00   Max.   :74.00   Max.   :97.00   Max.   :94.00
##       MCV1
##  Min.   :82.00
##  1st Qu.:90.00
##  Median :92.00
##  Mean   :91.24
##  3rd Qu.:92.00
##  Max.   :98.00
```

```
str(usVaccines)
```

```
##  Time-Series [1:38, 1:5] from 1980 to 2017: 83 84 83 84 84 85 88 88 89 81 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:5] "DTP1" "HepB_BD" "Pol3" "Hib3" ...
```

2. Checking for skewness

```
skewness(usVaccines)
```

```
## [1] -1.659623
```

There is barely any skewness in the dataset

3. Checking for outliers:

```
diagnose_outlier(as.data.frame(usVaccines))
```

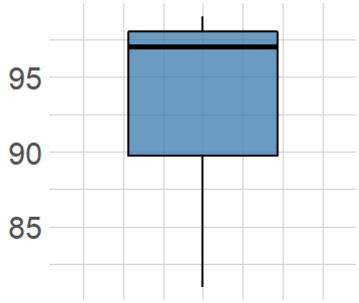
variables <chr>	outliers_cnt <int>	outliers_ratio <dbl>	outliers_mean <dbl>	with_mean <dbl>	without_mean <dbl>
DTP1	0	0.000000	NaN	94.05263	94.05263
HepB_BD	0	0.000000	NaN	34.21053	34.21053
Pol3	6	15.789474	56.66667	87.15789	92.87500
Hib3	1	2.631579	52.00000	89.21053	90.21622
MCV1	12	31.578947	91.58333	91.23684	91.07692

5 rows

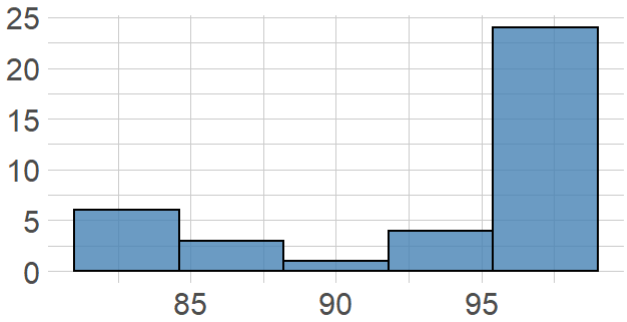
```
plot_outlier(as.data.frame(usVaccines))
```

Outlier Diagnosis Plot (DTP1)

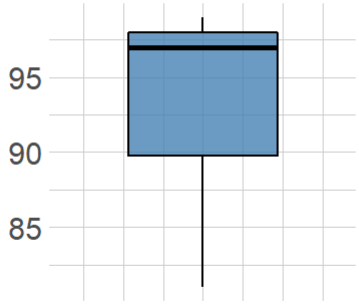
With outliers



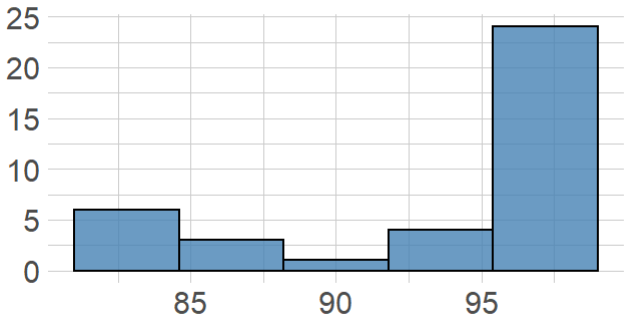
With outliers



Without outliers

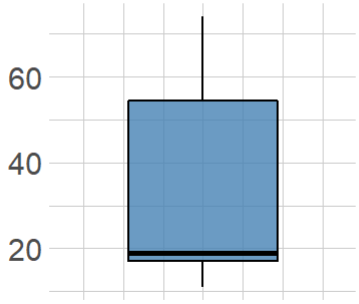


Without outliers

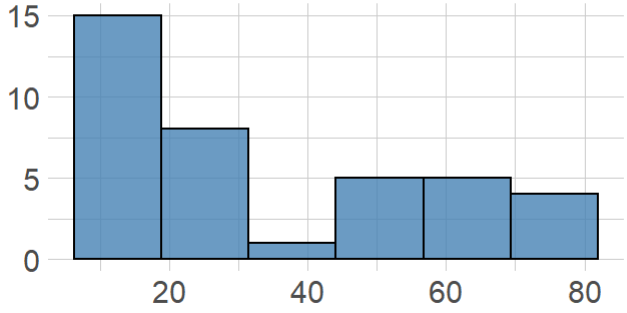


Outlier Diagnosis Plot (HepB_BD)

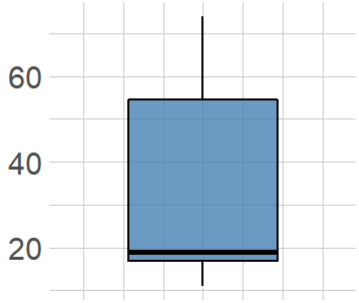
With outliers



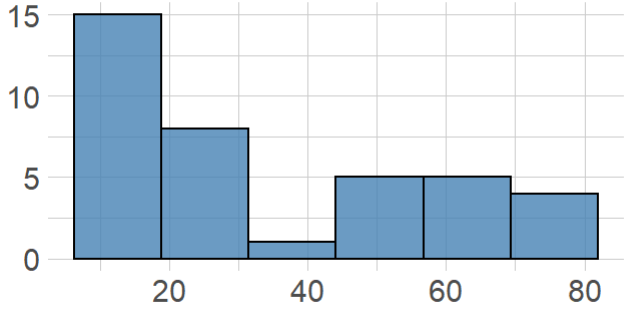
With outliers



Without outliers

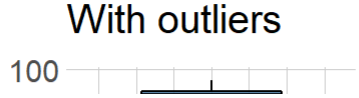


Without outliers

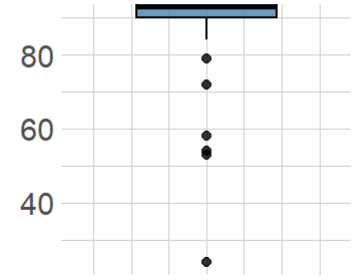


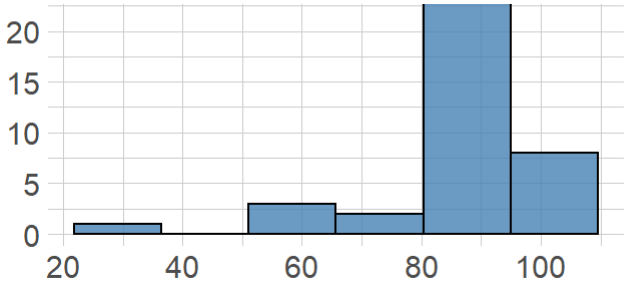
Outlier Diagnosis Plot (Pol3)

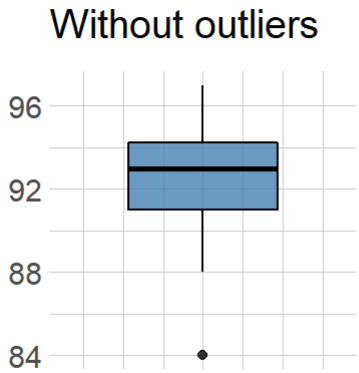
With outliers

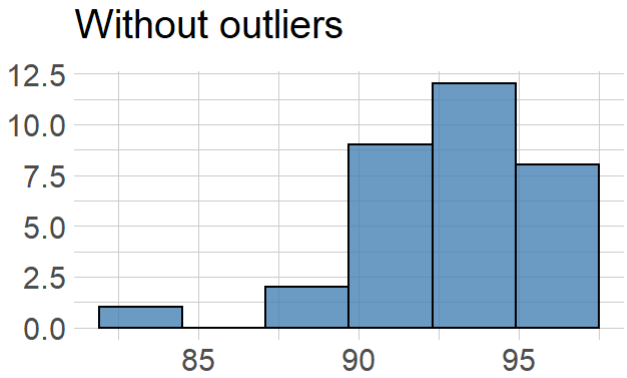


With outliers

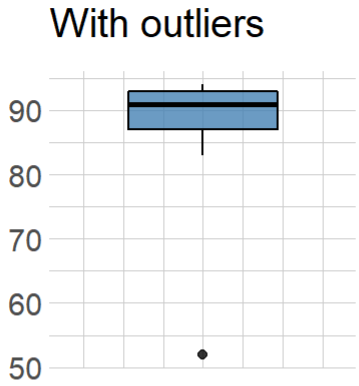


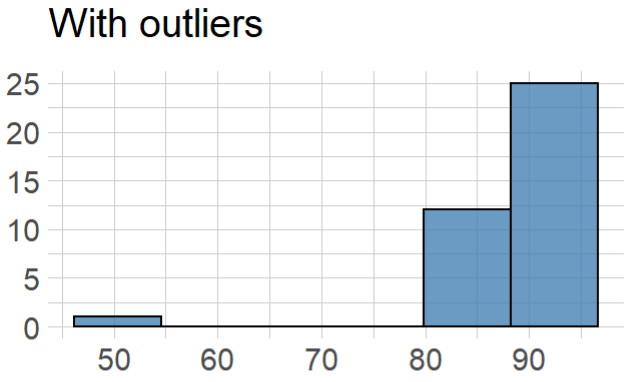


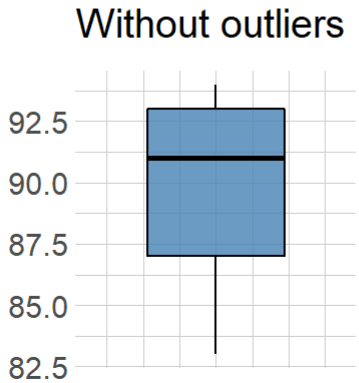


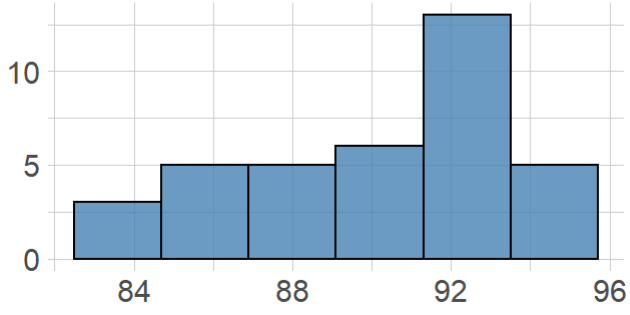


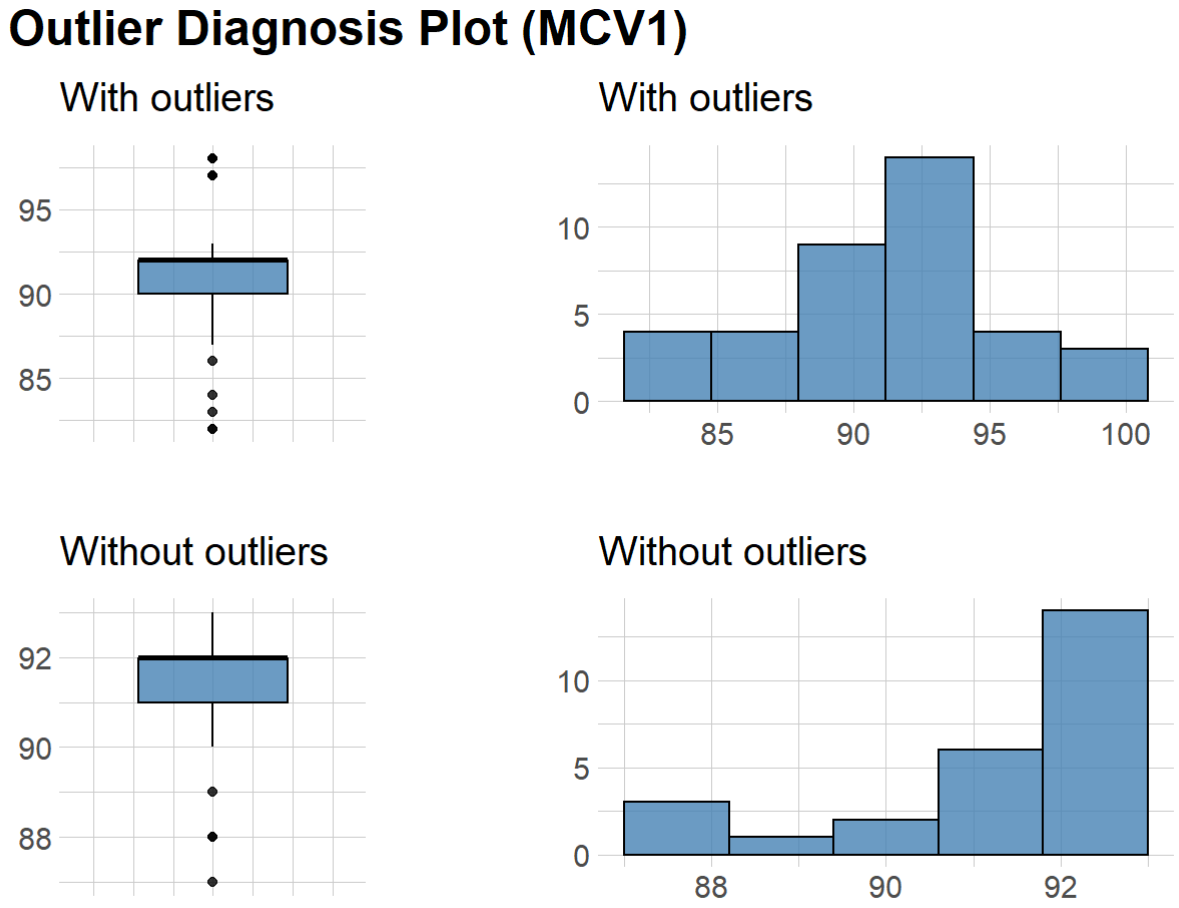
Outlier Diagnosis Plot (Hib3)











We can see that there are outliers but removing them won't make much of a difference in our descrptive statistics so we wont be making any changes and move ahead.

3. Checking for NA's

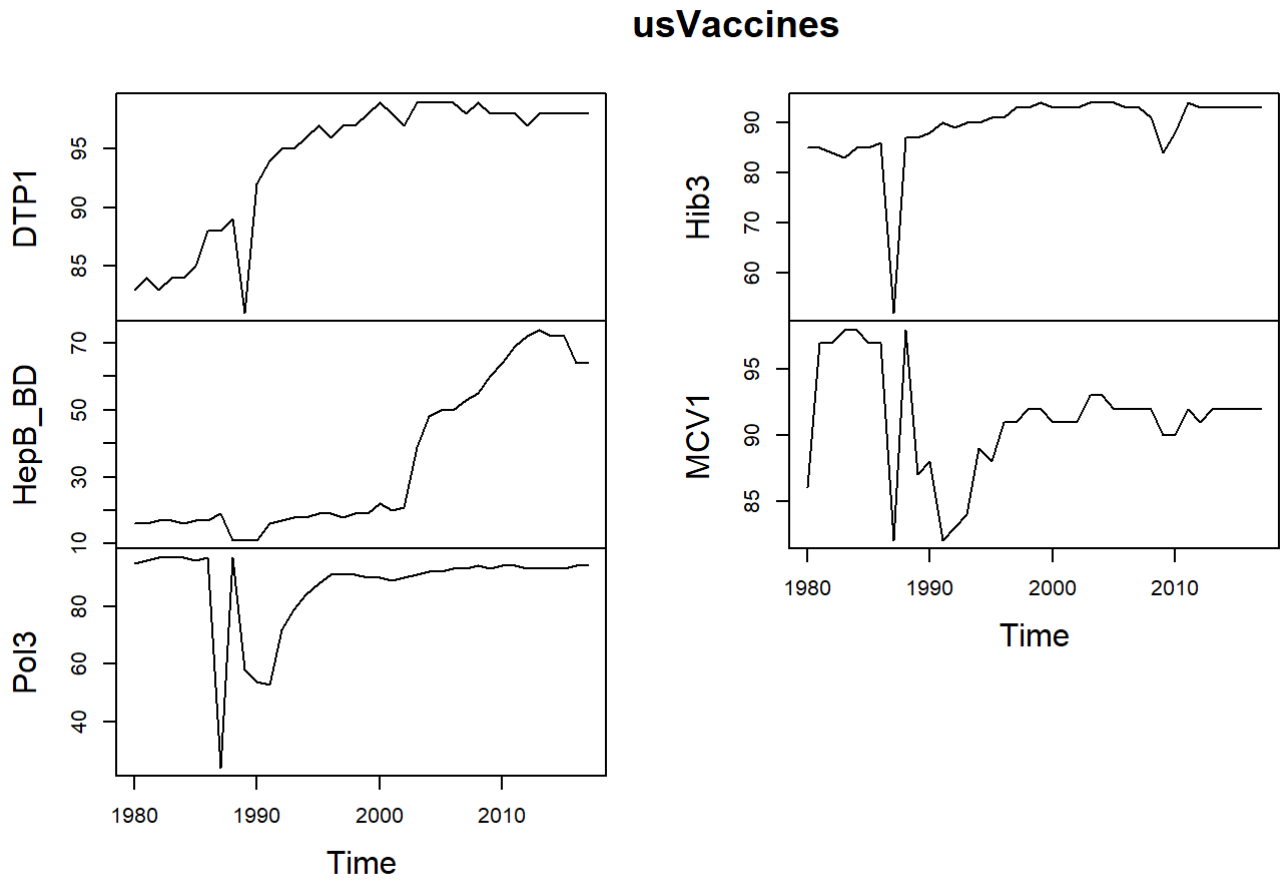
```
colSums(is.na(data.frame(usVaccines)))
```

```
##    DTP1 HepB_BD    Pol3    Hib3    MCV1
##       0       0       0       0       0
```

There are no NA's.

4. Checking the time series plot

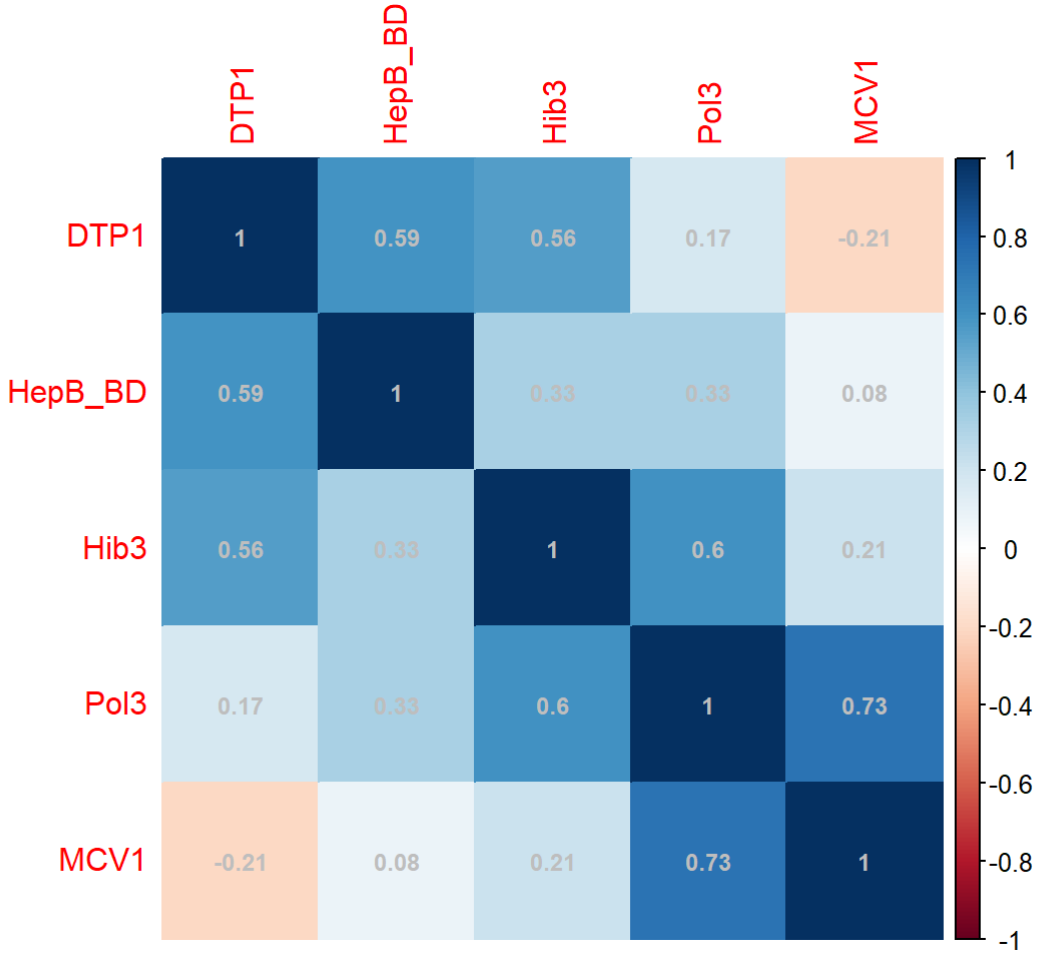
```
plot(usVaccines)
```



We can see that DTP1 was increasing steadily from 1980 but saw a sudden drop at around 1990 and then saw a good growth trend till 2010. Same is the case with HiB3. For HepB_BD we can see that it was constant from 1980 to around 2001 and then it just a good jump in the next decade. MCV1 saw high variability in vaccination rates from 1980 to start of 1990 with it's lowest at around 1988. Then it showed a pretty stead group rate till 2010 with no sudden spikes. Pol3 saw a high vaccination rate from 1980 to around 1987 then saw a sudden drop around 1987 and then peaked saw another drop at around the beginning of 1990 and then increased and got steady by the end of the time series. Basically all of them seemed to have dropped at around 1987.

5. Checking the correlation

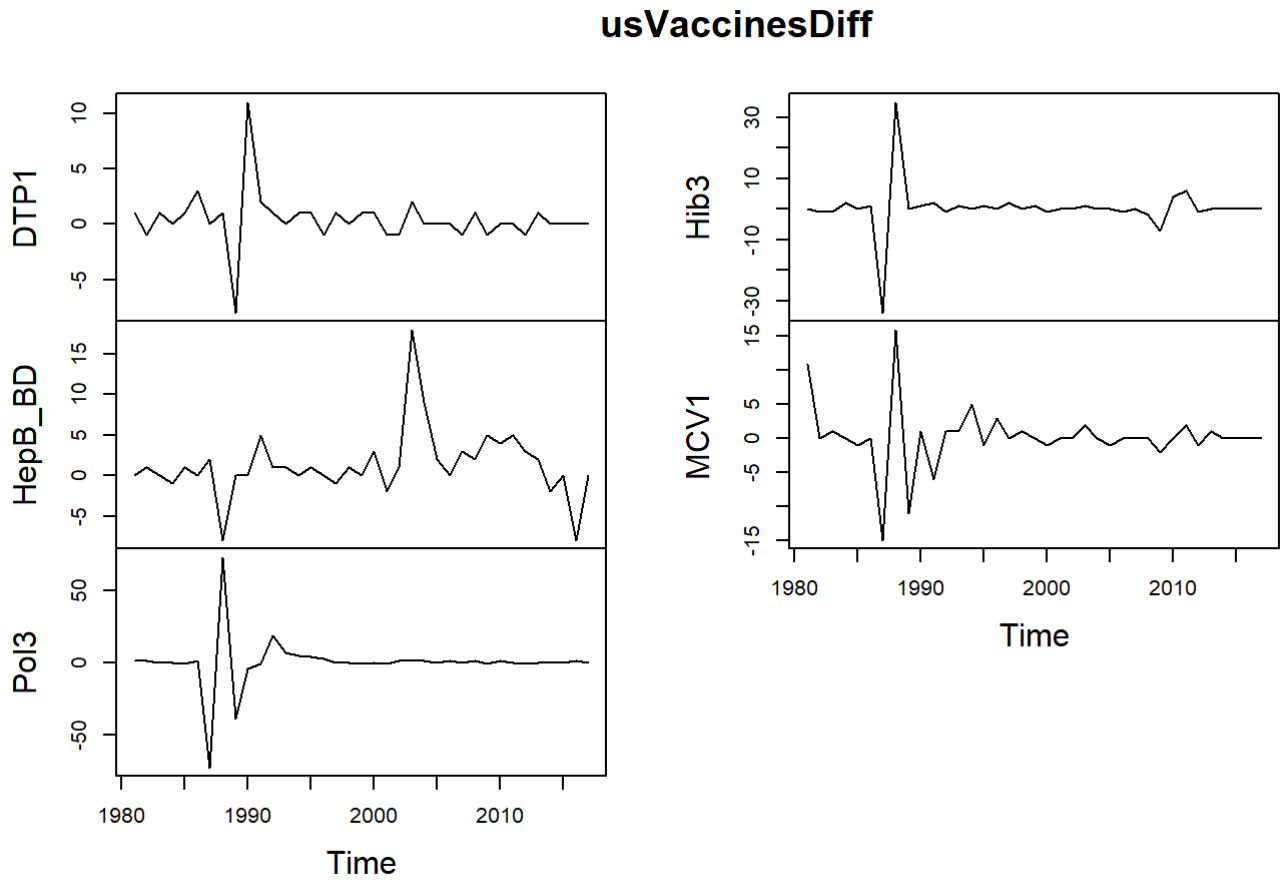
```
corrplot(cor(usVaccines), method = "color", addCoef.col="grey",
         order = "AOE",number.cex=0.75)
```



We can see that MCV1 and P0l3 are having the highest correlation in the series with the value of 0.73 then pol3 and hib3 then hepb_bd and dtp and then hib3 and dtp. We must remove these trend before proceeding with any substantive analysis

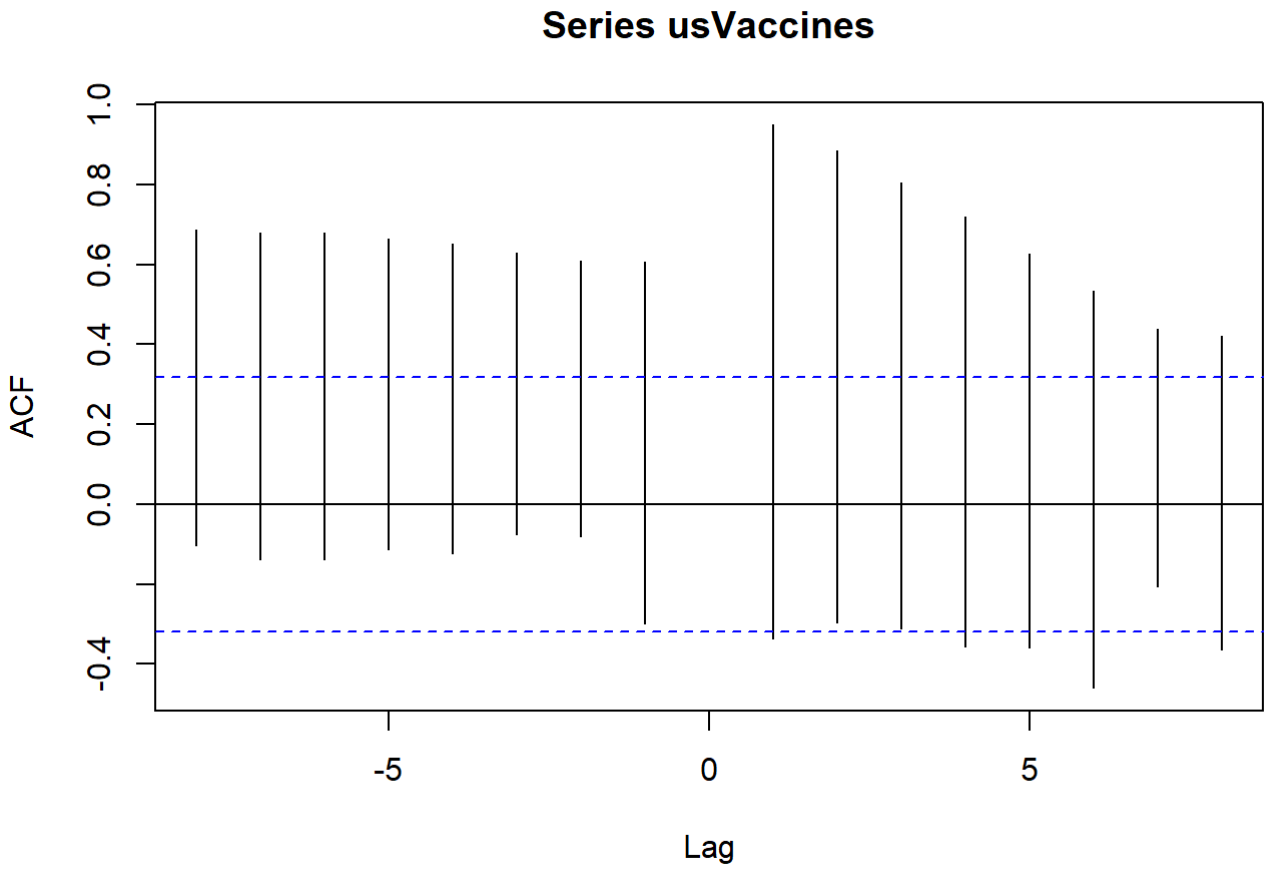
6. Doing differencing to remove trends before analysis

```
usVaccinesDiff <- diff(usVaccines)
plot(usVaccinesDiff)
```

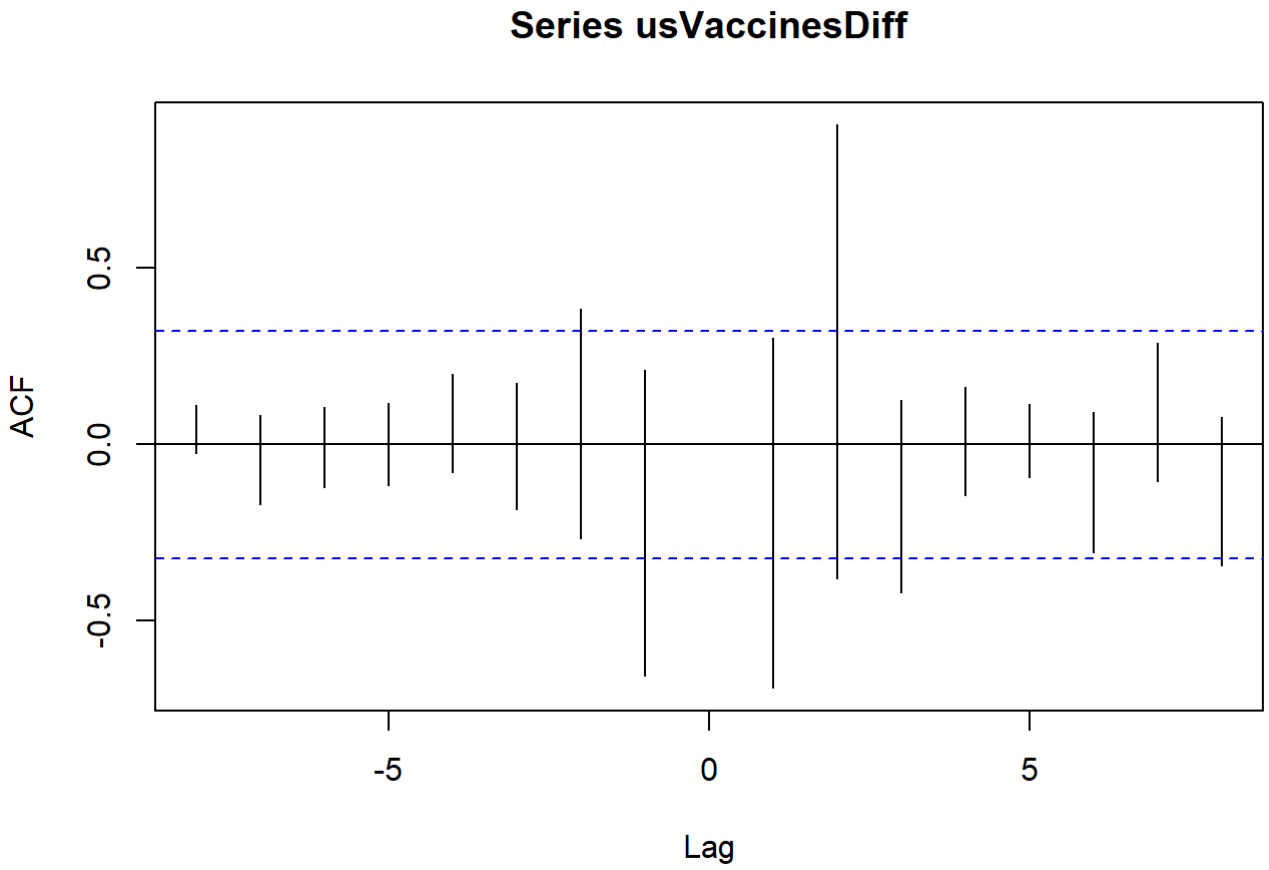


7. Examining the auto-correlation function (ACF) graph for the time series

```
acf(usVaccines)
```




```
acf(usVaccinesDiff)
```



We can see that 16 auto correlations are significant in the original data set and post differencing we can see that around 7 are significant. Since a stationary variable typically contains no significant or very few lagged correlations the data seems almost stationary and we also don't see pattern of positive and negative autocorrelations hence no sinusoidal pattern is still present at a low level in these data.

Now lets answer the questions finally!!

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 will be analyzing vaccination rates in California schools and school districts using various statistical tools using Time series data from the World Health Organization reporting vaccination rates in the U.S. for five common vaccines and California public school districts from 2017 data collection along with data from about 7,381 individual schools. Given the world's recent exposure to COVID-19, we know now more than ever how vaccinations can curb the spread of virus and saves lives, and in this report we will analyze and see how different factors like poverty, religion, medical factors affect vaccinations and analyze the vaccination rates.

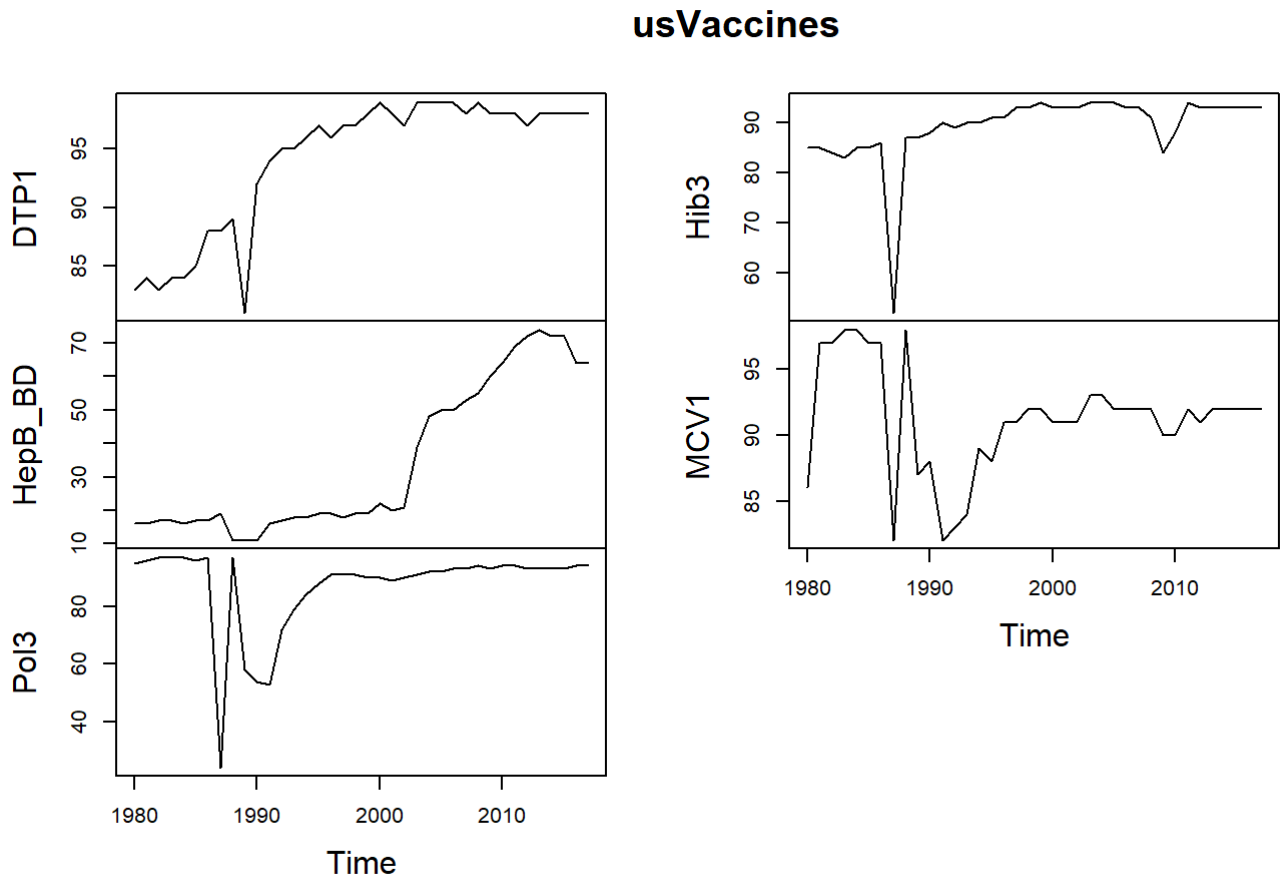
2. *Descriptive Overview of U.S. Vaccinations*

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.

```
usVaccines_latestData <- window(usVaccines, start = 2007, end = 2017)
```

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

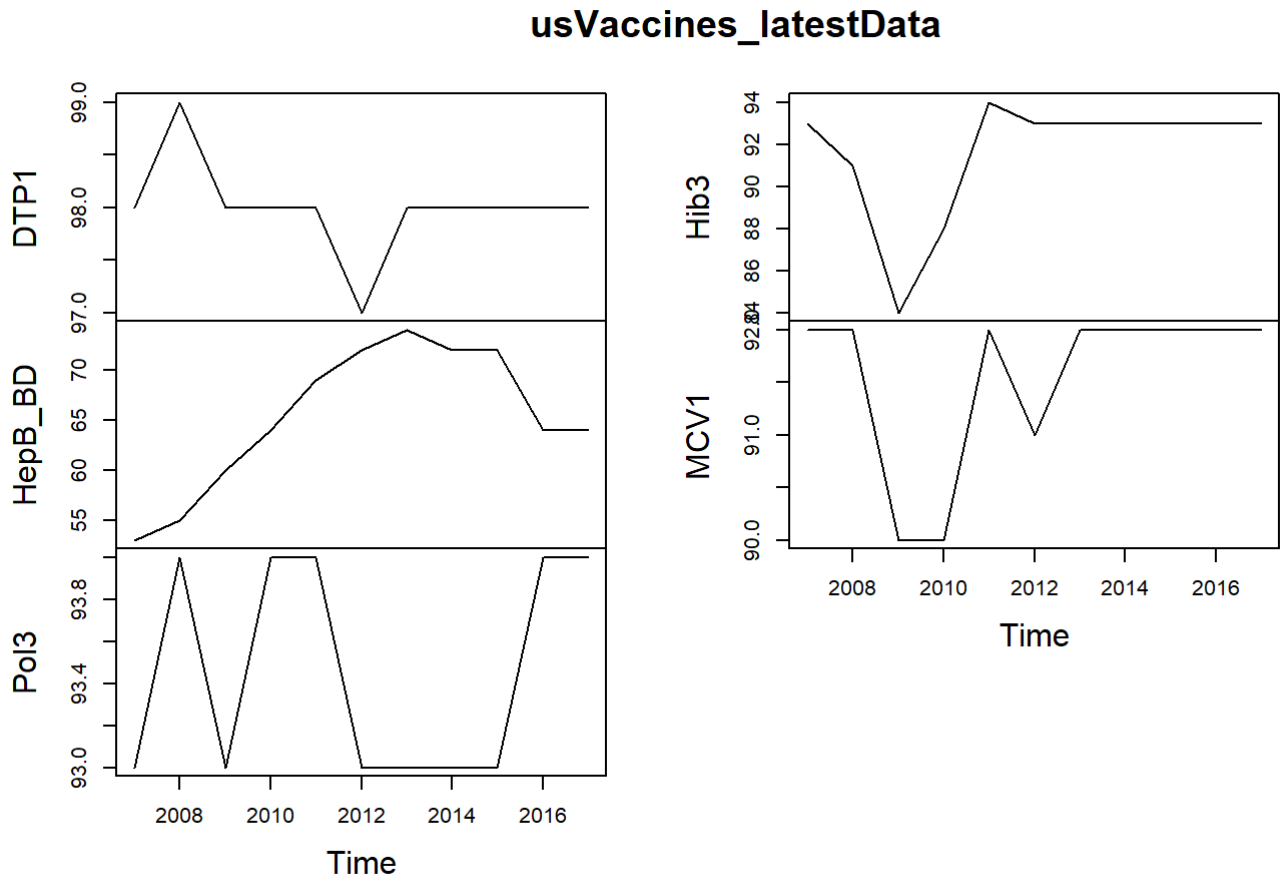
```
plot(usVaccines)
```



We can see that DTP1 was increasing steadily from 1980 but saw a sudden drop at around 1990 and then saw a good growth trend till 2010. Same is the case with HiB3. For HepB_BD we can see that it was constant from 1980 to around 2001 and then it just a good jump in the next decade. MCV1 saw high variability in vaccination rates from 1980 to start of 1990 with it's lowest at around 1988. Then it showed a pretty stead group rate till 2010 with no sudden spikes. Pol3 saw a high vaccination rate from 1980 to around 1987 then saw a sudden drop around 1987 and then peaked saw another drop at around the beginning of 1990 and then increased and got steady by the end of the time series. Basically all of them seemed to have dropped at around 1987.

Checking the same now only for latest data:

```
plot(usVaccines_latestData)
```

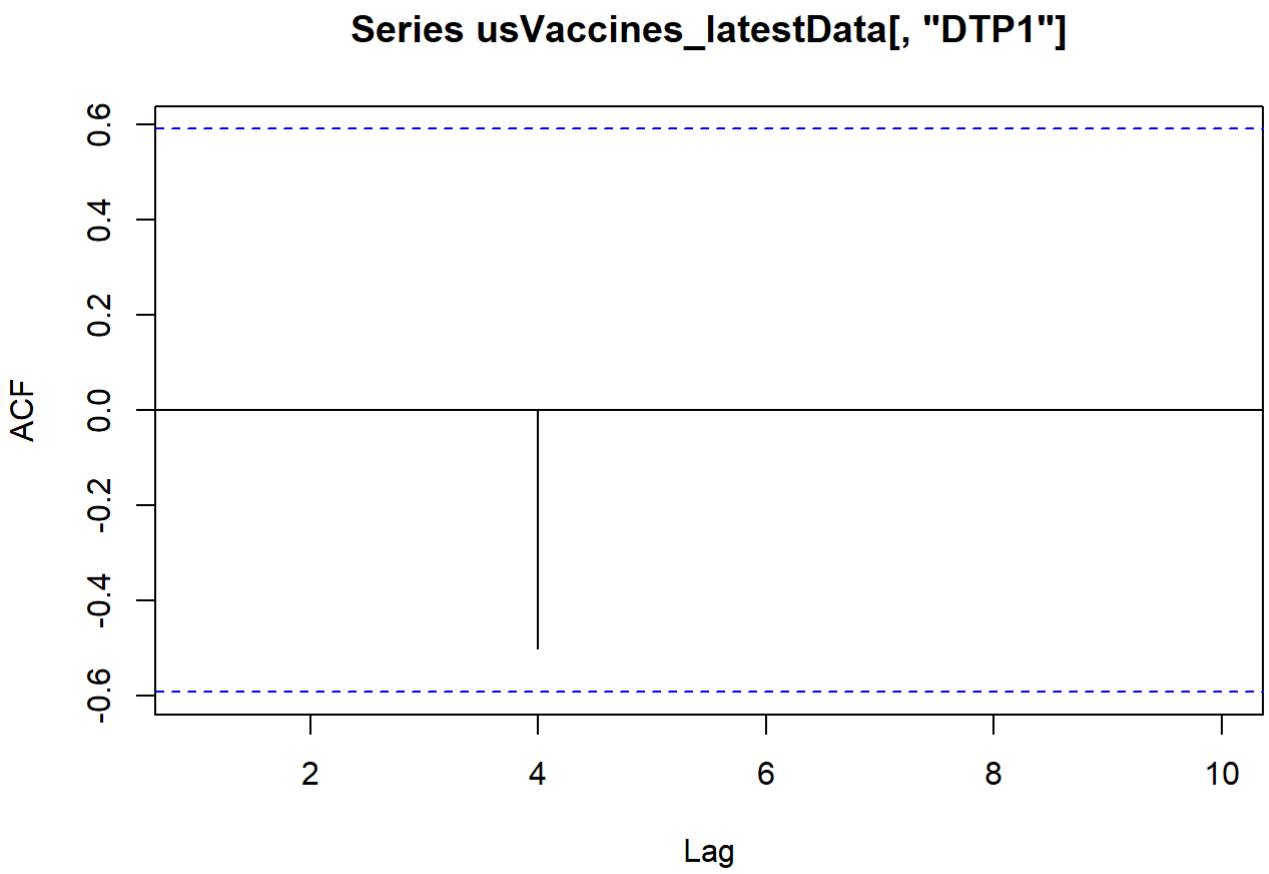


For DTP1 we can see that there is some variability where it increases at around 2008 the drops at around 2009 , remains constant till around 2011 then drops again at 2012 and post its increase in 2013 it remains constant till 2017. For Hib3 we can see that it also decreases in around 2009, then increases at around 2011 and then remains constant. For HEPB_BD, we see an increasing trend till 2013 and then it staggers alot till end of 2015 and becomes constant post that. For MCV1, there is variability till 2012 with highs and lows and then it increases in mid 2013 and remains constant till 2017. Pol3 has the most variability as its going up and down again and again and see's low rates of vaccination in between 2012 to 2015 and then increases and becomes almost constant till 2017.

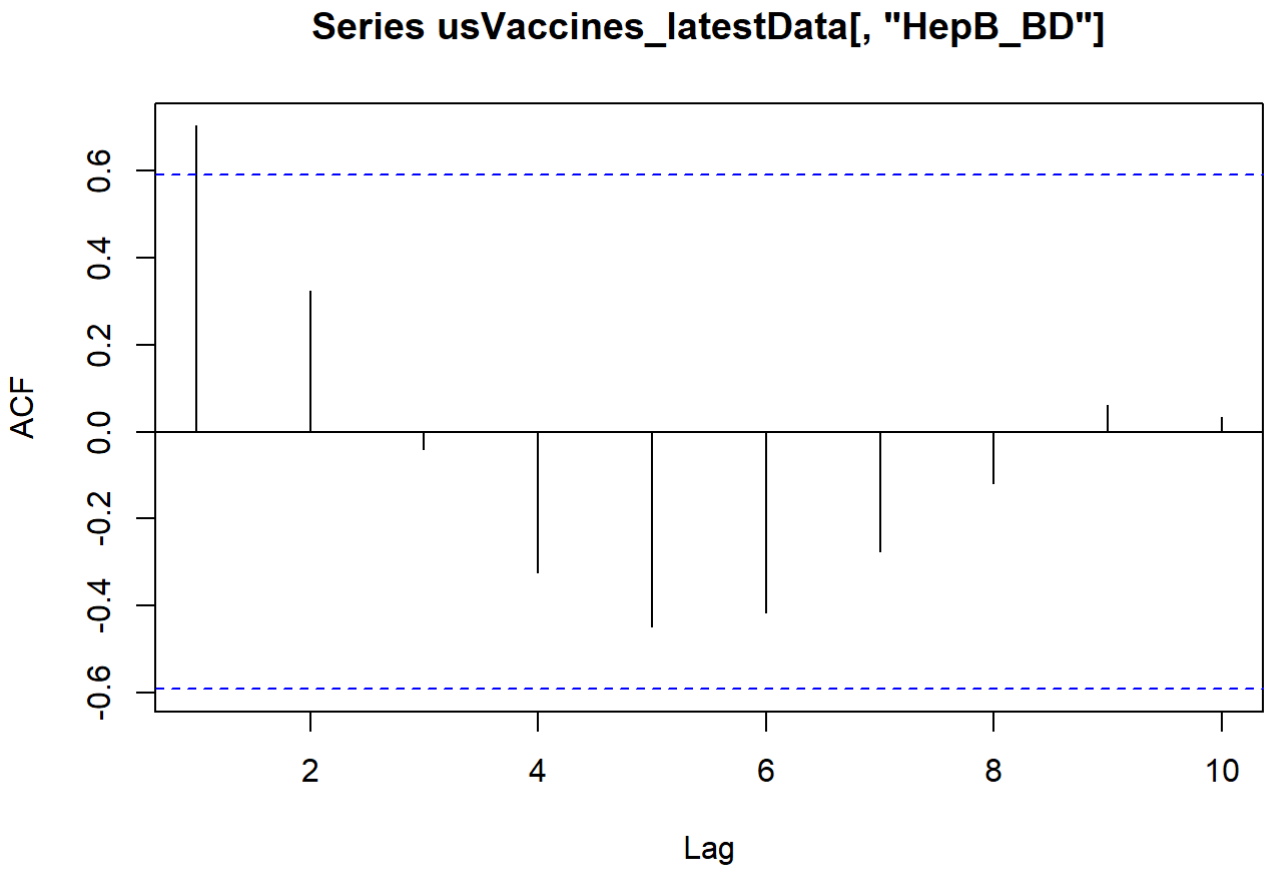
b. *Are there notable trends or cyclical variation in U.S. vaccination rates?*

Assumption: Since the staff members are only interested in recent vaccination rates we will use that dataset instead of the original data set.

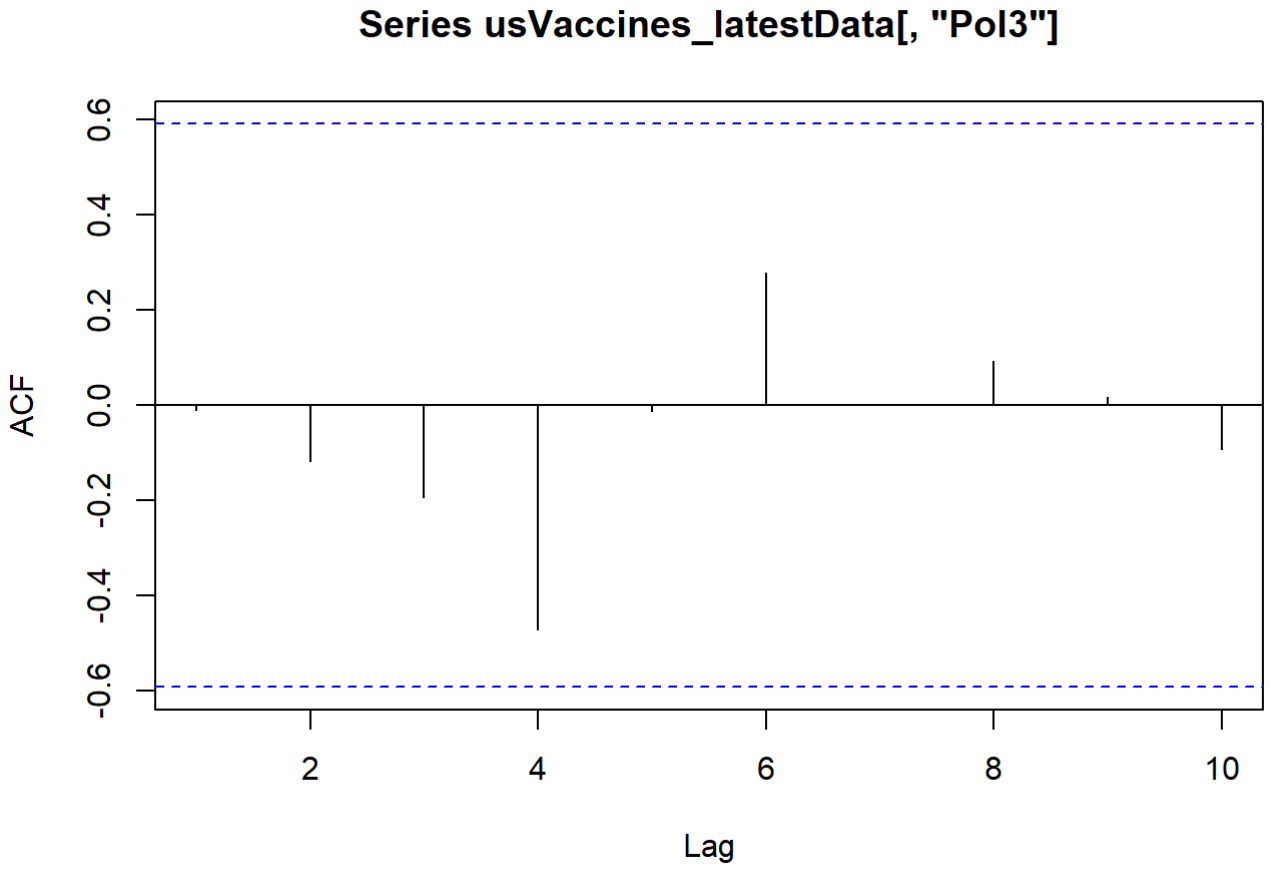
```
acf(usVaccines_latestData[,"DTP1"])
```



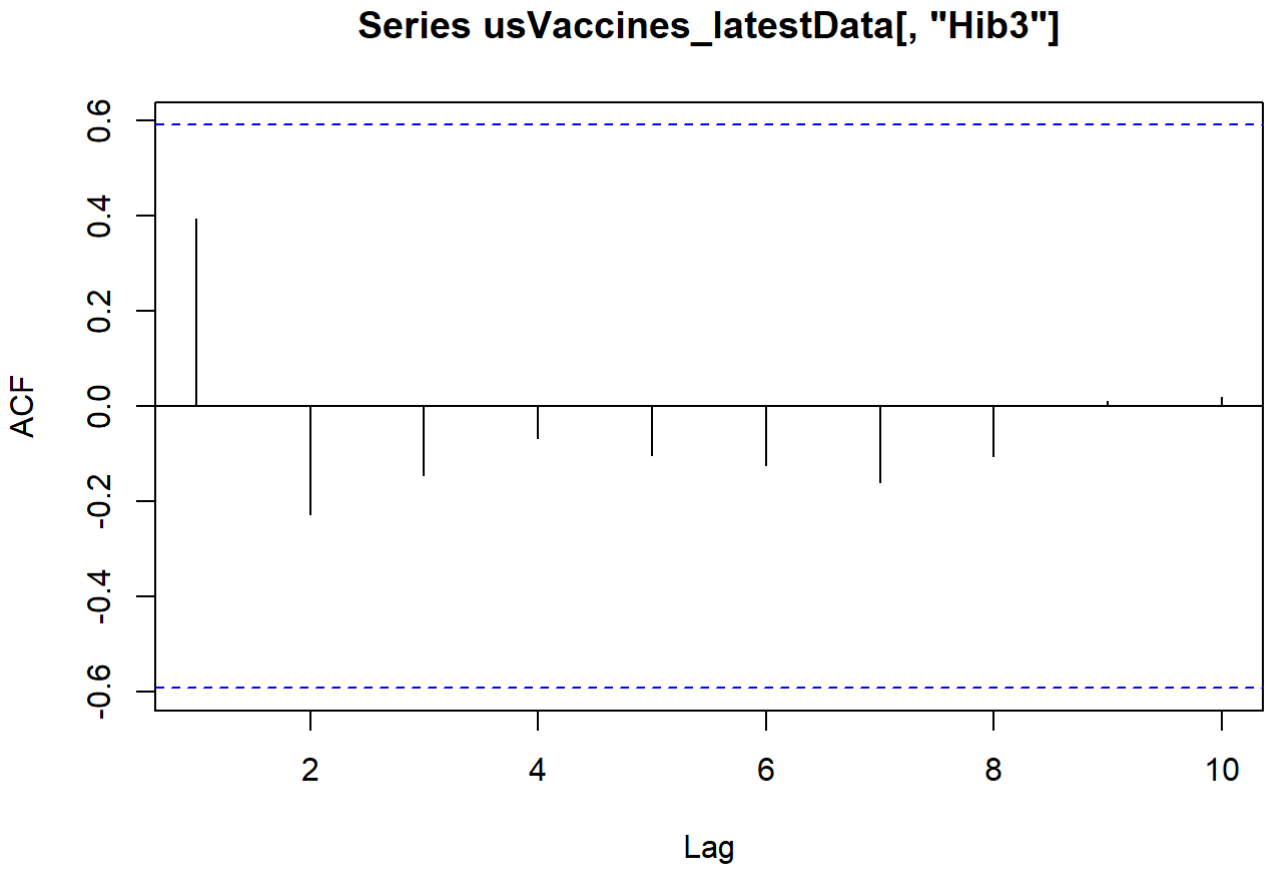

```
acf(usVaccines_latestData[,"HepB_BD"])
```



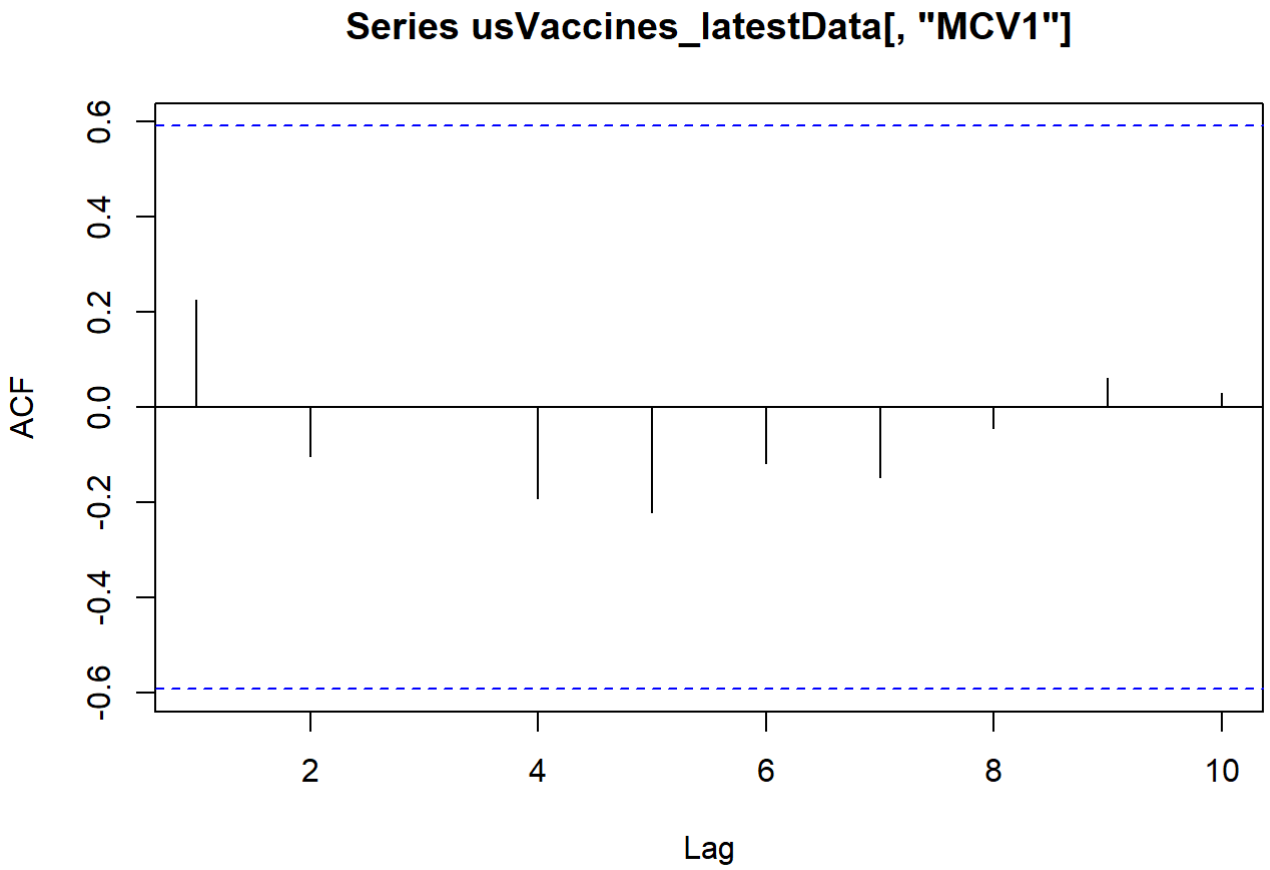
```
acf(usVaccines_latestData[,"Pol3"])
```



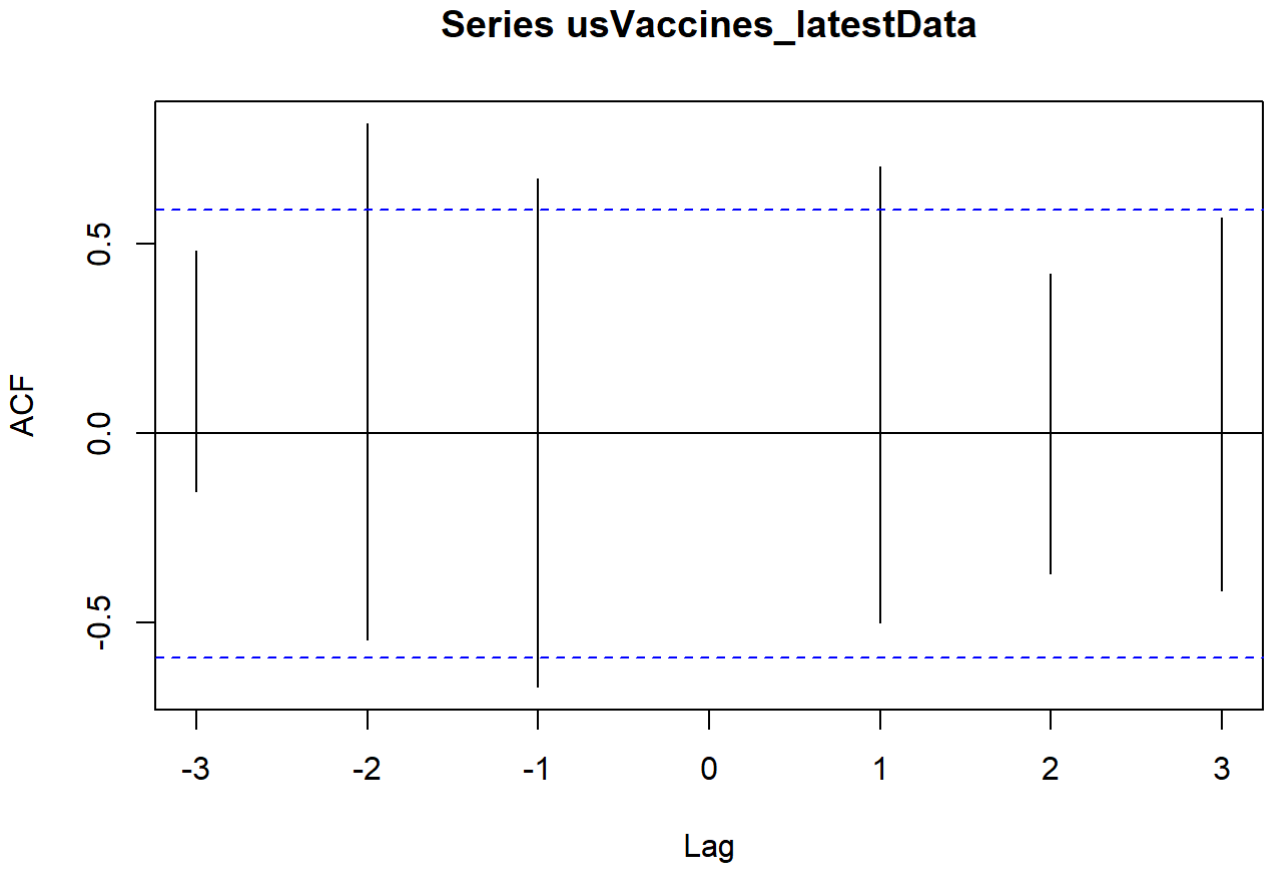

```
acf(usVaccines_latestData[,"Hib3"])
```



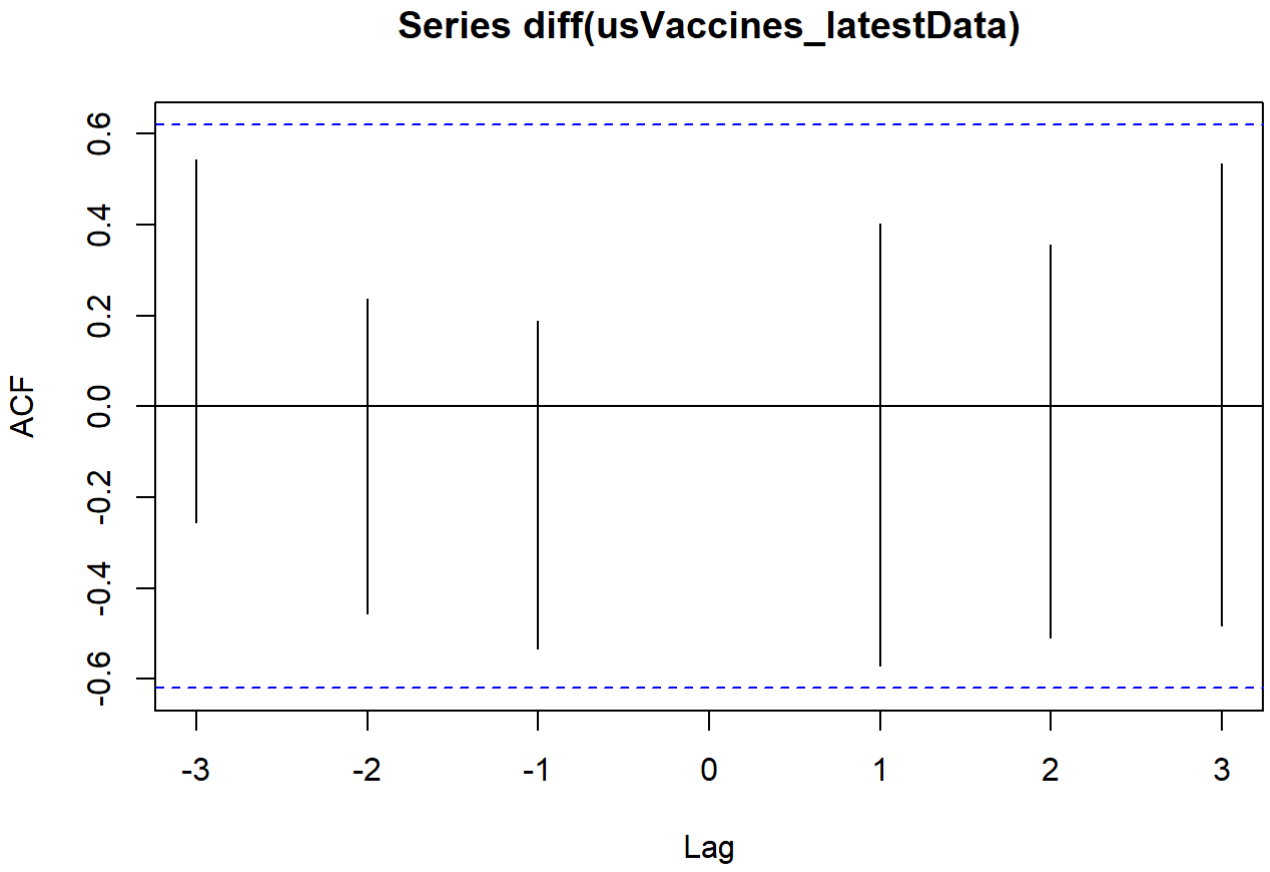

```
acf(usVaccines_latestData[,"MCV1"])
```



```
acf(usVaccines_latestData)
```




```
acf(diff(usVaccines_latestData))
```



For DTP1 we see no significant auto correlations and no cyclic variation as well, for HepB_BD again we don't see any auto correlations that are significant (not counting the perfect autocorrelation at lag = 0), same is the case for Pol3, Hib3 and MCV1 where there are no significant auto correlations and there is no strong cyclic pattern. Checking the lags of the whole data set we see 3 out of 6 auto correlations being significant but when we do differencing on the data set we see there are no significant auto correlations which is a good thing. We can further check this with the adf test

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines_latestData[, "DTP1"]
## Dickey-Fuller = -1.1492, Lag order = 2, p-value = 0.8965
## alternative hypothesis: stationary
```

```
adf.test(diff(usVaccines_latestData[,"DTP1"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_latestData[, "DTP1"])
## Dickey-Fuller = -1.0316, Lag order = 2, p-value = 0.9159
## alternative hypothesis: stationary
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines_latestData[, "HepB_BD"]
## Dickey-Fuller = -0.66013, Lag order = 2, p-value = 0.9617
## alternative hypothesis: stationary
```

```
adf.test(diff(usVaccines_latestData[,"HepB_BD"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_latestData[, "HepB_BD"])
## Dickey-Fuller = -1.6654, Lag order = 2, p-value = 0.6999
## alternative hypothesis: stationary
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines_latestData[, "Pol3"]
## Dickey-Fuller = -1.4899, Lag order = 2, p-value = 0.7667
## alternative hypothesis: stationary
```

```
adf.test(diff(usVaccines_latestData[,"Pol3"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_latestData[, "Pol3"])
## Dickey-Fuller = -1.2212, Lag order = 2, p-value = 0.8691
## alternative hypothesis: stationary
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines_latestData[, "Hib3"]
## Dickey-Fuller = -1.1869, Lag order = 2, p-value = 0.8821
## alternative hypothesis: stationary
```

```
adf.test(diff(usVaccines_latestData[,"Hib3"]))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_latestData[, "Hib3"])
## Dickey-Fuller = -2.6274, Lag order = 2, p-value = 0.3334
## alternative hypothesis: stationary
```

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  usVaccines_latestData[, "MCV1"]
## Dickey-Fuller = -2.7309, Lag order = 2, p-value = 0.294
## alternative hypothesis: stationary
```

```
adf.test(diff(usVaccines_latestData[,"MCV1"]))
```

```
## Warning in adf.test(diff(usVaccines_latestData[, "MCV1"])): p-value smaller than
## printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(usVaccines_latestData[, "MCV1"])
## Dickey-Fuller = -18.728, Lag order = 2, p-value = 0.01
## alternative hypothesis: stationary
```

We can see that for HepB_BD the test fails to reject the null hypothesis that the data is non stationary as the p value is above the alpha level threshold of 0.05. For the other variables, the values are significant hence the data is stationary. Hence we have mostly removed trend and cyclicity in the data.

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)?*

```
#selecting only the years that start showing constant trend
usVaccines_latestData_const <- window(usVaccines_latestData,start=2010,
                                      end=2017)
usVaccines_latestData_const
```

```
## Time Series:
## Start = 2010
## End = 2017
## Frequency = 1
##      DTP1 HepB_BD Pol3 Hib3 MCV1
## 2010   98      64   94   88   90
## 2011   98      69   94   94   92
## 2012   97      72   93   93   91
## 2013   98      74   93   93   92
## 2014   98      72   93   93   92
## 2015   98      72   93   93   92
## 2016   98      64   94   93   92
## 2017   98      64   94   93   92
```

```
#checking the mean
summary(usVaccines_latestData_const)
```

```
##       DTP1          HepB_BD           Pol3           Hib3           MCV1
##  Min.   :97.00   Min.   :64.00   Min.   :93.0   Min.   :88.0   Min.   :90.00
##  1st Qu.:98.00   1st Qu.:64.00   1st Qu.:93.0   1st Qu.:93.0   1st Qu.:91.75
##  Median :98.00   Median :70.50   Median :93.5   Median :93.0   Median :92.00
##  Mean   :97.88   Mean   :68.88   Mean   :93.5   Mean   :92.5   Mean   :91.62
##  3rd Qu.:98.00   3rd Qu.:72.00   3rd Qu.:94.0   3rd Qu.:93.0   3rd Qu.:92.00
##  Max.   :98.00   Max.   :74.00   Max.   :94.0   Max.   :94.0   Max.   :92.00
```

```
# checking the mean as a whole
mean(usVaccines_latestData_const)
```

```
## [1] 88.875
```

We can see that the mean for DTP1 is 97.88, HEPB_BD is 68.88, Pol3 is 93.5, Hib3 is 92.5 and MCV1 is 91.62. The overall vaccination mean is 88.9. Hence, Hepatitis B Birth Dose is the lowest.

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?*

```
#checking the mean based on districts

#aggregate((districts_log %>%  select(WithDTP, WithPolio, WithMMR,
#                                      WithHepB)),
#          list(districts_log$DistrictName)
#          , FUN=mean)
```

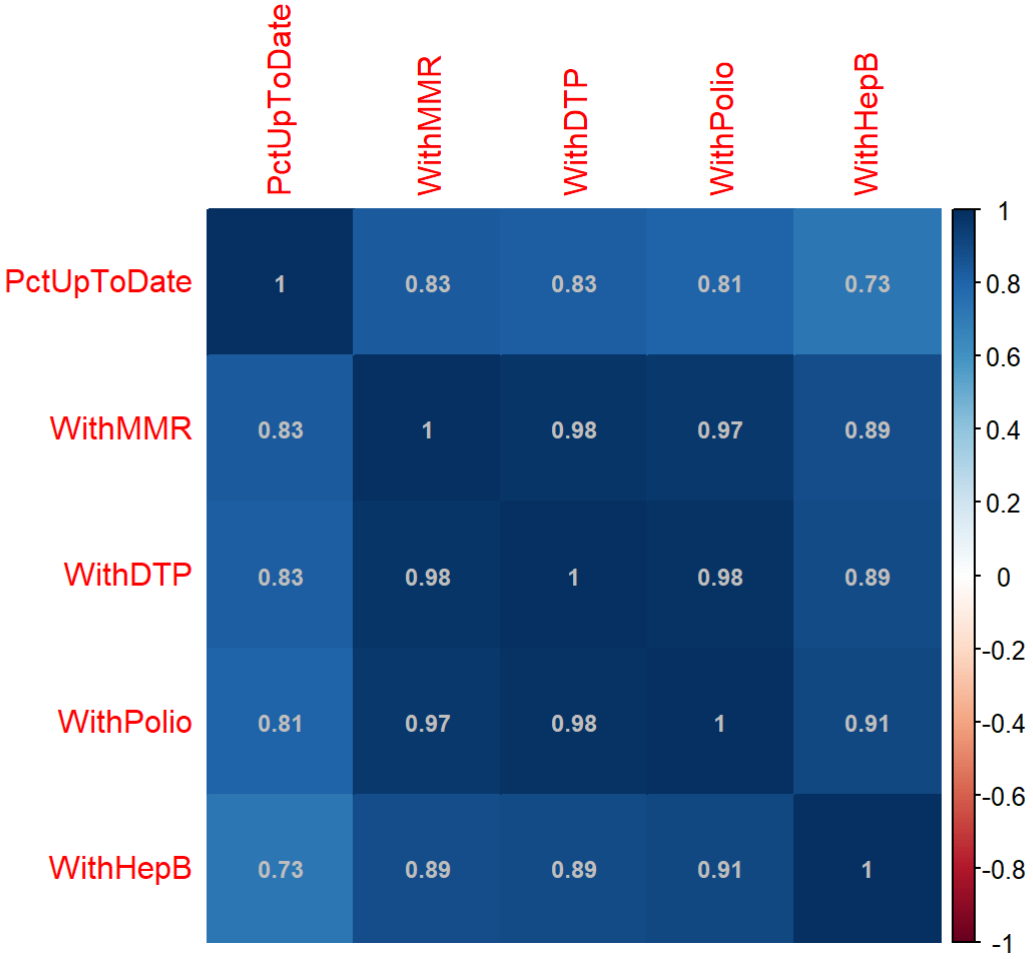
```
summary(districts_log %>%  dplyr::select(WithDTP, WithPolio, WithMMR,
                                         WithHepB))
```

```
##     WithDTP        WithPolio        WithMMR          WithHepB
##  Min.   : 23.0   Min.   : 23.0   Min.   : 23.00   Min.   : 23.00
##  1st Qu.: 86.0   1st Qu.: 87.0   1st Qu.: 86.00   1st Qu.: 90.00
##  Median : 93.0   Median : 94.0   Median : 94.00   Median : 96.00
##  Mean   : 89.8   Mean   : 90.2   Mean   : 89.79   Mean   : 92.26
##  3rd Qu.: 97.0   3rd Qu.: 97.0   3rd Qu.: 97.00   3rd Qu.: 98.00
##  Max.   :100.0   Max.   :100.0   Max.   :100.00   Max.   :100.00
```

We can see that the overall mean is 89.8 for WithDTP, 90.2 for WithPolio, 89.79 for WithMMR and 92.26 for WithHepB.

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?*

```
corrplot(cor(districts_log %>%  dplyr::select(WithDTP, WithPolio, WithMMR,
                                               WithHepB, PctUpToDate)),
         method = "color",addCoef.col="grey",order = "AOE",
         number.cex=0.75)
```



Looking at the correlation matrix above, it is visible that children who take one vaccine are very highly likely to take the other vaccines as well and when we check this with how how up to date students are with their vaccines we can see high correlation there as well.

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

```
tabledf <- data.frame(Type = c("U.S. vaccination levels(recent years only)",
                               "Californian vaccination levels"),
                      DTP1 = c(98,89.8),
                       HepB_BD = c(68,92.26),
                       Pol3 =c(93.5,90.2),
                       MMR = c(91.62,89.79))
tabledf
```

Type <chr>	DT… <dbl>	HepB_BD <dbl>	Pol3 <dbl>	MMR <dbl>
U.S. vaccination levels(recent years only)	98.0	68.00	93.5	91.62
Californian vaccination levels	89.8	92.26	90.2	89.79

2 rows

Using t-test to compare the means:

```
#t.test((tabledf$Type == "U.S. vaccination levels(recent years only)"),
#       (tabledf$Type == "Californian vaccination levels"))
compare_ts_districts_DTP <- t.test(usVaccines_latestData_const[,"DTP1"],
                                   districts_log[,"WithDTP"])
compare_ts_districts_DTP
```

```
##
##  Welch Two Sample t-test
##
## data:  usVaccines_latestData_const[, "DTP1"] and districts_log[, "WithDTP"]
## t = 18.557, df = 459.73, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  7.223716 8.934856
## sample estimates:
## mean of x mean of y
##  97.87500  89.79571
```

A Welch's unequal variances independent sample t-test was performed to compare Californian vaccination levels to U.S. vaccination levels for DTP. The mean rate for the recent years for U.S. vaccination 98; for California vaccination level, 89.79. The t-test found that the difference was statistically significant at p<<0.05, t(495) = 18.557, p = 2.2e-16(the degrees of freedom is adjusted to account for the unequal variances of the groups). The 95% confidence interval for the difference was 7.2237 to 8.934, which does not include 0. Since the CI doesn't contain 0, we don't think it's likely that the real difference is 0. Since the CI is narrow we have less uncertainty. In summary, it seems that the overall U.S Vaccination rate for DTP was better than California vaccination rate which makes sense as all the states data is included in the us vaccination dataset.

Doing the Bayesian version of the t-test

```
bestout <- BESTmcmc(usVaccines_latestData_const[,"DTP1"],
                                districts_log[,"WithDTP"])
```

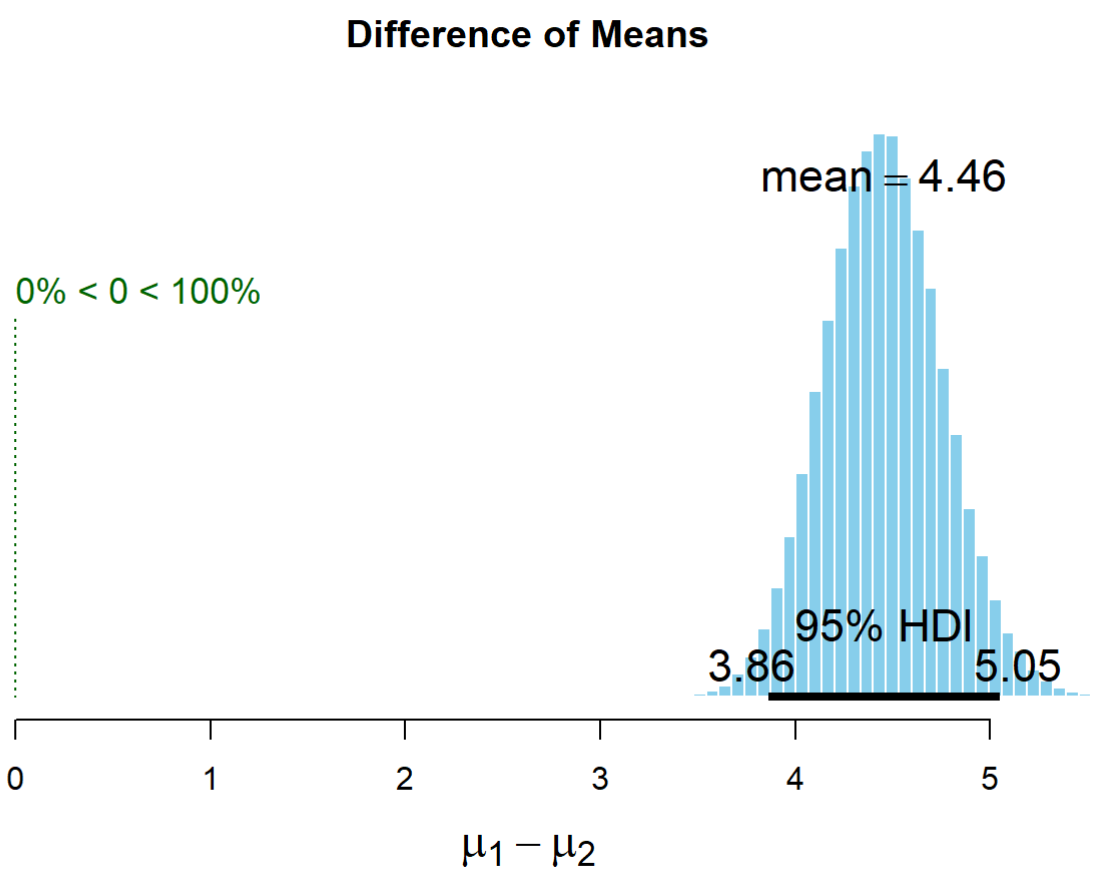
```
## Waiting for parallel processing to complete...done.
```

```
bestout
```

mu1 <dbl>	mu2 <dbl>	nu <dbl>	sigma1 <dbl>	sigma2 <dbl>
98.00077	93.35883	1.724948	0.02168879	5.110460
98.00196	94.02236	1.775094	0.02927149	4.912909
98.01194	93.20322	1.971077	0.03329052	4.837822
98.01690	93.73565	2.028323	0.03485518	5.242821
98.02657	93.64304	1.980548	0.03093456	5.303374
98.02481	93.33045	1.871437	0.03446651	5.415412
98.01893	93.60857	1.805470	0.04958744	5.390424
98.01700	93.88447	1.843220	0.04997215	4.967852
97.99693	93.71029	1.629947	0.04012340	5.130377
98.00182	93.57852	2.024773	0.03807988	5.100454

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

```
plot(bestout)
```



A

Bayesian t-test was performed using the BESTmcmc function in the R BEST package to to compare Californian vaccination levels to U.S. vaccination levels for DTP.10002 simulations were saved. The Rhats for parameters were all reported as 1, suggesting that the sampling converged successfully and the results are interpretable. The MCMC sampling found a mean difference of 4.45 between the vaccination levels between the 2 datasets. The 95% HDI for the difference was 3.86 to 5.04 (i.e., there's a 95% chance that the true difference lies within this range). None of the samples had a difference of less than 0, meaning that the probability that the difference is 0 is extremely low. In summary, the analysis provides very strong evidence that the overall U.S Vaccination rate was better than California vaccination rate for DTP which makes sense as all the states data is included in the us vaccination dataset.

Doing the same for HepB_BD, Pol3 and MMR

```
compare_ts_districts_HepB_BD <- t.test(usVaccines_latestData_const[,"HepB_BD"],
                                       districts_log[,"WithHepB"])
compare_ts_districts_HepB_BD
```

```
##
##  Welch Two Sample t-test
##
## data:  usVaccines_latestData_const[, "HepB_BD"] and districts_log[, "WithHepB"]
## t = -15.079, df = 7.8918, p-value = 4.244e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -26.97307 -19.80264
## sample estimates:
## mean of x mean of y
##  68.87500  92.26286
```

A Welch's unequal variances independent sample t-test was performed to compare Californian vaccination levels to U.S. vaccination levels for HepB. The mean rate for the recent years for U.S. vaccination 68.87; for California vaccination level, 92.26. The t-test found that the difference was statistically significant at p<<0.05, t(7.89) = -15.079, p = 4.244e-07 (the degrees of freedom is adjusted to account for the unequal variances of the groups). The 95% confidence interval for the difference was -26.97 to -19.80, which does not include 0. Since the CI doesn't contain 0, we don't think it's likely that the real difference is 0. Since the CI is narrow we have less uncertainty. In summary, it seems that the overall U.S Vaccination rate for HepB was worse than California vaccination rate which could be possible since california got a mandate for children to get vaccination for HEPB in 1997 while teh rest of the states didn't.

```
bestout_WithHepB <- BESTmcmc(usVaccines_latestData_const[,"HepB_BD"],
                             districts_log[,"WithHepB"])
```

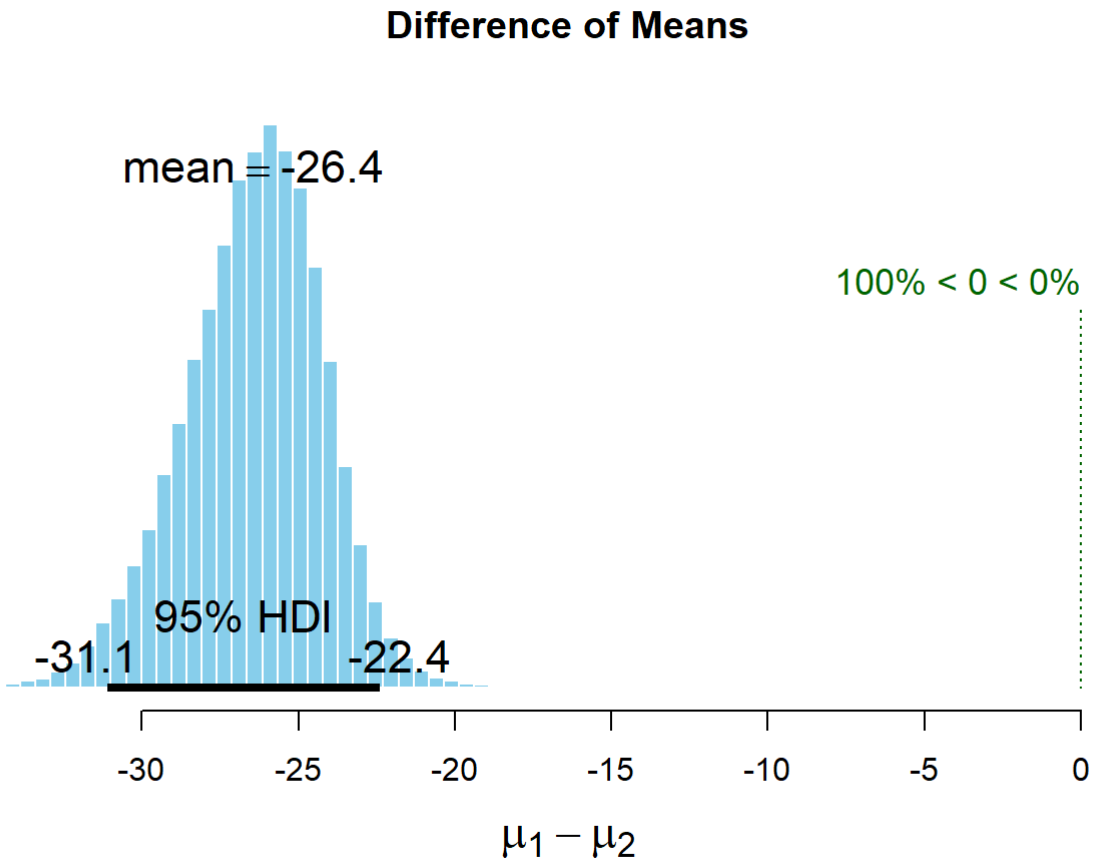
```
## Waiting for parallel processing to complete...done.
```

```
bestout_WithHepB
```

mu1 <dbl>	mu2 <dbl>	nu <dbl>	sigma1 <dbl>	sigma2 <dbl>
71.48846	95.40938	1.855072	5.5495922	4.078823
68.90628	95.94215	1.660890	4.8392152	4.128336
72.28642	95.81574	1.494814	3.1108197	3.903924
66.52304	95.78447	1.524124	3.9267450	3.636946
72.01612	95.84201	1.634051	1.5439791	3.728559
70.49176	96.13915	1.505371	3.4563267	3.948507
69.74709	96.01653	1.603124	3.1202186	3.491430
70.53193	95.75544	1.662980	3.2722292	3.775369
70.86036	95.93889	1.667529	2.9043024	3.735294
71.17461	95.96144	1.610443	4.6245043	3.580558

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

```
plot(bestout_WithHepB)
```



A Bayesian t-test was performed using the BESTmcmc function in the R BEST package to to compare Californian vaccination levels to U.S. vaccination levels for hepB_BD simulations were saved. The Rhats for parameters were all reported as 1, suggesting that the sampling converged successfully and the results are interpretable. The MCMC sampling found a mean difference of -26.4 between the vaccination levels between the 2 datasets. The 95% HDI for the difference was -30.9 to -22.3 (i.e., there's a 95% chance that the true difference lies within this range). None of the samples had a difference of greater than 0, meaning that the probability that the difference is 0 is extremely low. In summary, the analysis provides very strong evidence that the overall U.S Vaccination rate was worse than California vaccination rate for HepB_BD. This could be possible since california got a mandate for children to get vaccination for HEPB in 1997 while the rest of the states didn't.

```
compare_ts_districts_Pol3 <- t.test(usVaccines_latestData_const[,"Pol3"],
                                    districts_log[,"WithPolio"])
compare_ts_districts_Pol3
```

```
##
##  Welch Two Sample t-test
##
## data:  usVaccines_latestData_const[, "Pol3"] and districts_log[, "WithPolio"]
## t = 7.2169, df = 193.34, p-value = 1.182e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.395026 4.196402
## sample estimates:
## mean of x mean of y
##  93.50000  90.20429
```

A Welch's unequal variances independent sample t-test was performed to compare Californian vaccination levels to U.S. vaccination levels for Polio. The mean rate for the recent years for U.S. vaccination 93.5; for California vaccination level, 90.20. The t-test found that the difference was statistically significant at p<<0.05, t(193.34) = 7.2169, p = 1.182e-11 (the degrees of freedom is adjusted to account for the unequal variances of the groups). The 95% confidence interval for the difference was 2.395026 to 4.196402, which does not include 0. Since the CI doesn't contain 0, we don't think it's likely that the real difference is 0. Since the CI is narrow we have less uncertainty. In summary, it seems that the overall U.S Vaccination rate for polio was better than California vaccination rate which makes sense as all the states data is included in the us vaccination dataset.

```
bestout_polio <- BESTmcmc(usVaccines_latestData_const[,"Pol3"],
                                districts_log[,"WithPolio"])
```

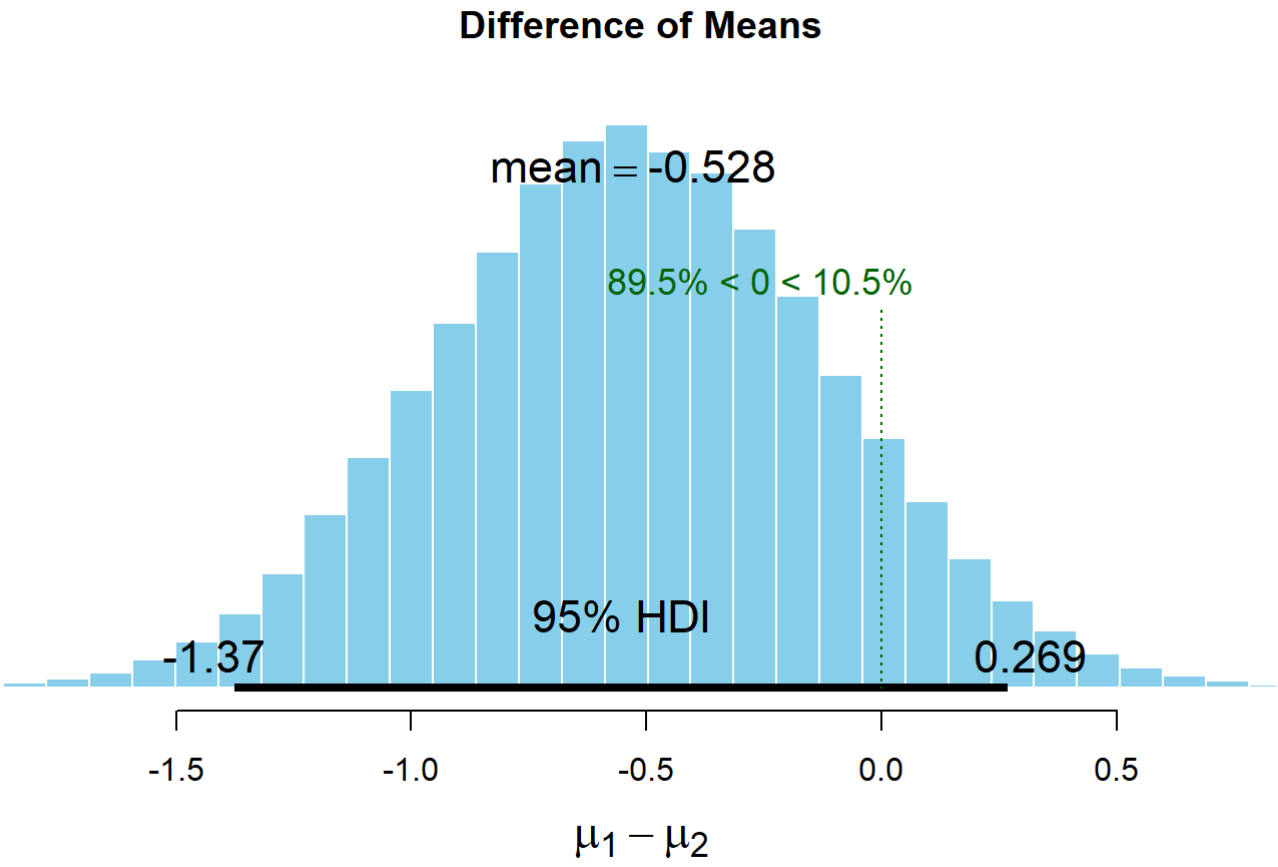
```
## Waiting for parallel processing to complete...done.
```

```
bestout_polio
```

mu1 <dbl>	mu2 <dbl>	nu <dbl>	sigma1 <dbl>	sigma2 <dbl>
93.45839	93.62604	1.945088	1.49236768	5.105042
93.51252	93.99225	1.914769	0.22409631	5.163441
93.18374	93.77014	1.997511	0.68655827	5.113165
93.68045	93.64069	2.247894	1.00038398	5.032987
93.69206	93.68974	2.232305	0.56186331	5.581919
93.35475	93.78409	2.200906	0.56607667	5.583571
93.84429	93.81817	2.409680	0.52434016	5.783784
93.39616	94.16780	2.328589	0.34583109	5.472709
93.46919	93.64777	2.328352	0.59625964	5.494670
93.48911	93.66470	2.304625	0.40091365	5.906922

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

```
plot(bestout_polio)
```



A Bayesian t-test was performed using the BESTmcmc function in the R BEST package to to compare Californian vaccination levels to U.S. vaccination levels for Polio simulations were saved. The Rhats for parameters were all reported as 1, suggesting that the sampling converged successfully and the results are interpretable. The MCMC sampling found a mean difference of -0.53 between the vaccination levels between the 2 datasets. The 95% HDI for the difference was -1.36 to 0.279 (i.e., there's a 95% chance that the true difference lies within this range). Since the HDI contains 0, there is a chance that there is no credible difference between the two groups. We can also see an expression 89.5% < 0 < 10.5% - This expression shows the proportion of mean differences in the MCMC run that were negative vs the the proportion that were positive. Here we can see that 10.5% of the mean differences in the distribution were positive meaning Californian vaccination levels were just slightly higher than the U.S. vaccination levels. Basically the means are very close to each other.

Since we are getting two different results from the frequentist and bayesian approach and since bayesian approach doesn't work that well on small datasets like ours and that our point estimate is 0.53 suggesting barely any difference we will go ahead with the frequentist test answer and conclude that overall U.S Vaccination rate for polio was slightly better than California vaccination rate.

```
compare_ts_districts_MMR <- t.test(usVaccines_latestData_const[,"MCV1"],
                                   districts_log[,"WithMMR"])
compare_ts_districts_MMR
```

```
##
##  Welch Two Sample t-test
##
## data:  usVaccines_latestData_const[, "MCV1"] and districts_log[, "WithMMR"]
## t = 3.6552, df = 87.284, p-value = 0.0004383
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8385246 2.8371897
## sample estimates:
## mean of x mean of y
##  91.62500  89.78714
```

A Welch's unequal variances independent sample t-test was performed to compare Californian vaccination levels to U.S. vaccination levels for MMR. The mean rate for the recent years for U.S. vaccination 91.62; for California vaccination level, 89.787. The t-test found that the difference was statistically significant at p<<0.05, t(87.284) = 3.6552, p = 0.0004383 (the degrees of freedom is adjusted to account for the unequal variances of the groups). The 95% confidence interval for the difference was 0.8385 to 2.837, which does not include 0. Since the CI doesn't contain 0, we don't think it's likely that the real difference is 0. Since the CI is narrow we have less uncertainty. In summary, it seems that the overall U.S Vaccination rate for MMR was better than California vaccination rate which makes sense as all the states data is included in the us vaccination dataset.

```
bestout_MMR <- BESTmcmc(usVaccines_latestData_const[,"MCV1"],
                                districts_log[,"WithMMR"])
```

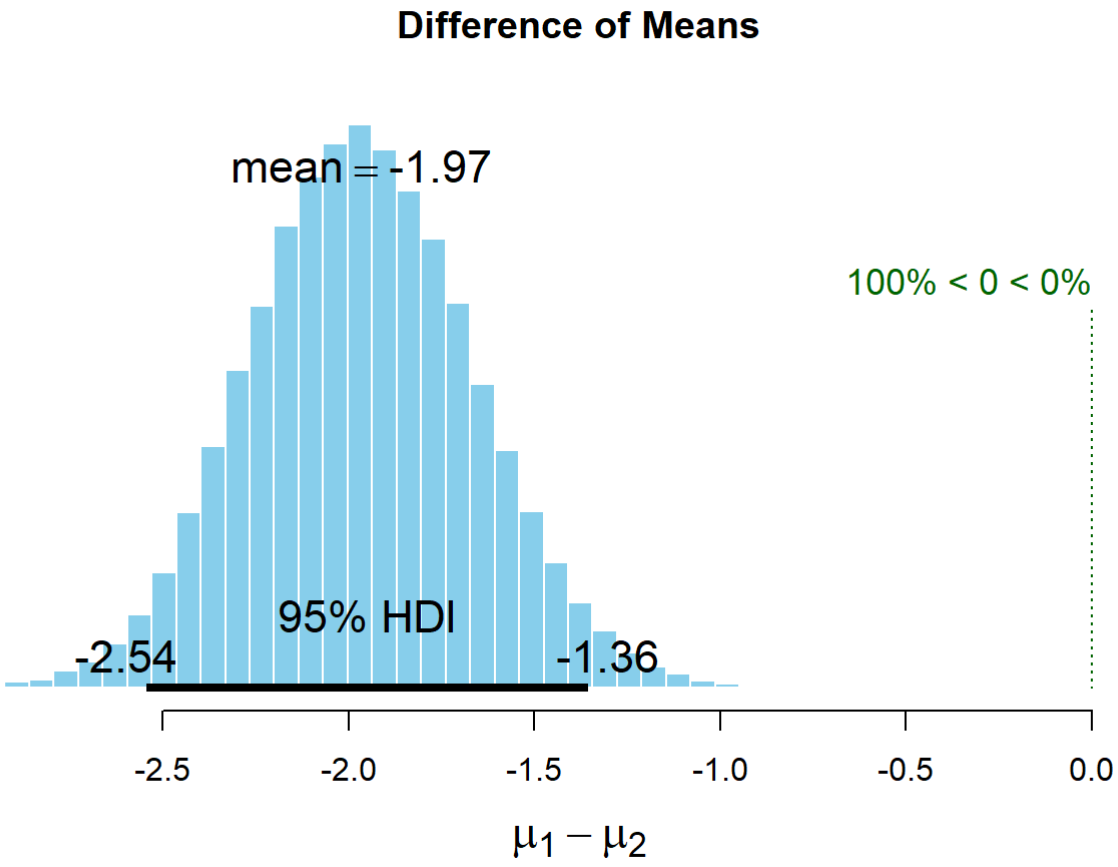
```
## Waiting for parallel processing to complete...done.
```

```
bestout_MMR
```

mu1 <dbl>	mu2 <dbl>	nu <dbl>	sigma1 <dbl>	sigma2 <dbl>
91.98412	93.75163	1.781251	0.02406509	5.424603
92.01585	93.61960	1.789077	0.03658535	5.147975
92.00457	94.02385	1.776714	0.01318238	5.304200
91.99786	93.91457	1.737575	0.01602195	5.334118
92.00259	93.86568	1.850047	0.05442732	5.082546
92.03846	93.85399	1.668765	0.04919695	4.859960
92.01684	94.17743	1.677318	0.04806617	4.861371
92.02203	93.91278	1.460431	0.02822388	4.874925
92.02638	93.77164	1.714555	0.04561643	5.031509
92.02230	93.72053	1.866661	0.03409147	5.183090

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

```
plot(bestout_MMR)
```



A Bayesian t-test was performed using the BESTmcmc function in the R BEST package to to compare Californian vaccination levels to U.S. vaccination levels for MMR simulations were saved. The Rhats for parameters were all reported as 1, suggesting that the sampling converged successfully and the results are interpretable. The MCMC sampling found a mean difference of -1.97 between the vaccination levels between the 2 datasets. The 95% HDI for the difference was -2.56 to -1.37 (i.e., there's a 95% chance that the true difference lies within this range). None of the samples had a difference of greater than 0, meaning that the probability that the difference is 0 is extremely low. In summary, the analysis provides very strong evidence that the overall U.S Vaccination rate was worse than California vaccination rate for MMR.

Since we are getting opposite results from the frequentist and Bayesian approach and since bayesian approach showed us that our point estimate is 1.97 suggesting barely any difference we will go ahead with the frequentist test answer and conclude that the overall U.S Vaccination rate for MMR was better than California vaccination rate.

4. *Comparison of public and private schools (i.e., from the All Schools data)*

a. *What proportion of public schools reported vaccination data?*

Using the cleaned dataset

```
public_schools_count <- nrow(subset(schools_sqrt , PUBLIC..PRIVATE == 'PUBLIC'
                                    &  REPORTED == 'Y'))

100*(public_schools_count / nrow(schools_sqrt))
```

```
## [1] 79.97708
```

79.97 schools were public which reported their vaccination data.

Checking proportion how many public schools reported in just the public schools domain

```
100*(public_schools_count / nrow(subset(schools_sqrt , PUBLIC..PRIVATE == 'PUBLIC')))
```

```
## [1] 100
```

All the public schools reported their vaccination data.

Using the original dataset to see what was the actual proportion

```
public_schools_count_og <- nrow(subset(schools , PUBLIC..PRIVATE == 'PUBLIC'
                                  &  REPORTED == 'Y'))

100*(public_schools_count_og / nrow(schools))
```

```
## [1] 75.65371
```

75.65 schools were public which reported their vaccination data.

Checking proportion how many public schools reported in just the public schools domain

```
100*(public_schools_count_og / nrow(subset(schools_sqrt , PUBLIC..PRIVATE == 'PUBLIC')))
```

```
## [1] 100
```

97.41% of the public schools reported their vaccination data.

b. *What proportion of private schools reported vaccination data?*

Using the cleaned dataset

```
private_schools_count <- nrow(subset(schools_sqrt ,
                                   PUBLIC..PRIVATE == 'PRIVATE'
                                  &  REPORTED == 'Y'))

100*(private_schools_count / nrow(schools_sqrt))
```

```
## [1] 20.00859
```

Only 20 percent of the schools which were private reported their vaccination data.

Checking proportion how many public schools reported in just the public schools domain

```
100*(private_schools_count / nrow(subset(schools_sqrt , PUBLIC..PRIVATE == 'PRIVATE')))
```

```
## [1] 99.92847
```

```
subset(schools_sqrt , PUBLIC..PRIVATE == 'PRIVATE' &  REPORTED == 'N')
```

	SCHOOL.C…	PUBLIC..PRIVATE	Public.School.District.ID	PUBLIC.SCHOOL.DISTRIC
	<int>	<chr>	<chr>	<chr>
5674	6143788	PRIVATE	Private	Private

1 row | 1-6 of 19 columns

Only one school in Pleasanton city didn't report their vaccination records.

Using the origial data set

```
private_schools_count_og <- nrow(subset(schools ,
                                  PUBLIC..PRIVATE == 'PRIVATE'
                                 &  REPORTED == 'Y'))

100*(private_schools_count_og / nrow(schools))
```

```
## [1] 18.92697
```

Only 18.92 percent of the schools which were private reported their vaccination data.

Checking proportion how many public schools reported in just the public schools domain

```
100*(private_schools_count_og / nrow(subset(schools , PUBLIC..PRIVATE == 'PRIVATE')))
```

```
## [1] 84.71801
```

84.72% reported their vaccination data for private schools.

c. *Was there any credible difference in reporting between public and private schools?*

Not really, checking both with cleaned and uncleaned dataset, both show high proportions for reporting their data.

Let's validate this with a chi squared test

First for original data

```
private_schools_count_og_N <- nrow(subset(schools,
                                    PUBLIC..PRIVATE == 'PRIVATE'
                                   &  REPORTED == 'N'))
public_schools_count_og_N <- nrow(subset(schools,
                                    PUBLIC..PRIVATE == 'PUBLIC'
                                   &  REPORTED == 'N'))

SchoolOg <- matrix(c(public_schools_count_og, public_schools_count_og_N,
                     private_schools_count_og, private_schools_count_og_N),
                   ncol=2, byrow=TRUE)
colnames(SchoolOg) <- c('NotReported','Reported')
rownames(SchoolOg) <- c('Public','Private')
SchoolOg <- as.table(SchoolOg)
addmargins(SchoolOg)
```

```
##         NotReported Reported  Sum
## Public         5584      148 5732
## Private        1397      252 1649
## Sum            6981      400 7381
```

```
chisqOut <- chisq.test(SchoolOg)
chisqOut
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  SchoolOg
## X-squared = 400.49, df = 1, p-value < 2.2e-16
```

The reported value of chi-square is 400.49 on 1 degree of freedom. The df is 1 as its a 2x2 table and df was calculated using the formula of (2-1)*(2-1). The chi-squared value is very high (for 1 df and alpha level of 0.05 the chi-squared critical value is 3.84) and hence we can reject the null hypothesis and can say that the two public and private schools are non independent and that there is no credible difference.

Now checking for the cleaned data:

```
private_schools_count_N <- nrow(subset(schools_sqrt,
                                    PUBLIC..PRIVATE == 'PRIVATE'
                                   &  REPORTED == 'N'))
public_schools_count_N <- nrow(subset(schools_sqrt,
                                    PUBLIC..PRIVATE == 'PUBLIC'
                                   &  REPORTED == 'N'))

SchoolClean <- matrix(c(public_schools_count, public_schools_count_N,
                     private_schools_count, private_schools_count_N),
                   ncol=2, byrow=TRUE)
colnames(SchoolClean) <- c('NotReported','Reported')
rownames(SchoolClean) <- c('Public','Private')
SchoolOg <- as.table(SchoolClean)
addmargins(SchoolClean)
```

```
##         NotReported Reported  Sum
## Public         5584        0 5584
## Private        1397        1 1398
## Sum            6981        1 6982
```

```
chisqOutClean <- chisq.test(SchoolClean)
```

```
## Warning in chisq.test(SchoolClean): Chi-squared approximation may be incorrect
```

```
chisqOutClean
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  SchoolClean
## X-squared = 0.56124, df = 1, p-value = 0.4538
```

The reported value of chi-square is 0.5612 on 1 degree of freedom. The df is 1 as its a 2x2 table and df was calculated using the formula of (2-1)*(2-1). The chi-squared value is very low (for 1 df) and we can see that P-value is 0.45 which is greater than alpha level of 0.05 and hence we fail to reject the null hypothesis and can say that the two public and private schools are independent for the cleaned dataset and that there is credible difference.

Overall for the original dataset we can see that there is credible difference.

d. *Does the proportion of students with up-to-data vaccinations vary from county to county?*

```
total_up_to_date <- sum(schools_sqrt$UP_TO_DATE)
by_county <- schools_sqrt %>%
   group_by(COUNTY) %>%
   summarise(Proportion.Percent = (sum(UP_TO_DATE)/total_up_to_date)*100)
by_county[order(by_county$Proportion.Percent, decreasing=TRUE),]
```

COUNTY <chr>	**Proportion.Percent** <dbl>
LOS ANGELES	23.964019294
SAN DIEGO	8.094405485
ORANGE	7.935851398
RIVERSIDE	6.582396945
SAN BERNARDINO	6.343416872
SANTA CLARA	5.030906558
ALAMEDA	3.643401622
FRESNO	3.491741191
SACRAMENTO	3.427609301

COUNTY <chr>	Proportion.Percent <dbl>
KERN	3.125124034

1-10 of 58 rows Previous **1** 2 3 4 5 6 Next

As we can see in the table above, the proportion of students with up-to-data vaccinations vary from county to county and LA has the highest rate of up to date vaccinations which is 23.9% and then the second one in San Diego with 8% (possibly due to the vaccination mandate in 1997).

Using Anova frequentist and bayesian to test this further:

```
aov_schools_sqrt <- schools_sqrt
aov_schools_sqrt$COUNTY <- as.factor(aov_schools_sqrt$COUNTY)
county_aov <- aov(UP_TO_DATE ~ COUNTY, data = aov_schools_sqrt)
summary(county_aov)
```

```
##               Df   Sum Sq Mean Sq F value Pr(>F)
## COUNTY        57   960203   16846   10.47 <2e-16 ***
## Residuals   6924 11139890    1609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that we get significant results and that we can reject the null hypothesis that students with up to date vaccinations do not vary from county to county. Looking at the results we see - F(57,6924) = 10.47, p<<0.05. Hence, we can say that yes that atleast some county see proportion of students with up to date vaccinations.

Bayesian Approach:

```
bayesaov <- anovaBF(UP_TO_DATE ~ COUNTY, data = aov_schools_sqrt) # Calc Bayes Factors
mcmcaov <- posterior(bayesaov,iterations=10000) # Run mcmc iterations
#summary(mcmcaov)
bayesaov
```

```
## Bayes factor analysis
## --------------
## [1] COUNTY : 7.109797e+86 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

Seeing that our value is 7.109797e+86 we can say that we have strong evidence that the proportion varies by county. In conclusion, we can see that the proportions vary by county but can't be sure which of the county and we can see that using the group by results that was shown at the beginning of this question.

5. *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.

Looking at the analysis so far, with or without outlier in the school data set we could see a very high proportion of children being vaccinated and as we just saw above, in the whole of the dataset of schools that we are provided, two major cities of California are topping the charts for highest vaccination rate. Comparing the California school districts rate with that of the whole of U.S we can see that California is doing a great job in terms of vaccinating children.

6. *Inferential reporting about districts*

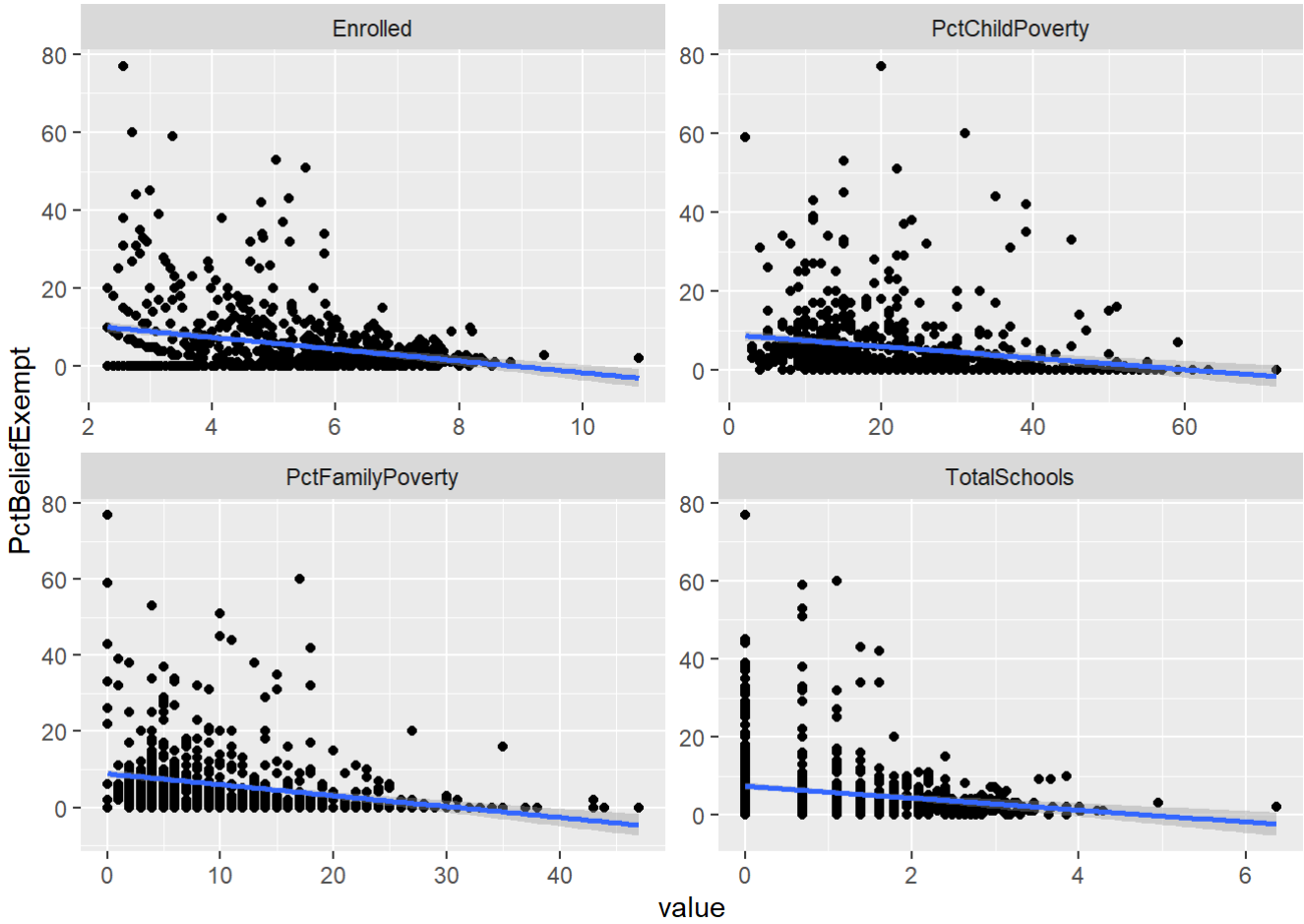
For every item below except question c, use PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors. Explore the data and transform variables as necessary to improve prediction and/or interpretability. Be sure to include appropriate diagnostics and modify your analyses as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

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

```
# creating a new df of the cleaned data for better readability
districts_log_belief <- subset(districts_log, select = c(PctChildPoverty,
                                                         PctFamilyPoverty,
                                                         Enrolled,
                                                         TotalSchools,
                                                         PctBeliefExempt))

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

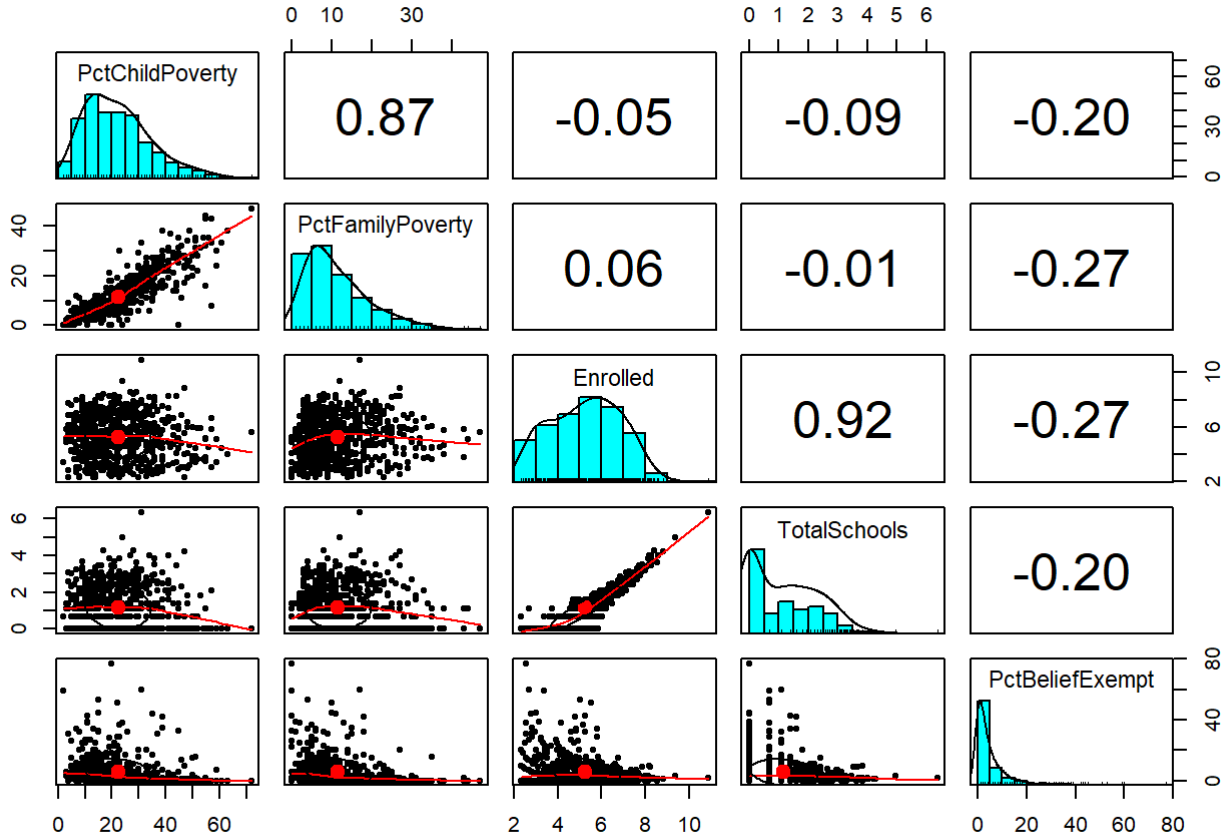
```
## `geom_smooth()` using formula 'y ~ x'
```



Looking at the above graphs we can see that there is slight negative correlation in between the variables with respect to percentage of students with belief exceptions.

Let's check the pairs plot to examine plots and correlation:

```
pairs.panels(districts_log_belief)
```



We can see that PctChildPoverty, PctFamilyPoverty and Enrolled are almost normal and TotalSchools is slightly skewed (but alot better than before doing the log transformation) and PctBeliefExempt is right skewed. We can also see that there is high correlation between Percentage of children in district living below the poverty line and Percentage of families in district living below the poverty line which makes sense as they can be inter related i.e they must be children of the families staying in districts below the poverty line. There is also high correlation between totalschools and enrolled students which makes sense as there is interloping of the districts with a child being enrolled and in the district of the school. Rest of the correlations are low which is great for linear modelling.

```
belief_lm <- lm(PctBeliefExempt ~ ., data=districts_log_belief)
```

Lets check multicolinarity in the model:

```
vif(belief_lm)
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##         4.237053         4.283179         6.492899         6.380393
```

Values in the range of 4 to 5 are regarded as being moderate to high for VIF

```
vif(lm(PctBeliefExempt ~ . - TotalSchools, data=districts_log_belief))
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled
##         4.224727         4.225850         1.046572
```

Looks like removing totalschools reduces VIF for enrolled

```
vif(lm(PctBeliefExempt ~ . - PctChildPoverty, data=districts_log_belief))
```

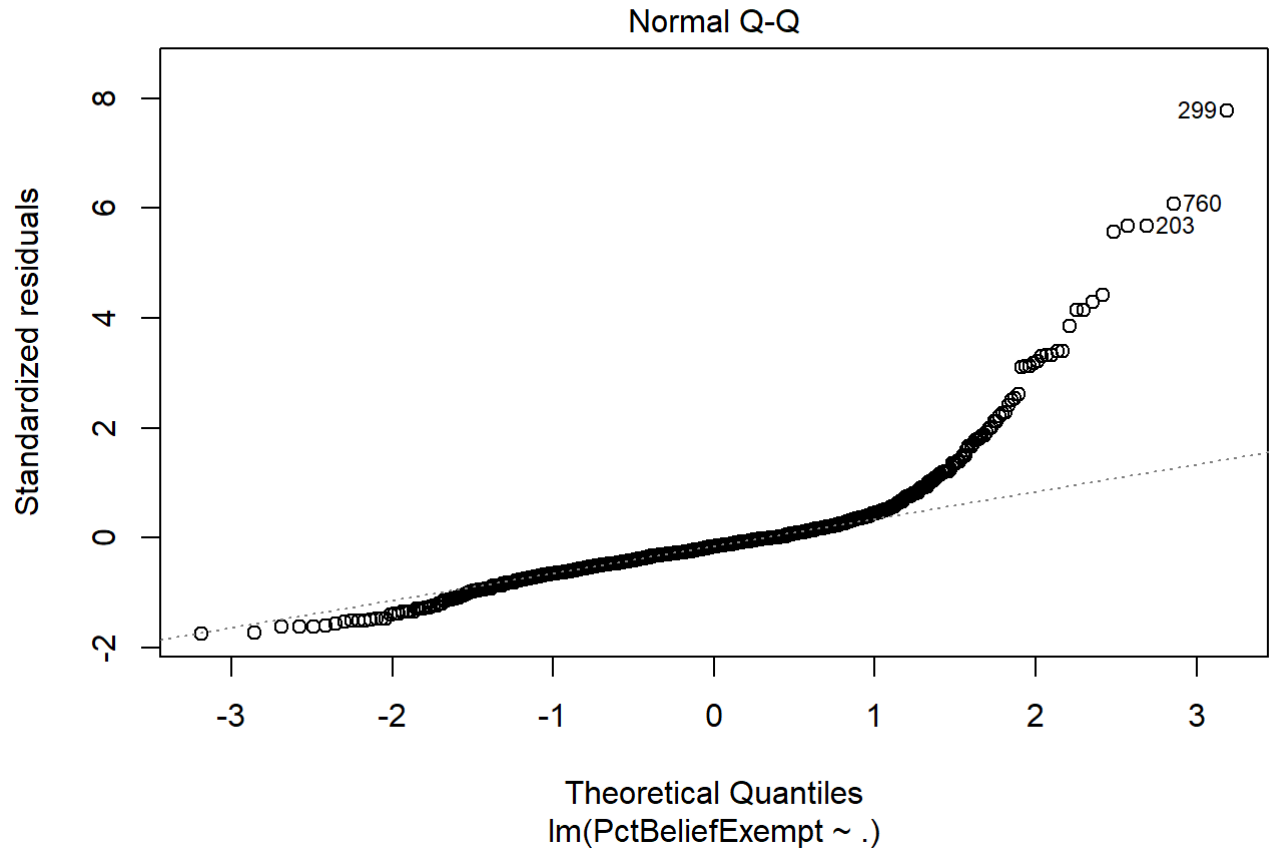
```
## PctFamilyPoverty         Enrolled     TotalSchools
##         1.023323         6.381273         6.361831
```

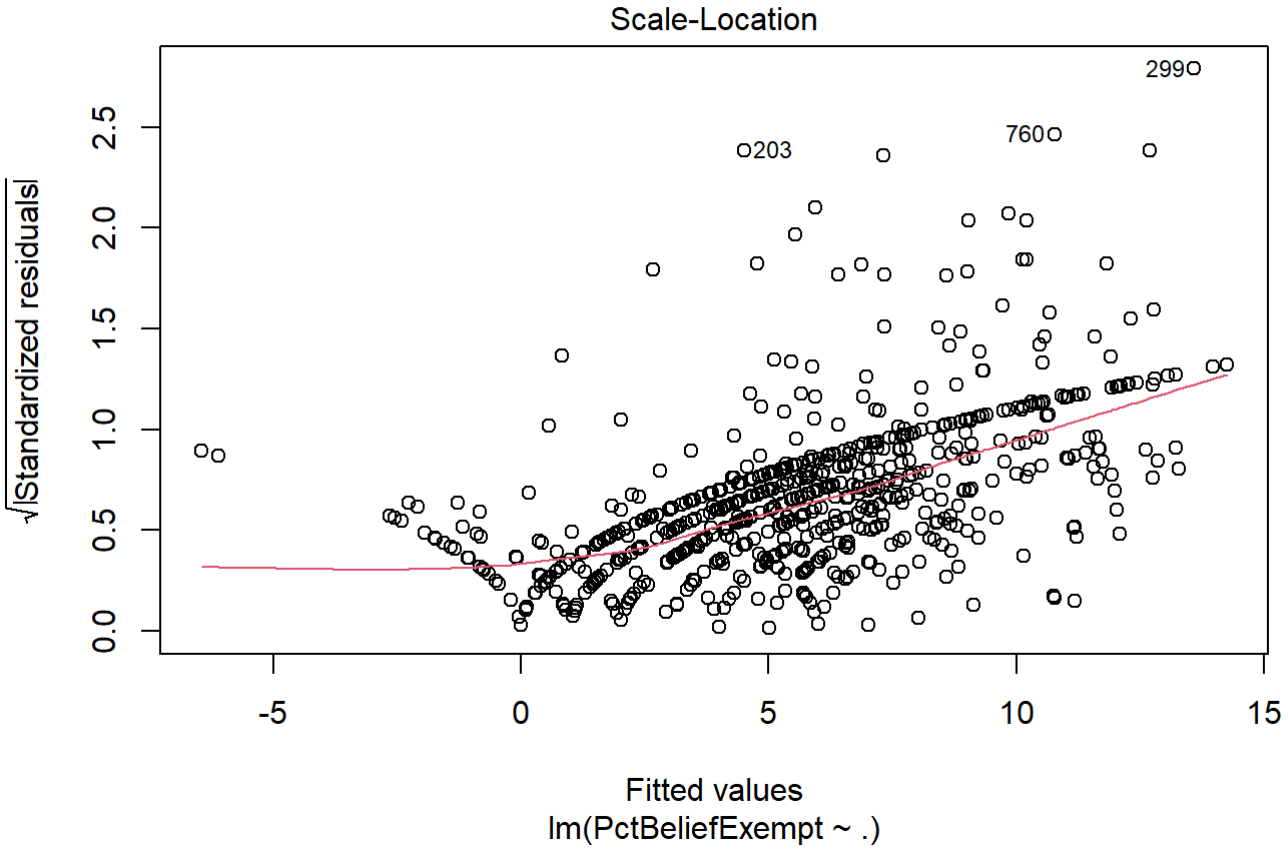
Looks like removing PctChildPoverty reduces VIF for PctFamilyPoverty

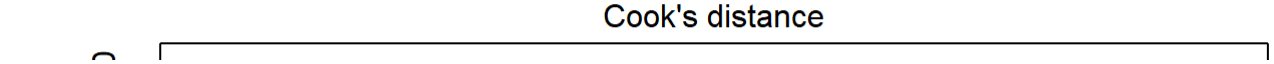
Since the VIF for these columns is below 10 and looking at model where we removed columns to check for VIF it seems that since TotalSchools and Enrolled are highly correlated as shown in the pairs plot and Child and Family poverty columns are highly correlated we get a moderate VIF of value 6. We will go with the original model with all four columns as the VIF is moderate.

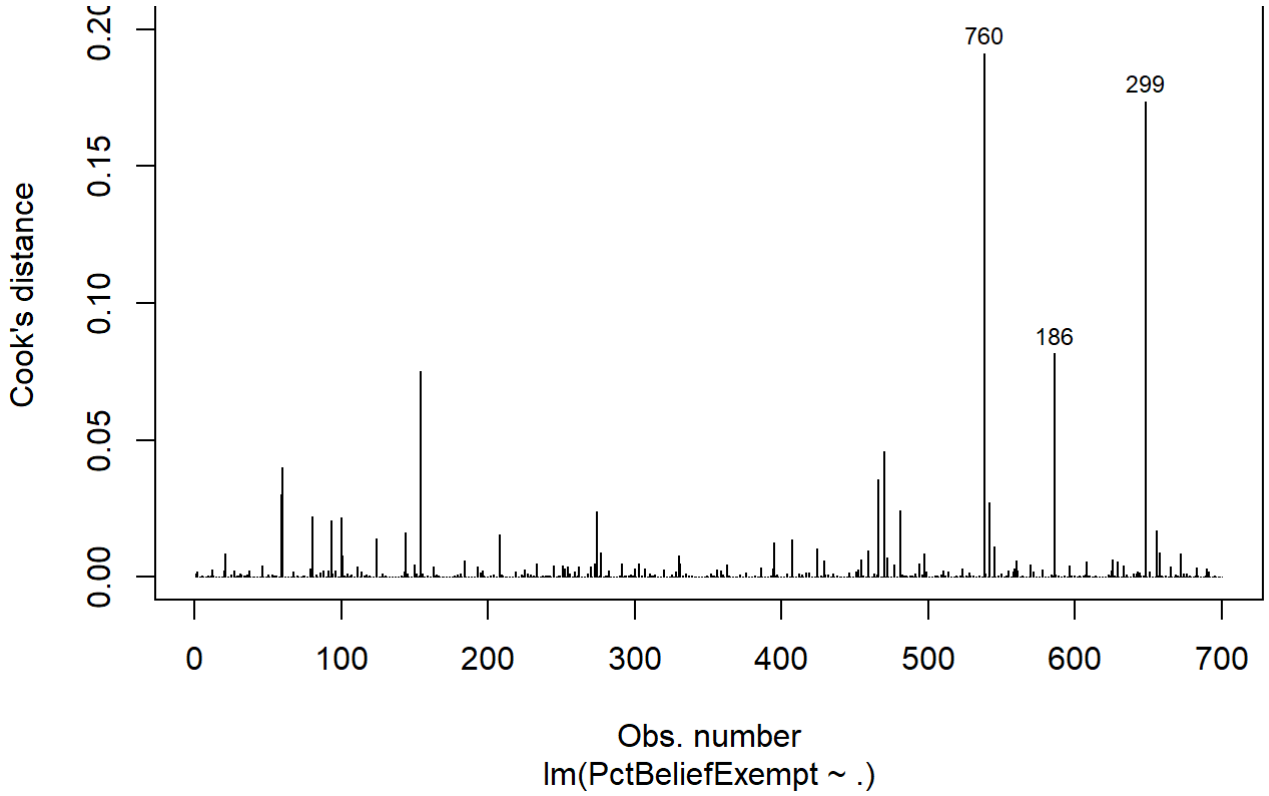
Checking the residual plots

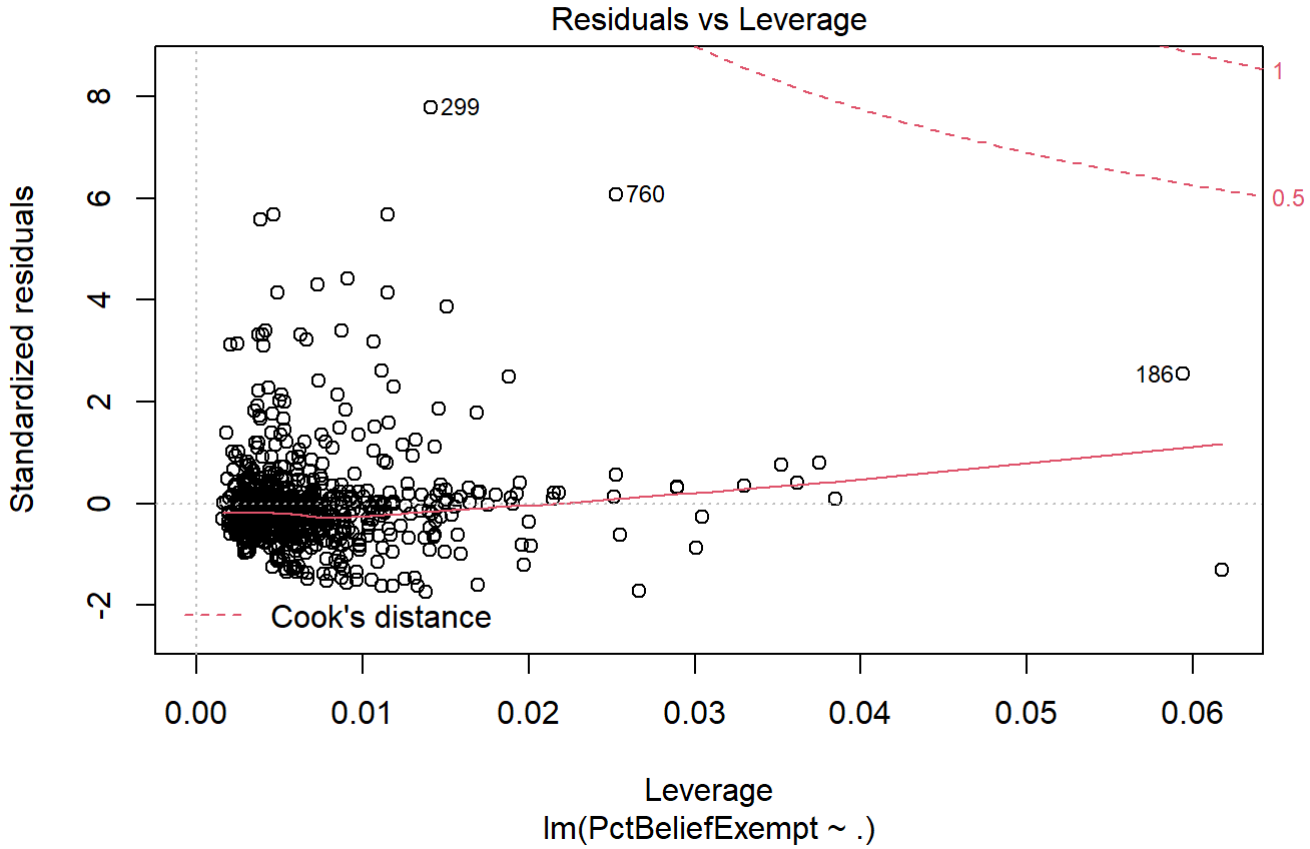
```
plot(belief_lm, which=2:5)
```


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(PctBeliefExempt ~ .)
299
760
203


Scale-Location
√|Standardized residuals|
Fitted values
lm(PctBeliefExempt ~ .)
299
760
203


Cook's distance





We can see at the Cook's distance graph that observation number 760 only has 0.20 influence on the regression line, 186 has 0.10 and 299 has around 0.18. Looking at the Residuals vs Leverage graph we can see that nothing is in the Cook's distance which is good. The q-q plot tells us how normally distributed the residuals are which is a basic assumption of the linear regression. Looking at the graph we can see that at the end we have more variability than the model can account for but overall the model is not bad. Now checking our model summary

```
summary(belief_lm)
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ ., data = districts_log_belief)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.250  -3.980  -1.272   1.467  63.411
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      20.273906   2.017310  10.050  < 2e-16 ***
## PctChildPoverty   0.001632   0.052491   0.031 0.975208
## PctFamilyPoverty -0.259860   0.078542  -3.309 0.000986 ***
## Enrolled         -2.618940   0.495613  -5.284 1.69e-07 ***
## TotalSchools      1.785245   0.681047   2.621 0.008950 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.209 on 695 degrees of freedom
## Multiple R-squared:  0.1458, Adjusted R-squared:  0.1409
## F-statistic: 29.65 on 4 and 695 DF,  p-value: < 2.2e-16
```

We can see that the median of -1.272 is close to 0. We can also see that the f-statistic is F(4,695) = 29.65, the r-squared and adjusted r squared is 0.1458 and 0.1409 respectively which isn't a huge value but the p-value is significant. Looking at the columns we can see that PctFamilyPoverty significantly predicts PctBeliefExempt (b = -0.259, t(695)=-3.309, p<.001) Enrolled significantly predicts PctBeliefExempt (b = -2.618, t(695)=-5.284, p<.001) TotalSchools significantly predicts PctBeliefExempt (b = -1.78, t(695)=2.62, p<.01) PctChildPPoverty is not significant as p value of 0.975 is greater than the alpha level of 0.05.

To see which predictors have the biggest impact on the result, we will compare standardized coefficients, i.e., those based on standardized variables:

```
summary(lm.beta(belief_lm))
```

```
##
## Call:
## lm(formula = PctBeliefExempt ~ ., data = districts_log_belief)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.250  -3.980  -1.272   1.467  63.411
##
## Coefficients:
##                   Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      20.273906     0.000000   2.017310  10.050  < 2e-16 ***
## PctChildPoverty   0.001632     0.002243   0.052491   0.031 0.975208
## PctFamilyPoverty -0.259860    -0.240055   0.078542  -3.309 0.000986 ***
## Enrolled         -2.618940    -0.472054   0.495613  -5.284 1.69e-07 ***
## TotalSchools      1.785245     0.232131   0.681047   2.621 0.008950 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.209 on 695 degrees of freedom
## Multiple R-squared:  0.1458, Adjusted R-squared:  0.1409
## F-statistic: 29.65 on 4 and 695 DF,  p-value: < 2.2e-16
```

The significant variables are the percentage of family poverty, the number of enrolled students and number of different schools in the district. It means the more poverty in the area, less belief exception, more enrolled students less belief exception. and one unit change in school can lead to belief expection to increase. Since enrolled and total schools have log values so every 1 standard deviation increase in the log of number of students enrolled in the district will have 2.61 standard deviation decrease in the percentage of belief exception. Similarly every 1 standard deviation increase in the log of number of total schools in the district will have 1.78 standard deviation increase in the percentage of belief exception and 1 SD increase in percentage of families in district living below the poverty line will have 0.259 decrease in the percentage of belief exception.

Doing the Bayesian Test for the same:

```
belief_lmBF <- lmBF(PctBeliefExempt ~ ., data=districts_log_belief,
                    posterior=TRUE, iterations=10000, rnd.seed=772)
summary(belief_lmBF)
```

```
##
## 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                5.625884 0.30666 0.0030666      0.0031375
## PctChildPoverty   0.001432 0.05216 0.0005216      0.0005216
## PctFamilyPoverty -0.253510 0.07796 0.0007796      0.0007796
## Enrolled         -2.557157 0.48878 0.0048878      0.0048878
## TotalSchools      1.740225 0.67086 0.0067086      0.0067086
## sig2             67.422823 3.59895 0.0359895      0.0359895
## g                 0.102222 0.17655 0.0017655      0.0018713
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%       50%     75%    97.5%
## mu                5.01655  5.41922  5.626598  5.8317  6.22770
## PctChildPoverty  -0.10297 -0.03337  0.001371  0.0370  0.10345
## PctFamilyPoverty -0.40341 -0.30551 -0.254187 -0.2014 -0.09873
## Enrolled         -3.51355 -2.88659 -2.559565 -2.2189 -1.60994
## TotalSchools      0.41354  1.27959  1.734451  2.2011  3.05790
## sig2             60.75167 64.95763 67.296823 69.7922 74.88466
## g                 0.02246  0.04384  0.067502  0.1102  0.37497
```

In the output displayed above, we have parameter estimates for the B-weights of each of our predictions (the column labeled “Mean”). In the second section, we have the 2.5% and 97.5% boundaries of the HDI for each of the B-weights.These boundaries mark the edges of the central region of the posterior distribution for each B-weight. So PctChildPoverty predictor has a lower bound of -0.099 to upper bound at 0.1027, since the HDI contains 0 the observed differences could be due to chance. The PctFamilyPoverty predictor has a lower bound of -0.406 to upper bound of -0.1024, Since the HDI does not contain 0 we have credible evidence that there is a difference in between the variables. The Enrolled predictor has HDI from -3.52 to -1.60 and totalschools has HDI from 0.44 to 3.04 and since both these preictors dont span 0 we have credible evidence that the difference in between the variables. Also we can see that the means from our bayesian test match the estimates we got from the frequentist model.

```
# running the same model without the iterations to get bayes factor value
belief_lmBF_out <- lmBF(PctBeliefExempt ~
                          PctChildPoverty+PctFamilyPoverty+Enrolled+TotalSchools
                        ,data=districts_log_belief,
                        rnd.seed=772)
belief_lmBF_out
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFamilyPoverty + Enrolled + TotalSchools : 3.797184e+19 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

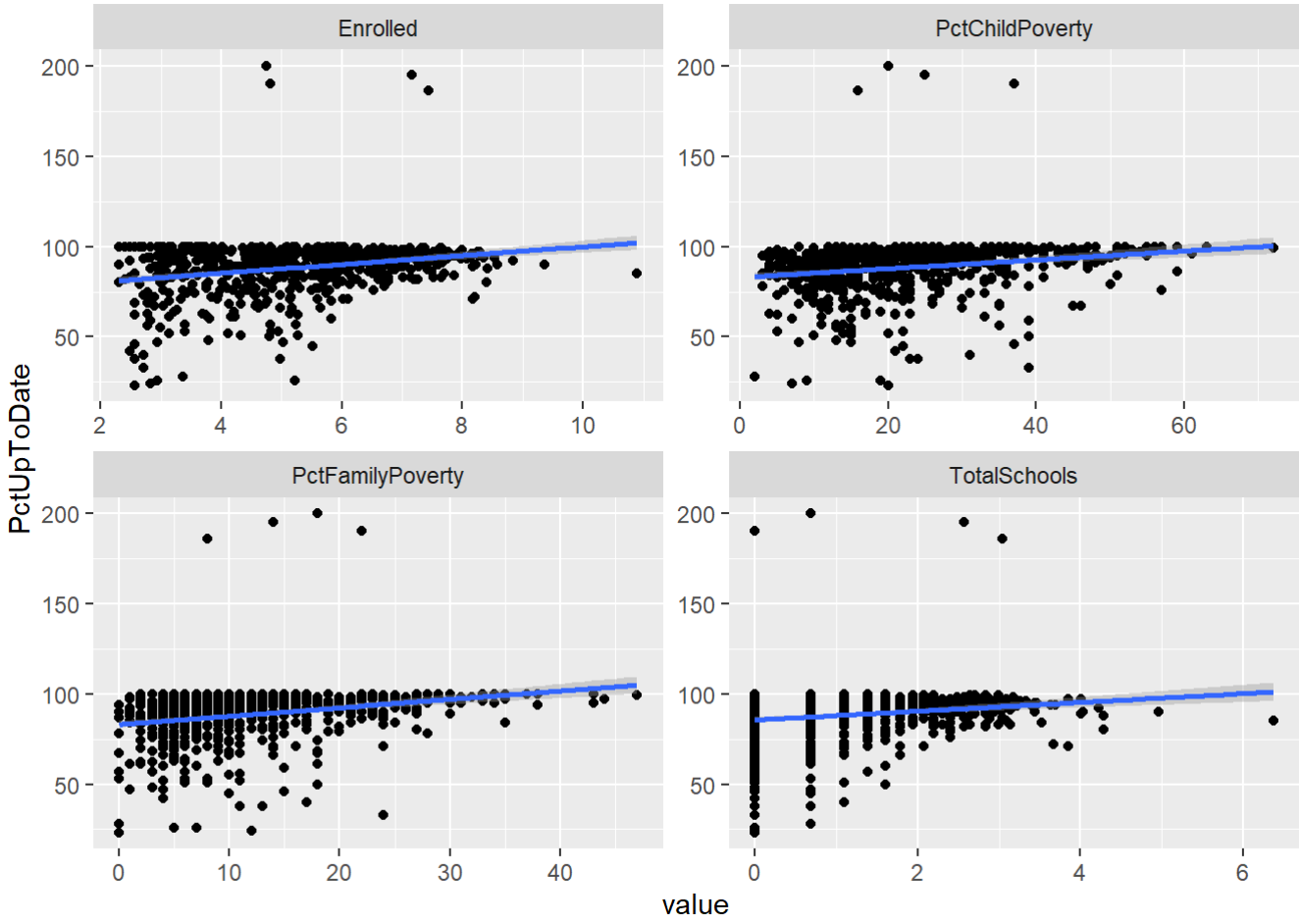
We get a very high bayes factor showing us that our results are significant. In conclusion, a linear regression was performed to estimate the percentage of all enrolled students with belief exceptions with use of PctChildPoverty, PctFamilyPoverty, Enrolled , and TotalSchools as the four predictors. In the data cleaning process to remove skewness we took log transformation on Enrolled and Total schools columns. In the bivariate analysis we saw that the remaining or little skewness didn't cause any major issues and that the data was linear to carry out linear regression. Using the result from both the Bayesian and frequentist approach we got evidence that PctFamilyPoverty , Enrolled and TotalSchools are good predictors for PctBeliefExempt.

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

```
# creating a new df of the cleaned data for better readability
districts_log_uptodate <- subset(districts_log, select = c(PctChildPoverty,
                                                           PctFamilyPoverty,
                                                           Enrolled,
                                                           TotalSchools,
                                                           PctUpToDate))

districts_log_uptodate %>% 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")
```

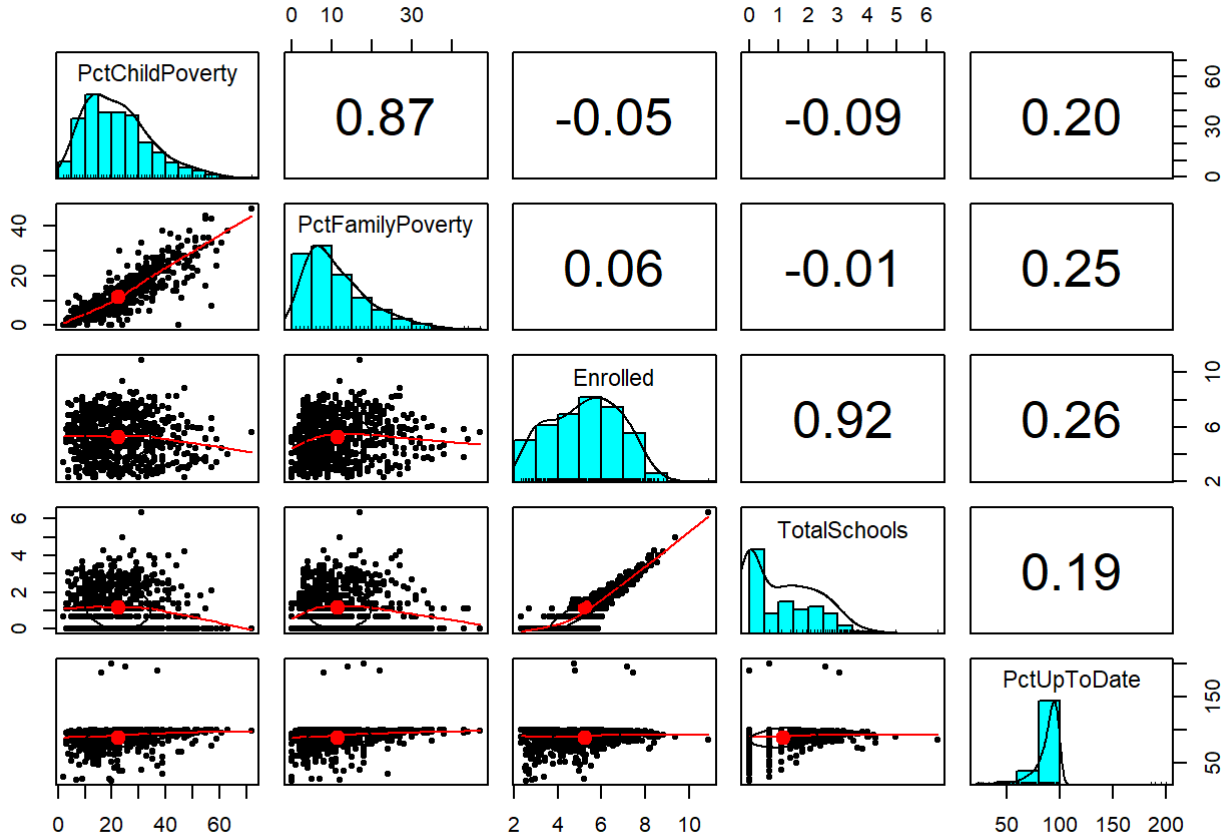
```
## `geom_smooth()` using formula 'y ~ x'
```



Looking at the above graphs we can see that there is an almost a positive correlation in between the variables with respect to percentage of students with completely up-to-date vaccines.

Let's check the pairs plot to examine plots and correlation:

```
pairs.panels(districts_log_uptodate)
```



We can see that PctChildPoverty, PctFamilyPoverty, PctUpToDate and Enrolled are almost normal and TotalSchools is slightly skewed (but alot better than before doing the log transformation). We can also see that there is high correlation between Percentage of children in district living below the poverty line and Percentage of families in district living below the poverty line which makes sense as they can be inter related i.e they must be children of the families staying in districts below the poverty line. There is also high correlation between totalschools and enrolled students which makes sense as there is interloping of the districts with a child being enrolled and in the district of the school. Rest of the correlations are low which is great for linear modelling.

```
uptodate_lm <- lm(PctUpToDate ~ ., data=districts_log_uptodate)
```

Lets check multicolinarity in the model:

```
vif(uptodate_lm)
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##         4.237053         4.283179         6.492899         6.380393
```

Values in the range of 4 to 5 are regarded as being moderate to high for VIF

```
vif(lm(PctUpToDate ~ . - TotalSchools, data=districts_log_uptodate))
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled
##         4.224727         4.225850         1.046572
```

Looks like removing totalschools reduces VIF for enrolled

```
vif(lm(PctUpToDate ~ . - PctChildPoverty, data=districts_log_uptodate))
```

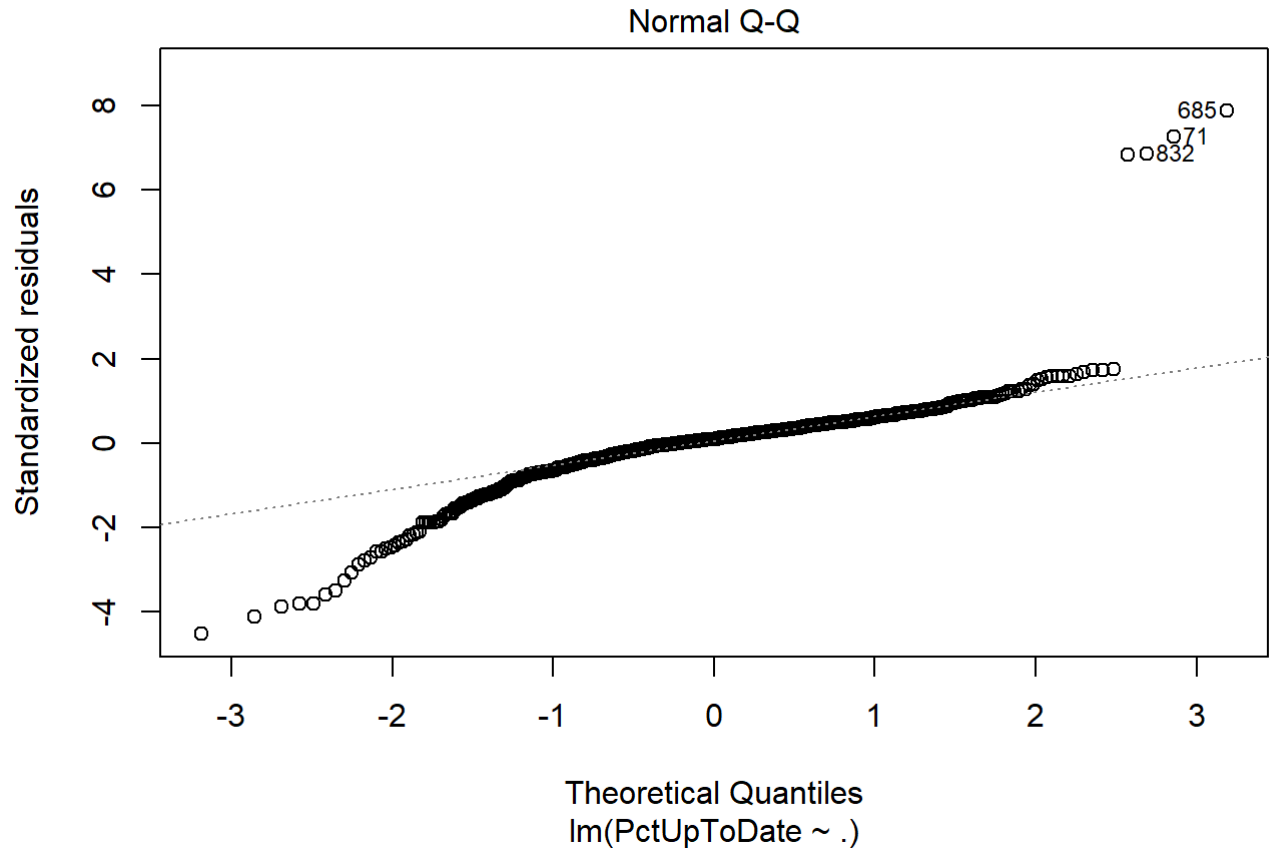
```
## PctFamilyPoverty         Enrolled     TotalSchools
##         1.023323         6.381273         6.361831
```

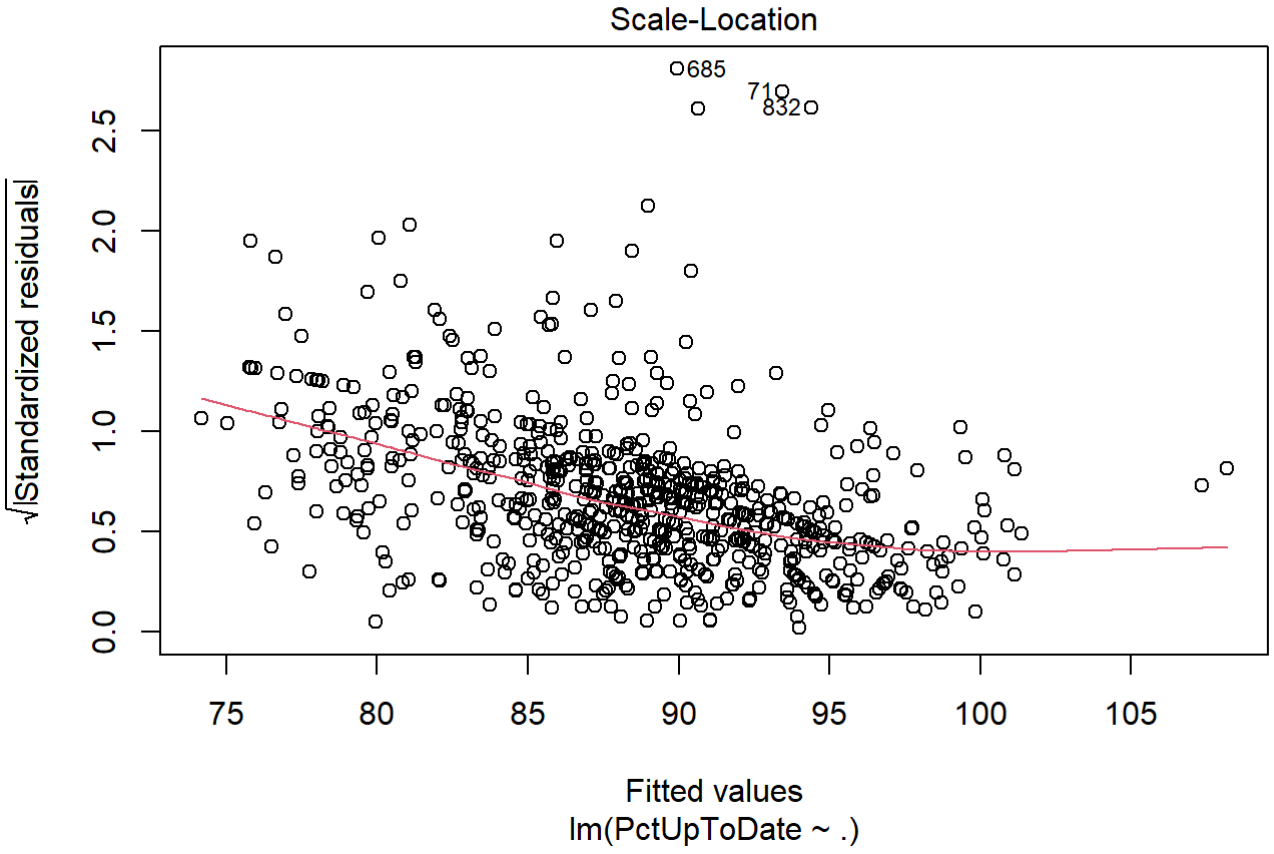
Looks like removing PctChildPoverty reduces VIF for PctFamilyPoverty

Since the VIF for these columns is below 10 and looking at model where we removed columns to check for VIF it seems that since TotalSchools and Enrolled are highly correlated as shown in the pairs plot and Child and Family poverty columns are highly correlated we get a moderate VIF of value 6. We will go with the original model with all four columns as the VIF is moderate.

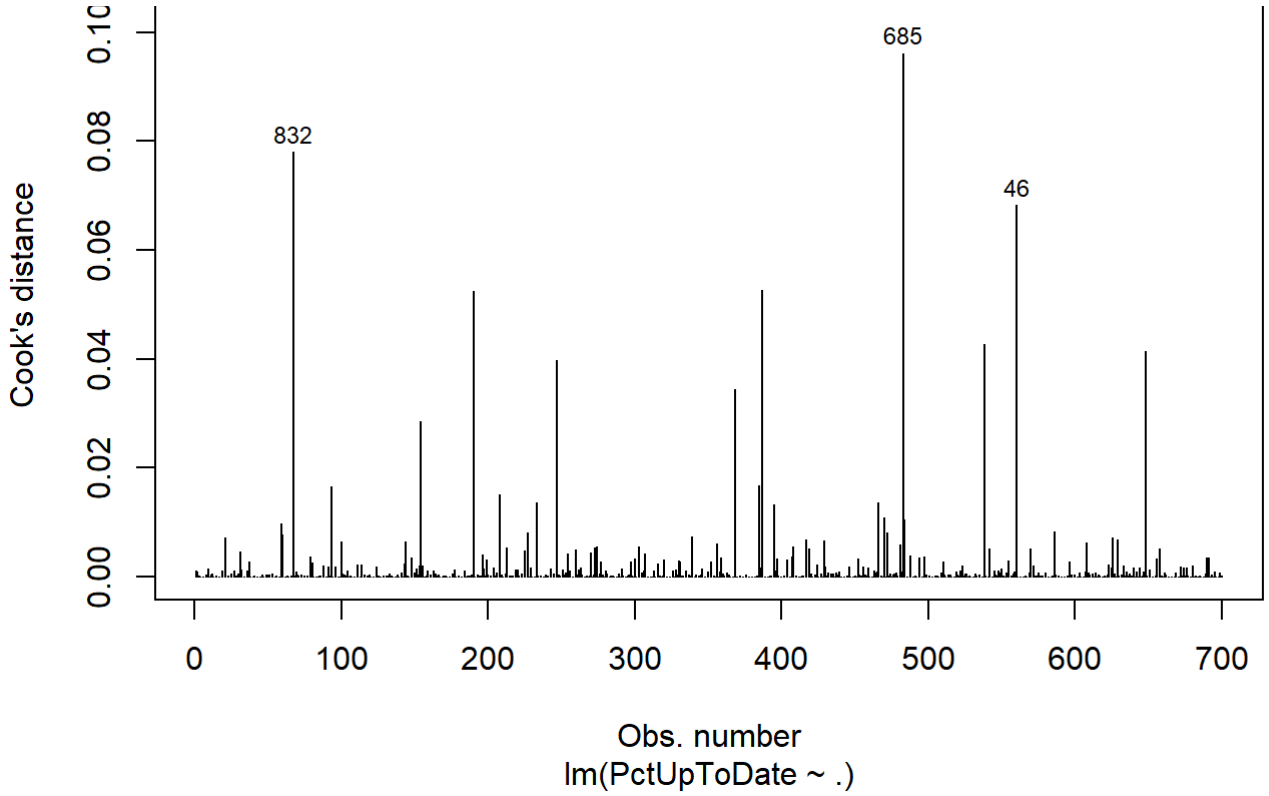
Checking the residual plots

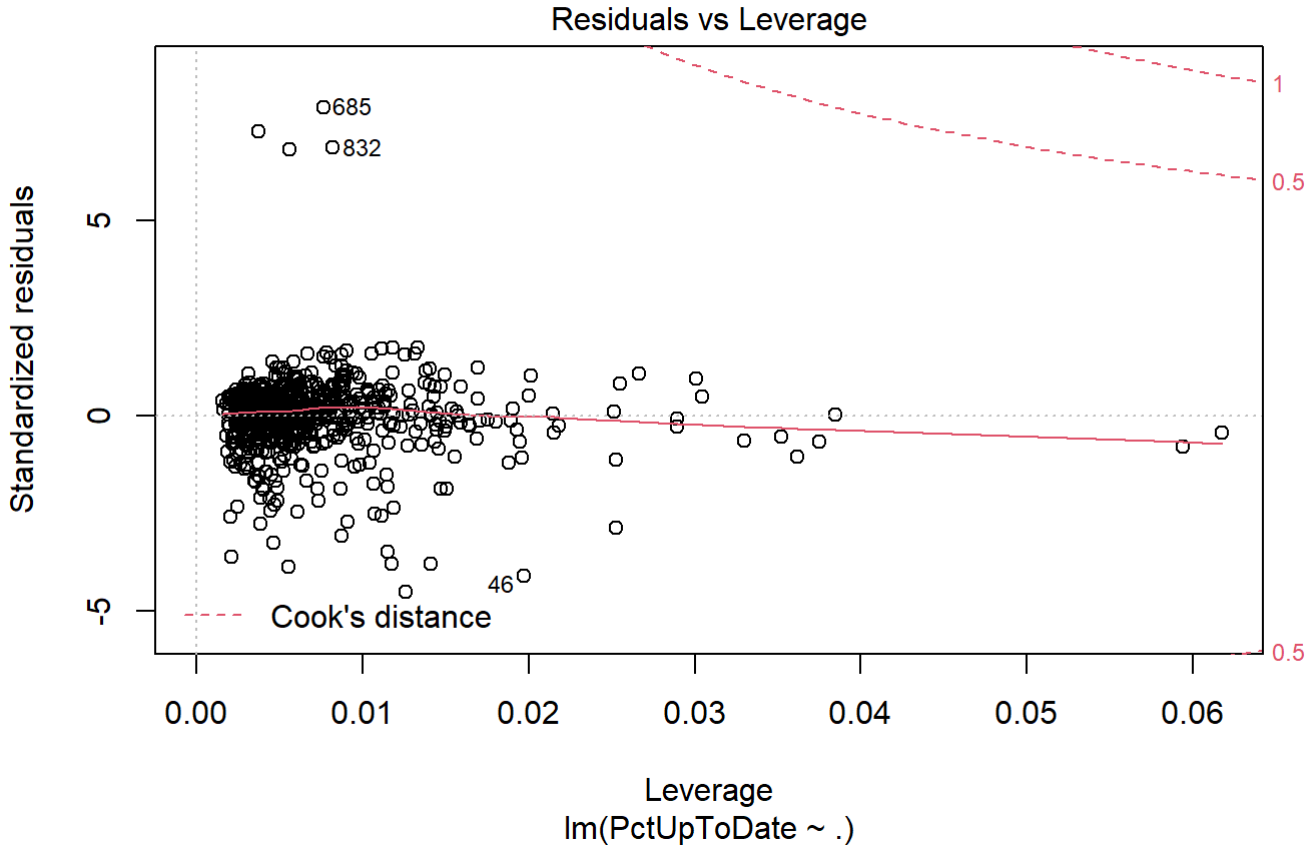
```
plot(uptodate_lm, which=2:5)
```





Cook's distance





We can see at the Cook's distance graph that observation number 832 only has 0.08 influence on the regression line, 685 has 0.10 and 46 has around 0.07. Looking at the Residuals vs Leverage graph we can see that nothing is in the Cook's distance which is good. The q-q plot tells us how normally distributed the residuals are which is a basic assumption of the linear regression. Looking at the graph we can see that at the start we have more variability than the model can account for but overall the model is good.

Now checking our model summary

```
summary(uptodate_lm)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ ., data = districts_log_uptodate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.987  -4.633   1.312   6.185 110.053
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      63.93271    3.43860  18.593  < 2e-16 ***
## PctChildPoverty   0.03017    0.08947   0.337  0.73604
## PctFamilyPoverty  0.37267    0.13388   2.784  0.00552 **
## Enrolled          4.39497    0.84480   5.202 2.59e-07 ***
## TotalSchools     -3.10330    1.16088  -2.673  0.00769 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.99 on 695 degrees of freedom
## Multiple R-squared:  0.133,  Adjusted R-squared:  0.128
## F-statistic: 26.66 on 4 and 695 DF,  p-value: < 2.2e-16
```

We can see that the median of 1.312 is close to 0. We can also see that the f-statistic is F(4,695) = 26.66, the r-squared and adjusted r squared is 0.133 and 0.128 respectively which isn't a huge value but the p-value is significant. Looking at the columns we can see that PctFamilyPoverty significantly predicts PctUpToDate (b = 0.03017, t(695)=2.784, p<.01) Enrolled significantly predicts PctUpToDate (b = 4.39, t(695)= 5.202, p<.001) TotalSchools significantly predicts PctUpToDate (b = -3.103, t(695)=-2.673, p<.01) PctChildPPoverty is not significant as p value of 0.736 is greater than the alpha level of 0.05.

To see which predictors have the biggest impact on the result, we will compare standardized coefficients, i.e., those based on standardized variables:

```
summary(lm.beta(uptodate_lm))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ ., data = districts_log_uptodate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.987  -4.633   1.312   6.185 110.053
##
## Coefficients:
##                  Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)      63.93271      0.00000    3.43860  18.593  < 2e-16 ***
## PctChildPoverty   0.03017      0.02452    0.08947   0.337  0.73604
## PctFamilyPoverty  0.37267      0.20347    0.13388   2.784  0.00552 **
## Enrolled          4.39497      0.46821    0.84480   5.202 2.59e-07 ***
## TotalSchools     -3.10330     -0.23849    1.16088  -2.673  0.00769 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.99 on 695 degrees of freedom
## Multiple R-squared:  0.133,  Adjusted R-squared:  0.128
## F-statistic: 26.66 on 4 and 695 DF,  p-value: < 2.2e-16
```

The significant variables are the percentage of family poverty, the number of enrolled students and number of different schools in the district. It means the more poverty in the area, more percentage of students with completely up-to-date vaccines, more enrolled students more up to date vaccination and one unit change in school can lead to decrease in up to date vaccination which makes sense as more new students in the new school and it will take some time for all of them to be vaccinated. Since enrolled and total schools have log values so every 1 standard deviation increase in the log of number of students enrolled in the district will have 4.39 standard deviation increase in the percentage of up to date vaccination. Similarly every 1 standard deviation increase in the log of number of total schools in the district will have 3.10 standard deviation decrease in the percentage of up to date vaccination and 1 SD increase in percentage of families in district living below the poverty line will have 0.37 increase in the percentage of up to date vaccinations.

Doing the Bayesian Test for the same:

```
belief_lmBF <- lmBF(PctBeliefExempt ~ ., data=districts_log_belief,
                    posterior=TRUE, iterations=10000, rnd.seed=772)
summary(belief_lmBF)
```

```
##
## 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                5.627973 0.31556 0.0031556      0.0031556
## PctChildPoverty   0.002792 0.05141 0.0005141      0.0005141
## PctFamilyPoverty -0.255237 0.07743 0.0007743      0.0007743
## Enrolled         -2.557965 0.49452 0.0049452      0.0051589
## TotalSchools      1.746810 0.68350 0.0068350      0.0068350
## sig2             67.434387 3.64214 0.0364214      0.0378860
## g                 0.099088 0.17551 0.0017551      0.0017551
##
## 2. Quantiles for each variable:
##
##                      2.5%      25%       50%      75%   97.5%
## mu                5.02126  5.41270  5.624617  5.84206  6.2560
## PctChildPoverty  -0.09968 -0.03192  0.002821  0.03845  0.1031
## PctFamilyPoverty -0.40721 -0.30800 -0.255036 -0.20300 -0.1040
## Enrolled         -3.52564 -2.89288 -2.557838 -2.22123 -1.6093
## TotalSchools      0.41653  1.28774  1.751908  2.21208  3.0901
## sig2             60.66436 64.93456 67.283089 69.78647 75.0324
## g                 0.02246  0.04364  0.067504  0.11027  0.3604
```

In the output displayed above, we have parameter estimates for the B-weights of each of our predictions (the column labeled “Mean”). In the second section, we have the 2.5% and 97.5% boundaries of the HDI for each of the B-weights.These boundaries mark the edges of the central region of the posterior distribution for each B-weight. So PctChildPoverty predictor has a lower bound of -0.100 to upper bound at 0.100, since the HDI contains 0 the observed differences could be due to chance. The PctFamilyPoverty predictor has a lower bound of -0.405 to upper bound of -0.09761, Since the HDI does not contain 0 we have credible evidence that there is a difference in between the variables. The Enrolled predictor has HDI from -3.51 to -1.58 and totalschools has HDI from 0.42 to 3.06 and since both these preictors dont span 0 we have credible evidence that the difference in between the variables. Also we can see that the means from our bayesian test don’t match the estimates we got from the frequentist model.

```
# running the same model without the iterations to get bayes factor value
uptodate_lmBF_out <- lmBF(PctUpToDate ~
                          PctChildPoverty+PctFamilyPoverty+Enrolled+TotalSchools
                        ,data=districts_log_uptodate,
                        rnd.seed=772)
uptodate_lmBF_out
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty + PctFamilyPoverty + Enrolled + TotalSchools : 2.452817e+17 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

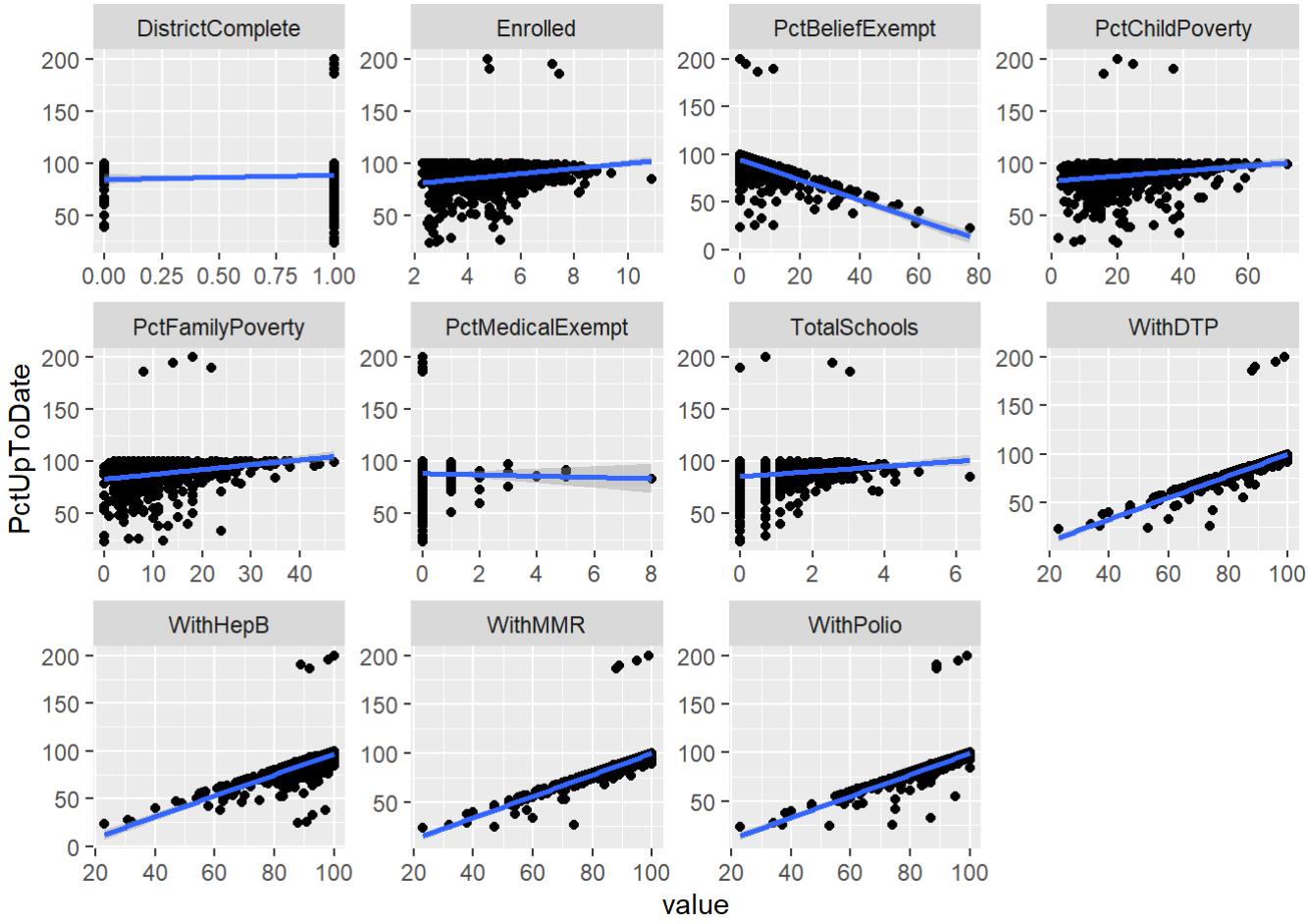
We get a very high bayes factor of 2.45 e+17 showing us that our results are significant. In conclusion, a linear regression was performed to estimate the percentage of all enrolled students with up to date vaccination with use of PctChildPoverty, PctFamilyPoverty, Enrolled, and TotalSchools as the four predictors. In the data cleaning process to remove skewness we took log transformation on Enrolled and Total schools columns. In the bivariate analysis we saw that the remaining or little skewness didn't cause any major issues and that the data was linear to carry out linear regression. Using the result from both the Bayesian and frequentist approach we got evidence that PctFamilyPoverty , Enrolled and TotalSchools are good predictors for PctUpToDate.

c. *Using any set or combination 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?*

Checking with the cleaned model where PctFreeMeal was removed.

```
districts_log %>% pivot_longer(-c(PctUpToDate,DistrictName), 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")
```

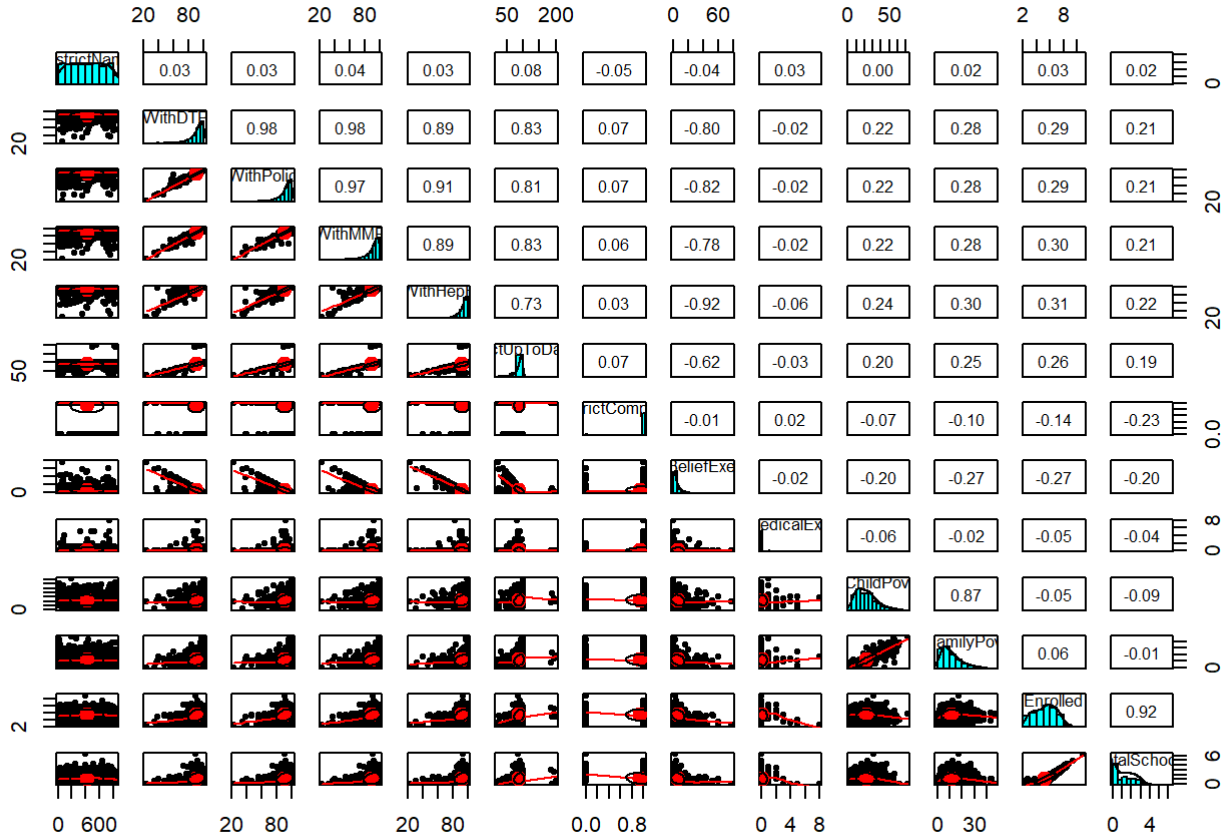
```
## `geom_smooth()` using formula 'y ~ x'
```



WithDTP, WithHepB, WithMMR, WithPolio show a good positive correlation, Enrolled, PctChildPoverty, PctFamilyPoverty, TotalSchools show sub par or an almost positive correlation, PctBeliefExempt shows a negative correlation and PctMedicalExempt shows an almost about to happen negative correlation with respect to percentage of students with completely up-to-date vaccines.

Let's check the pairs plot to examine plots and correlation:

```
pairs.panels(districts_log)
```



We can see that PctChildPoverty, PctFamilyPoverty, PctUpToDate and Enrolled are almost normal and TotalSchools is slightly skewed (but alot better than before doing the log transformation). We can also see that there is high correlation between Percentage of children in district living below the poverty line and Percentage of families in district living below the poverty line which makes sense as they can be inter related i.e they must be children of the families staying in districts below the poverty line. There is also high correlation between totalschools and enrolled students which makes sense as there is interloping of the districts with a child being enrolled and in the district of the school. Rest of the correlations are low which is great for linear modelling.

Let's create models and find which model has the least multicolinarity and then analyze it further

```
vif(lm(PctUpToDate ~ .-DistrictName, data=districts_log))
```

```
##           WithDTP         WithPolio           WithMMR         WithHepB
##         41.239411         32.374276         24.311894         13.440201
## DistrictComplete  PctBeliefExempt PctMedicalExempt  PctChildPoverty
##          1.136615          7.083332          1.057158          4.284271
## PctFamilyPoverty          Enrolled      TotalSchools
##          4.409052          7.200366          7.027245
```

We are removing column PctChildPoverty as from our previous models and bi variate analysis we can see that it is having multicolinarity with PctFamilyPoverty and TotalSchools as it has multi colinarity with Enrolled.

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName, data=districts_log))
```

```
##          WithDTP        WithPolio          WithMMR         WithHepB
##        41.206415        32.365186        24.305864        13.297409
## DistrictComplete  PctBeliefExempt PctMedicalExempt PctFamilyPoverty
##         1.058160         7.047888         1.050270         1.119531
##         Enrolled
##         1.144201
```

Checking dropping which column can give us least amount of multicolinarity for the group dtp, polio, mmr and hepb

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP
         , data=districts_log))
```

```
##        WithPolio          WithMMR         WithHepB DistrictComplete
##        18.514422        16.401231        13.222778         1.055128
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         7.034568         1.050232         1.117587         1.144170
```

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithMMR
         , data=districts_log))
```

```
##          WithDTP        WithPolio         WithHepB DistrictComplete
##        27.805469        31.697260        12.512217         1.055342
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         6.750636         1.049224         1.119080         1.143079
```

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithPolio
         , data=districts_log))
```

```
##          WithDTP          WithMMR         WithHepB DistrictComplete
##        23.572024        23.804259        12.936572         1.057502
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         7.008121         1.049791         1.119315         1.144166
```

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithHepB
         , data=districts_log))
```

```
##          WithDTP        WithPolio          WithMMR DistrictComplete
##        40.975148        31.486928        22.870640         1.058155
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         3.102735         1.008402         1.111109         1.138765
```

Looking at the above analysis it looks like removing HepB PctBeliefExempt reduces and removing WithDTP reduces the multicolinarity best from polio, mmr and hepb variable

Checking if removing both these columns gives us a good model that doesnt have multicolinarity

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP -WithHepB
         , data=districts_log))
```

```
##        WithPolio          WithMMR DistrictComplete  PctBeliefExempt
##        18.078561        15.420808         1.055125         3.101614
## PctMedicalExempt PctFamilyPoverty         Enrolled
##         1.008319         1.109716         1.138641
```

Since we still don't get a value less than 10 for WithPolio and WithMMR we will remove Polio and we get the least multicolinarity and since we already saw that one child getting one vaccine is very likely to get all the other vaccines, we can just use WithMMR to see if getting a shot of a vaccine is a good predictor for having up to date vaccinations.

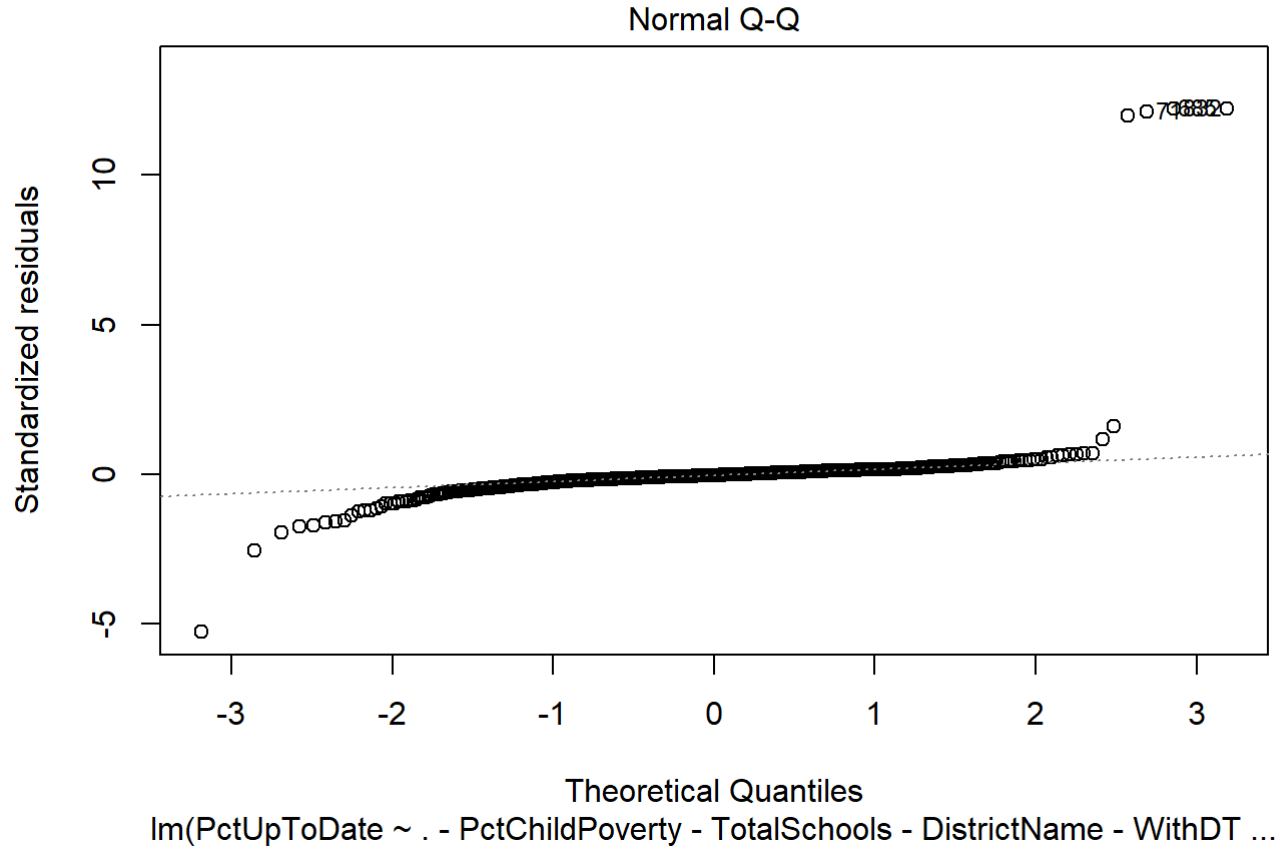
```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP -WithHepB -WithPolio
         , data=districts_log))
```

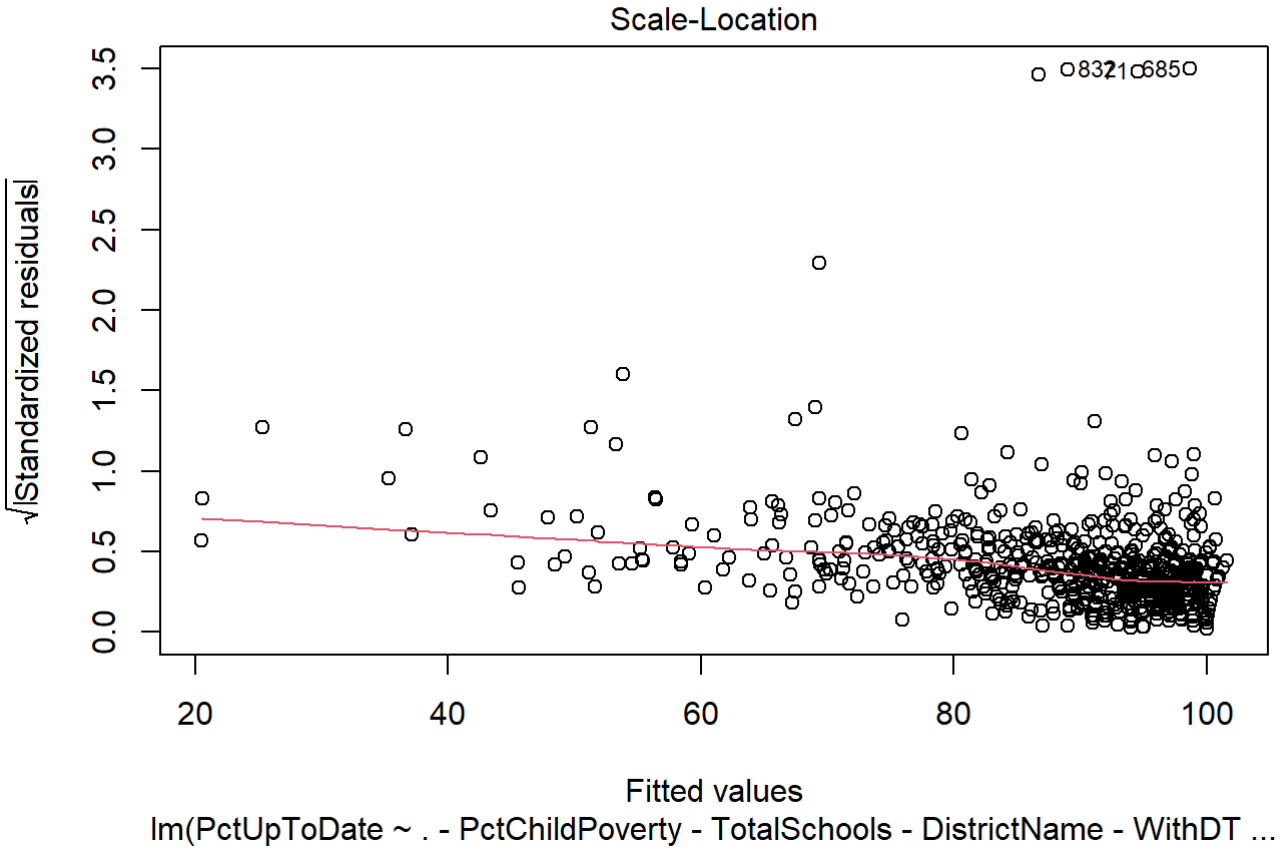
```
##          WithMMR DistrictComplete  PctBeliefExempt PctMedicalExempt
##         2.730953         1.048359         2.644949         1.008256
## PctFamilyPoverty         Enrolled
##         1.108645         1.138429
```

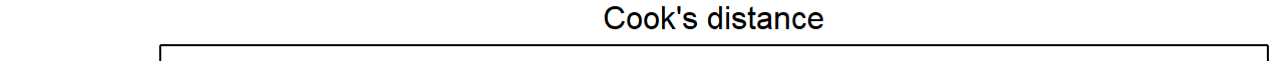
```
# saving the model in a varibale to carry out further analysis
lm_highR2 <- lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP -WithHepB -WithPolio
         , data=districts_log)
```

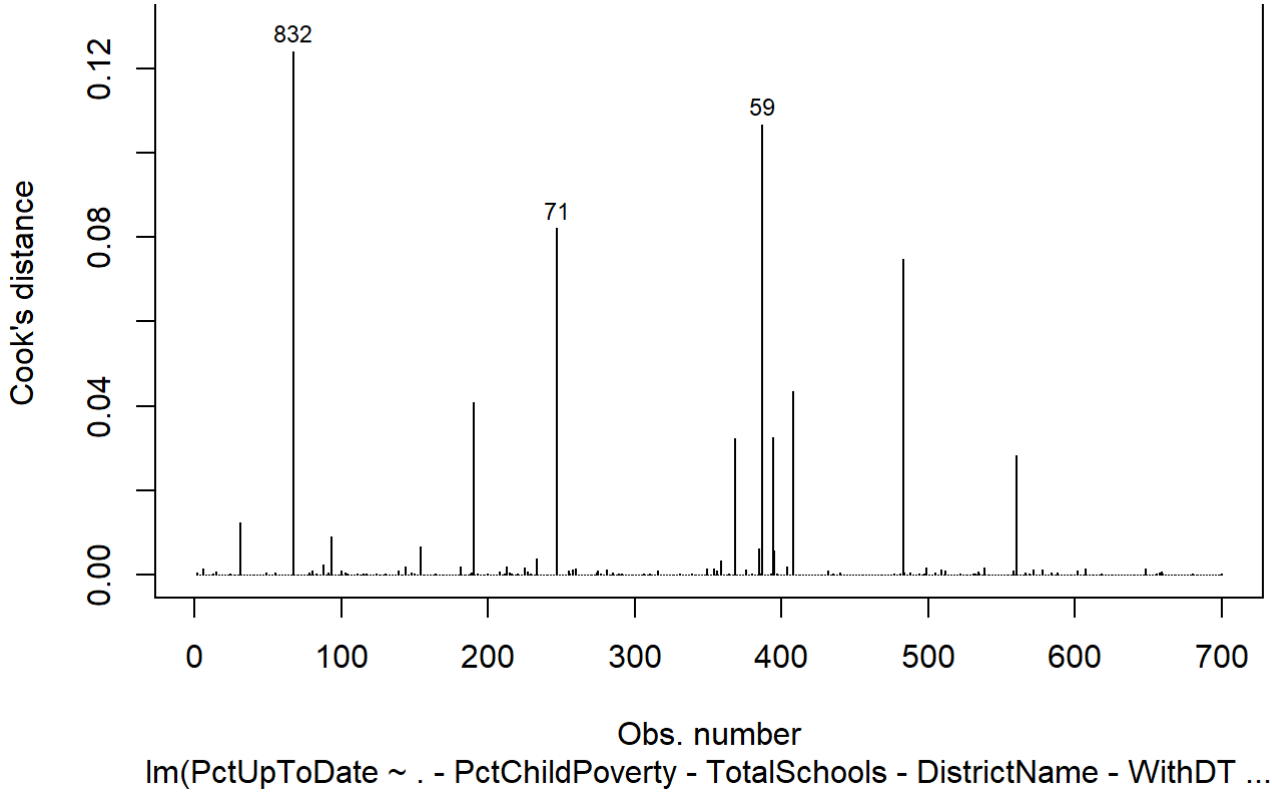
Checking the residual plots

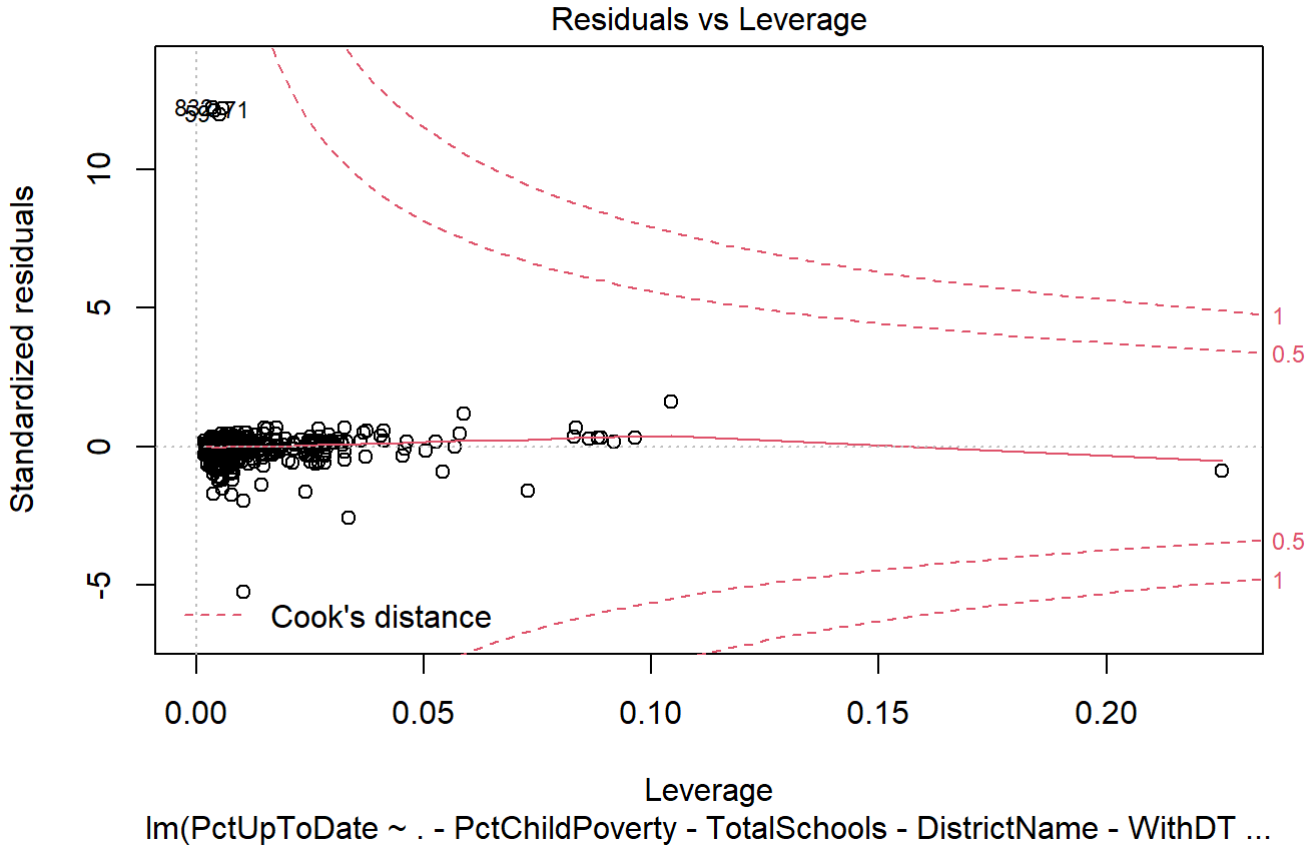
```
plot(lm_highR2, which=2:5)
```


Normal Q-Q
Standardized residuals
10
5
0
-5
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ . - PctChildPoverty - TotalSchools - DistrictName - WithDT ...


Scale-Location
√|Standardized residuals|
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
40
60
80
100
Fitted values
lm(PctUpToDate ~ . - PctChildPoverty - TotalSchools - DistrictName - WithDT ...


Cook's distance





We can see at the Cook's distance graph that observation number 832 only has 0.12 influence on the regression line, 71 has 0.09 and 59 has around 0.10. Looking at the Residuals vs Leverage graph we can see that nothing is in the Cook's distance which is good. The q-q plot tells us how normally distributed the residuals are which is a basic assumption of the linear regression. Looking at the graph we can see that at the model does not have any such variability than the model can't account for and the model look pretty good.

Now checking our model summary

```
summary(lm_highR2)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ . - PctChildPoverty - TotalSchools -
##     DistrictName - WithDTP - WithHepB - WithPolio, data = districts_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.396  -1.338  -0.181   0.942 101.291
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -21.67099    4.44622  -4.874 1.36e-06 ***
## WithMMR                1.17884    0.04571  25.791  < 2e-16 ***
## DistrictCompleteTRUE   1.59579    1.33617   1.194   0.2328
## PctBeliefExempt        0.16723    0.05758   2.904   0.0038 **
## PctMedicalExempt      -0.05684    0.49762  -0.114   0.9091
## PctFamilyPoverty       0.05982    0.04035   1.482   0.1387
## Enrolled               0.21126    0.20958   1.008   0.3138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.29 on 693 degrees of freedom
## Multiple R-squared:  0.6966, Adjusted R-squared:  0.6939
## F-statistic: 265.1 on 6 and 693 DF,  p-value: < 2.2e-16
```

We can see that the median of -0.181 is very close to 0 and comparing it our previous models it is the best we have got. We can also see that the f-statistic is F(6,693) = 265.1, the r-squared and adjusted r squared is 0.696 and 0.694 respectively which is a good value and the p-value of 2.2e-16 of the overall model which is significant at alpha level of 0.05. Looking at the columns we can see that WithMMR significantly predicts PctUpToDate (b = 1.178, t(693)=25.791, p<.001) PctBeliefExempt significantly predicts PctUpToDate (b = 0.167, t(693)= 2.904, p<.01). Rest of the columns are not significant at alpha level of 0.05

Now we will again check if including PctFreeMeal gives a better R2 and more columns that can be good predictors and then check which predictors have the biggest impact on the result by comparing standardized coefficents.

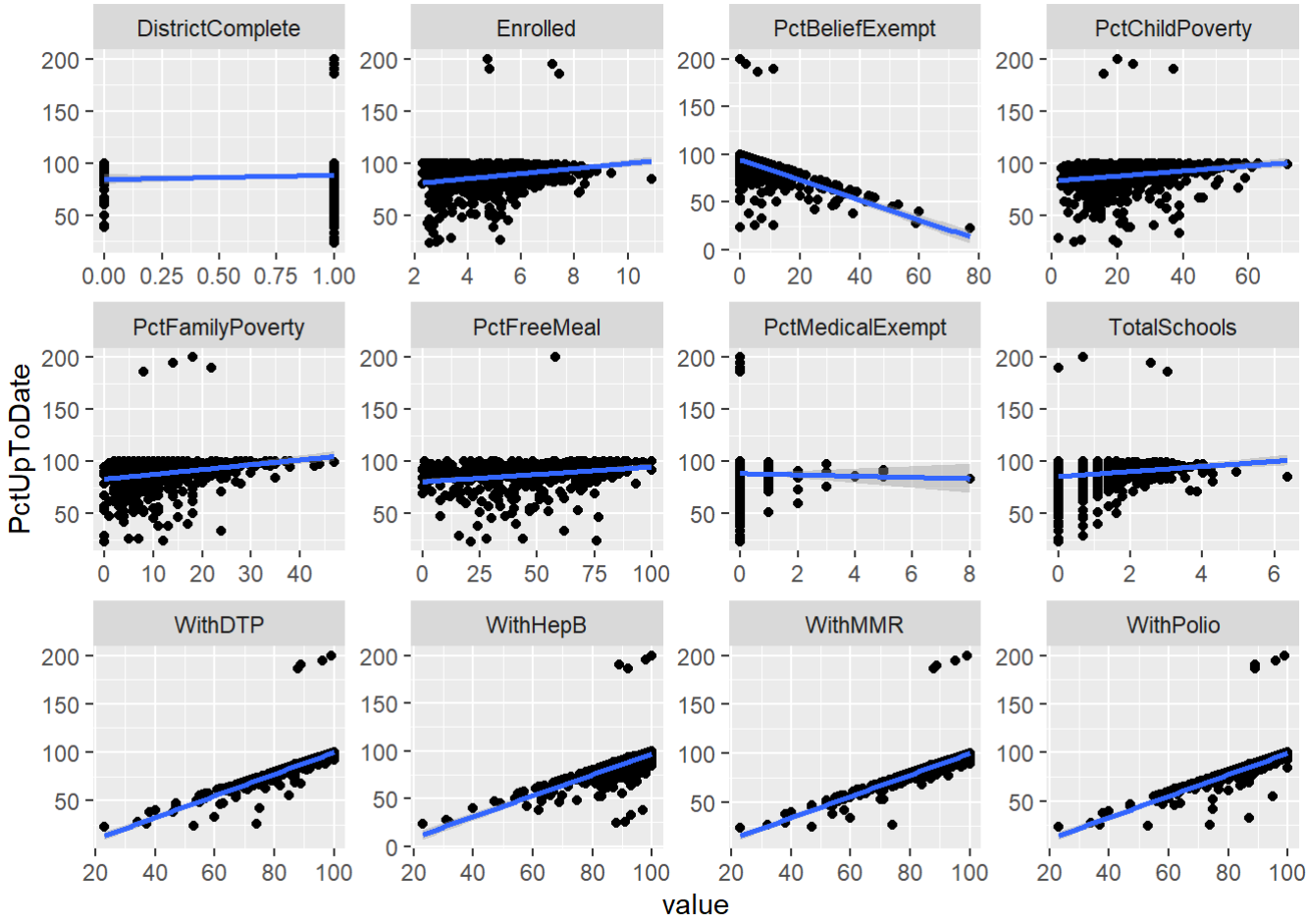
Checking with the a new model where PctFreeMeal is included and we also have log transformation

```
districts_log_PctFreeMeal <- districts_log
districts_log_PctFreeMeal$PctFreeMeal <- districts$PctFreeMeal
dim(districts_log_PctFreeMeal)
```

```
## [1] 700  14
```

```
# creating a new df of the cleaned data for better readability
districts_log_PctFreeMeal %>% pivot_longer(-c(PctUpToDate,DistrictName), 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")
```

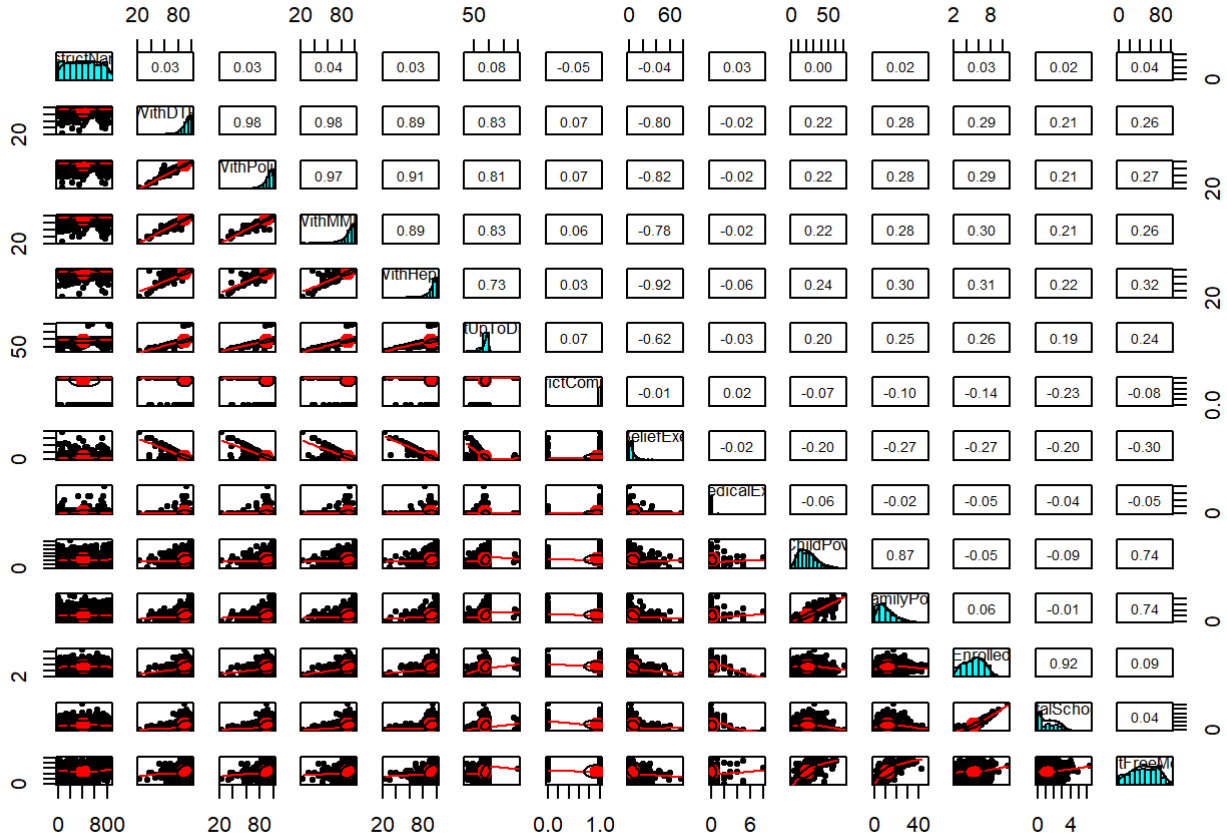
```
## `geom_smooth()` using formula 'y ~ x'
```



WithDTP, WithHepB, WithMMR, WithPolio show a good positive correlation, Enrolled, PctChildPoverty, PctFamilyPoverty, TotalSchools, PctFreeMeal show sub par or an almost positive correlation, PctBeliefExempt shows a negative correlation and PctMedicalExempt shows an almost about to happen negative correlation with respect to percentage of students with completely up-to-date vaccines.

Let's check the pairs plot to examine plots and correlation:

```
pairs.panels(districts_log_PctFreeMeal)
```



We can see that PctChildPoverty, PctFamilyPoverty, PctUpToDate and Enrolled are almost normal and TotalSchools is slightly skewed (but alot better than before doing the log transformation). We can also see that there is high correlation between Percentage of children in district living below the poverty line and Percentage of families in district living below the poverty line which makes sense as they can be inter related i.e they must be children of the families staying in districts below the poverty line. There is also high correlation between totalschools and enrolled students which makes sense as there is interloping of the districts with a child being enrolled and in the district of the school. PctFreeMeal and rest of the columns have low correlations which is great for linear modelling.

Let's create models and find which model has the least multicolinarity and then analyze it further

```
vif(lm(PctUpToDate ~ .-DistrictName, data=districts_log_PctFreeMeal))
```

```
##           WithDTP         WithPolio           WithMMR         WithHepB
##         38.873234         31.230591         23.819986         17.684619
## DistrictComplete  PctBeliefExempt PctMedicalExempt  PctChildPoverty
##          1.141456          8.237278          1.096159          4.443070
## PctFamilyPoverty          Enrolled      TotalSchools      PctFreeMeal
##          4.540019          7.356433          7.058019          2.582600
```

We are removing column PctChildPoverty as from our previous models and bi variate analysis we can see that it is having multicolinarity with PctFamilyPoverty and TotalSchools as it has multi colinarity with Enrolled.

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName,
       data=districts_log_PctFreeMeal))
```

```
##          WithDTP        WithPolio          WithMMR         WithHepB
##        38.857884        31.217534        23.811965        17.530826
## DistrictComplete  PctBeliefExempt PctMedicalExempt PctFamilyPoverty
##         1.047255         8.162044         1.091678         2.298341
##         Enrolled      PctFreeMeal
##         1.177192         2.348234
```

Checking dropping which column can give us least amount of multicolinarity for the group dtp, polio, mmr and hepb

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP
         , data=districts_log_PctFreeMeal))
```

```
##        WithPolio          WithMMR         WithHepB DistrictComplete
##        19.360691        15.285964        17.378587         1.042936
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         8.106379         1.091069         2.287536         1.177021
##      PctFreeMeal
##         2.337266
```

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithMMR
         , data=districts_log_PctFreeMeal))
```

```
##          WithDTP        WithPolio         WithHepB DistrictComplete
##        24.944611        30.771332        16.575275         1.042005
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         7.769698         1.088696         2.294799         1.175612
##      PctFreeMeal
##         2.345719
```

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithPolio
         , data=districts_log_PctFreeMeal))
```

```
##          WithDTP          WithMMR         WithHepB DistrictComplete
##        24.099132        23.471612        16.238757         1.047223
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         8.161357         1.088537         2.298336         1.177086
##      PctFreeMeal
##         2.348215
```

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithHepB
         , data=districts_log_PctFreeMeal))
```

```
##          WithDTP        WithPolio          WithMMR DistrictComplete
##        38.520440        28.916719        22.514049         1.045287
##  PctBeliefExempt PctMedicalExempt PctFamilyPoverty         Enrolled
##         3.634536         1.021102         2.296870         1.176251
##      PctFreeMeal
##         2.342449
```

Looking at the above analysis it looks like removing HepB PctBeliefExempt reduces and removing WithDTP reduces the multicolinarity best from polio, mmr and hepb variable

Checking if removing both these columns gives us a good model that doesnt have multicolinarity

```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP -WithHepB
         , data=districts_log_PctFreeMeal))
```

```
##        WithPolio          WithMMR DistrictComplete  PctBeliefExempt
##        17.917837        14.527414         1.041461         3.633106
## PctMedicalExempt PctFamilyPoverty         Enrolled      PctFreeMeal
##         1.021102         2.286707         1.176146         2.329837
```

Since we still don't get a value less than 10 for WithPolio and WithMMR we will remove Polio and we get the least multicolinarity and since we already saw that one child getting one vaccine is very likely to get all the other vaccines, we can just use WithMMR to see if getting a shot of a vaccine is a good predictor for having up to date vaccinations.

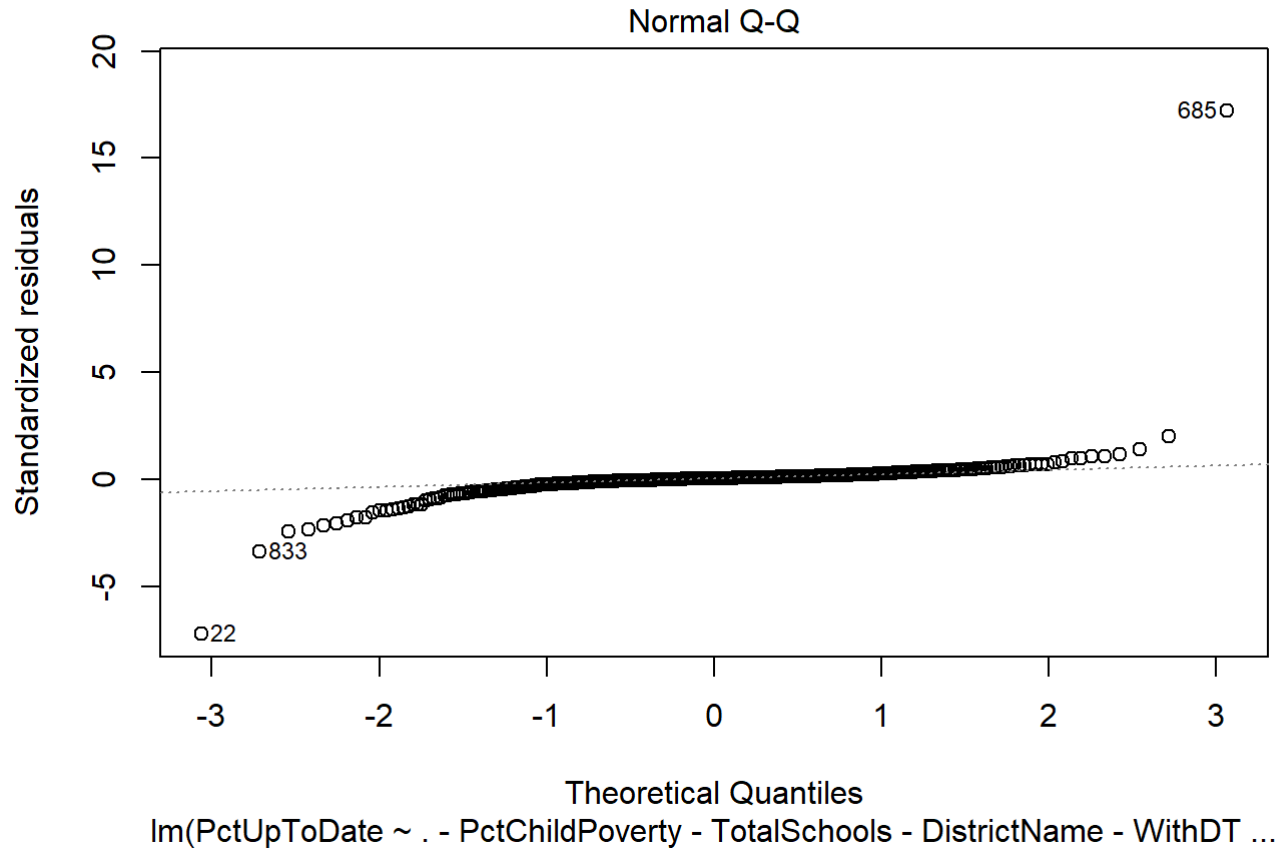
```
vif(lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP -WithHepB -WithPolio
         , data=districts_log_PctFreeMeal))
```

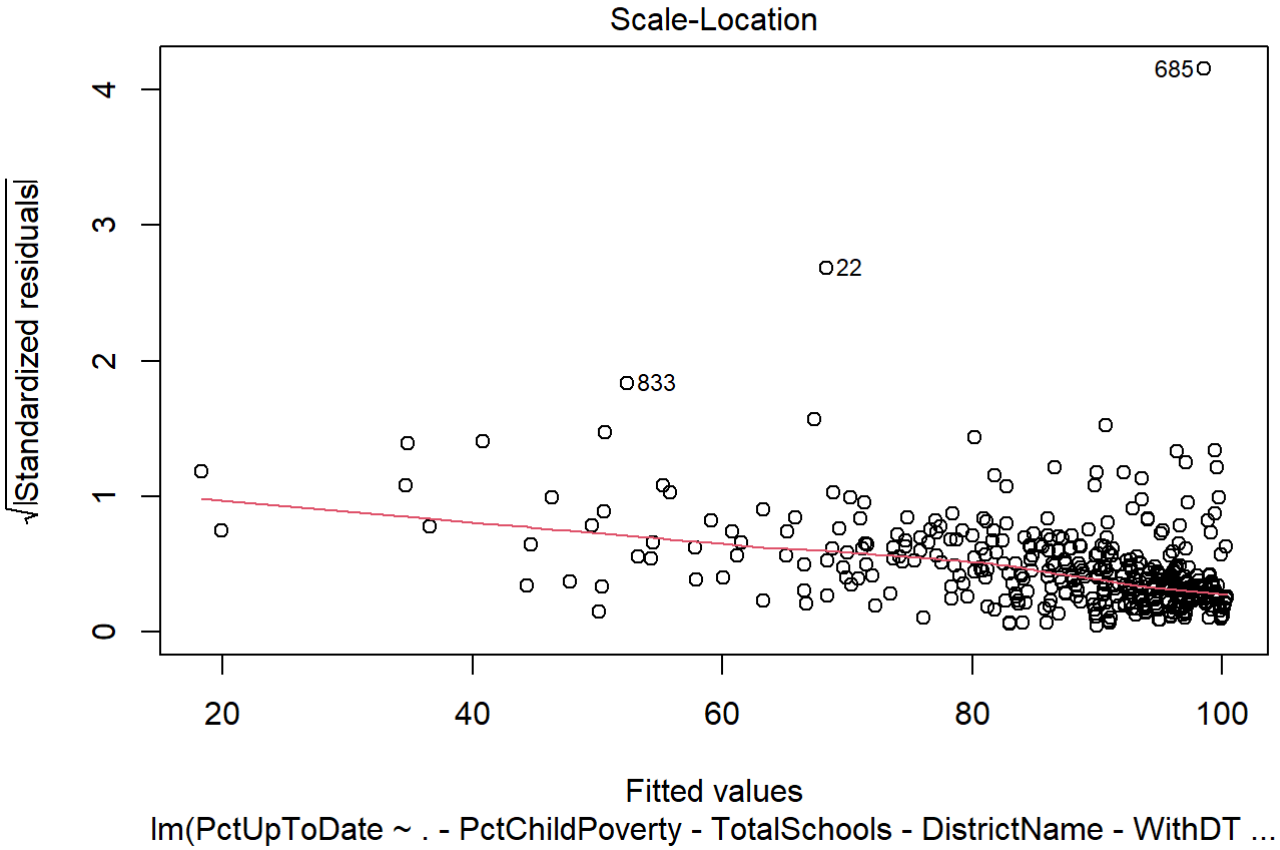
```
##          WithMMR DistrictComplete  PctBeliefExempt PctMedicalExempt
##         3.018487         1.036296         2.949675         1.020654
## PctFamilyPoverty         Enrolled      PctFreeMeal
##         2.278646         1.175082         2.326807
```

```
# saving the model in a varibale to carry out further analysis
lm_highR2_FreeMeal <- lm(PctUpToDate ~ .-PctChildPoverty -TotalSchools -DistrictName
       -WithDTP -WithHepB -WithPolio
         , data=districts_log_PctFreeMeal)
```

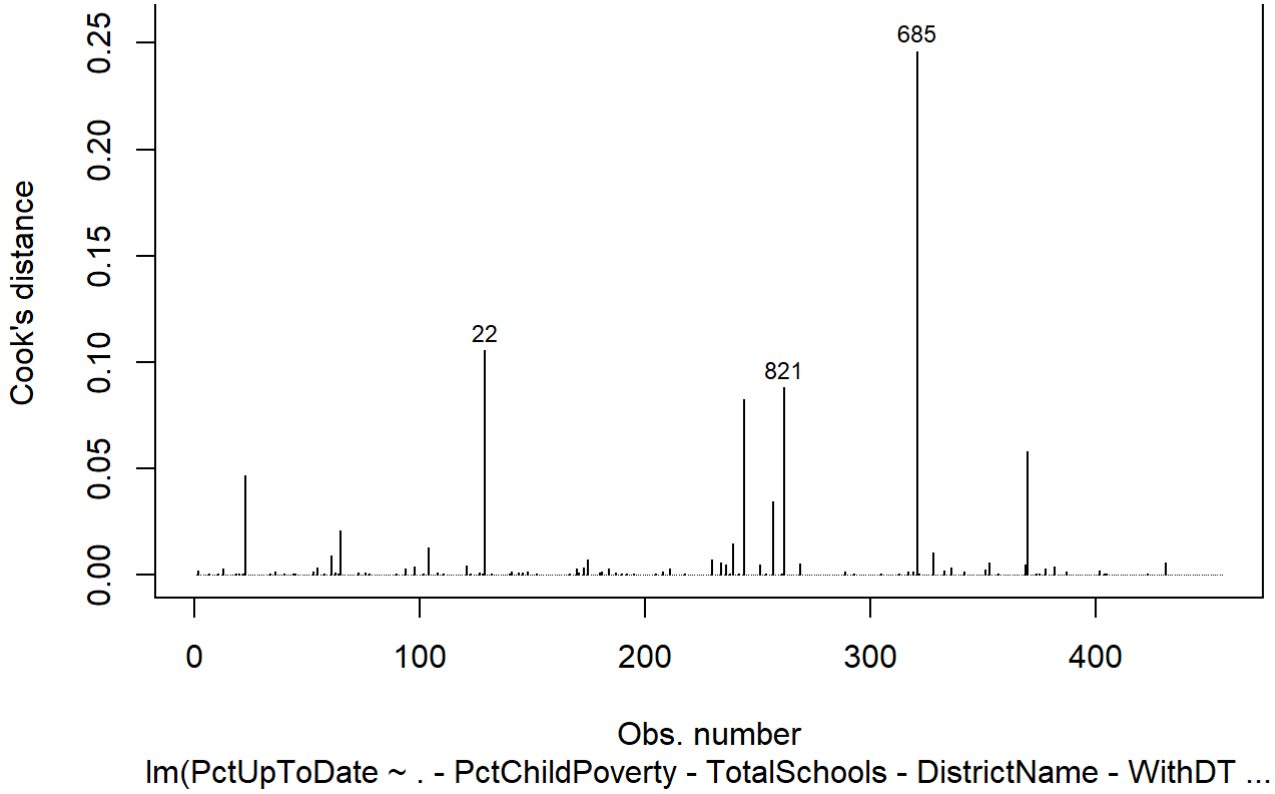
Checking the residual plots

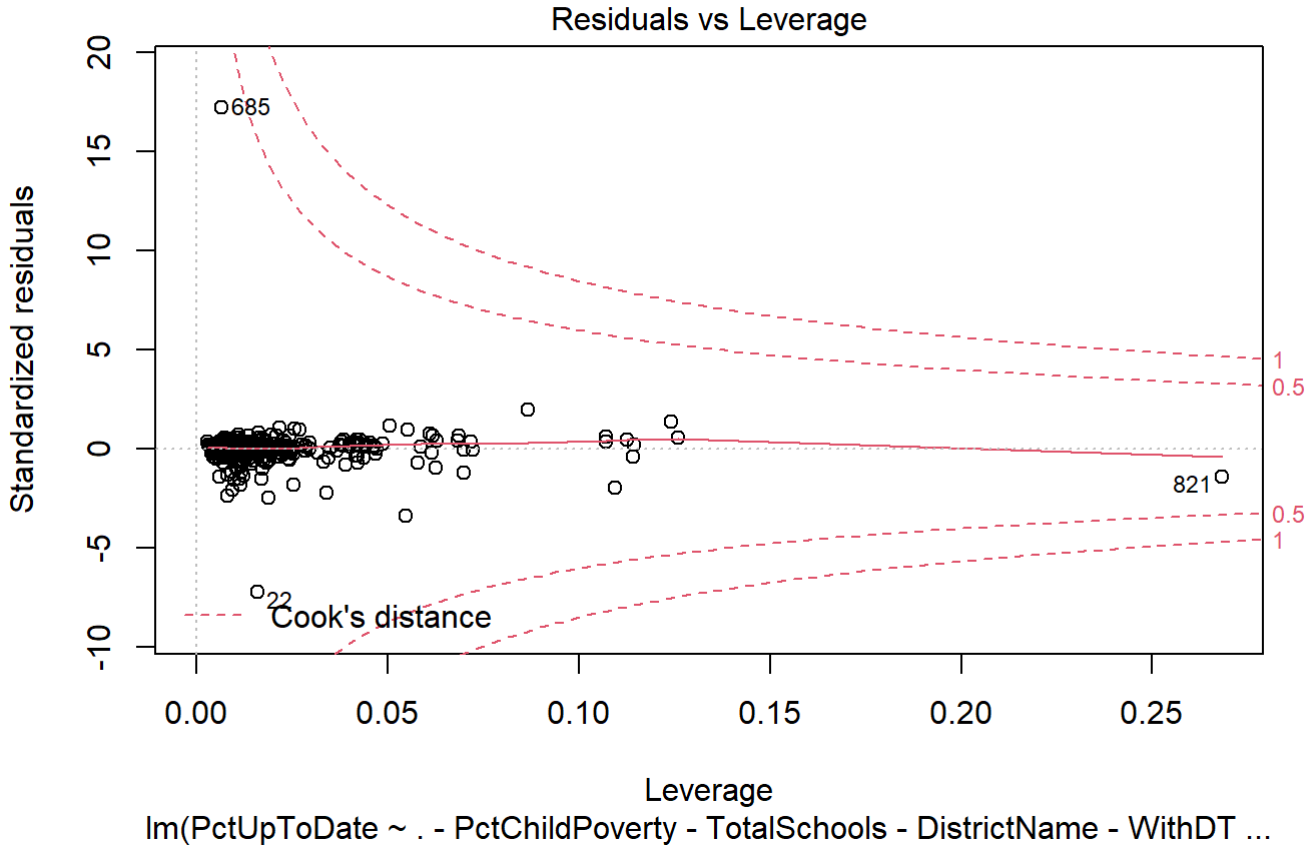
```
plot(lm_highR2_FreeMeal, which=2:5)
```





Cook's distance





We can see at the Cook's distance graph that observation number 685 only has 0.25 influence on the regression line, 22 has 0.13 and 821 has around 0.10. Looking at the Residuals vs Leverage graph we can see that nothing is in the Cook's distance which is good. The q-q plot tells us how normally distributed the residuals are which is a basic assumption of the linear regression. Looking at the graph we can see that at the model does have alittle variability at the lower end that the model can account for but the model look pretty good.

Now checking our model summary

```
summary(lm_highR2_FreeMeal)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ . - PctChildPoverty - TotalSchools -
##     DistrictName - WithDTP - WithHepB - WithPolio, data = districts_log_PctFreeMeal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.318  -0.617   0.272   0.988 101.464
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -24.702268   3.964177  -6.231 1.07e-09 ***
## WithMMR                1.226508   0.040251  30.472  < 2e-16 ***
## DistrictCompleteTRUE   1.353700   1.192991   1.135    0.257
## PctBeliefExempt        0.196036   0.048846   4.013 7.02e-05 ***
## PctMedicalExempt      -0.011053   0.389643  -0.028    0.977
## PctFamilyPoverty       0.026844   0.054771   0.490    0.624
## Enrolled              -0.067731   0.186078  -0.364    0.716
## PctFreeMeal            0.005147   0.017127   0.301    0.764
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.906 on 448 degrees of freedom
##   (244 observations deleted due to missingness)
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8337
## F-statistic:   327 on 7 and 448 DF,  p-value: < 2.2e-16
```

We can see that the median of -0.272 is very close to 0. We can also see that the f-statistic is F(7,448) = 327, the r-squared and adjusted r squared is 0.8363 and 0.8337 respectively which is a good value and the p-value of 2.2e-16 of the overall model which is significant at alpha level of 0.05. Looking at the columns we can see that WithMMR significantly predicts PctUpToDate (b = 1.226, t(448)=30.4472, p<.001) PctBeliefExempt significantly predicts PctUpToDate (b = 0.196, t(448)= 4.013, p<.001). Rest of the columns are not significant at alpha level of 0.05

Since we got a better R2 in this new model and the columns that were significant were same in the model (PctMedicalExempt was significant at more alpha levels in the model with PctFreeModel than without) so we will go ahead and choose the model with PctFreeModel column included and check beta weights to see which predictors have the biggest impact on the result, we will compare standardized coefficients, i.e., those based on standardized variables:

```
summary(lm.beta(lm_highR2_FreeMeal))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ . - PctChildPoverty - TotalSchools -
##     DistrictName - WithDTP - WithHepB - WithPolio, data = districts_log_PctFreeMeal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.318  -0.617   0.272   0.988 101.464
##
## Coefficients:
##                       Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)          -2.470e+01    0.000e+00  3.964e+00  -6.231 1.07e-09 ***
## WithMMR               1.227e+00    1.012e+00  4.025e-02  30.472  < 2e-16 ***
## DistrictCompleteTRUE  1.354e+00    2.208e-02  1.193e+00   1.135    0.257
## PctBeliefExempt       1.960e-01    1.318e-01  4.885e-02   4.013 7.02e-05 ***
## PctMedicalExempt     -1.105e-02   -5.478e-04  3.896e-01  -0.028    0.977
## PctFamilyPoverty      2.684e-02    1.414e-02  5.477e-02   0.490    0.624
## Enrolled             -6.773e-02   -7.542e-03  1.861e-01  -0.364    0.716
## PctFreeMeal           5.147e-03    8.763e-03  1.713e-02   0.301    0.764
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.906 on 448 degrees of freedom
##   (244 observations deleted due to missingness)
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8337
## F-statistic:   327 on 7 and 448 DF,  p-value: < 2.2e-16
```

The significant variables are the percentage of students in the district with the MMR vaccine and percentage of all enrolled students with belief exceptions. It means the if a student is vaccinated (since we are using MMR as a placeholder for all vaccines to remove multicolinarity we can assume that a student vaccinated by any of the vaccine i.e Polio, HepB, DTP) more up to date they are with their vaccination, more percentage of students with completely up-to-date vaccines, more enrolled students with belief exception more up to date vaccination and one unit change in MMR vaccine can lead to increase in up to date vaccination which makes sense as more students are up to date with thier vaccination. Similarly every 1 standard deviation increase in the percent of students with belief exception will have 1.960e-01 standard deviation increase in the percentage of up to date vaccination.

Doing the Bayesian Test for both the models:

```
lm_highR2_lmBF <- lmBF(PctUpToDate ~ .-PctChildPoverty -TotalSchools
                       -DistrictName
                       -WithDTP -WithHepB -WithPolio,
                       data=districts_log,
                       posterior=TRUE, iterations=10000)

summary(lm_highR2_lmBF)
```

```
##
## 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                                88.39813 0.31532 0.0031532      0.0031532
## WithMMR-WithMMR                    1.17367 0.04564 0.0004564      0.0004564
## DistrictComplete-DistrictComplete  1.58400 1.32969 0.0132969      0.0132969
## PctBeliefExempt-PctBeliefExempt    0.16611 0.05687 0.0005687      0.0005687
## PctMedicalExempt-PctMedicalExempt -0.06077 0.50098 0.0050098      0.0050098
## PctFamilyPoverty-PctFamilyPoverty  0.05957 0.04069 0.0004069      0.0004069
## Enrolled-Enrolled                  0.20591 0.20749 0.0020749      0.0020105
## sig2                              68.98281 3.66576 0.0366576      0.0366576
## g_continuous                       0.47195 0.37410 0.0037410      0.0037410
##
## 2. Quantiles for each variable:
##
##                                     2.5%      25%      50%      75%   97.5%
## mu                                87.78644 88.18429 88.39496 88.61156 89.0225
## WithMMR-WithMMR                    1.08619  1.14262  1.17372  1.20383  1.2636
## DistrictComplete-DistrictComplete -1.00166  0.67902  1.59168  2.47661  4.1670
## PctBeliefExempt-PctBeliefExempt    0.05534  0.12697  0.16599  0.20438  0.2792
## PctMedicalExempt-PctMedicalExempt -1.04087 -0.39915 -0.05812  0.28081  0.8964
## PctFamilyPoverty-PctFamilyPoverty -0.01904  0.03226  0.05913  0.08719  0.1396
## Enrolled-Enrolled                 -0.20374  0.06619  0.20855  0.34369  0.6093
## sig2                              62.18949 66.41608 68.85735 71.41298 76.4598
## g_continuous                       0.14772  0.26286  0.37246  0.55482  1.3861
```

In the output displayed above, we have parameter estimates for the B-weights of each of our predictions (the column labeled “Mean”). In the second section, we have the 2.5% and 97.5% boundaries of the HDI for each of the B-weights.These boundaries mark the edges of the central region of the posterior distribution for each B-weight. So WIthMMR predictor has a lower bound of 1.08 to upper bound at 1.262, and PctBeliefExempt has HDI from 0.05 to 0.27. Since the HDI for these 2 columns does not contains 0 we have credible evidence that there is a difference in between the variables. The DistrictComplete predictor has a lower bound of -1.057 to upper bound of 4.2051, PctMedicalExempt has -1.05 to 0.91, PctFamilyPoverty has -0.02 to 0.132 and enrolled has -0.197 to 0.6238. Since the HDI for these columns contains 0, the observed differences could be due to chance.

Also we can see that the means from our bayesian test are close to the estimates we got from the frequentist model.

```
# running the same model without the iterations to get bayes factor value
lm_highR2_out <- lmBF(PctUpToDate ~ .-PctChildPoverty -TotalSchools
                      -DistrictName
                      -WithDTP -WithHepB -WithPolio,
                      data=districts_log)
lm_highR2_out
```

```
## Bayes factor analysis
## --------------
## [1] . - PctChildPoverty - TotalSchools - DistrictName - WithDTP - WithHepB - WithPolio :
7.484836e+171 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

We get a very high bayes factor of 7.484836e+171 showing us that our results are significant.

Running Bayesian for data frame with PctFreeMeal column:

```
#lm_highR2_lmBF_FreeMeal <- lmBF(PctUpToDate ~ .-PctChildPoverty -TotalSchools
#                        -DistrictName
#                        -WithDTP -WithHepB -WithPolio,
#                        data=districts_log_PctFreeMeal,
#                        posterior=TRUE, iterations=10000)
```

We can see that we get an Error in checkFormula(formula, data, analysis = “lm”) : Predictors must not contain missing values when we run the above model hence we can’t do a baysian model for this.

In conclusion, a linear regression was performed to estimate the percentage of all enrolled students with up to date vaccination with use of WithMMR, DistrictComplete, PctBeliefExempt, PctMedicalExempt, PctFamilyPoverty and Enrolled as the predictors. In the data cleaning process to remove skewness we took log transformation on Enrolled and Total schools columns. In the bivariate analysis we saw that there was barely any visible skewness and there were no major issues leading to the conclusion that the data is linear to carry out linear regression. Using the result from both the Bayesian and frequentist approach we got evidence that WithMMR (or any vaccination) and PctBeliefExempt are good predictors for PctUpToDate and we get a good R2 to represents the proportion of about 69% variation in PctUpToDate for model without the PctFreeMeal column and 83% for model with PctMealFree column. We could see that we get similar results from the bayesian test for the model without the PctFreeMeal and we could’t run the model with PctFreeMeal column included as that has null values(which is why it was dropped originally in the data cleaning step). Overall, using both frequentist and bayesian method we reached the same conclusion that MMR and PctBeliefExemption are good predictors for PctUpToDate column.

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

```
districts_log_interaction <- districts_log
```

Centering the variables before beginning with the prediction

```
districts_log_interaction$PctUpToDate <- scale(districts_log_interaction$PctUpToDate,
                                               center=T, scale=F)
districts_log_interaction$PctChildPoverty <- scale(districts_log_interaction$PctChildPoverty,
                                               center=T, scale=F)

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

Conducting linear regression on the centered data:

```
lm_interaction <- lm(PctUpToDate ~ PctChildPoverty * Enrolled,
                  data=districts_log_interaction)
```

Lets check multicolinarity in the model:

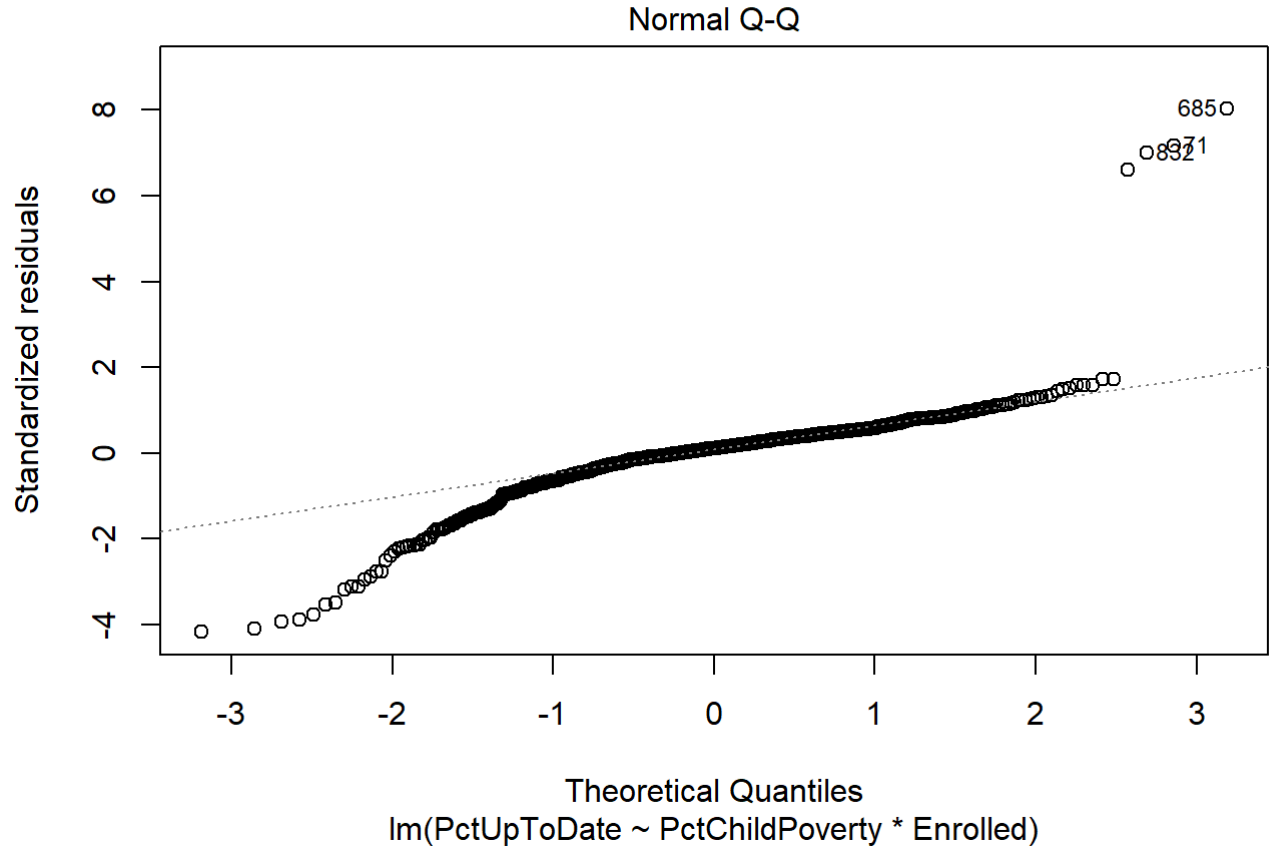
```
vif(lm_interaction)
```

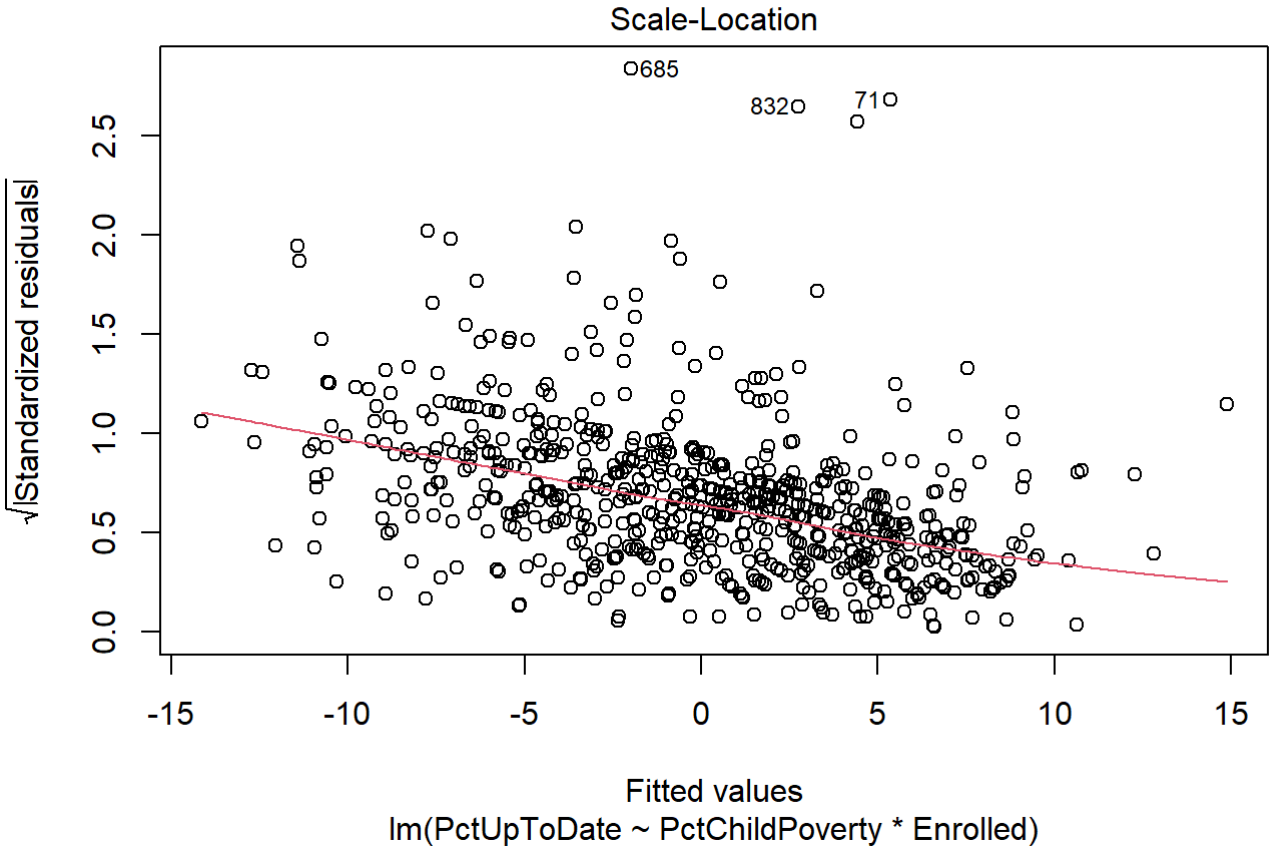
```
##          PctChildPoverty                 Enrolled PctChildPoverty:Enrolled
##                 1.022717                 1.004499                 1.022192
```

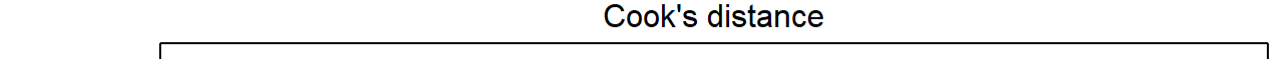
Since the VIF for these columns is below 5 we can conclude the the output of VIF looks good and that there we dont see any hint of multicolinarity and go ahead and check the residual plots.

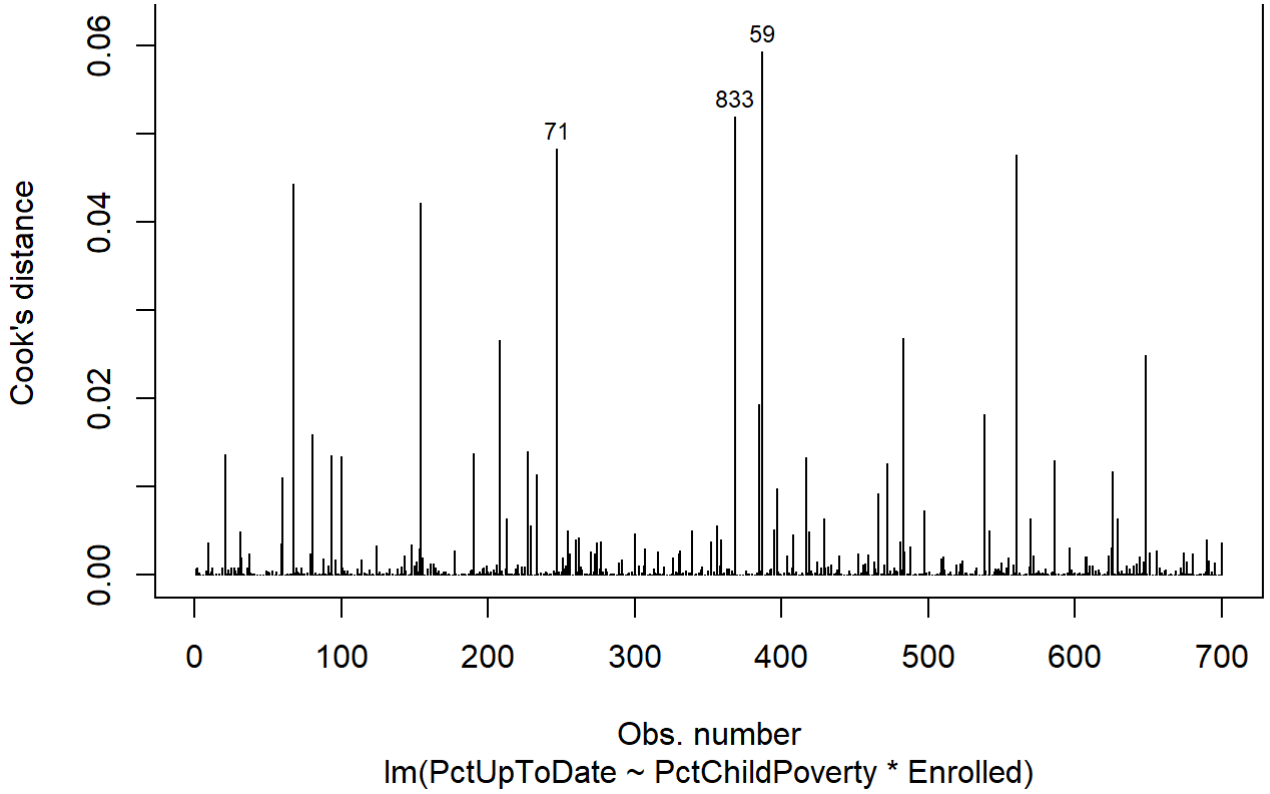
Checking the residual plots

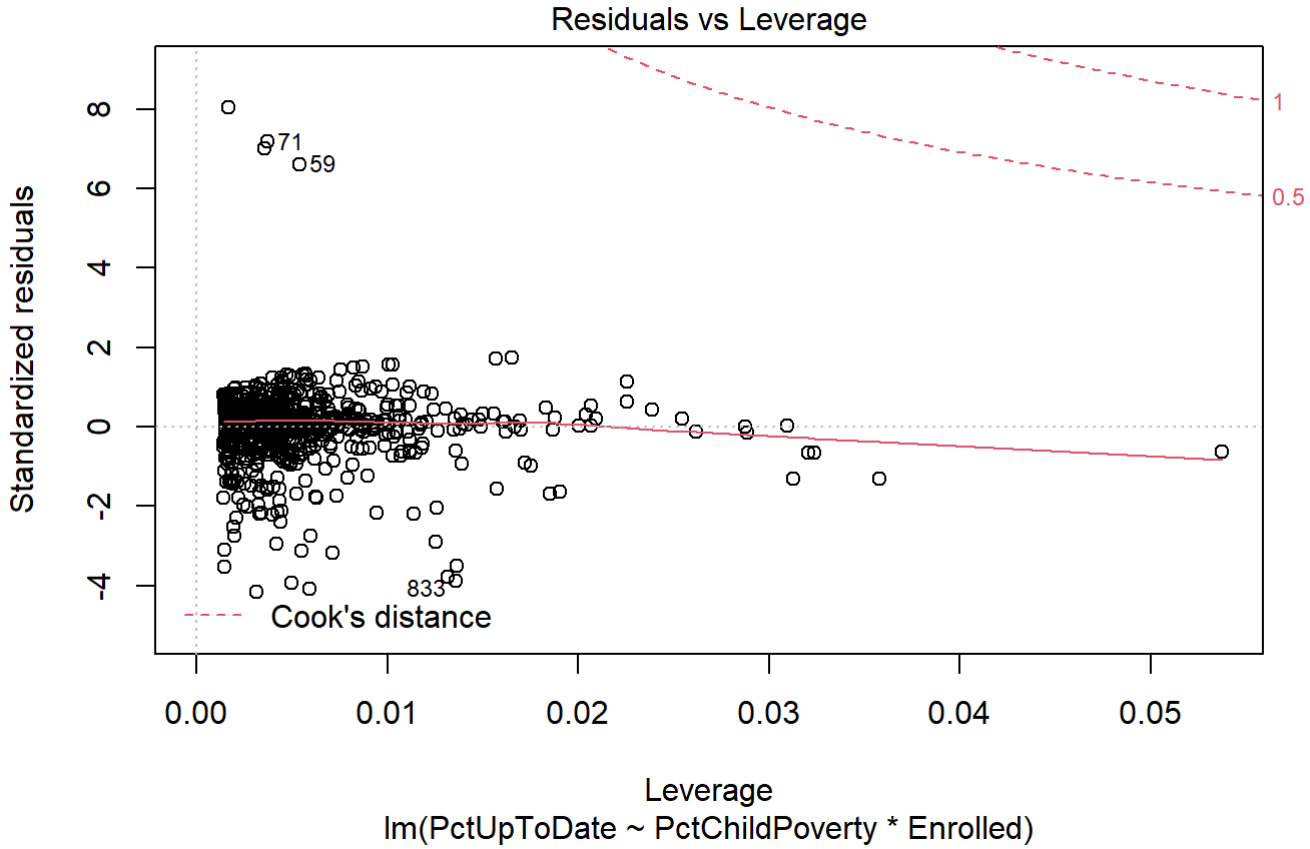
```
plot(lm_interaction, which=2:5)
```


Normal Q-Q
Standardized residuals
685
832
71
-4
-2
0
2
4
6
8
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(PctUpToDate ~ PctChildPoverty * Enrolled)


Scale-Location
√|Standardized residuals|
685
832
71
0.0
0.5
1.0
1.5
2.0
2.5
-15
-10
-5
0
5
10
15
Fitted values
lm(PctUpToDate ~ PctChildPoverty * Enrolled)


Cook's distance





We can see at the Cook's distance graph that observation number 833 only has 0.05 influence on the regression line, 71 has 0.05 and 59 has around 0.06. Looking at the Residuals vs Leverage graph we can see that nothing is in the Cook's distance but there is one outlier that is influencing teh regression line but its not alot which is good. The q-q plot tells us how normally distributed the residuals are which is a basic assumption of the linear regression. Looking at the graph we can see that at the start we have more variability but the model stabilizes as the graph progresses and that the model can account for variability, overall the model is good.

Now checking our model summary

```
summary(lm_interaction)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty * Enrolled, data = districts_log_interaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.854  -4.205   1.598   6.392 113.587
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.03609    0.53524  -0.067    0.946
## PctChildPoverty           0.25296    0.04441   5.696 1.81e-08 ***
## Enrolled                  2.55866    0.33571   7.622 8.22e-14 ***
## PctChildPoverty:Enrolled -0.03507    0.02995  -1.171    0.242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 696 degrees of freedom
## Multiple R-squared:  0.1138, Adjusted R-squared:  0.1099
## F-statistic: 29.78 on 3 and 696 DF,  p-value: < 2.2e-16
```

We can see that the median of 1.598 is close to 0. We can also see that the f-statistic is F(3,696) = 29.78, the r-squared and adjusted r squared is 0.113 and 0.109 respectively which isn't a huge value but the p-value is significant for the whole model. We will first examine the interaction term because the interpretation of the interaction may supersede the interpretation of the main effects. So F(3,696)=29.78 is not statistically significant at alpha level of 0.05.Hence there is no statistically significant interaction as we can see that the p value is 0.242 > 0.05. Hence we fail to reject the null hypothesis which is that there is no significant interaction effect of independent variables PctChildPoverty and Enrolled on dependent variable PctUpToDate. The main effects of PctChildPoverty and Enrolled have values of [PctChildPoverty]:(b = 0.25296, t(696)=5.696, p<.001) and [Enrolled]: (b = 2.55866, t(696)=7.622, p<.001). Since they are statistically significant at alpha level of 0.05, we reject the null hypothesis which is that there is no significant main effect of independent variables PctChildPoverty and Enrolled on dependent variable PctFreeMeal. In statistics, an omnibus test is any statistical test that tests for the significance of several parameters in a model at once. The omnibus statistics here tells us that the main effects are statistically significant as their p-value is lower than the 0.05 threshold, but the interaction effect is not significant as the p-value of 0.242 >> 0.05.

To see which predictors have the biggest impact on the result, we will compare standardized coefficients, i.e., those based on standardized variables:

```
summary(lm.beta(lm_interaction))
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctChildPoverty * Enrolled, data = districts_log_interaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.854  -4.205   1.598   6.392 113.587
##
## Coefficients:
##                          Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)              -0.03609      0.00000    0.53524  -0.067    0.946
## PctChildPoverty           0.25296      0.20555    0.04441   5.696 1.81e-08 ***
## Enrolled                  2.55866      0.27258    0.33571   7.622 8.22e-14 ***
## PctChildPoverty:Enrolled -0.03507     -0.04225    0.02995  -1.171    0.242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 696 degrees of freedom
## Multiple R-squared:  0.1138, Adjusted R-squared:  0.1099
## F-statistic: 29.78 on 3 and 696 DF,  p-value: < 2.2e-16
```

The significant variables are the percentage of children in district living below the poverty line and percentage of all enrolled students. It means the if a student is living in district living below poverty line is more up to date with their vaccination, more enrolled students are more up to date vaccination. We can see that the interaction term is not significant.

Doing the Bayesian Test for the same:

```
interaction_lmBF <- lmBF(PctUpToDate ~ PctChildPoverty * Enrolled,
                  data=districts_log_interaction,
                 posterior=TRUE, iterations=10000, rnd.seed=772)
summary(interaction_lmBF)
```

```
##
## 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                          -0.008248  0.53612 0.0053612      0.0053612
## PctChildPoverty              0.247327  0.04449 0.0004449      0.0004449
## Enrolled                     2.499168  0.33276 0.0033276      0.0033276
## PctChildPoverty.&.Enrolled  -0.033995  0.02980 0.0002980      0.0002927
## sig2                       200.214843 10.69667 0.1069667      0.1078401
## g                            0.127719  0.22740 0.0022740      0.0022740
##
## 2. Quantiles for each variable:
##
##                                2.5%       25%        50%       75%     97.5%
## mu                          -1.08237  -0.36551  -0.006692   0.35635   1.03179
## PctChildPoverty              0.15780   0.21776   0.247561   0.27745   0.33360
## Enrolled                     1.85410   2.27212   2.497179   2.71996   3.16091
## PctChildPoverty.&.Enrolled  -0.09183  -0.05405  -0.034261  -0.01376   0.02442
## sig2                       180.09544 192.79675 199.839936 207.19135 222.10038
## g                            0.02152   0.04610   0.073716   0.13188   0.55517
```

In the output displayed above, we have parameter estimates for the B-weights of each of our predictions (the column labeled “Mean”). In the second section, we have the 2.5% and 97.5% boundaries of the HDI for each of the B-weights.These boundaries mark the edges of the central region of the posterior distribution for each B-weight. So PctChildPoverty predictor has a lower bound of 0.16 to upper bound at 0.33, and Enrolled has HDI of 1.85 to 3.15; since the HDI does not contain 0 we have credible evidence that there is a difference in between the variables. The HDI for the interaction term is -0.09 to 0.02 and since it spans 0 the observed differences could be due to chance.

Also we can see that the means from our bayesian test match the estimates we got from the frequentist model.

```
# running the same model without the iterations to get bayes factor value
interaction_lmBF_out <- lmBF(PctUpToDate ~ PctChildPoverty * Enrolled,
                  data=districts_log_interaction,
                  rnd.seed=772)
interaction_lmBF_out
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty * Enrolled : 1.016188e+15 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

We get a very high bayes factor of 1.016188e+15 showing there are very strong odds in the favor of the alternative hypothesis and we can reject the null hypothesis which suggests that interpect only model is better. In conclusion, a linear regression was performed to estimate the percentage of all enrolled students with up to date vaccination with use of PctChildPoverty, and Enrolled as the predictors. In the data cleaning process to remove skewness we took log transformation on Enrolled and Total schools columns. In the bivariate analysis we saw that the remaining or little skewness didn't cause any major issues and that the data was linear to carry out linear regression. Using the result from both the Bayesian and frequentist approach we got evidence that PctFamilyPoverty , Enrolled are good predictors for PctUpToDate but the interaction term isn't a good predictor.

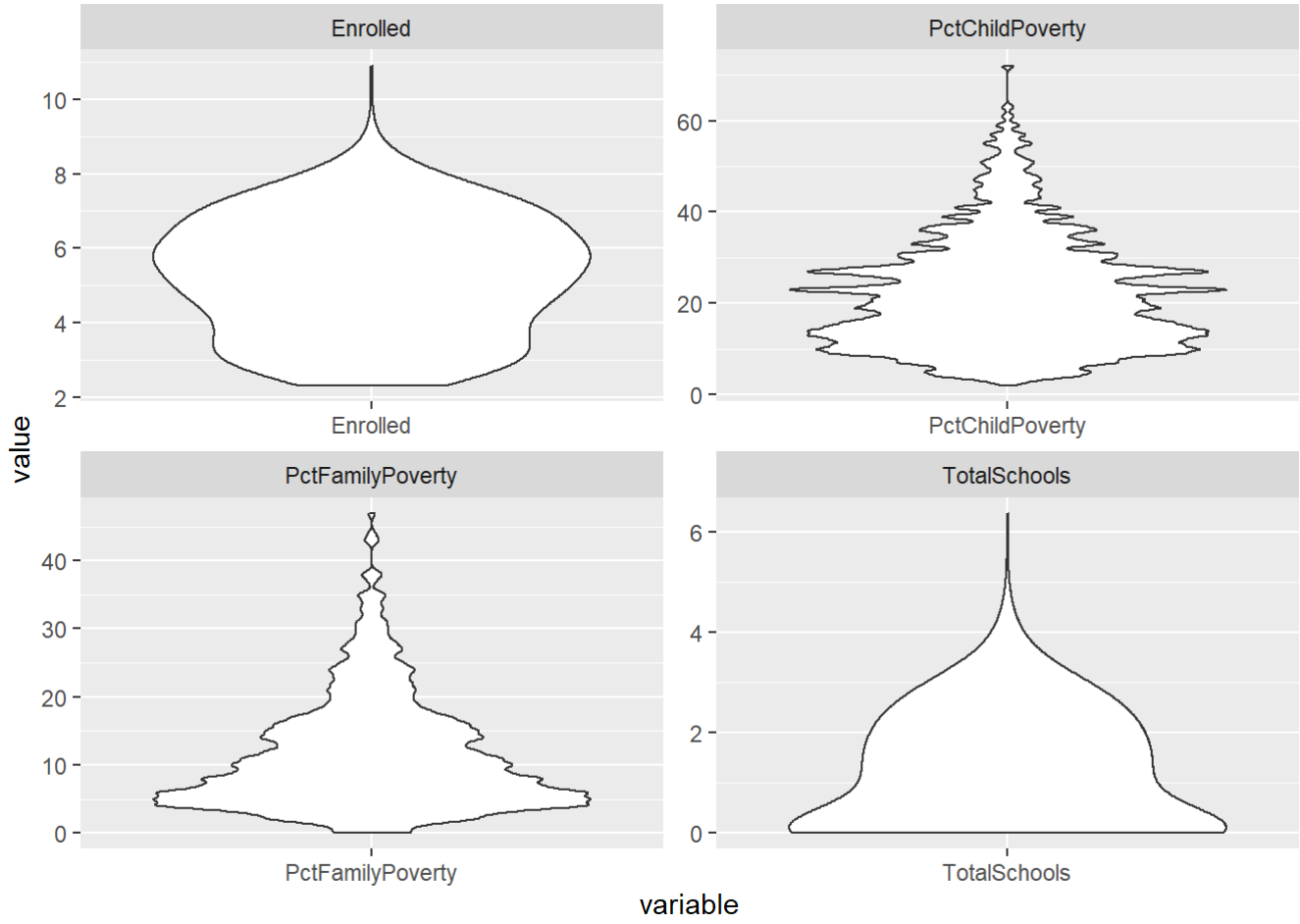
e. *Which, if any, of the four predictor variables predict whether or not a district's reporting was complete?*

```
# creating a new df of the cleaned data for better readability
# also converting logical DistrictComplete to factor
districts_log_districtComplete <- subset(districts_log, select = c(PctChildPoverty,
                                                                   PctFamilyPoverty,
                                                                   Enrolled,
                                                                   TotalSchools,
                                                                   DistrictComplete))
districts_log_districtComplete$DistrictComplete <- as.integer(districts_log_districtComplete
$DistrictComplet)
str(districts_log_districtComplete)
```

```
## 'data.frame':    700 obs. of  5 variables:
##  $ PctChildPoverty : num  29 23 4 41 27 20 24 32 44 27 ...
##  $ PctFamilyPoverty: num  10 11 3 22 14 10 13 18 27 14 ...
##  $ Enrolled        : num  3.74 2.83 7.75 6.42 4.16 ...
##  $ TotalSchools    : num  0 0 3.04 2.08 0 ...
##  $ DistrictComplete: int  1 1 1 0 1 0 1 1 1 1 ...
```

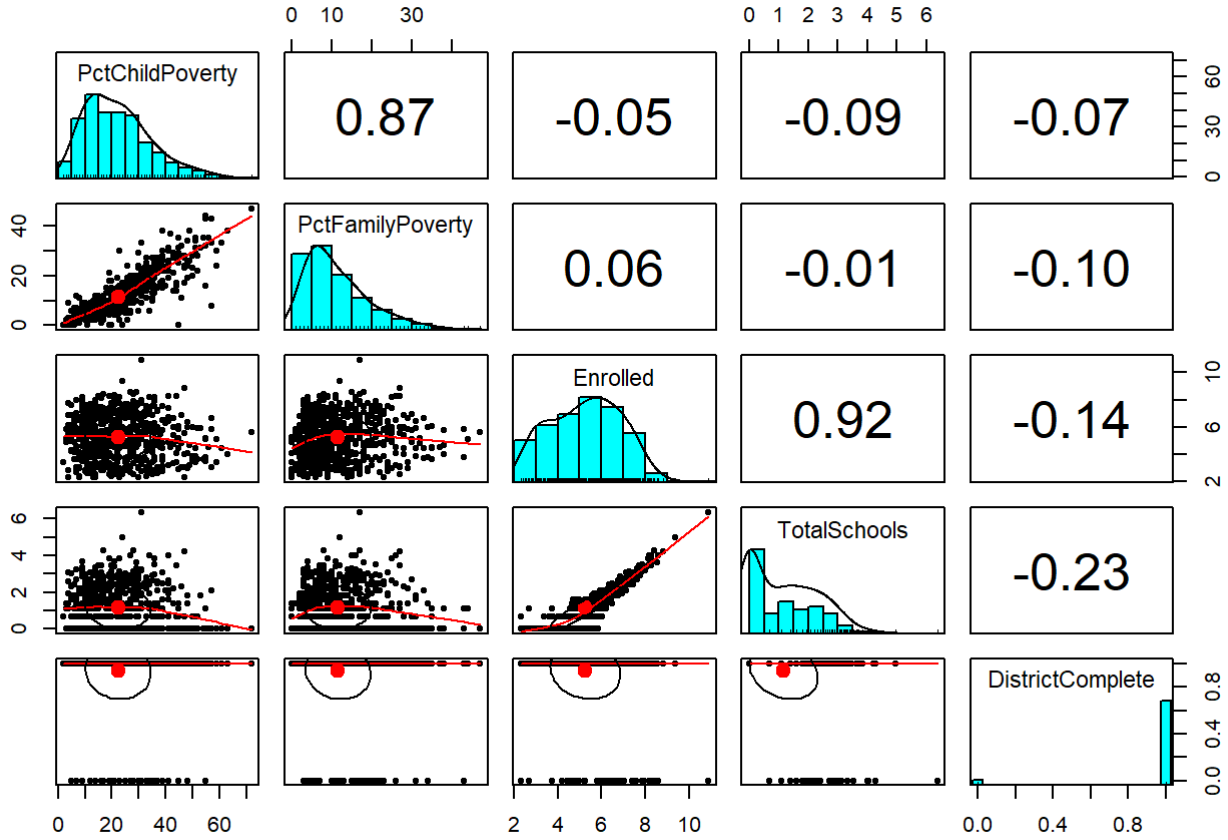
```
districts_log_districtComplete %>% pivot_longer(cols=-c(DistrictComplete),
                                                names_to="variable",
                                                values_to="value",
                                                values_drop_na = TRUE) %>%

  ggplot(aes(x=variable, y=value)) +
  geom_violin(bw=.5) +
  facet_wrap( ~ variable, scales="free")
```



We can high variability in PctChildPoverty and PctFamilyPoverty and low variability in TotalSchools and Enrolled.

```
pairs.panels(districts_log_districtComplete)
```



We can see that PctChildPoverty, PctFamilyPoverty are right skewed, Enrolled are almost normal and TotalSchools is slightly skewed (but alot better than before doing the log transformation). We can also see that there is high correlation between Percentage of children in district living below the poverty line and Percentage of families in district living below the poverty line which makes sense as they can be inter related i.e they must be children of the families staying in districts below the poverty line. There is also high correlation between totalschools and enrolled students which makes sense as there is interloping of the districts with a child being enrolled and in the district of the school. Rest of the correlations are low which is great for linear modelling.

```
districtsComplete_glm <- glm(formula = DistrictComplete ~ .,
                             family = binomial(link="logit"),
                             data = districts_log_districtComplete)
```

Checking heteroskadacity

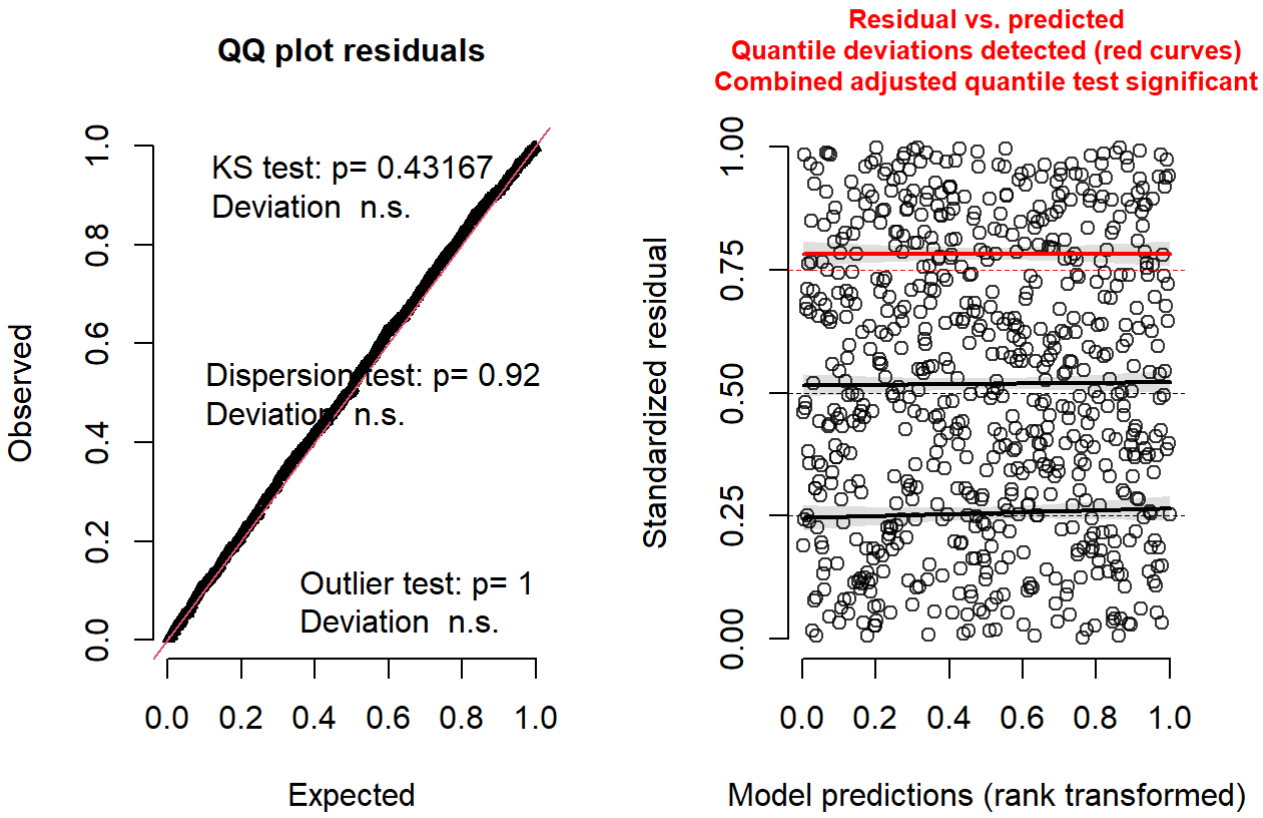
```
vif(districtsComplete_glm)
```

```
##  PctChildPoverty PctFamilyPoverty         Enrolled     TotalSchools
##         4.548039         4.579030        15.578199        15.511391
```

Checking the model performance

```
simulationOutput <- simulateResiduals(fittedModel = districtsComplete_glm, n = 250)

plot(simulationOutput)
```



Looking at the qq plot are normally distributed and no deviations are visible and the residual vs predicted plot des show a deviation at 0.75.

Since we got a high vif and devaition in the dharma plot let's check create another model and check if we can get better values

```
vif(glm(formula = DistrictComplete ~ . -TotalSchools,
                            family = binomial(link="logit"),
                            data = districts_log_districtComplete))
```

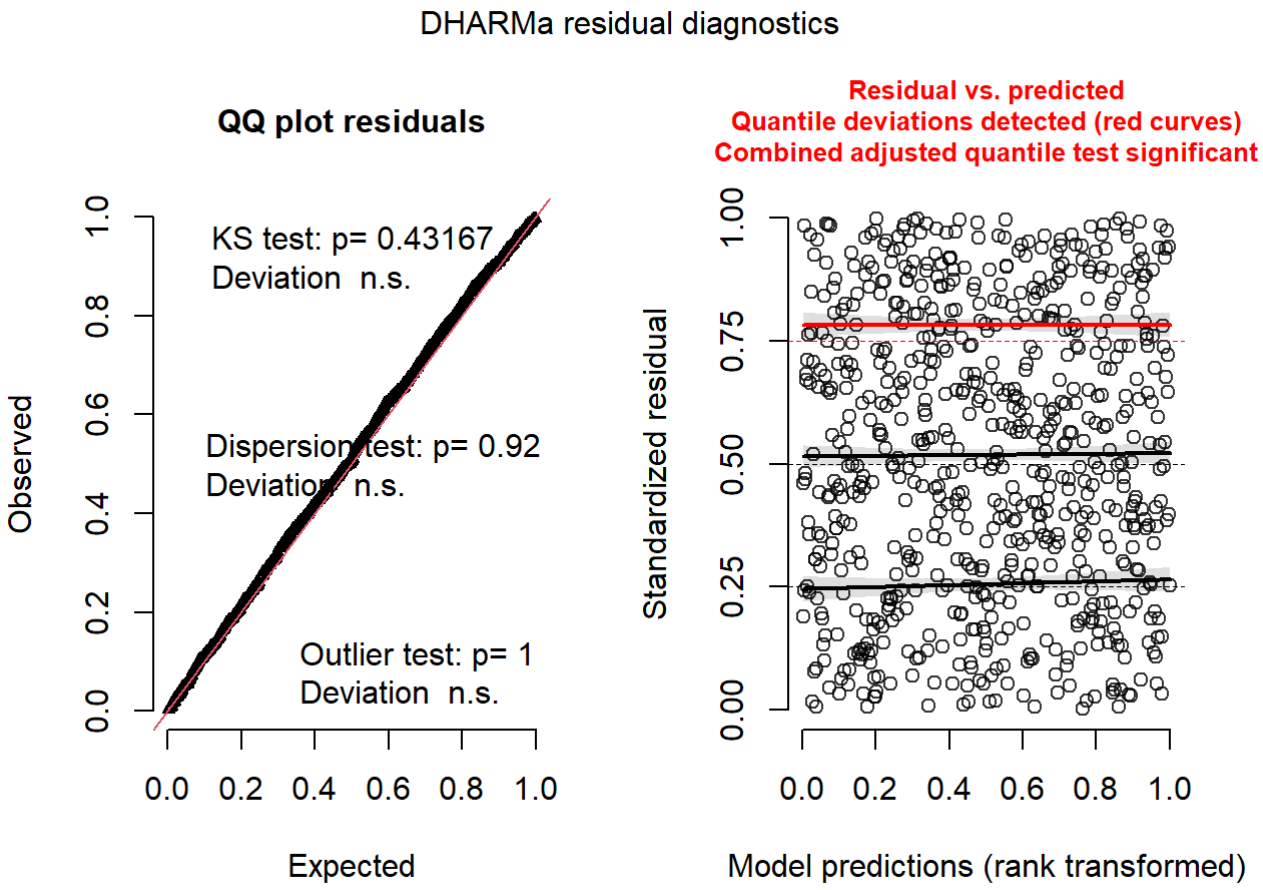
```
##  PctChildPoverty PctFamilyPoverty         Enrolled
##         6.038376         6.013165         1.010628
```

```
vif(glm(formula = DistrictComplete ~ . -Enrolled,
                            family = binomial(link="logit"),
                            data = districts_log_districtComplete))
```

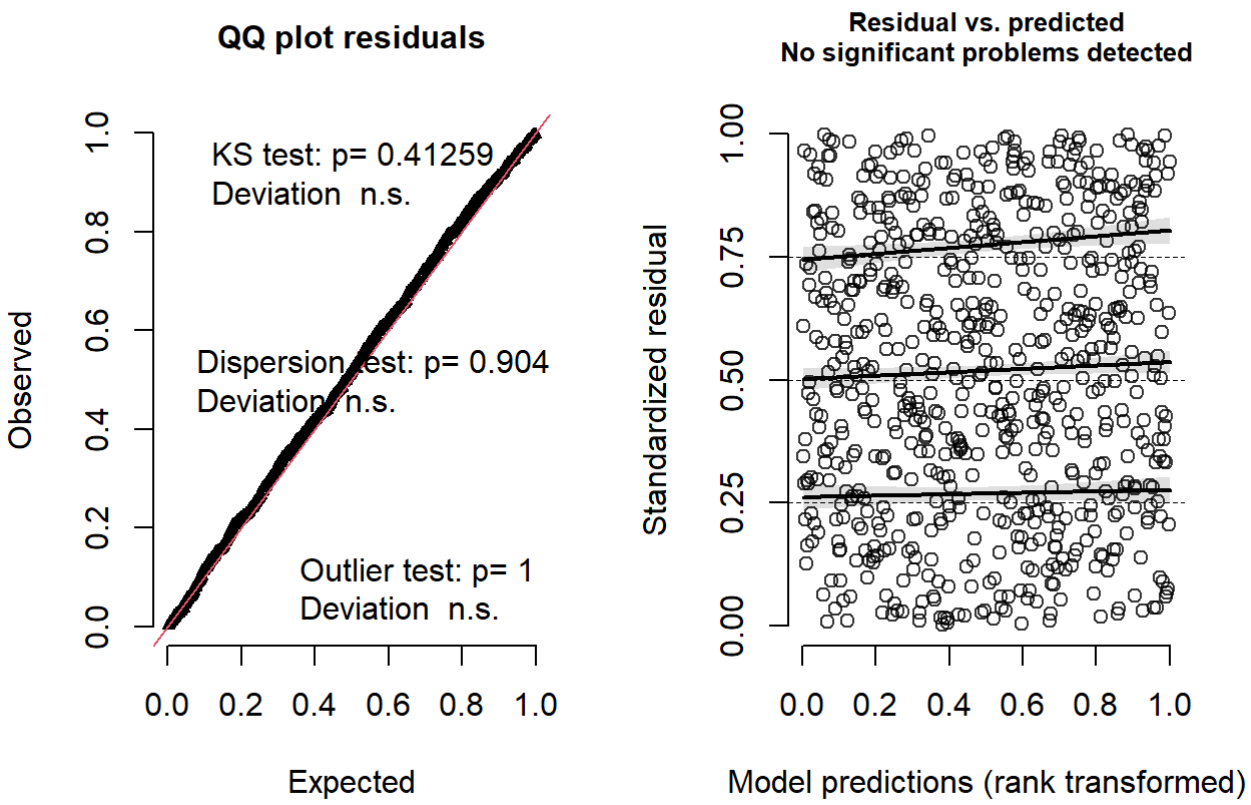
```
##  PctChildPoverty PctFamilyPoverty     TotalSchools
##         5.962513         5.945463         1.014070
```

```
# Since Removing ENrolled gives a better vif result and reduces
# multi collinarity more
districtsComplete_glm2 <- glm(formula = DistrictComplete ~ . -Enrolled,
                            family = binomial(link="logit"),
                            data = districts_log_districtComplete)
```

```
simulationOutput <- simulateResiduals(fittedModel = districtsComplete_glm,
                                      n=250)
plot(simulationOutput)
```




```
simulationOutput2 <- simulateResiduals(fittedModel = districtsComplete_glm2,
                                      n=250)
plot(simulationOutput2)
```



For the first model we see that qq plot are normally ditributed but the residual vs predicted plot see one deviation. Looking at the qq plot and residual vs predicted plot we can see that the values are normally distributed and no deviations are visible so we can go ahead and look at the summary of the model.

Checking the model accuracy

```
model_performance(districtsComplete_glm)
```

	AIC <dbl>	BIC <dbl>	R2_Tjur <dbl>	RMSE <dbl>	Sigma <dbl>	Log_loss <dbl>	Score_log <dbl>	Score_spheric <db
1	259.9391	282.6945	0.1712321	0.2194429	0.5996871	0.1785279	-Inf	0.00142857

1 row

```
model_performance(districtsComplete_glm2)
```

	AIC <dbl>	BIC <dbl>	R2_Tjur <dbl>	RMSE <dbl>	Sigma <dbl>	Log_loss <dbl>	Score_log <dbl>	Score_spheri <d
1	290.7433	308.9477	0.07749715	0.2310299	0.6373699	0.2019595	-Inf	0.001428

1 row

Since we get Tjur's R2 is having 0.171 for model 1 and 0.077 for model 2, since model 1 has a better R2 we will use that model for analysis.

```
summary(districtsComplete_glm)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ ., family = binomial(link = "logit"),
##     data = districts_log_districtComplete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0469   0.1205   0.2279   0.3538   1.6278
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.96007    1.20472  -1.627   0.1037
## PctChildPoverty   0.01754    0.03139   0.559   0.5763
## PctFamilyPoverty -0.08229    0.04387  -1.876   0.0607 .
## Enrolled          1.87705    0.35146   5.341 9.26e-08 ***
## TotalSchools     -3.24282    0.52353  -6.194 5.86e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 323.23  on 699  degrees of freedom
## Residual deviance: 249.94  on 695  degrees of freedom
## AIC: 259.94
##
## Number of Fisher Scoring iterations: 7
```

Here we see that TotalSchools is an excellent predictors for whether or not district's reporting was complete. The median value that we see is 0.2575 which is close to zero. Here there is a z-value which is Wald's Z test and is conceptually similar to a t-test. We can see that TotalSchools reports - (b= 1.87705, t(695)=5.341,p<.001) and Enrolled reports - (b= -3.24, t(695)=-6.194,p<.001) which shows that we can reject the null hypothesis that the coefficient on totalschools are 0 in the population. The variable PctChildPoverty reports (b=0.01754, t(695)=0.559, p>0.05) and PctFamilyPoverty reports (b=-0.08229, t(695)=-1.876, p>0.05) and for these two predictors we fail to reject the null. The bottom part of the output contains the omnibus test. There is a section Null Deviance, which is a model representing the null hypothesis. Then there is residual deviance, which represents what the model is like with predictors in it. The difference between the null deviance and the residual deviance model is the omnibus test and that is distributed as chi-squared. Since there is no chi-squared in the output we can do it manually and see that the value would be 323.23 - 249.94 = 73.29. AIC stands for Akaike information criterion and is the measure of stress on the model and we get a value of 259.94 but this is not important to use now as AIC is used to compare models and since we are only looking at one model here there is nothing to compare to.

Converting log odds to odds and the Confidence interval

```
exp(coef(districtsComplete_glm))
```

```
##      (Intercept)  PctChildPoverty PctFamilyPoverty         Enrolled
##       0.14084894       1.01769642       0.92100599       6.53417961
##     TotalSchools
##       0.03905349
```

```
exp(confint(districtsComplete_glm))
```

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

```
##                      2.5 %     97.5 %
## (Intercept)      0.01295616  1.4968940
## PctChildPoverty  0.95890342  1.0843360
## PctFamilyPoverty 0.84378917  1.0025368
## Enrolled         3.36761705 13.4719170
## TotalSchools     0.01309476  0.1030349
```

Looking at the above 2 outputs , we an interpret that one unit change in PctChildPoverty leads to 1.01769642 chance of whether or not district's reporting was complete ; every unit increase in PctFamilyPoverty leads to 0.92100599 chance ; every unit log increase in TotalSchools leads to 0.03905349 increase in chance of whether or not district's reporting was complete and every unit log increase in Enrolled leads to 6.53417961 increase in chance of whether or not district's reporting was complete. Looking at the CI we can see that we have a lower bound of 0.95890342 to upper bound of 1.0843360 for PctChildPoverty, PctFamilyPoverty has 0.8437891 to 1.0025368 , TotalSchools has 0.01309476 to 0.1030349 and Enrolled has 3.36761705 to 13.4719170. Since none of the CI don't coincide with 0 it means that it's unlikely that the true value is 0

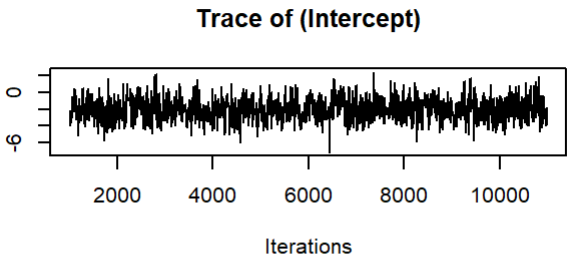
Doing the Bayesian version of the test:

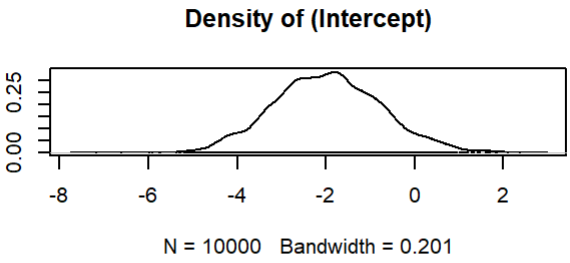
```
districtsComplete_glm.bayes <- MCMClogit(formula = DistrictComplete ~.,
                                         data=districts_log_districtComplete,
                                         rnd.seed=772)
summary(districtsComplete_glm.bayes)
```

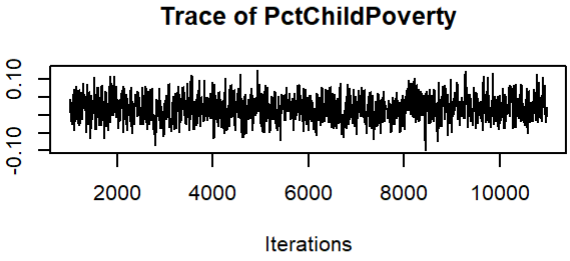
```
##
## Iterations = 1001:11000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                      Mean      SD  Naive SE Time-series SE
## (Intercept)      -1.96705 1.19625 0.0119625       0.048108
## PctChildPoverty   0.01789 0.03089 0.0003089       0.001239
## PctFamilyPoverty -0.08316 0.04329 0.0004329       0.001720
## Enrolled          1.91226 0.34656 0.0034656       0.014352
## TotalSchools     -3.31890 0.51754 0.0051754       0.020819
##
## 2. Quantiles for each variable:
##
##                      2.5%       25%      50%      75%     97.5%
## (Intercept)      -4.26948 -2.783116 -1.97454 -1.15240  0.460003
## PctChildPoverty  -0.04126 -0.002774  0.01719  0.03855  0.081162
## PctFamilyPoverty -0.16976 -0.112317 -0.08158 -0.05303 -0.001577
## Enrolled          1.23055  1.695666  1.90699  2.15022  2.587017
## TotalSchools     -4.37810 -3.648540 -3.30099 -2.97163 -2.320364
```

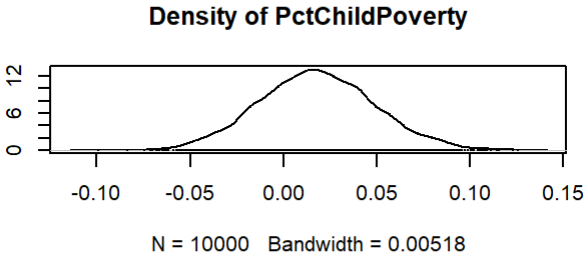
We can see that the point estimates for the intercept and the coefficients are quite similar to the output from the traditional logistic regression. Next we can see SD column which corresponds to the standard error in the output from the traditional logistic regression (because in this analysis it is a standard deviation of a sampling distribution). The second part of the output displays quantiles for each coefficient, including the 2.5% and 97.5% quantiles. The region in between the 2.5% and the 97.5% quantiles for each coefficient is the highest density interval (HDI) for the given coefficient. We can see PctChildPoverty predictor has a lower bound of -4.26948 to upper bound of 0.460003 and it coincides with 0 so it is possible that the observed difference is due to chance. Comparing it with out frequentist model we can see that we reach the same conclusion that PctChildPoverty predictor is not significant. Looking at the -0.16976 predictor we have a lower bound of -0.16976 to upper bound of -0.001577, ENrolled has 1.23055 to 2.587017 and TotalSchools has -4.37810 to -2.320364. Since the HDI doesn't coincide with 0 for these 3 columns we have credible evidence that there is a difference in between the variables. Except for PctFamilyPoverty we can see that the results match with our frequentist model.

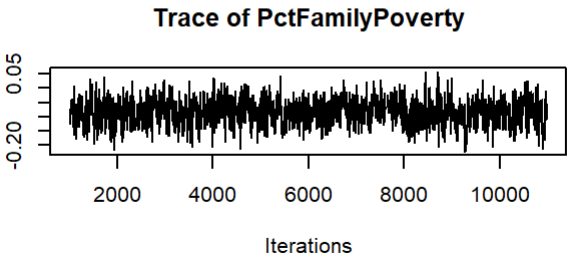
```
plot(districtsComplete_glm.bayes)
```

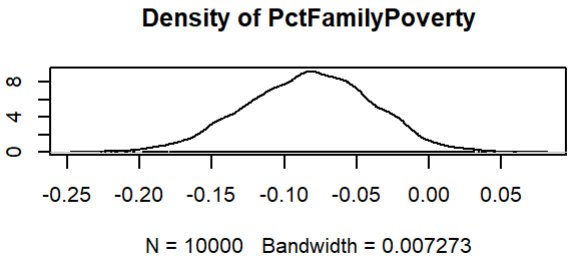


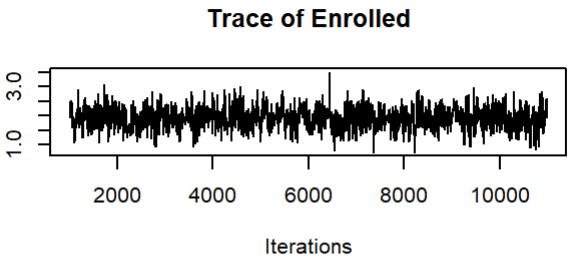


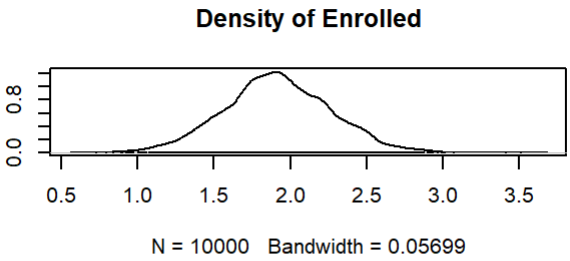


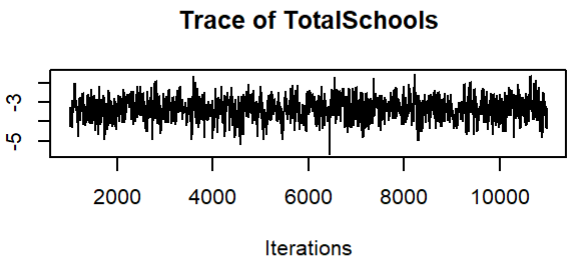


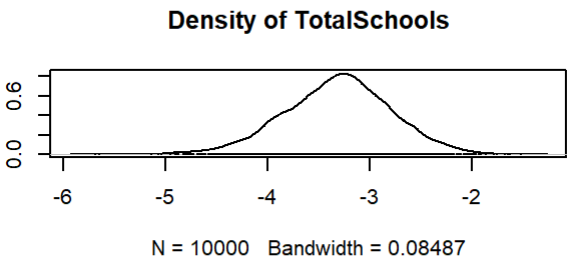












Looking at the density plots for all the columns we don’t see any distinctive patterns and the density plot looks pretty smooth.

In conclusion, using the result from both the Bayesian and frequentist approach we got evidence that totalSchools and Enrolled are good predictors for DistrictComplete.

7. *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.

We were analyzing vaccination rates to see how different factors like poverty, religion, medical factors affect vaccinations and analyze the vaccination rates. Considering all the models we built so far, be it frequentist or Bayesian we saw that our results from both were matching reinforcing our confidence in our results and helping us with our recommendations which we will give over. First we saw that if a student took one vaccine, there is a high chance that they took all the other vaccines like Polio, MMR, DTP, HepB. We also saw that having rules mandated in states affects the vaccination rate in a positive way and that belonging to a district which is below poverty line, there is a boost in vaccination probably. We can see that in the recent years the vaccination rate was pretty much constant.

We got the below observations from our analysis/conclusion: 1. public schools were the major ones that reported their vaccination data and since they are funded by the state, the state legislator's office should allocate financial assistance to school districts as they will not only help eradicate or avoid an epidemic but also help students which come from low economic backgrounds to get the necessary vaccinations. 2. Since the state would be funding these schools they can even get compliance on reporting of vaccination records from atleast the public schools. 3. Families in the below poverty district affected negatively meaning , the income didn't affect their will to get their children vaccinated which could be because they go in state funded schools so they don't have to pay for the vaccination. 4. Families who enrolled their children in school didn't want to take the belief exempt as maybe since the parents are educated they know that vaccination is important. 5. Families below poverty line preferred to get the vaccination and same is the case with enrolled students where we saw that such children had their vaccination up to date. 6. We could see that the interaction between PctChildPoverty and Enrolled was not significant.

Recommendations and Further Analysis: 1. The state should make reporting vaccination data a rule so that even private schools report their data 2. We saw that mandating vaccination in California really helped in increasing the vaccination rate, so mandating vaccination for children in every public and private school would be really beneficial.But we can't be sure until we analyze the state wise data for the whole of the United States and compare the vaccination rate between states where vaccination is not mandated and where it is like in California. 3. It would be good to know if state schools pay for the vaccination or not as the whole assumption here along with the conclusion of the model are that the state is paying for public school vaccinations. 4. Parents in private schools must have to fill in their childrens vaccination as a compulsary task for admission.