IST 772 Quantitative Reasoning– Final Examination

Tamilselvan Tamilmani

2021-12-22

** Introduction ** We are tasked with analyses and then write up a technical report for a scientifically knowledgeable staff member in a state legislator's office for the vaccine data in district 19 schools. The 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.
We will begin with exploratory analysis and come up with statistical analysis to help improve the vaccination rate and reporting compliance.

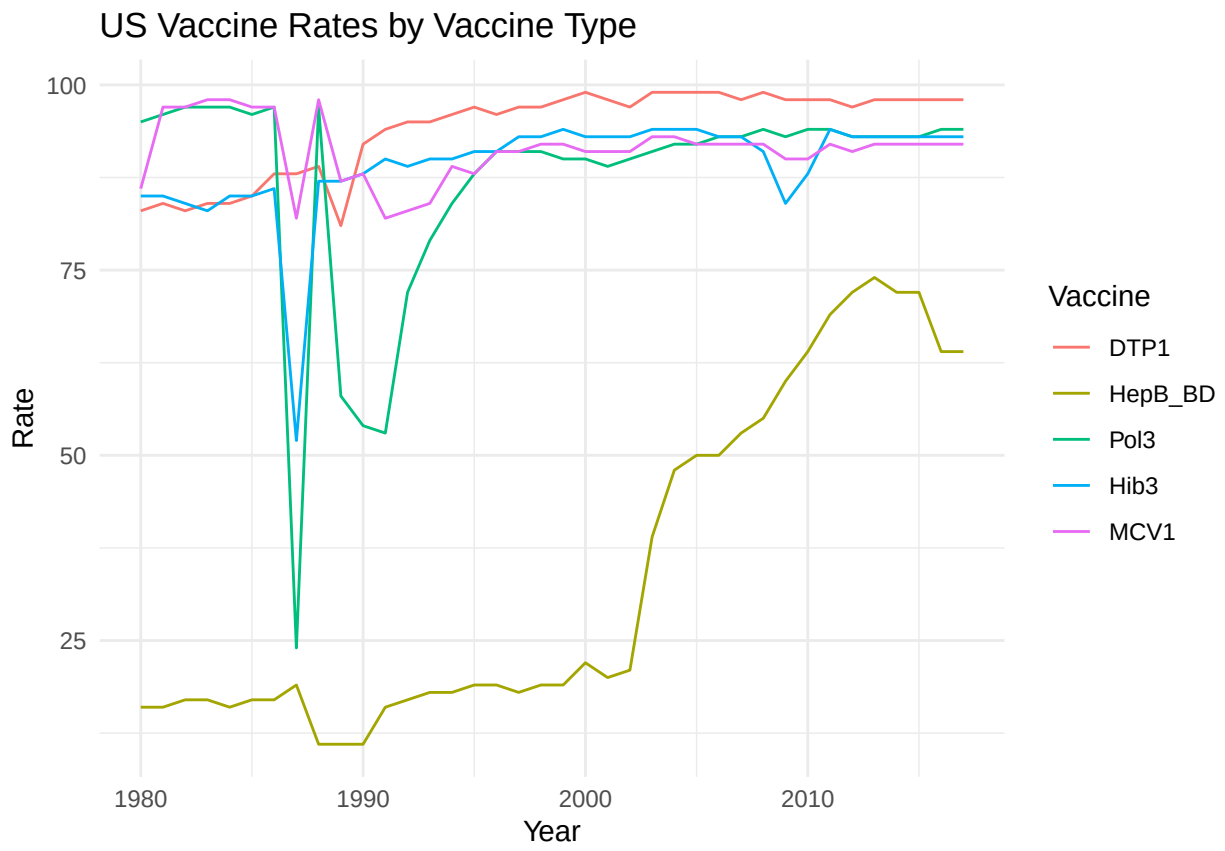
** Questions **
** Question 1 **
1. How have U.S. vaccination rates varied over time? Are vaccination rates increasing or decreasing? Which vaccination has the highest rate at the conclusion of the time series? Which vaccination has the lowest rate at the conclusion of the time series? Which vaccine has the greatest volatility?

```
set.seed(202112)
load("districts19.RData")
load("allSchoolsReportStatus.RData")
load("usVaccines.RData")

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
```

```
usvaccineDF <- data.frame(usVaccines)
usvaccineDF$year <- 1980:2017
library(reshape2)
usvaccineDF_melted<-melt(usvaccineDF,id.vars="year")
colnames(usvaccineDF_melted) <- c("Year","Vaccine","Rate")
library(ggplot2)
```

```
ggplot(usvaccineDF_melted, aes(x=Year, y=Rate,group=Vaccine, color=Vaccine)) +
  geom_line()  + ggtitle("US Vaccine Rates by Vaccine Type") +
  theme_minimal()
```




```
library(changepoint)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Successfully loaded changepoint package version 2.2.2
##  NOTE: Predefined penalty values changed in version 2.2.  Previous penalty values with a postfix 1 i
```

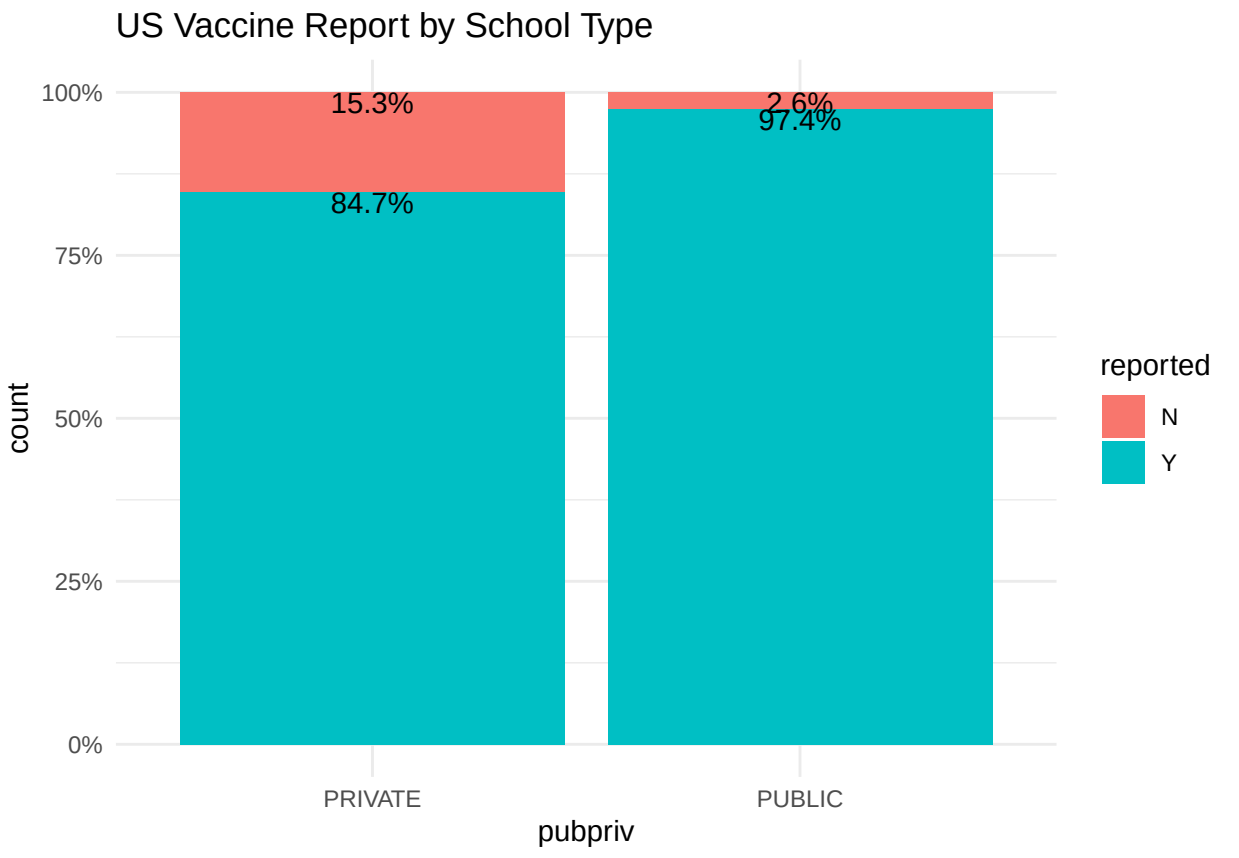
```
for (v in names(usvaccineDF)){
  #print(v)
  cp <- cpt.var(diff(usvaccineDF[[v]]),class=TRUE)
  print(paste(v,":",cpts(cp)))
}
```

```
## [1] "DTP1 : 10"
## [1] "HepB_BD : "
## [1] "Pol3 : 16"
## [1] "Hib3 : 8"
## [1] "MCV1 : 16"
## [1] "year : "
```

The plot shows the vaccine rates of individual vaccines over the years.
The US Vaccine rates gradually increased over time, except a sharp drop around late 80's
DTP1 - First dose of Diphtheria/Pertussis/Tetanus has the highest rate as of 2017
HepB_BD - Hepatitis B, Birth Dose has lowest rate as of 2017
Pol3 - Polio third dose and MCV1 - Measles first dose has large number of change points at 16, but Pol3 has the greatest volatility, since it has the largest range.

** Question 2 **
2. What proportion of public schools reported vaccination data? What proportion of private schools reported vaccination data? Was there any credible difference in overall reporting proportions between public and private schools?

```
library(scales)
pct_format = scales::percent_format(accuracy = .1)
ggplot(allSchoolsReportStatus, aes(x=pubpriv,fill=reported)) +
  geom_bar(position="fill", stat="count" )  +
  geom_text(aes(label =pct_format( ..count.. / tapply(..count.., ..x.., sum)[as.character(..x..)])), sta
  scale_y_continuous(labels = percent) + ggtitle("US Vaccine Report by School Type") +
  theme_minimal()
```



```
#Is there a difference, using chi.square test for categorical variable
pub_vs_private<-table(allSchoolsReportStatus$reported,allSchoolsReportStatus$pubpriv)
pub_vs_private
```

```
##
##     PRIVATE PUBLIC
##   N     252    148
##   Y    1397   5584
```

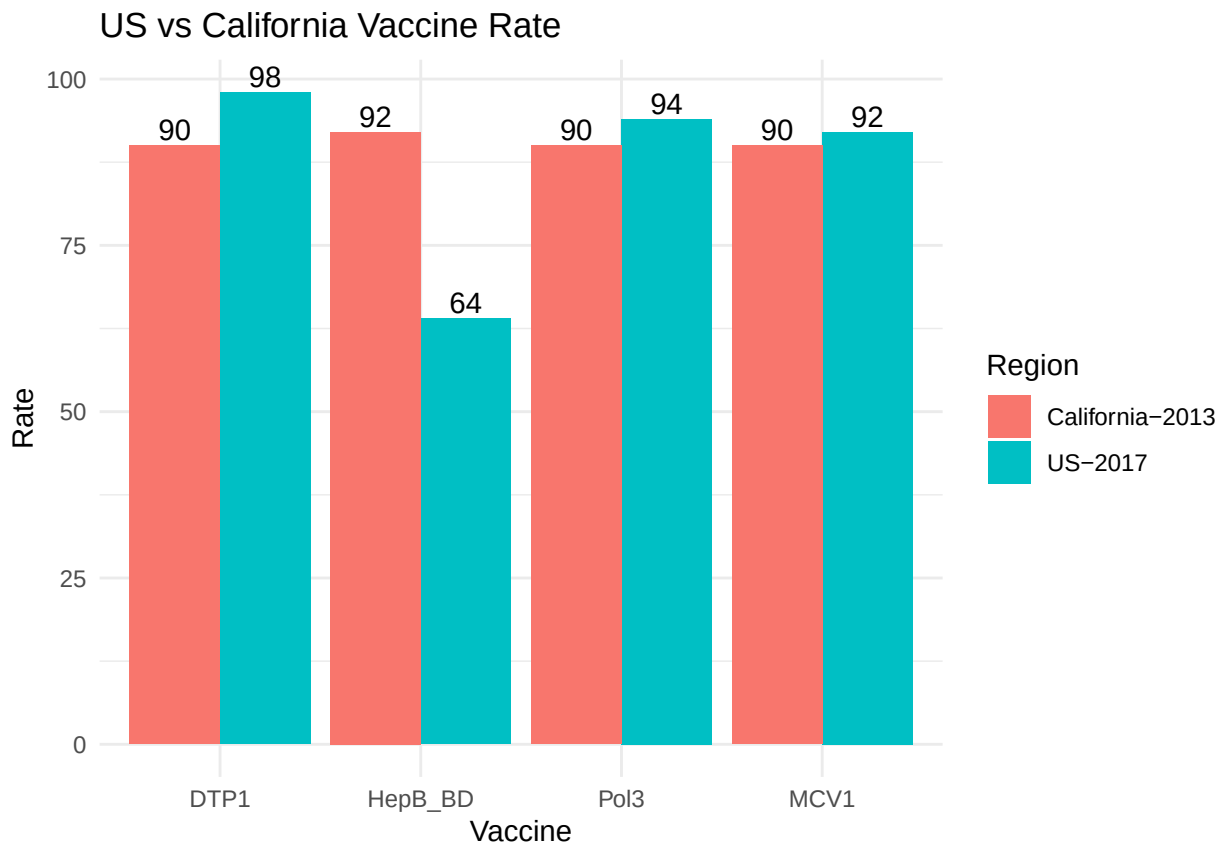
```
chisq.test(pub_vs_private)
```

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

The plot shows the distribution of vaccine reporting among public and private schools.
97.4% of public schools reported vaccine data.
84.7% of private schools reported vaccine data.
The p-value of chi square test on the public vs private vaccine reporting is very low, so we can reject the null hypothesis of no difference between the public vs private reporting(non independence), thus favoring the alternate hypothesis of there is a difference in reporting(independence) between public and private schools. So we can conclude there is a credible difference in the reporting between public and private schools.

** Question 3 **
3. What are 2013 vaccination rates for individual vaccines (i.e., DOT, Polio, MMR, and HepB) in California public schools? How do these rates for individual vaccines in California districts compare with overall US vaccination rates (make an informal comparison to the final observations in the time series)?

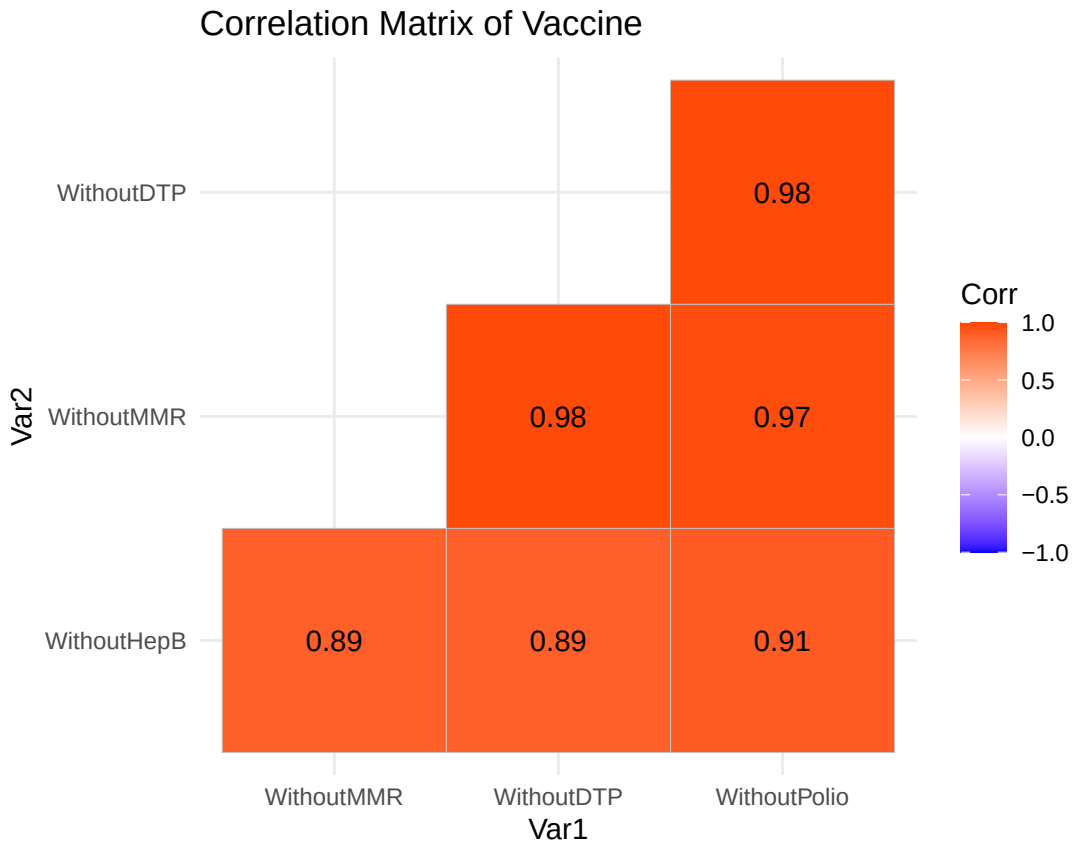
```
calf_2013_rate<- c(round(100 - mean(districts$WithoutDTP)), round(100 - mean(districts$WithoutHepB)),
                   round(100 - mean(districts$WithoutPolio)), round(100-mean(districts$WithoutMMR)), "Ca
us_2017_rate <- usvaccineDF[usvaccineDF$year==2017,]

df <- subset(us_2017_rate,select=-c(Hib3,year))
df$Region<-c("US-2017")
df <- rbind(df,calf_2013_rate)
df_melted<-melt(df,id.vars="Region")
colnames(df_melted) <- c("Region","Vaccine","Rate")
df_melted$Rate<- as.integer(df_melted$Rate)
ggplot(df_melted,aes(x=Vaccine, y=Rate, fill=Region)) +
  geom_bar(stat="identity",position="dodge") +
  geom_text(aes(label=Rate), position=position_dodge(width=0.9), vjust=-0.25) +
  ggtitle("US vs California Vaccine Rate") + theme_minimal()
```



The Plot show the comparison of vaccine rates between California in 2013 and overall US in 2017.
The individual vaccine rates of DOT, Polio, MMR, and HepB are 90,92,90 and 90 respectively in California public schools.
California is leading in HepB vaccine than overall US even before 3 years, and lagging on the remaining three vaccines.

** Question 4 **
4. Among districts, how are the vaccination rates for individual vaccines related? In other words, if students are missing one vaccine are they missing all of the others?

```
library(ggcorrplot)
ggcorrplot(cor( districts[,c(2:5)]),lab = TRUE,
           hc.order = TRUE, type = "lower",
           p.mat=cor_pmat(districts[,c(2:5)]),
           colors = c("blue", "white", "orangered"))  +
  ggtitle("Correlation Matrix of Vaccine") + theme_minimal()
```

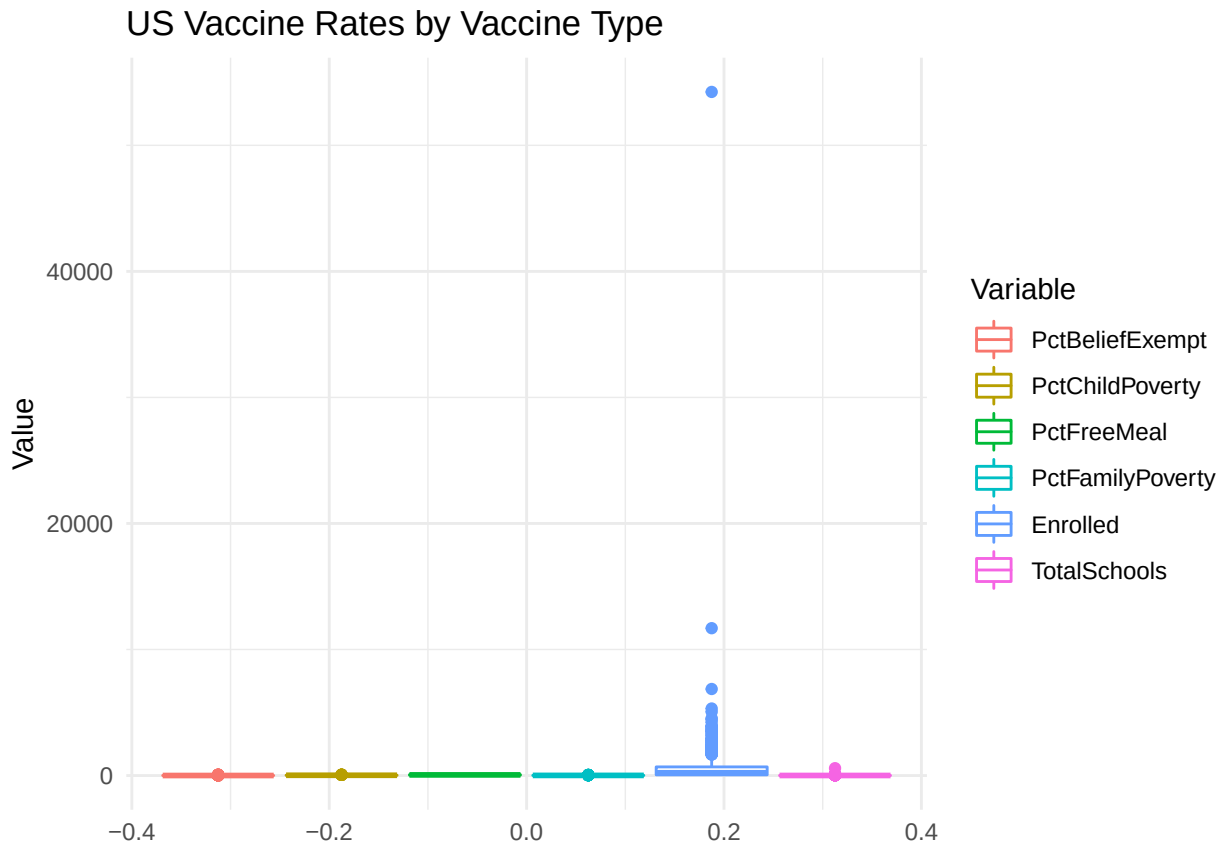


We can use correlation matrix to compare numeric variables.The correlation among the vaccine rates are very high and their p values are also high, so its highly likely students are missing all the vaccines if they miss any one. ** EDA & Data Preparation ** (For all of these analyses, use PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools as predictors. Transform variables as necessary to improve prediction and/or interpretability. In general, if there is a Bayesian version of an analysis available, you are expected to run that analysis in addition to the frequentist version of the analysis.)

```
districts_melted<-melt(districts[,8:13])
```

```
## No id variables; using all as measure variables
```

```
colnames(districts_melted) <- c("Variable","Value")
ggplot(districts_melted, aes( y=Value,group=Variable, color=Variable)) +
  geom_boxplot()  + ggtitle("US Vaccine Rates by Vaccine Type") +
  theme_minimal()
```




```
summary(districts)
```

```
##  DistrictName         WithoutDTP     WithoutPolio      WithoutMMR
##  Length:700         Min.   : 0.00   Min.   : 0.000   Min.   : 0.00
##  Class :character   1st Qu.: 3.00   1st Qu.: 3.000   1st Qu.: 3.00
##  Mode  :character   Median : 6.00   Median : 6.000   Median : 6.00
##                     Mean   :10.12   Mean   : 9.691   Mean   :10.12
##                     3rd Qu.:13.00   3rd Qu.:13.000   3rd Qu.:14.00
##                     Max.   :77.00   Max.   :77.000   Max.   :77.00
##   WithoutHepB      PctUpToDate     DistrictComplete PctBeliefExempt
##  Min.   : 0.000   Min.   : 23.00   Mode :logical    Min.   : 0.00
##  1st Qu.: 2.000   1st Qu.: 84.00   FALSE:41         1st Qu.: 1.00
##  Median : 4.000   Median : 92.00   TRUE :659        Median : 2.00
##  Mean   : 7.644   Mean   : 87.98                    Mean   : 5.54
##  3rd Qu.:10.000   3rd Qu.: 96.00                    3rd Qu.: 7.00
##  Max.   :77.000   Max.   :100.00                    Max.   :77.00
##  PctChildPoverty  PctFreeMeal     PctFamilyPoverty    Enrolled
##  Min.   : 2.00   Min.   :  0.00   Min.   : 0.00    Min.   :   10.0
##  1st Qu.:13.00   1st Qu.: 31.00   1st Qu.: 5.75    1st Qu.:   55.0
##  Median :21.00   Median : 50.00   Median :10.00    Median :  219.5
##  Mean   :22.45   Mean   : 49.18   Mean   :11.57    Mean   :  641.5
##  3rd Qu.:30.00   3rd Qu.: 70.00   3rd Qu.:16.00    3rd Qu.:  686.2
##  Max.   :63.00   Max.   :100.00   Max.   :44.00    Max.   :54238.0
##   TotalSchools
##  Min.   :  1.000
```

```
##  1st Qu.:  1.000
##  Median :  3.000
##  Mean   :  7.396
##  3rd Qu.:  8.000
##  Max.   :582.000
```

```
Q <- quantile(districts$Enrolled, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(districts$Enrolled)
up <-  Q[2]+1.5*iqr
low<- Q[1]-1.5*iqr
outlier_removed<- subset(districts, districts$Enrolled > low & districts$Enrolled < up)
summary(outlier_removed)
```

```
##  DistrictName         WithoutDTP     WithoutPolio     WithoutMMR
##  Length:636         Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  Class :character   1st Qu.: 3.00   1st Qu.: 3.00   1st Qu.: 3.00
##  Mode  :character   Median : 7.00   Median : 6.00   Median : 6.00
##                     Mean   :10.47   Mean   :10.06   Mean   :10.51
##                     3rd Qu.:14.00   3rd Qu.:13.00   3rd Qu.:14.00
##                     Max.   :77.00   Max.   :77.00   Max.   :77.00
##   WithoutHepB      PctUpToDate     DistrictComplete PctBeliefExempt
##  Min.   : 0.000   Min.   : 23.00   Mode :logical    Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 84.00   FALSE:29         1st Qu.: 0.000
##  Median : 4.000   Median : 92.00   TRUE :607        Median : 3.000
##  Mean   : 7.981   Mean   : 87.59                    Mean   : 5.862
##  3rd Qu.:10.000   3rd Qu.: 96.00                    3rd Qu.: 7.000
##  Max.   :77.000   Max.   :100.00                    Max.   :77.000
##  PctChildPoverty  PctFreeMeal     PctFamilyPoverty    Enrolled
##  Min.   : 2.00   Min.   :  0.00   Min.   : 0.00    Min.   :  10.00
##  1st Qu.:13.00   1st Qu.: 30.00   1st Qu.: 5.00    1st Qu.:  44.75
##  Median :21.00   Median : 50.00   Median : 9.00    Median : 169.50
##  Mean   :22.41   Mean   : 48.68   Mean   :11.46    Mean   : 340.13
##  3rd Qu.:30.00   3rd Qu.: 69.00   3rd Qu.:15.25    3rd Qu.: 484.75
##  Max.   :63.00   Max.   :100.00   Max.   :44.00    Max.   :1595.00
##   TotalSchools
##  Min.   : 1.000
##  1st Qu.: 1.000
##  Median : 2.000
##  Mean   : 4.222
##  3rd Qu.: 6.000
##  Max.   :23.000
```

The enrolled students have outlier in it, so we removed the outlier by using the IQR method described in https://www.r-bloggers.com/2020/01/how-to-remove-outliers-in-r/ . The Schools also have outliers in them, by removing the corresponding enrolled students the schools are also got rectified.

** Question 5 **
5. What variables predict whether or not a district's reporting was complete?

```
glm5_1<- glm(DistrictComplete~PctChildPoverty+PctFreeMeal+PctFamilyPoverty+Enrolled+TotalSchools, data =
summary(glm5_1)
```

```
##
```

```
## Call:
## glm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, family = binomial(),
##     data = outlier_removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5701   0.1537   0.2197   0.3062   1.5682
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       4.591826   0.635352   7.227 4.93e-13 ***
## PctChildPoverty   0.037037   0.033129   1.118    0.264
## PctFreeMeal      -0.019823   0.012310  -1.610    0.107
## PctFamilyPoverty -0.063415   0.040326  -1.573    0.116
## Enrolled          0.009781   0.002187   4.473 7.73e-06 ***
## TotalSchools     -0.798517   0.154302  -5.175 2.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 235.76  on 635  degrees of freedom
## Residual deviance: 194.64  on 630  degrees of freedom
## AIC: 206.64
##
## Number of Fisher Scoring iterations: 7
```

```
exp(coef(glm5_1))
```

```
##      (Intercept)  PctChildPoverty      PctFreeMeal PctFamilyPoverty
##       98.6744548        1.0377318        0.9803720        0.9385541
##         Enrolled     TotalSchools
##        1.0098288        0.4499960
```

```
library(BaylorEdPsych)
PseudoR2(glm5_1)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##       0.17440414       0.11502058       0.06260352       0.20211736
## McKelvey.Zavoina           Effron            Count        Adj.Count
##       0.33577306       0.09618136       0.95125786      -0.06896552
##              AIC    Corrected.AIC
##     206.63873267     206.77227798
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif(glm5_1)
```

```
##  PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled
##         4.279861         1.949058         3.675201        15.296711
##     TotalSchools
##        15.392057
```

Both Enrolled and Totalschools are highly correlated due to collinearity. So we will combine them by normalizing the enrolled stutdents to enrolled per school. Child Poverty and Family Povery are again correlated, so we will remove the child povery which has higher vcf

```
outlier_removed$Enrolled_norm <- outlier_removed$Enrolled / outlier_removed$TotalSchools
glm5_2<- glm(DistrictComplete~PctFreeMeal+PctFamilyPoverty+Enrolled_norm,
             data = outlier_removed, family = binomial())
summary(glm5_2)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctFreeMeal + PctFamilyPoverty +
##     Enrolled_norm, family = binomial(), data = outlier_removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8781   0.2049   0.2621   0.3415   0.7320
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.050060   0.568368   5.366 8.03e-08 ***
## PctFreeMeal      -0.010786   0.010734  -1.005  0.31498
## PctFamilyPoverty -0.027718   0.026695  -1.038  0.29911
## Enrolled_norm     0.014841   0.005533   2.682  0.00731 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 235.76  on 635  degrees of freedom
## Residual deviance: 222.49  on 632  degrees of freedom
## AIC: 230.49
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(glm5_2))
```

```
##      (Intercept)      PctFreeMeal PctFamilyPoverty    Enrolled_norm
##       21.1166156        0.9892723        0.9726623        1.0149512
```

```
PseudoR2(glm5_2)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##       0.05628497       0.01386814       0.02064783       0.06666215
## McKelvey.Zavoina           Effron            Count        Adj.Count
##       0.14539294       0.02036979               NA               NA
##              AIC    Corrected.AIC
##     230.48597265     230.54936409
```

```
vif(glm5_2)
```

```
##      PctFreeMeal PctFamilyPoverty    Enrolled_norm
##         1.726192         1.727606         1.005246
```

Free meal and Family poverty are some what correlated, so we will remove free meal since it has high vif.

```
glm5_3<- glm(DistrictComplete~PctFreeMeal+Enrolled_norm,
             data = outlier_removed, family = binomial())
summary(glm5_3)
```

```
##
## Call:
## glm(formula = DistrictComplete ~ PctFreeMeal + Enrolled_norm,
##     family = binomial(), data = outlier_removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8384   0.2070   0.2683   0.3453   0.6305
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.082506   0.578259   5.331 9.79e-08 ***
## PctFreeMeal   -0.017745   0.008327  -2.131  0.03308 *
## Enrolled_norm  0.014524   0.005516   2.633  0.00846 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 235.76  on 635  degrees of freedom
## Residual deviance: 223.52  on 633  degrees of freedom
## AIC: 229.52
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(glm5_3))
```

```
##   (Intercept)   PctFreeMeal Enrolled_norm
##     21.812995      0.982411      1.014630
```

```
PseudoR2(glm5_3)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##       0.05189386       0.01796040       0.01905242       0.06151133
## McKelvey.Zavoina           Effron            Count        Adj.Count
##       0.14204697       0.01698787               NA               NA
##              AIC    Corrected.AIC
##     229.52120028     229.55917496
```

```
vif(glm5_3)
```

```
##   PctFreeMeal Enrolled_norm
##      1.001986      1.001986
```

```
library(MCMCpack)
```

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

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

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

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

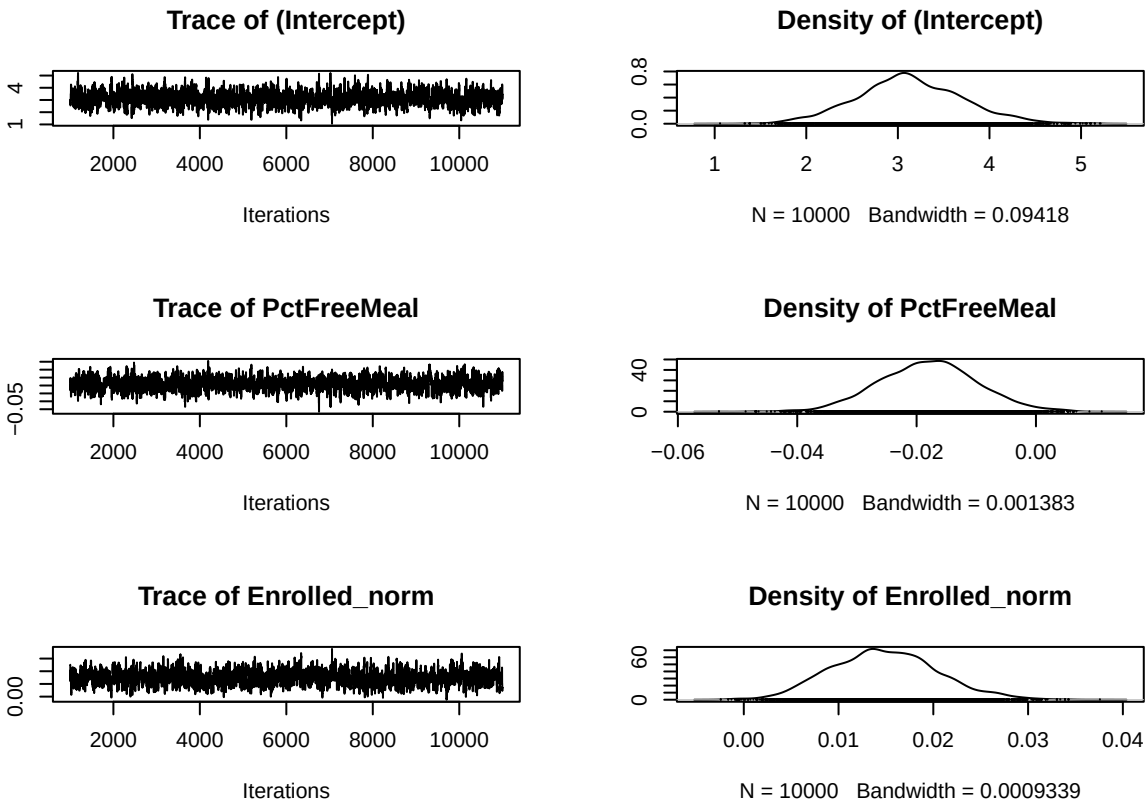
```
## ##
## ## Support provided by the U.S. National Science Foundation
```

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

```
bayes_glm5_3<- MCMClogit(DistrictComplete~PctFreeMeal+Enrolled_norm, data = outlier_removed)
summary(bayes_glm5_3)
```

```
##
## 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)    3.12819 0.574174 5.742e-03      0.0184684
## PctFreeMeal   -0.01803 0.008270 8.270e-05      0.0002710
## Enrolled_norm  0.01484 0.005559 5.559e-05      0.0001818
##
## 2. Quantiles for each variable:
##
##                    2.5%      25%      50%      75%     97.5%
## (Intercept)    1.997120  2.75399  3.10485  3.50516  4.296901
## PctFreeMeal   -0.034025 -0.02363 -0.01786 -0.01260 -0.001747
## Enrolled_norm  0.004724  0.01085  0.01469  0.01851  0.026538
```

```
plot(bayes_glm5_3)
```



Out of the three models we select the third one which has lower AIC score after eliminating the collinear variables.
PctFreeMeal and Enrolled students per School predicts the Districts reporting is complete or not. The frequentist method gives us a very low r square of 7% (Nagelkerke), makes us not very confident in our model. The Percent Free Meal is significant with p-value .036, and also the HDI does not cross zero , -0.037(2.5%) to -0.002(97.5%), both the frequentist and Bayesian confirms the significance.
The Enrolled per School is also significant with p-value .011, and also the HDI does not cross zero , 0.005(2.5%) to 0.029(97.5%), both the frequentist and Bayesian confirms the significance.
Further the trace of the variables have no outliers, indicating the mcmc converged. And both frequentist and Bayesian agree on the coefficients at -0.02 for PctFreeMeal and .02 for Enrolled per School.

** Question 6 **
6. What variables predict the percentage of all enrolled students with completely up-to-date vaccines?

```
glm6_1<- glm(PctUpToDate~PctChildPoverty+PctFreeMeal+PctFamilyPoverty+Enrolled+TotalSchools, data = outl
summary(glm6_1)
```

```
##
## Call:
## glm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
##     Enrolled + TotalSchools, family = gaussian(), data = outlier_removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -62.675   -4.115    2.417    7.486   20.323
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      79.723887   1.265946  62.976  < 2e-16 ***
## PctChildPoverty  -0.001448   0.081225  -0.018  0.98578
## PctFreeMeal       0.066323   0.029900   2.218  0.02690 *
## PctFamilyPoverty  0.219635   0.113081   1.942  0.05255 .
## Enrolled          0.019260   0.003876   4.969  8.7e-07 ***
## TotalSchools     -1.041962   0.349053  -2.985  0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 146.7208)
##
##     Null deviance: 107834  on 635  degrees of freedom
## Residual deviance:  92434  on 630  degrees of freedom
## AIC: 4985.6
##
## Number of Fisher Scoring iterations: 2
```

```
PseudoR2(glm6_1)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##     1.428104e-01     1.426806e-01     1.000000e+00     1.000000e+00
## McKelvey.Zavoina           Effron            Count        Adj.Count
##               NA     1.428104e-01     4.716981e-03    -7.106599e-02
##              AIC    Corrected.AIC
##     9.244609e+04     9.244622e+04
```

```
vif(glm6_1)
```

```
##  PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled
##         4.235801         2.388627         3.744828         9.886842
##     TotalSchools
##         9.884949
```

```
glm6_2<- glm(PctUpToDate~PctFreeMeal+Enrolled_norm, data = outlier_removed, family = gaussian())
vif(glm6_2)
```

```
##   PctFreeMeal Enrolled_norm
##      1.013247      1.013247
```

```
PseudoR2(glm6_2)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##     1.268940e-01     1.268198e-01     1.000000e+00     1.000000e+00
## McKelvey.Zavoina           Effron            Count        Adj.Count
##               NA     1.268940e-01     4.716981e-03    -7.106599e-02
##              AIC    Corrected.AIC
##     9.415641e+04     9.415645e+04
```

```
summary(glm6_2)
```

```
##
## Call:
## glm(formula = PctUpToDate ~ PctFreeMeal + Enrolled_norm, family = gaussian(),
##     data = outlier_removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -62.529   -3.689    2.800    7.383   22.162
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.58527    1.26315  60.630  < 2e-16 ***
## PctFreeMeal    0.11387    0.01961   5.807 1.01e-08 ***
## Enrolled_norm  0.07589    0.01097   6.920 1.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 148.7368)
##
##     Null deviance: 107834  on 635  degrees of freedom
## Residual deviance:  94150  on 633  degrees of freedom
## AIC: 4991.3
##
## Number of Fisher Scoring iterations: 2
```

```
lm6_2<-lm(PctUpToDate~PctFreeMeal+Enrolled_norm, data = outlier_removed)
summary(lm6_2)
```

```
##
## Call:
## lm(formula = PctUpToDate ~ PctFreeMeal + Enrolled_norm, data = outlier_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.529  -3.689   2.800   7.383  22.162
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.58527    1.26315  60.630  < 2e-16 ***
## PctFreeMeal    0.11387    0.01961   5.807 1.01e-08 ***
## Enrolled_norm  0.07589    0.01097   6.920 1.11e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.2 on 633 degrees of freedom
## Multiple R-squared:  0.1269, Adjusted R-squared:  0.1241
## F-statistic:    46 on 2 and 633 DF,  p-value: < 2.2e-16
```

```
library(BayesFactor)
```

```
## Loading required package: Matrix

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

```
bayes_glm6_2<- regressionBF(PctUpToDate~PctChildPoverty+PctFreeMeal+PctFamilyPoverty+Enrolled_norm, data
summary(bayes_glm6_2)
```

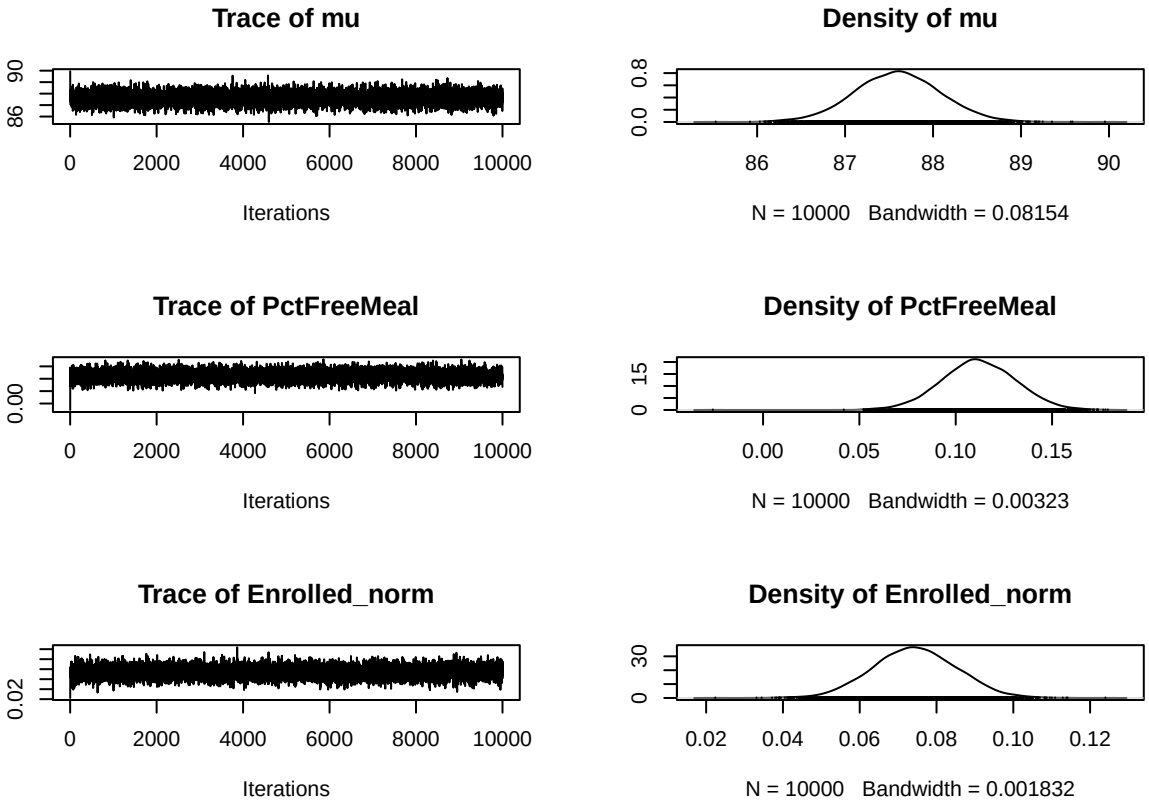
```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty                                                 : 42263.42     ±0%
## [2] PctFreeMeal                                                     : 26836164     ±0%
## [3] PctFamilyPoverty                                                : 8417333      ±0%
## [4] Enrolled_norm                                                   : 18963713207  ±0%
## [5] PctChildPoverty + PctFreeMeal                                   : 4078650      ±0%
## [6] PctChildPoverty + PctFamilyPoverty                              : 963856.4     ±0%
## [7] PctChildPoverty + Enrolled_norm                                 : 4.072524e+14 ±0.01%
## [8] PctFreeMeal + PctFamilyPoverty                                  : 48186059     ±0%
## [9] PctFreeMeal + Enrolled_norm                                     : 2.051801e+16 ±0.01%
## [10] PctFamilyPoverty + Enrolled_norm                               : 2.038377e+15 ±0.01%
## [11] PctChildPoverty + PctFreeMeal + PctFamilyPoverty               : 12556532     ±0%
## [12] PctChildPoverty + PctFreeMeal + Enrolled_norm                  : 4.685417e+15 ±0%
## [13] PctChildPoverty + PctFamilyPoverty + Enrolled_norm             : 3.871342e+14 ±0%
## [14] PctFreeMeal + PctFamilyPoverty + Enrolled_norm                 : 1.259119e+16 ±0.01%
## [15] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled_norm : 1.796828e+15 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

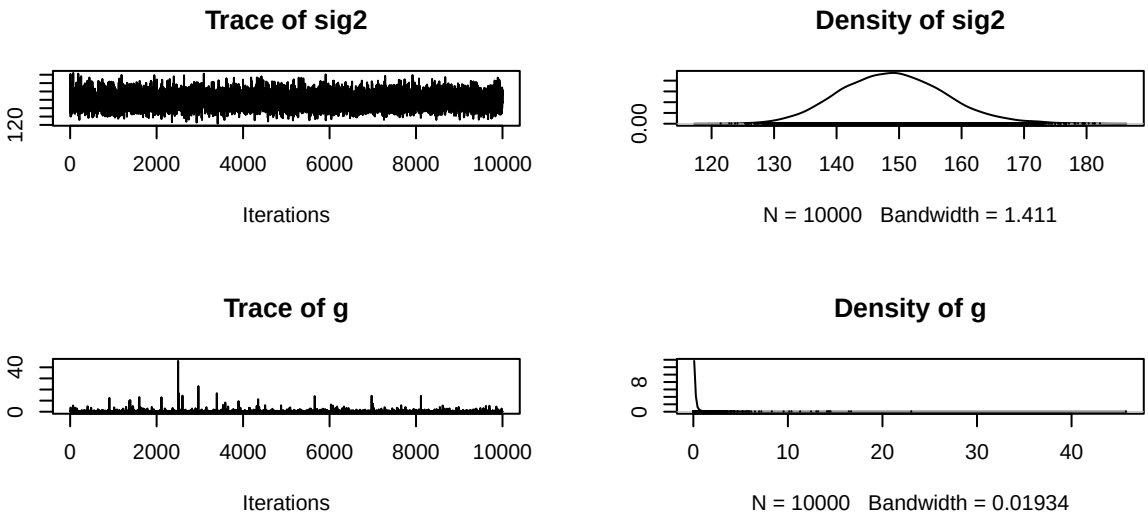
```
bayes_glm6_2_final<-lmBF(PctUpToDate~PctFreeMeal+Enrolled_norm, data = outlier_removed,posterior=TRUE, i
summary(bayes_glm6_2_final)
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                    Mean      SD  Naive SE Time-series SE
## mu             87.59580 0.48693 0.0048693      0.0048693
## PctFreeMeal     0.11193 0.01923 0.0001923      0.0001923
## Enrolled_norm   0.07436 0.01090 0.0001090      0.0001090
## sig2          149.09171 8.39686 0.0839686      0.0839686
## g               0.24932 0.81796 0.0081796      0.0089048
```

```
##
## 2. Quantiles for each variable:
##
##                     2.5%       25%       50%       75%     97.5%
## mu             86.62948  87.26879  87.59491  87.91920  88.55752
## PctFreeMeal     0.07389   0.09910   0.11177   0.12499   0.14963
## Enrolled_norm   0.05336   0.06703   0.07423   0.08174   0.09555
## sig2          133.48392 143.22621 148.86930 154.59817 166.75043
## g               0.02772   0.06324   0.10960   0.21754   1.24343
```

```
plot(bayes_glm6_2_final)
```





We tried linear modeling , it didn't give a good R square value, So we tried with GLM and normal distribution, which gave a high pseudo R square and low AIC making us more confident in our model. Out of the two models we select the second one which has lower AIC score after eliminating the collinear variables.
PctFreeMeal and Enrolled students per School predicts the Percentage of Students up to date with vaccines.
The frequentist method gives us a very high r square of 100% (Nagelkerke), makes us very confident in our model, although the Adjusted McFadden is in line with the LM model at 11%
The Percent Free Meal is very significant with p-value 1.15e-07, and Enrolled per School is also significant with p-value 3.10e-09.
And Bayesian method also picked PctFreeMeal + Enrolled_norm as the predictors with highest factor st 2.77478e+12.
The HDI intervals are also not crossing zero, giving us high confidence for the coefficients and are in line with frequentist estimates at 0.12 and 0.07 for PctFreeMeal and Enrolled students per School respectively.
Further the trace of the variables have no outliers, indicating the mcmc converged.

** Question 7 **
7. What variables predict the percentage of all enrolled students with belief exceptions?

```
glm7_1<- glm(PctBeliefExempt~PctChildPoverty+PctFreeMeal+PctFamilyPoverty+Enrolled+TotalSchools, data =
summary(glm7_1)
```

```
##
## Call:
## glm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
##     PctFamilyPoverty + Enrolled + TotalSchools, family = gaussian(),
##     data = outlier_removed)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -13.108   -4.360   -1.683    1.728   64.732
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.474353   0.896462  12.800  < 2e-16 ***
## PctChildPoverty   0.121285   0.057519   2.109   0.0354 *
## PctFreeMeal      -0.096838   0.021173  -4.574 5.77e-06 ***
## PctFamilyPoverty -0.192876   0.080077  -2.409   0.0163 *
## Enrolled         -0.010870   0.002745  -3.960 8.35e-05 ***
## TotalSchools      0.542614   0.247177   2.195   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 73.574)
##
##     Null deviance: 54330  on 635  degrees of freedom
## Residual deviance: 46352  on 630  degrees of freedom
## AIC: 4546.6
##
## Number of Fisher Scoring iterations: 2
```

```
PseudoR2(glm7_1)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##        0.1468476        0.1465899        0.9999964        0.9999964
## McKelvey.Zavoina           Effron            Count        Adj.Count
##               NA        0.1468476        0.1430818       -0.1571125
##              AIC    Corrected.AIC
##    46363.6204088    46363.7539542
```

```
vif(glm7_1)
```

```
##  PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled
##         4.235801         2.388627         3.744828         9.886842
##     TotalSchools
##         9.884949
```

```
glm7_2<- glm(PctBeliefExempt~PctFreeMeal+Enrolled_norm, data = outlier_removed, family = gaussian)
summary(glm7_2)
```

```
##
## Call:
## glm(formula = PctBeliefExempt ~ PctFreeMeal + Enrolled_norm,
##     family = gaussian, data = outlier_removed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -13.026   -4.380   -1.923    1.229   65.789
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.810820   0.896419  15.407  < 2e-16 ***
## PctFreeMeal   -0.095346   0.013915  -6.852 1.73e-11 ***
## Enrolled_norm -0.045966   0.007783  -5.906 5.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 74.90793)
##
##     Null deviance: 54330  on 635  degrees of freedom
## Residual deviance: 47417  on 633  degrees of freedom
## AIC: 4555
##
## Number of Fisher Scoring iterations: 2
```

```
PseudoR2(glm7_2)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##        0.1272433        0.1270960        0.9999810        0.9999810
## McKelvey.Zavoina           Effron            Count        Adj.Count
##               NA        0.1272433        0.1493711       -0.1486200
##              AIC    Corrected.AIC
##    47422.7179284    47422.7559031
```

```
vif(glm7_2)
```

```
##   PctFreeMeal Enrolled_norm
##      1.013247      1.013247
```

```
bayes_glm7_2<- regressionBF(PctBeliefExempt~PctChildPoverty+PctFreeMeal+PctFamilyPoverty+Enrolled_norm,
summary(bayes_glm7_2)
```

```
## Bayes factor analysis
## --------------
## [1] PctChildPoverty                                                : 828.5981     ±0.01%
## [2] PctFreeMeal                                                    : 12532822080  ±0%
## [3] PctFamilyPoverty                                               : 1867517      ±0%
## [4] Enrolled_norm                                                  : 46567829     ±0%
## [5] PctChildPoverty + PctFreeMeal                                  : 4072760286   ±0%
## [6] PctChildPoverty + PctFamilyPoverty                             : 419617       ±0%
## [7] PctChildPoverty + Enrolled_norm                                : 20997950912  ±0%
## [8] PctFreeMeal + PctFamilyPoverty                                 : 2488777965   ±0%
## [9] PctFreeMeal + Enrolled_norm                                    : 2.323886e+16 ±0.01%
## [10] PctFamilyPoverty + Enrolled_norm                              : 2.022349e+12 ±0.01%
## [11] PctChildPoverty + PctFreeMeal + PctFamilyPoverty              : 33415534503  ±0.01%
## [12] PctChildPoverty + PctFreeMeal + Enrolled_norm                 : 6.118625e+15 ±0.01%
## [13] PctChildPoverty + PctFamilyPoverty + Enrolled_norm            : 284317833088 ±0%
## [14] PctFreeMeal + PctFamilyPoverty + Enrolled_norm                : 3.352769e+15 ±0%
## [15] PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled_norm : 6.081803e+15 ±0.01%
##
## Against denominator:
##   Intercept only
```

```
## ---
## Bayes factor type: BFlinearModel, JZS
```

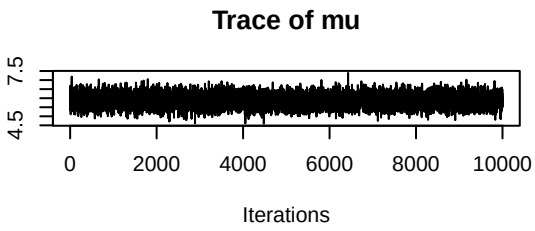
```
bayes_glm7_2[9]
```

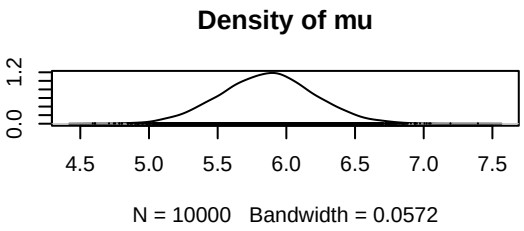
```
## Bayes factor analysis
## --------------
## [1] PctFreeMeal + Enrolled_norm : 2.323886e+16 ±0.01%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

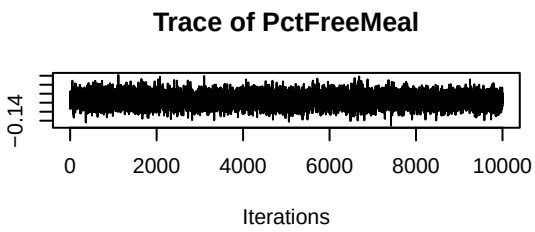
```
bayes_glm7_2_final<-lmBF(PctBeliefExempt~PctFreeMeal+Enrolled_norm, data = outlier_removed,posterior=TRU
summary(bayes_glm7_2_final)
```

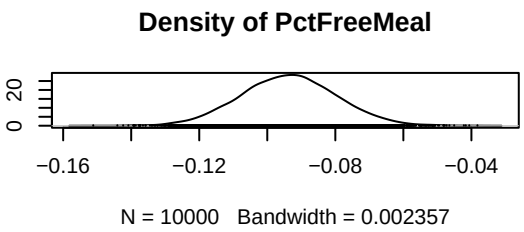
```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                    Mean       SD  Naive SE Time-series SE
## mu              5.86464 0.342193 3.422e-03      3.298e-03
## PctFreeMeal    -0.09369 0.014029 1.403e-04      1.403e-04
## Enrolled_norm -0.04512 0.007804 7.804e-05      7.804e-05
## sig2          75.09314 4.182570 4.183e-02      4.263e-02
## g              0.31729 3.491764 3.492e-02      3.492e-02
##
## 2. Quantiles for each variable:
##
##                    2.5%      25%      50%      75%    97.5%
## mu              5.19608  5.63395  5.86812  6.09015  6.53958
## PctFreeMeal    -0.12086 -0.10313 -0.09368 -0.08426 -0.06603
## Enrolled_norm -0.06029 -0.05039 -0.04508 -0.03994 -0.02978
## sig2          67.27460 72.19279 74.99185 77.89071 83.53192
## g              0.02803  0.06453  0.11232  0.22102  1.21915
```

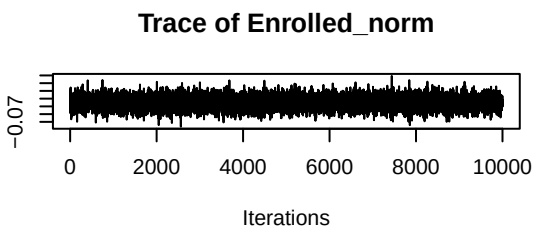
```
plot(bayes_glm7_2_final)
```

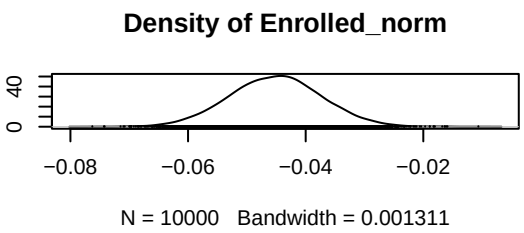

Trace of mu
4.5
7.5
0
2000
4000
6000
8000
10000
Iterations

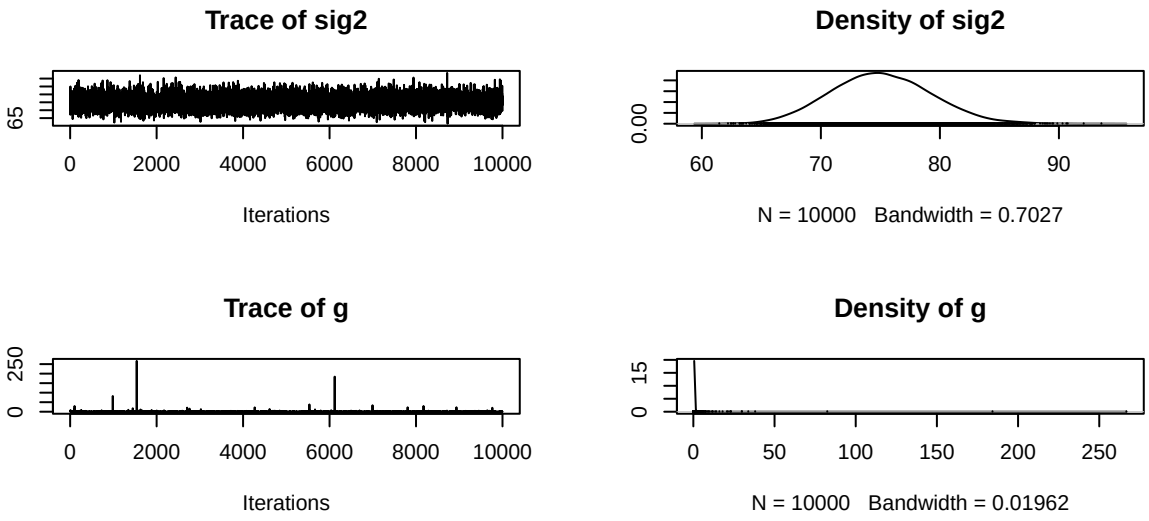

Density of mu
0.0
1.2
4.5
5.0
5.5
6.0
6.5
7.0
7.5
N = 10000 Bandwidth = 0.0572


Trace of PctFreeMeal
−0.14
0
2000
4000
6000
8000
10000
Iterations


Density of PctFreeMeal
0
20
−0.16
−0.12
−0.08
−0.04
N = 10000 Bandwidth = 0.002357


Trace of Enrolled_norm
−0.07
0
2000
4000
6000
8000
10000
Iterations


Density of Enrolled_norm
0
40
−0.08
−0.06
−0.04
−0.02
N = 10000 Bandwidth = 0.001311



Out of the two models we selected the second one which has lower AIC score after eliminating the collinear variables.
PctFreeMeal and Enrolled students per School predicts the Percentage of Students with belief exemptions.
The frequentist method gives us a very high r square of 99.99% (Nagelkerke), makes us very confident in our model.
The Percent Free Meal is very significant with p-value 4.67e-10, and Enrolled per School is also significant with p-value 4.49e-07.
And Bayesian method also picked PctFreeMeal + Enrolled_norm as the predictors with highest factor st 4.663353e+12.
The HDI intervals are also not crossing zero, giving us high confidence for the coefficients and are in line with frequentist estimates at -0.1 and -0.04 for PctFreeMeal and Enrolled students per School respectively.
Further the trace of the variables have no outliers, indicating the mcmc converged.

** Question 8 **
8. What's the big picture, based on all of the foregoing analyses? 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. What have you learned from the data and analyses that might inform this question?

```
g1<-ggplot(outlier_removed,aes(y=PctBeliefExempt,x=PctFreeMeal)) +
  geom_point() + theme_minimal() +
  geom_smooth(method="glm")

g2<-ggplot(outlier_removed,aes(y=PctBeliefExempt,x=Enrolled_norm)) +
  geom_point() + theme_minimal() +
  geom_smooth(method="glm")
```

```
library(patchwork)
```

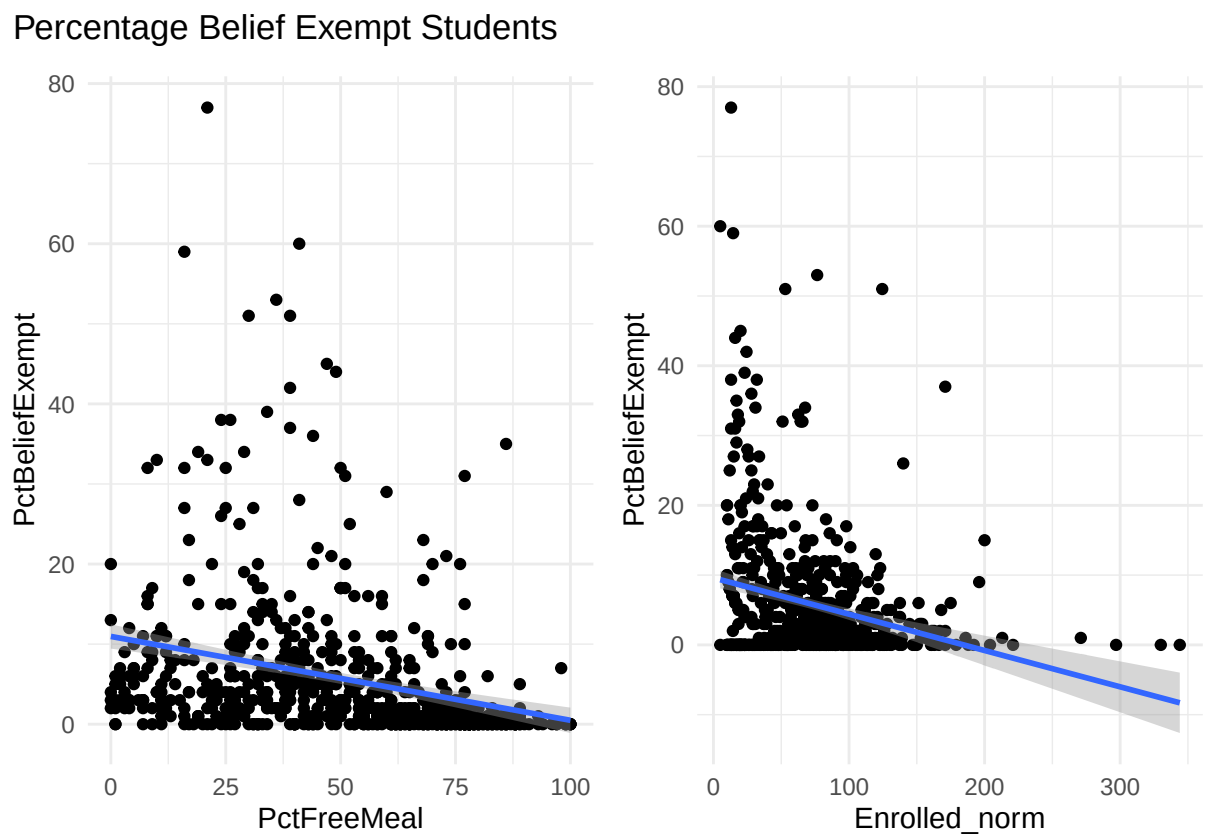
```
##
## Attaching package: 'patchwork'
```

```
## The following object is masked from 'package:MASS':
##
##     area
```

```
g1 + g2  +  plot_annotation(title = "Percentage Belief Exempt Students")
```

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

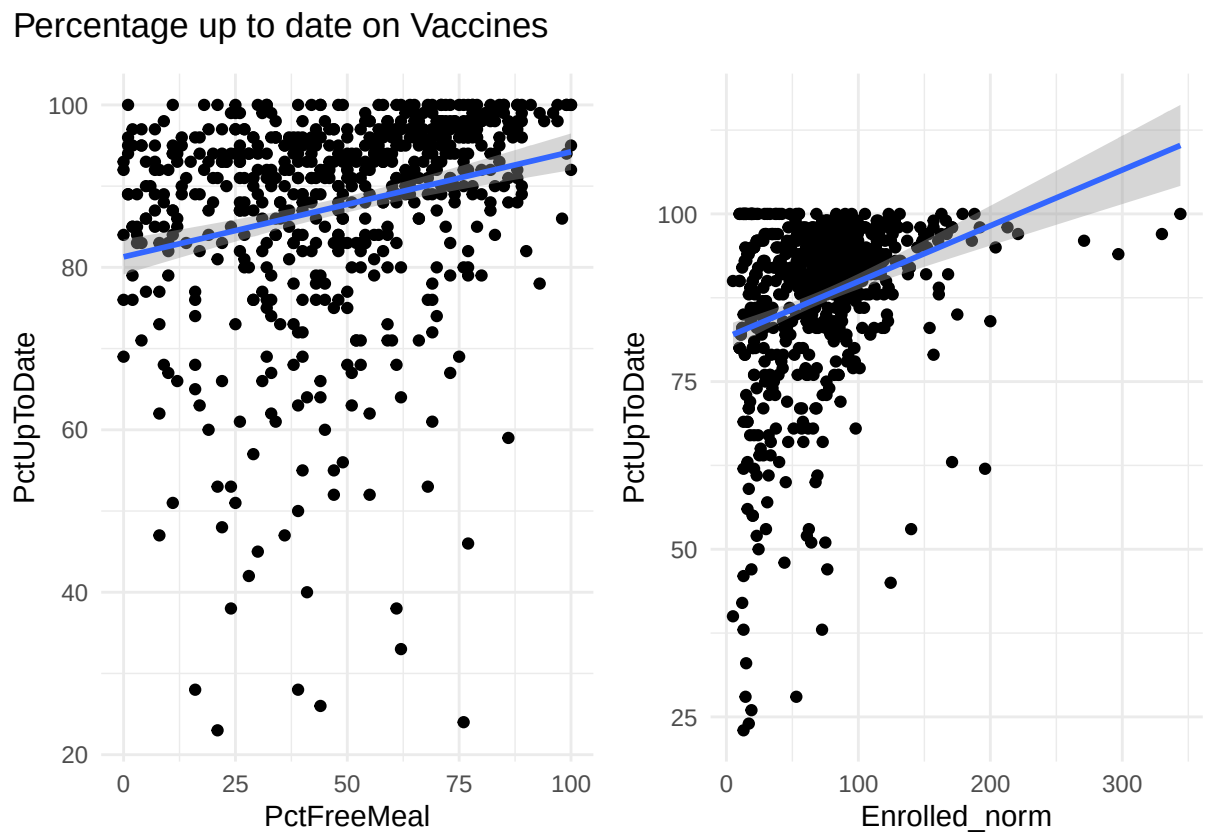
```
## `geom_smooth()` using formula 'y ~ x'
```



The figure shows the change in percent Belief exemptions with respect to Free Meals and Enrolled students. The percentage of belief exempt students go down with increase in percent Free Meal and it is significant , so we advice the state department to increase the funding for free meal programs.
The lower the enrolled students the higher the belief exemptions, so given the size of enrolled students we can guess this might belong to rural areas, we need to verify the school locations and concentrate on increasing the vaccine awareness.

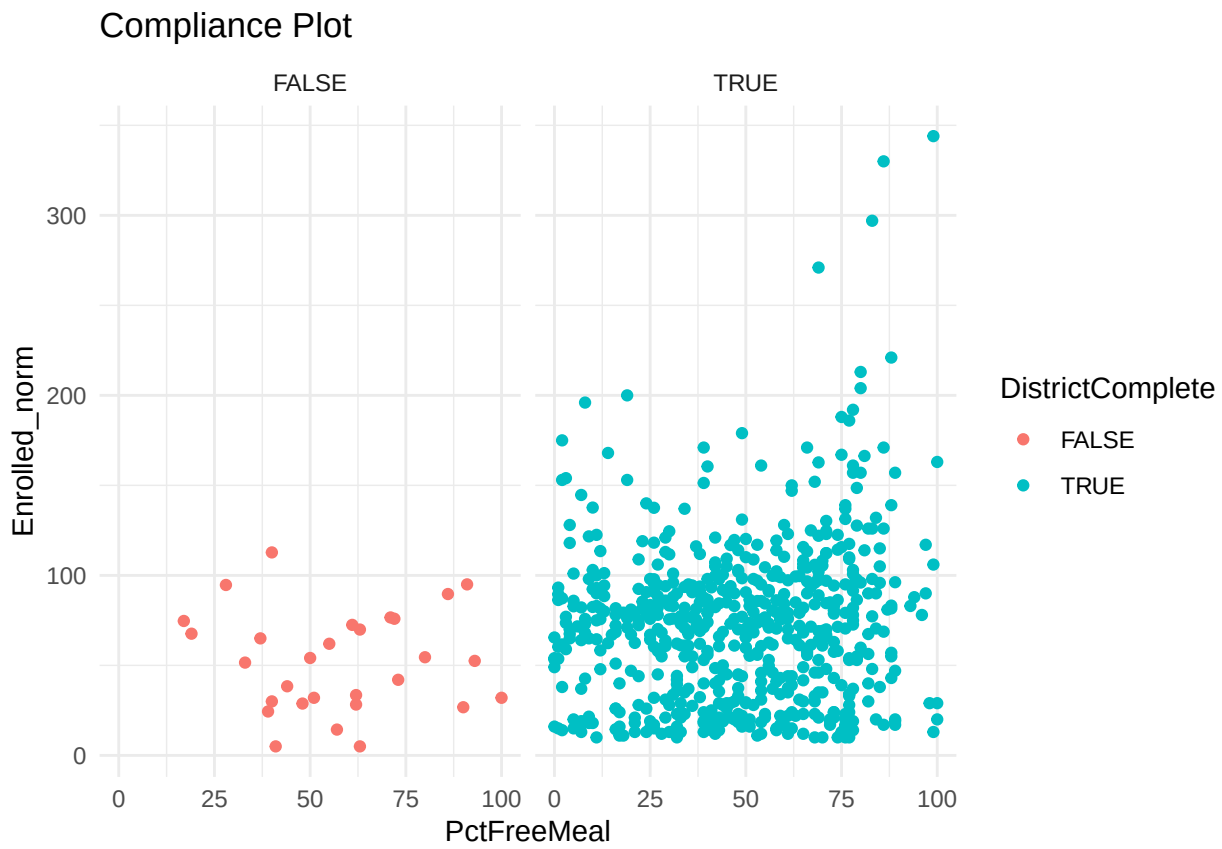
```
g3<-ggplot(outlier_removed,aes(y=PctUpToDate,x=PctFreeMeal)) +
  geom_point() + theme_minimal() +
  geom_smooth(method="glm")

g4<-ggplot(outlier_removed,aes(y=PctUpToDate,x=Enrolled_norm)) +
  geom_point() + theme_minimal() +
  geom_smooth(method="glm")

g3 + g4  +  plot_annotation(title = "Percentage up to date on Vaccines")
```

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



The figure shows the chnage in percent up to date on vaccines with respect to Free Meals and Enrolled students. The percentage up to date go up with increase in percent Free Meal and it is significant , so we advice the state department to increase the funding for free meal programs to improve the continued vaccination. The lower the enrolled students the lower the percent up to date, so given the size of enrolled students we can guess this might belong to rural areas, we need to verify the school locations and concentrate on increasing the vaccine awareness.

```
p<-ggplot(outlier_removed,
          aes(x=PctFreeMeal,y=Enrolled_norm,color=DistrictComplete)) +
  geom_point()  +theme_minimal()
p+facet_wrap(vars(DistrictComplete )) + ggtitle("Compliance Plot")
```



The figure shows the district complete on vaccination with respect to Free Meals and Enrolled students. The Districts complaint go up with increase in percent Free Meal and it is significant , so we advice the state department to increase the funding for free meal programs to improve the continued vaccination. The lower the enrolled students the lower the district complaint, so given the size of enrolled students we can guess this might belong to rural areas, we need to verify the school locations and concentrate on increasing the vaccine awareness. ** Conclusion ** From the analysis we can clearly see investing in free meals at small to medium schools will increase the overall compliance and increase the vaccine adoption both in terms of completeness and belief exemptions.