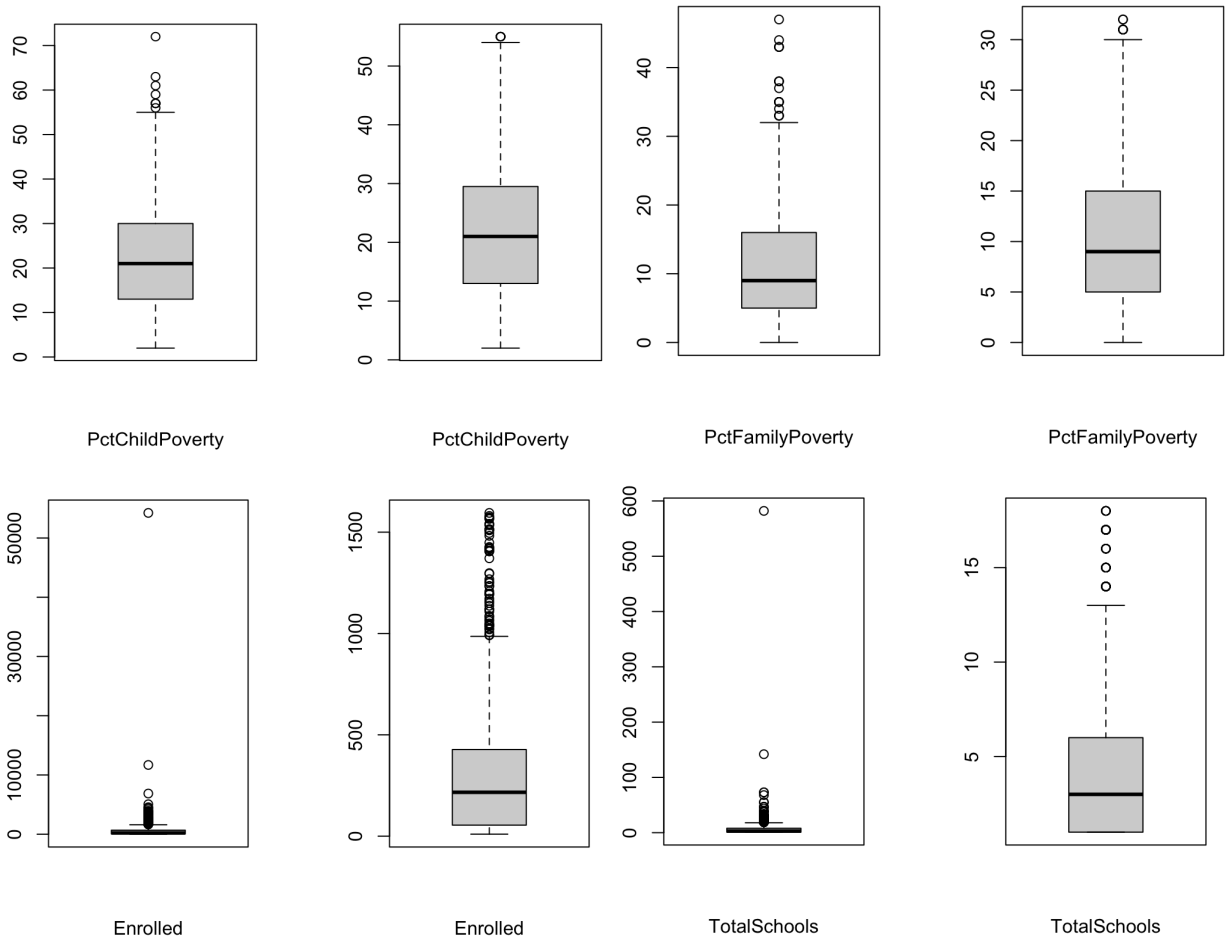
IST_772 Final

Suihin Wong(Henry)

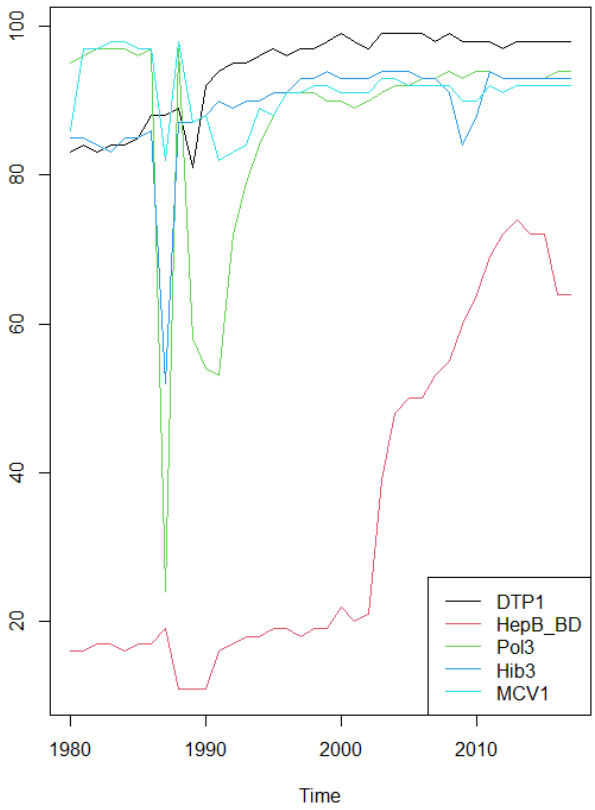
Data cleaning

Replace outlier with median to better represent the data. Below graph shows after outliers have been replaced with median. After replacing the outliers, it still shows some outliers in the boxplot. However, it is not outliers, we have non-normal distribution data at this point.



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?

As we can see from the graph, there was a drop in 1988. With the mean change point analysis, DTP1 change point occurred in 1990, Hib3 change point occurred in 1988, MCV1 change point occurred in 1987, Pol3 change point occurred in 1995 and HepB_BD change point occurred in 2004. All of the change points occurred when the vaccine rate increased as we can see from the graph below. DTP1 had the highest rate at the conclusion in 2017 and HepB_BD had the lowest rate at the conclusion in 2017. Also the HepB_BD vaccine had the greatest volatility, because it started from 11 in 1980 to 63 in 2017.



```
> cpt.mean(usVaccines[,"DTP1"])
Class 'cpt' : Changepoint Object
       ~~   : S4 class containing 12 slots with name
              cpttype date version data.set method t
m.est

Created on  : Tue May 25 17:44:21 2021

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in mean
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 10.91276
Minimum Segment Length : 1
Maximum no. of cpts   : 1
Changepoint Locations : 10
```

```
> cpt.mean(usVaccines[,"Hib3"])
Class 'cpt' : Changepoint Object
       ~~   : S4 class containing 12 slots with names
              cpttype date version data.set method test.
m.est

Created on  : Tue May 25 17:44:21 2021

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in mean
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 10.91276
Minimum Segment Length : 1
Maximum no. of cpts   : 1
Changepoint Locations : 8
```

```
> cpt.mean(usVaccines[,"Pol3"])
Class 'cpt' : Changepoint Object
       ~~   : S4 class containing 12 slots with names
              cpttype date version data.set method test.
m.est

Created on  : Tue May 25 17:44:21 2021

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in mean
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 10.91276
Minimum Segment Length : 1
Maximum no. of cpts   : 1
Changepoint Locations : 15
```

```
> cpt.mean(usVaccines[,"HepB_BD"])
Class 'cpt' : Changepoint Object
       ~~   : S4 class containing 12 slots with names
              cpttype date version data.set method test
m.est

Created on  : Tue May 25 17:44:21 2021

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in mean
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 10.91276
Minimum Segment Length : 1
Maximum no. of cpts   : 1
Changepoint Locations : 24
```

```
> cpt.mean(usVaccines[,"MCV1"])
Class 'cpt' : Changepoint Object
       ~~   : S4 class containing 12 slots with names
              cpttype date version data.set method tes
m.est

Created on  : Tue May 25 17:44:21 2021

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in mean
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 10.91276
Minimum Segment Length : 1
Maximum no. of cpts   : 1
Changepoint Locations : 7
```

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?

I used the Chi Square Test to compare the two categorical variables to see if there's a significant difference between the two variables. There are 5584 public schools reported vaccination data. There are 1397 private schools reported vaccination data. From the Chi Square Test, the p-value is less than 0.001, we can reject the null hypothesis. There is statistically significant between public and private schools in overall reporting proportion.

```
> table(schoolData[2:3])
         reported
pubpriv      Y
  PRIVATE 1397
  PUBLIC  5584
> chisq.test(table(schoolData[2:3]))

        Chi-squared test for given probabilities

data:  table(schoolData[2:3])
X-squared = 2511.2, df = 1, p-value < 2.2e-16
```

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)?

In 2013, there was an average 10.09% of students without the DTP vaccine, an average 9.689% of students without the Polio vaccine, an average 10.08% of students without MMR vaccine, and an average 7.614% of students without the HepB vaccine.

When comparing individual vaccines in California districts with overall US vaccination rates, we can see that DTP has 98% vaccination rates in the US and 89.91% in the California school districts. Polio has 93% vaccination rates in the US and 90.311% in the California school districts HepB has 74% vaccination rates in the US and 92.386% in the California school districts. MMR has 92% vaccination rates in the US and about 89.92% in the California school districts.

```
> usv_2013
Time Series:
Start = 2013
End = 2013
Frequency = 1
     DTP1 HepB_BD Pol3 Hib3 MCV1
2013   98      74   93   93   92
> summary(districts2[2:5])
   WithoutDTP     WithoutPolio      WithoutMMR     WithoutHepB
 Min.   : 0.00   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
 1st Qu.: 3.00   1st Qu.: 3.000   1st Qu.: 3.00   1st Qu.: 2.000
 Median : 7.00   Median : 6.000   Median : 6.00   Median : 4.000
 Mean   :10.09   Mean   : 9.689   Mean   :10.08   Mean   : 7.614
 3rd Qu.:13.25   3rd Qu.:13.000   3rd Qu.:13.25   3rd Qu.:10.000
 Max.   :66.00   Max.   :66.000   Max.   :68.00   Max.   :69.000
```

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?

From the correlation matrix, it shows that all of the individual vaccines have strong positive relationships with each other. If students are missing one vaccine, they have a high chance of missing all of the others vaccines as well.

```
> cor(districts[2:5])
             WithoutDTP WithoutPolio WithoutMMR WithoutHepB
WithoutDTP    1.0000000    0.9825210  0.9748848   0.8850896
WithoutPolio  0.9825210    1.0000000  0.9652914   0.8975436
WithoutMMR    0.9748848    0.9652914  1.0000000   0.8844034
WithoutHepB   0.8850896    0.8975436  0.8844034   1.0000000
```

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

In the base model, PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled these four variables are not statistically significant. We failed to reject the null hypothesis that the coefficients of the variables are different from zero. PctChildPoverty has a p-value of 0.89558 larger than the alpha level 0.05 with z value -0.131. PctFreeMeal has a p-value of 0.20154 larger than the alpha level 0.05 with z value -1.277. PctFamilyPoverty has a p-value of 0.9549 larger than the alpha level 0.05 with a z value of -0.057 and Enrolled has a p-value of 0.05045 larger than the alpha level 0.05 with a z value of 1.956. I want to investigate the Enrolled variable since the p-value is very close to the alpha level.

After removing the three not significant variables, we can bring down the AIC from 313.31 to 312.23 and the null deviance 317.75 on 699 degrees of freedom reduces to residual deviance 306.23 on 697 degrees of freedom. When I convert the log odds to odds, TotalSchools have 0.85:1change in the odds of the district's reporting being completed and the Enrolled variable is very close to 1. Let's look at the confidence interval, it is from 1.0000386 to 1.0022972 which is very close to 1 but does not overlap with 1 and it supports our result in the second model hypothesis test. In the chi-square test, it shows that the model with Enrolled and TotalSchools are significant with p-value less than the alpha level 0.05. This result also supports the frequentist test. With the Bayesian method in logistic regression model, from the HDI, we can see that 95% of the time HDI will include the true value for Enrolled and TotalSchools variables.

```
Call:
glm(formula = DistrictComplete ~ PctChildPoverty + PctFreeMeal +
    PctFamilyPoverty + Enrolled + TotalSchools, family = binomial(),
    data = districts2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6403   0.2635   0.3187   0.3683   1.0463

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.9184508  0.4860925   8.061 7.56e-16 ***
PctChildPoverty  -0.0034430  0.0262336  -0.131  0.89558
PctFreeMeal      -0.0134927  0.0105646  -1.277  0.20154
PctFamilyPoverty -0.0022258  0.0393557  -0.057  0.95490
Enrolled          0.0011064  0.0005656   1.956  0.05045 .
TotalSchools     -0.1494560  0.0411351  -3.633  0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 317.75  on 699  degrees of freedom
Residual deviance: 301.31  on 694  degrees of freedom
AIC: 313.31

Number of Fisher Scoring iterations: 6
```

```
> predict_glm <- glm(formula = DistrictComplete~TotalSchools+Enrolled, data=districts2, family=binomial())
> summary(predict_glm)

Call:
glm(formula = DistrictComplete ~ TotalSchools + Enrolled, family = binomial(),
    data = districts2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4573   0.3085   0.3159   0.3421   0.9345

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.1096486  0.2421389  12.842  < 2e-16 ***
TotalSchools -0.1530418  0.0411280  -3.721 0.000198 ***
Enrolled      0.0011406  0.0005699   2.002 0.045338 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 317.75  on 699  degrees of freedom
Residual deviance: 306.23  on 697  degrees of freedom
AIC: 312.23

Number of Fisher Scoring iterations: 5
```

```
> bayesLogit <- MCMClogit(formula = DistrictComplete~Enrolled+TotalSchools, data=districts2)
> summary(bayesLogit)

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)   3.124806 0.2448204 2.448e-03      7.994e-03
Enrolled      0.001202 0.0005707 5.707e-06      1.884e-05
TotalSchools -0.154674 0.0411728 4.117e-04      1.337e-03

2. Quantiles for each variable:

                  2.5%        25%       50%       75%     97.5%
(Intercept)   2.667e+00  2.9573293  3.120870  3.288279  3.623503
Enrolled      8.531e-05  0.0007951  0.001207  0.001578  0.002348
TotalSchools -2.335e-01 -0.1828461 -0.155697 -0.128043 -0.072879

> exp(coef(predict_glm))
 (Intercept) TotalSchools     Enrolled
  22.4131661    0.8580938    1.0011413
> exp(confint(predict_glm))
Waiting for profiling to be done...
                   2.5 %     97.5 %
(Intercept)  14.2593946 36.9600495
TotalSchools  0.7931737  0.9344966
Enrolled      1.0000386  1.0022972
> PseudoR2(predict_glm)
       McFadden     Adj.McFadden        Cox.Snell       Nagelkerke McKelvey.Zavoina           Effron
     0.03627334       0.01109667       0.01633092       0.04475745       0.04500150       0.00580834
          Count        Adj.Count              AIC    Corrected.AIC
             NA               NA     312.22851340     312.26299616


> anova(predict_glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: DistrictComplete

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                          699     317.75
TotalSchools  1    7.410       698     310.34 0.006486 **
Enrolled      1    4.116       697     306.23 0.042478 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

With the base model, it shows that PctChildPoverty, PctFamilyPoverty and TotalSchools are not statistically significant in the hypothesis test. PctChildPovery has a p-value of 0.633 with t value of 0.477. PctFamilyPoverty has a p-value of 0.172742 with t value of 1.365 and TotalSchools has p-value of 0.833598 with t value of 0.210. After removing the three variables, the adjusted R-squared 0.1182 which is very close to the base model. The adjusted R-squared is low because the data is not linear. The new model has F-score of 47.85on2 and 697 DF with p-value of 2.2e-16 which is less than the alpha level. This model is statistically significant. The VIF for the model is less than 5 which shows there is no multicollinearity in this linear model. With the MCMC technique, we can see that both of the variables don't overlap with 0 in the HDI

which supports the frequentist approach earlier. In the Bayes factor analysis, it shows that the odds are in favor of the alternative hypothesis which means that the model contains PctFreeMeal and Enrolled better than the model that only contains the y-intercept.

```
Call:
lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
    Enrolled + TotalSchools, data = districts2)

Residuals:
    Min      1Q  Median      3Q     Max
-63.337  -3.945   2.623   7.231  19.952

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      78.915188   1.129728  69.853  < 2e-16 ***
PctChildPoverty   0.034216   0.071758   0.477 0.633643
PctFreeMeal       0.081555   0.027285   2.989 0.002898 **
PctFamilyPoverty  0.151395   0.110925   1.365 0.172742
Enrolled          0.007680   0.001983   3.872 0.000118 ***
TotalSchools      0.036781   0.175009   0.210 0.833598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.56 on 694 degrees of freedom
Multiple R-squared:  0.1265,    Adjusted R-squared:  0.1202
F-statistic:  20.1 on 5 and 694 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = PctUpToDate ~ Enrolled + PctFreeMeal, data = districts2)

Residuals:
    Min      1Q  Median      3Q     Max
-64.856  -3.870   2.560   7.265  20.418

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 79.339124   1.032134  76.869  < 2e-16 ***
Enrolled     0.007949   0.001183   6.720 3.77e-11 ***
PctFreeMeal  0.123443   0.017632   7.001 5.99e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.58 on 697 degrees of freedom
Multiple R-squared:  0.1207,    Adjusted R-squared:  0.1182
F-statistic: 47.85 on 2 and 697 DF,  p-value: < 2.2e-16
```

```
> vif(predict_date)
    Enrolled TotalSchools
    2.805717     2.805717
```

```
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          87.93968 0.443838 4.438e-03      4.438e-03
PctFreeMeal  0.12156 0.017596 1.760e-04      1.801e-04
Enrolled     0.00781 0.001187 1.187e-05      1.221e-05
sig2       134.44391 7.219747 7.220e-02      7.220e-02
g            0.24445 0.830141 8.301e-03      8.301e-03

2. Quantiles for each variable:

                2.5%       25%       50%       75%     97.5%
mu          8.706e+01 8.764e+01 8.794e+01 8.824e+01  88.81443
PctFreeMeal 8.698e-02 1.098e-01 1.216e-01 1.334e-01   0.15620
Enrolled    5.471e-03 7.037e-03 7.815e-03 8.595e-03   0.01014
sig2        1.210e+02 1.295e+02 1.342e+02 1.391e+02 149.17834
g           2.743e-02 6.313e-02 1.099e-01 2.133e-01   1.12517

> date_BF <- lmBF(PctUpToDate~PctFreeMeal + Enrolled, data=districts2)
> date_BF
Bayes factor analysis
--------------
[1] PctFreeMeal + Enrolled : 1.278616e+17 ±0.01%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

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

We used PctChildPoverty + PctFreeMeal + PctFamilyPoverty + Enrolled+TotalSchools to predict PctBeliefExempt in our base Exponential Regression model. Only PctFreeMeal and Enrolled variables are statistically significant after removing the outlier. In variance inflation factor (VIF), it shows that PctFreeMeal and Enrolled have a score of 1 which are less than 5. There is no multicollinearity in this model. Also, after removing PctChildPoverty, PctFamilyPoverty and TotalSchools variables, the regression model has 0.2557 adjusted R-squared and the base model has around 0.2551 adjusted R-squared which is very close, but the new model is slightly better. PctFreeMeal has p-value 2e-16 with t value of -14.810 and Enrolled has p-value 6.83e-06 with t value of -4.533, it is all statistically significant that the coefficient is different from zero. P-value 2.2e-16 is less than the alpha level (0.05) shows that the model is statistically significant with F-Statistic 121 on 2 and 697 DF. The model is statistically significant. With the MCMC technique, we can see that both of the variables don't overlap with 0 in the HDI which

supports the frequentist approach earlier. In the bayes factor analysis, the Bayes factor shows the odds are strongly in favor of the alternative hypothesis. The model containing the X variables as predictor is better than the model only containing the y variable.

```
Call:
lm(formula = log(PctBeliefExempt + 1) ~ PctChildPoverty + PctFreeMeal +
    PctFamilyPoverty + Enrolled + TotalSchools, data = districts2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40557 -0.63260 -0.01668  0.55295  2.84465

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.4001947  0.0873995  27.462  < 2e-16 ***
PctChildPoverty   0.0048311  0.0055515   0.870  0.38447
PctFreeMeal      -0.0192616  0.0021108  -9.125  < 2e-16 ***
PctFamilyPoverty -0.0128448  0.0085815  -1.497  0.13490
Enrolled         -0.0004644  0.0001534  -3.027  0.00256 **
TotalSchools      0.0073312  0.0135393   0.541  0.58835
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8947 on 694 degrees of freedom
Multiple R-squared:  0.2604,    Adjusted R-squared:  0.2551
F-statistic: 48.88 on 5 and 694 DF,  p-value: < 2.2e-16



Call:
lm(formula = log(PctBeliefExempt + 1) ~ PctFreeMeal + Enrolled,
    data = districts2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40008 -0.63385 -0.02495  0.55797  2.89888

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.426e+00  7.973e-02  30.434  < 2e-16 ***
PctFreeMeal -2.017e-02  1.362e-03 -14.810  < 2e-16 ***
Enrolled    -4.142e-04  9.137e-05  -4.533 6.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8944 on 697 degrees of freedom
Multiple R-squared:  0.2578,    Adjusted R-squared:  0.2557
F-statistic:   121 on 2 and 697 DF,  p-value: < 2.2e-16


> vif(predict_exempt)
PctFreeMeal    Enrolled
   1.000253    1.000253
```

```
> summary(exempt_MCMC)

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           5.530614 0.2985004 2.985e-03      2.985e-03
PctFreeMeal -0.102821 0.0118663 1.187e-04      1.187e-04
Enrolled    -0.004876 0.0007961 7.961e-06      7.809e-06
sig2        61.223508 3.3035915 3.304e-02      3.304e-02
g            0.284012 1.0079620 1.008e-02      1.237e-02

2. Quantiles for each variable:

                2.5%       25%       50%       75%     97.5%
mu           4.944316  5.328392  5.530837  5.733946  6.112701
PctFreeMeal -0.125887 -0.110915 -0.102824 -0.094914 -0.079461
Enrolled    -0.006458 -0.005421 -0.004871 -0.004333 -0.003305
sig2        55.064961 58.926569 61.116822 63.363186 67.937008
g            0.030740  0.072269  0.124171  0.243862  1.269738

> exempt_BF <- lmBF(PctBeliefExempt~ PctFreeMeal + Enrolled, data=df)
> exempt_BF
Bayes factor analysis
--------------
[1] PctFreeMeal + Enrolled : 1.194841e+21 ±0.01%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

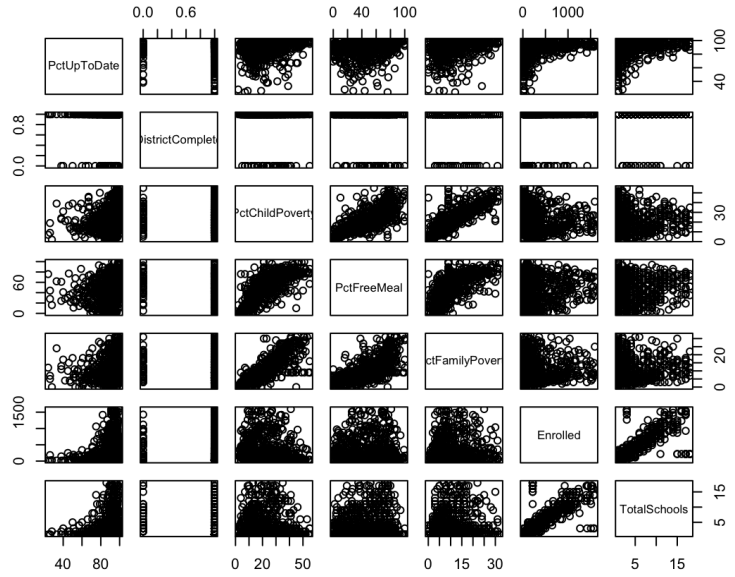
8. What's the big picture, based on all of the foregoing analyses? The staff member in the state legislator's office is interested to know how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance. What have you learned from the data and analyses that might inform this question?

In conclusion, as we can see from the correlation scatter plot, as the percentage of children in the district living below the poverty line increases, the percentage of families in the district living below the poverty line increases. Also, the percentage of children in the district eligible for free student meals increases because low income family children are usually eligible for free student meals. Total number of enrolled students in the district also has a positive linear relationship with the total number of different schools in the district. As the number of enrolled students increases in the district, the number of different schools in the district increases.

In a predictive analysis on improving vaccination rate, I used the number of enrolled students and total number of students eligible for free student meals in the district as a predictor because the two variables are statistically significant and there were trends in the scatter plot with the dependent variable. When the number of enrolled students and total number of students eligible for free student meals increases, the percentage of enrolled students with completely up

to date vaccines increases. For recommendation, if the government can spend more budgets to have more students enroll in the district and have more students eligible for free student meals, it would help with improving the vaccination rate.

In a predictive analysis on improving reporting compliance, I used the enrolled student and number of different schools in the district as a predictor because it showed statistically significant in the regression model. The log odds for TotalSchools is -0.153 and Enrolled is 0.00114. After converting from log odds to odds, a 0.858 : 1 change in odds when total number of schools increase 1 and 1.0014:1 in odds when total number of enrolled students in the district increase by 1. For recommendation, districts with more different schools have less reporting compliance and it could be due to lack of employment. The government can increase budget in district that have a high number of schools.



Appendix R Code

```
# data clearing
districts2 <- districts
summary(districts2)

# check outliers, replace outliers with median.
boxplot(districts2$PctChildPoverty,xlab = "PctChildPoverty")
outliers <- boxplot(districts2$PctChildPoverty, plot=FALSE)$out
districts2$PctChildPoverty[which(districts2$PctChildPoverty %in% outliers)] <-
median(districts2$PctChildPoverty)
boxplot(districts2$PctFreeMeal,xlab = "PctFreeMeal")

boxplot(districts2$PctFamilyPoverty,xlab = "PctFamilyPoverty")

outliers <- boxplot(districts2$PctFamilyPoverty, plot=FALSE)$out

districts2$PctFamilyPoverty[which(districts2$PctFamilyPoverty %in% outliers)] <-
median(districts2$PctFamilyPoverty)

boxplot(districts2$Enrolled,xlab = "Enrolled")

outliers <- boxplot(districts2$Enrolled, plot=FALSE)$out

districts2$Enrolled[which(districts2$Enrolled %in% outliers)] <- median(districts2$Enrolled)

boxplot(districts2$TotalSchools,xlab = "TotalSchools")

outliers <- boxplot(districts2$TotalSchools, plot=FALSE)$out

districts2$TotalSchools[which(districts2$TotalSchools %in% outliers)] <-
median(districts2$TotalSchools)
install.packages("nlme")
install.packages("car")
install.packages("ez")
install.packages("tseries")
install.packages("changepoint")
install.packages("BaylorEdPsych")
install.packages("MCMCpack")
install.packages("BayesFactor")
install.packages("bayesanova")
install.packages("mlr")
install.packages("drc")
install.packages("caret")
library("caret")
library("nlme")
library("car")
library("ez")
library("tseries")
library("changepoint")
library("BaylorEdPsych")
```

```
library("MCMCpack")
library("BayesFactor")
library("bayesanova")
library("mlr")
library("drc")

#1
plot.ts(usVaccines,frequency = 1)
ts.plot(usVaccines, col= 1:5)
legend("bottomright", colnames(usVaccines), col=1:5,lty=1)

view(usVaccines)

decUsVac <- decompose(ts(usVaccines[,"Hib3"],frequency = 1))

# change point analysis
plot(decUsVac)
cpt.var(diff(usVaccines[,"DTP1"]))
plot(cpt.var(diff(usVaccines[,"DTP1"])))
cpt.var(usVaccines[,"HepB_BD"])
plot(diff(usVaccines[,"HepB_BD"]))
cpt.var(diff(usVaccines[,"Pol3"]))
plot(cpt.var(diff(usVaccines[,"Pol3"])))
cpt.var(diff(usVaccines[,"Hib3"]))
plot(cpt.var(diff(usVaccines[,"Hib3"])))
cpt.var(diff(usVaccines[,"MCV1"]))
plot(cpt.var(diff(usVaccines[,"MCV1"])))
cpt.mean(usVaccines[,"DTP1"])
cpt.mean(usVaccines[,"HepB_BD"])
cpt.mean(usVaccines[,"Pol3"])
cpt.mean(usVaccines[,"Hib3"])
cpt.mean(usVaccines[,"MCV1"])

summary(usVaccines)

#2
nrow(allSchoolsReportStatus[which(allSchoolsReportStatus$pubpriv=="PUBLIC" &
allSchoolsReportStatus$reported == "Y"),])
nrow(allSchoolsReportStatus[which(allSchoolsReportStatus$pubpriv=="PRIVATE" &
allSchoolsReportStatus$reported == "Y"),])
schoolData <- allSchoolsReportStatus[which(allSchoolsReportStatus$reported == "Y"),]
schoolData$pubpriv <- as.factor(schoolData$pubpriv)
pubschool <- schoolData[schoolData$pubpriv == "PUBLIC",]
prischool <- schoolData[schoolData$pubpriv == "PRIVATE",]

chisq.test(table(schoolData[2:3]))
chisq.test(table(allSchoolsReportStatus[2:3]))
table(schoolData[2:3])
```

```
#3
summary(usVaccines)
head(usVaccines)
usv_2013 <- window(usVaccines, start =2013,end= 2013)
usv_2013
summary(districts2[2:5])

#4
#correlation matrix
cor(districts[2:5])

#5
# logistics regression model
colnames(districts2)
predict_glm <- glm(formula =
DistrictComplete~PctChildPoverty+PctFreeMeal+PctFamilyPoverty+TotalSchools+Enrolled,
data=districts2, family=binomial())
predict_glm <- glm(formula = DistrictComplete~TotalSchools+Enrolled, data=districts2,
family=binomial())
summary(predict_glm)
exp(coef(predict_glm))
exp(confint(predict_glm))
PseudoR2(predict_glm)
bayesLogit <- MCMClogit(formula = DistrictComplete~Enrolled+TotalSchools, data=districts2)
summary(bayesLogit)
anova(predict_glm, test="Chisq")
a <- predict(predict_glm,type="response")
table(districts2$DistrictComplete, a>0.5)
table(round(predict(predict_glm,type = "response")), districts2$DistrictComplete)

#6
col_name <-
c("PctUpToDate","PctChildPoverty","PctFreeMeal","PctFamilyPoverty","Enrolled","TotalSchools"
)
df <- districts2[col_name]
pairs(df)
predict_date <- lm(log(PctUpToDate)~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty +
Enrolled+TotalSchools, data=districts2)
predict_date <- lm(PctUpToDate~ PctFreeMeal +Enrolled, data=districts2)
summary(predict_date)
cor(districts2[9:13])
vif(predict_date)
date_MCMC <- lmBF(PctUpToDate~ PctFreeMeal + Enrolled, data=districts2,posterior =
TRUE,iterations=10000)
summary(date_MCMC)
date_BF <- lmBF(PctUpToDate~PctFreeMeal + Enrolled, data=districts2)
date_BF
```

```
#7
col_name <-
c("PctBeliefExempt","PctChildPoverty","PctFreeMeal","PctFamilyPoverty","Enrolled","TotalSchools")
df <- districts2[col_name]
pairs(df)
df$PctBeliefExempt <- log(df$PctBeliefExempt+1)
log(districts2$PctChildPoverty)
predict_exempt <- lm(log(PctBeliefExempt+1)~ PctFreeMeal + Enrolled, data=districts2)
summary(predict_exempt)
cor(districts[9:13])
vif(predict_exempt)
exempt_MCMC <- lmBF(PctBeliefExempt~ PctFreeMeal +  Enrolled, data=districts2,posterior = TRUE,iterations=10000)
summary(exempt_MCMC)
exempt_BF <- lmBF(PctBeliefExempt~ PctFreeMeal + Enrolled, data=df)
exempt_BF
```