IST772 Week 11: Vaccination Rates Analysis In the ever-evolving world of data analytics, ethical considerations play a pivotal role in shaping responsible and trustworthy practices. As we delve into the importance of ethics in data science, it becomes evident that ethical principles guide our actions when we create applications and interpret data.

Taking the context of our final project in IST-718 Big Data into account, our team embarked on a comprehensive assessment of credit risk, a domain where ethical considerations are of paramount importance. Our project, titled "Optimizing Credit Risk Assessment: Leveraging Customer Segmentation, Fraud Detection, and Default Risk Analysis," provides a unique opportunity to apply ethical considerations across various facets:

1. **Data Privacy and Security:** Given the sensitive nature of financial data, ensuring robust data privacy and security measures is an ethical imperative. We must safeguard customer information through anonymization and strong security protocols.
2. **Fairness in Credit Decisions:** Ethical concerns arise when algorithms exhibit bias in credit decisions. It's our ethical responsibility to eliminate bias and ensure fairness in approving or denying credit.
3. **Transparency and Explainability:** Transparency in our models is vital. Customers deserve clear explanations for credit decisions, and our ethical duty is to provide that transparency.
4. **Fraud Detection and Prevention:** Ethical fraud detection is about protecting both lenders and customers. Minimizing false positives in fraud detection is essential to avoid harming innocent customers.
5. **Customer Consent and Data Usage:** Ethical data usage requires obtaining informed consent and allowing customers to opt out if they choose not to participate.
6. **Responsible Use of Customer Segmentation:** Ethical customer segmentation ensures that it's used to understand behavior, not to discriminate against specific groups.
7. **Monitoring and Auditing:** Regular monitoring and auditing of models are ethical practices that help rectify biases and maintain fairness over time.
8. **Compliance with Regulatory Frameworks:** Adhering to regulations and industry standards is an ethical obligation to protect both lenders and customers.

In conclusion, the importance of ethics in data science cannot be overstated. In the context of our project, these ethical considerations are not just theoretical concepts but practical principles that guide our actions, ensuring that we build reliable, fair, and responsible systems for credit risk assessment.

Student: Sintia Stabel

Instructions: You have received three data sets, which are described below. Each dataset is in RData format, which means that you can simply open it with the open dialog on the Environment tab and it will read in as an R object. This will save time and effort in preparing your data.

Your goal for this final exam is to conduct the necessary analyses and then write up a technical report for a scientifically knowledgeable staff member in a state legislator's office. Thus, you should provide sufficient numeric and graphical detail that the staff member can create a comprehensive briefing for a legislator. You can assume that the staff member understands the concept of statistical significance. Your report should include a few graphics created by R, keeping in mind that you must provide some accompanying text to explain each graphic that you include in your report.

This exam is open book and open notes, but you may not receive assistance, help, coaching, guidance, or support from any human except your section instructor. Your section instructor will be available by email

throughout the exam period: If you are stuck on an R code problem, make sure to include your complete code in the email, preferably as a file attachment.

These three data sets all pertain to vaccinations. The first and second datasets are the same for everyone and are mainly included to provide context for interpretation of the results. Most of the substantive analyses occur in reference to the third dataset. This third dataset is different for every student and results will vary depending upon the sample the student received.

The datasets are:

- usVaccines.Rdata
- allSchoolsReportStatus.RData
- districtsX.RData

Here is a description of each dataset:

usVaccines.Rdata – 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; HepB_BD = Hepatitis B, Birth Dose; Pol3 = Polio third dose; Hib3 – Influenza third dose; MCV1 = Measles first dose)

allSchoolsReportStatus.RData – A list of California kindergartens and whether they reported vaccination data to the state in 2013

'data.frame': 7381 obs. of 3 variables:
$ name : Name of the school
$ pubpriv : "PUBLIC" or "PRIVATE"
$ reported: "Y" or "N"

districtsX.RData – (Where X is the number of your particular dataset) A sample of California public school districts from the 2013 data collection, along with specific numbers and percentages for each district:

'data.frame': 700 obs. of 13 variables:
$ DistrictName : Name of the district
$ WithoutDTP : Percentage of students without the DTP vaccine
$ WithoutPolio : Percentage of students without the Polio vaccine
$ WithoutMMR : Percentage of students without the MMR vaccine
$ WithoutHepB : Percentage of students without the Hepatitis B vaccine
$ PctUpToDate : Percentage of all enrolled students with completely up-to-date vaccines
$ DistrictComplete: Boolean indicating whether or not the district's reporting was complete
$ PctBeliefExempt : Percentage of all enrolled students with belief exceptions
$ PctChildPoverty : Percentage of children in the district living below the poverty line
$ PctFreeMeal : Percentage of children in the district eligible for free student meals

$ PctFamilyPoverty: num Percentage of families in the district living below the poverty line
$ Enrolled : Total number of enrolled students in the district
$ TotalSchools : Total number of different schools in the district

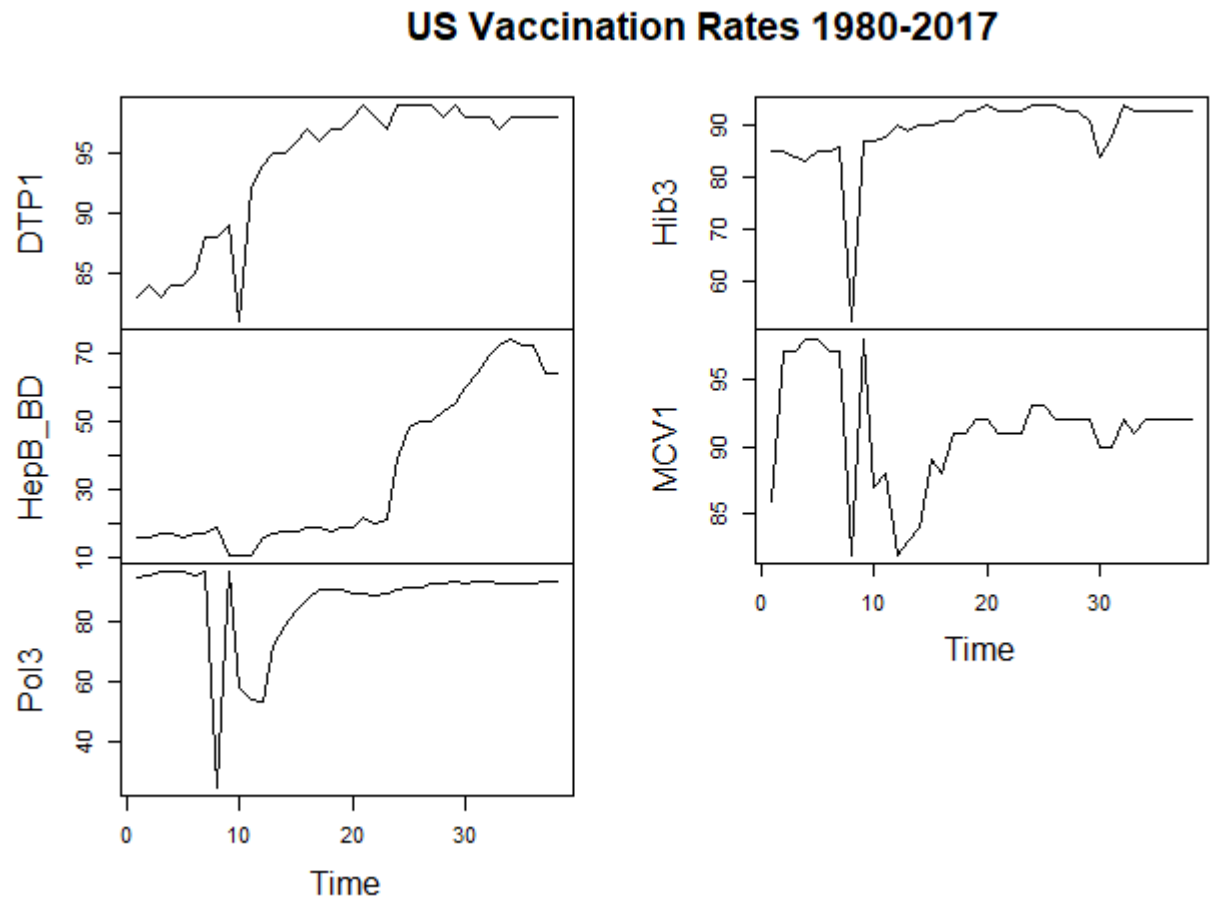
The research questions for you to explore with these three data sets are as follows:

Introductory/Descriptive Reports:

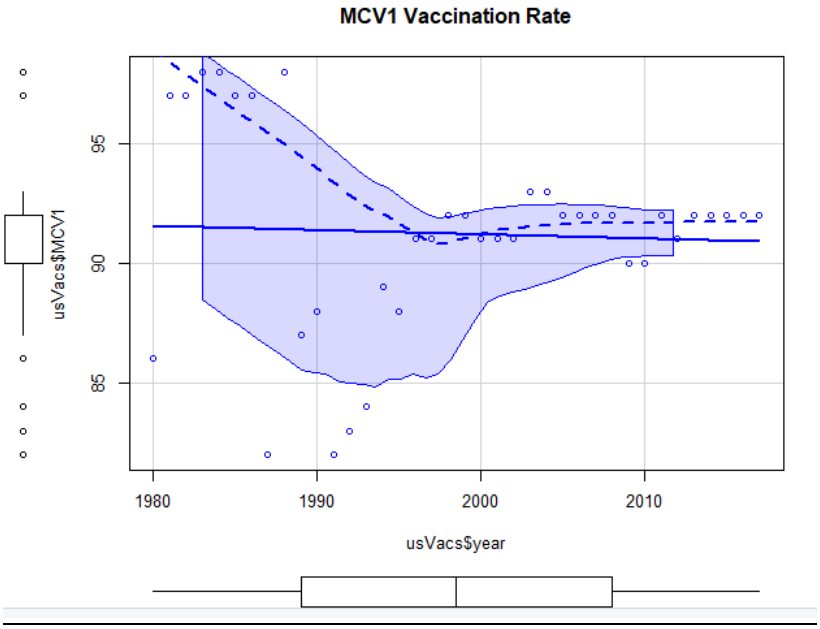
1. How have U.S. vaccination rates varied over time? Are vaccination rates increasing or decreasing? Which vaccination has the highest rate at the conclusion of the time series? Which vaccination has the lowest rate at the conclusion of the time series? Which vaccine has the greatest volatility?

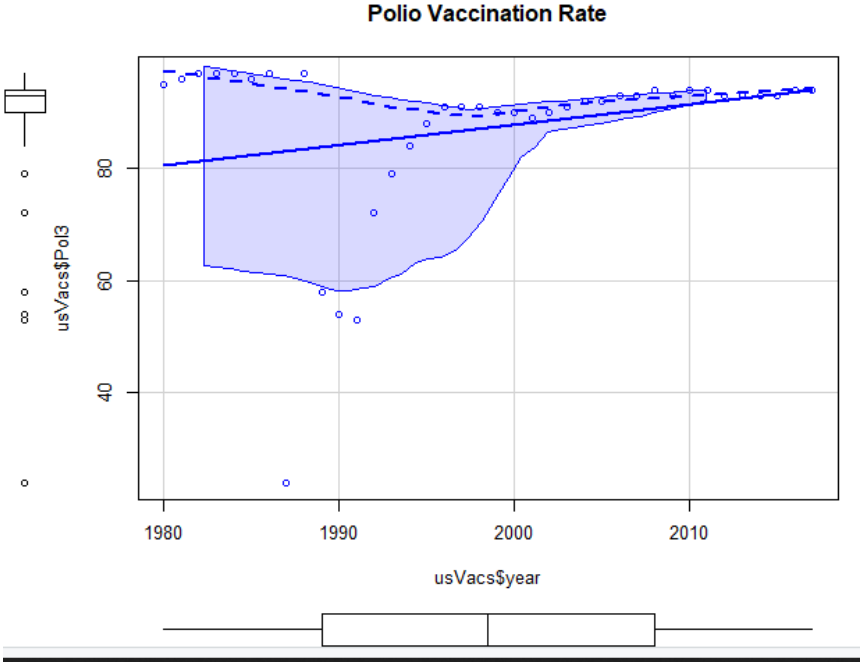
According to US vaccination rates reported from the World Health Organization for the years 1980 through 2017, there has been a clear upwards trend in rates for HepB_DB and DTP1 rates during the second third portion of the time series. Pol3, Hib3 and MCV have remained mostly unchanged for about the last two thirds of the same period. Looking at the side-by-side time series line charts, we also observe a significant drop in all vaccinations around 1987 through early 1990s. Scatterplots of POl3 and MCV1 also confirm the same observations.

Our time series plot also show that DTP1 has the highest rate at the end of the series, while HepB_BD has the lowest. The Summary function confirms that the DTP1 has the top rate and that HepB_BD has the lowest rate at the third and first quantiles respectively. When we remove the trend component of the time series and plot the diff function, we see that Hib3 is the vaccine with the widest variance, and therefore has the highest volatility.

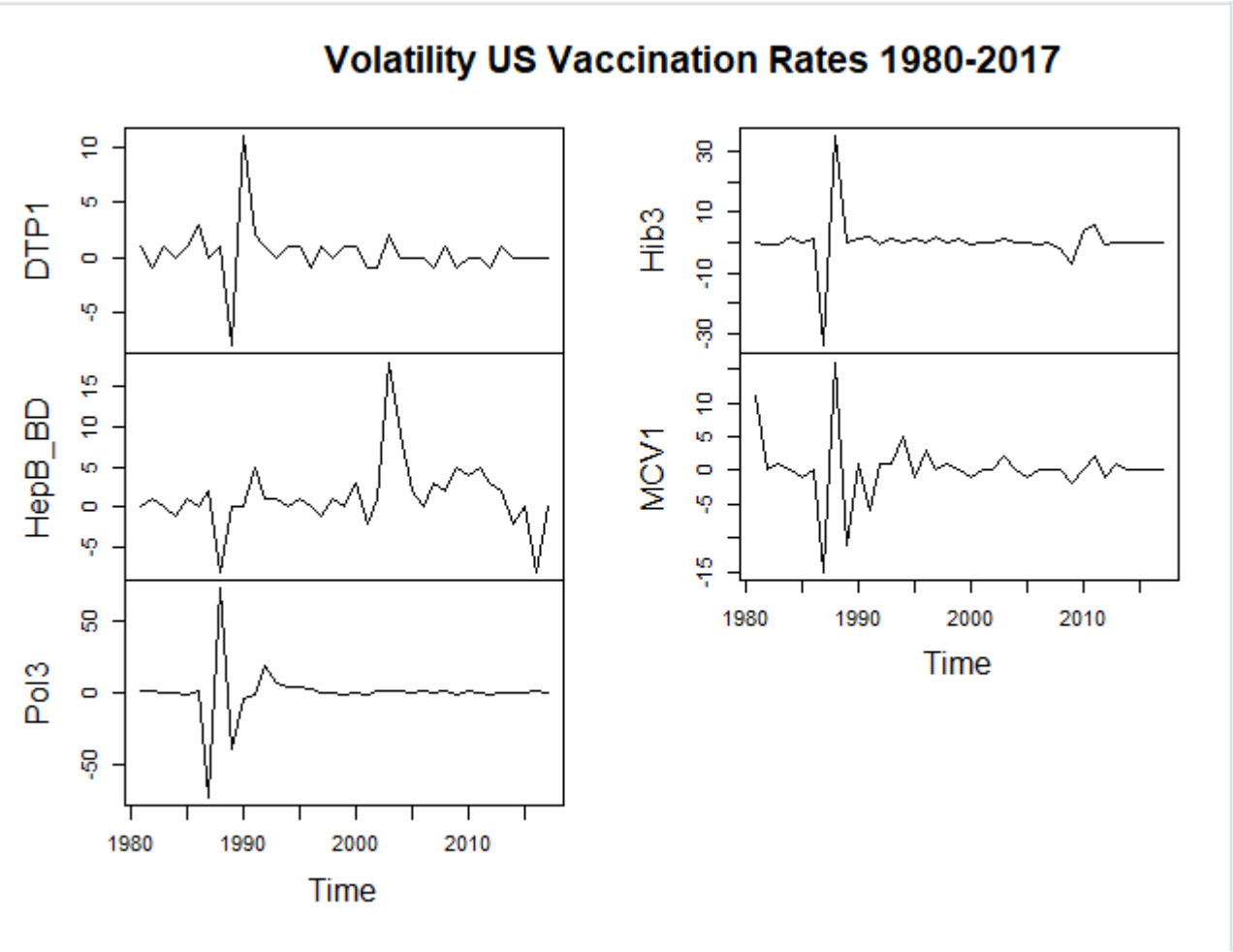


Time Series Scatterplot of MCV1 and Polio vaccination rates reported by WHO, showing sharp drops in vaccination rates around year 1987 and early 1990s:





Volatility rates for all reported vaccines using Diff function to remove the impact of the inherit trend from the time series:



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?

By sub-setting the data set, we can calculate the proportion of public and private schools that reported their vaccination data. Out of 6981 total reporting schools, about 80%, or 5584 public schools reported their vaccinations. As for the total private schools that reported, the total comes to 1397 or about 20%.

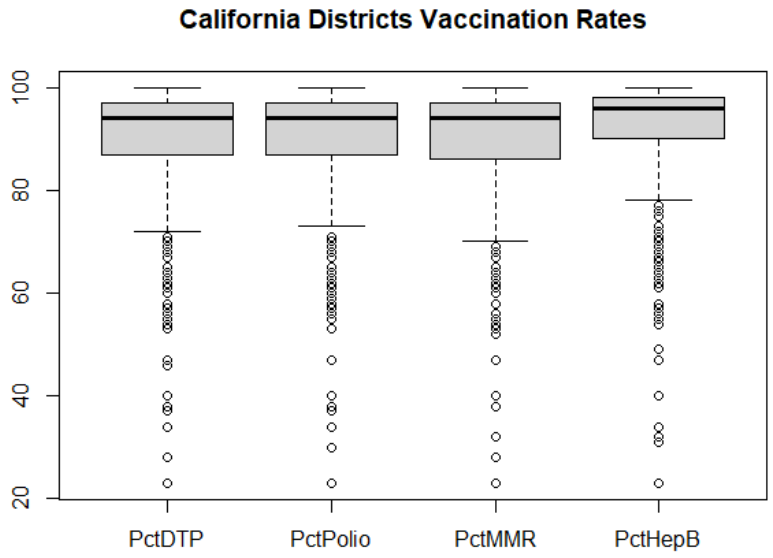
To determine if the difference in overall performance is credible, a frequentist t-test was performed to verify if the difference in means between the groups is not equal to zero.
The results of the t-test produced a statistically significant p-value of 2.2e-16, which is well below the standard alpha of 0.05. This also means that we can reject the null hypothesis in favor of the alternative hypothesis and say that the difference between the means of two groups is NOT equal to zero. As a result, our believe that there is a credible difference in the overall reporting is supported, because if we were to repeat this test over the long run, 95 times out of 100, the difference between the two means of the public schools that reported, and the private schools that reported, would very likely not be equal to zero. ** Note we were unable to run the Bayes BESTmcmc test due issues with R.

```
t = 13.944, df = 1835.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1091376 0.1448623
sample estimates:
mean of x mean of y
0.9741800 0.8471801
```

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)?

The 2013 vaccination rates for individual vaccines for California's districts are displayed in the box plot below. We can see that they range from low 20s to all the way up to 100% across all reported districts. To compare California districts' rates against the World Health Organization's (WHO) US rates for the same year, the mean rates for all 4 vaccines were calculated. California districts' average rate for DTP1 is 89.88% compared to 98% from reported by the WHO. California's average rate for HebB is 92.36% vs 74% for the WHO. California's Polio vaccine rate is 90.31% compared to 93%, and lastly, California's average rate for the MMR vaccine is 89.88% compared to 92% reported from the WHO.

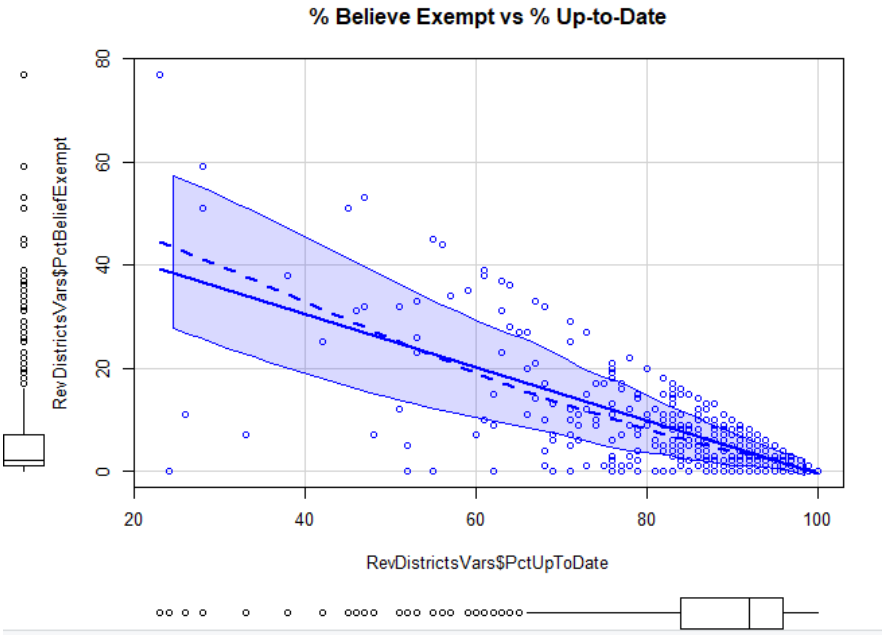
California districts' average percentage rates are overall slightly lower than the WHO's, except for HebB which is about 22% higher. During the last year of the time series stated by WHO, the following rates were reported: DPT1 98%, HebB 64%, Polio 94%, and 92% for MMR. These results are at about the same levels reported for the year 2013, except for HepB, which increased by 10%, from 64% to 74%.



```
> usVacs_2013 # results of  WHO US rates
2013
  year DTP1 HepB_BD Pol3 Hib3 MCV1
1 2013   98      74   93   93   92
> usVacs_2017 <- usVacs[38,]
> usVacs_2017
   year DTP1 HepB_BD Pol3 Hib3 MCV1
38 2017   98      64   94   93   92
> mean(vacc2013$PctDTP) # mean DTP
[1] 89.88429
> mean(vacc2013$PctHepB)# mean HepB
[1] 92.35571
> mean(vacc2013$PctPolio)# mean Polio
[1] 90.30857
> mean(vacc2013$PctMMR) # mean MMR
[1] 89.87714
```

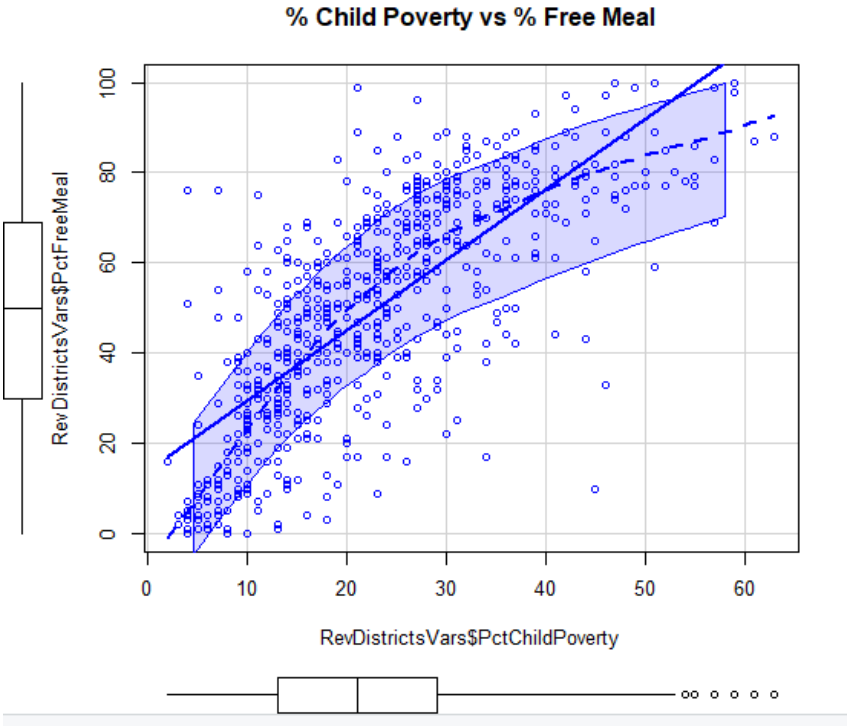
4. Among districts, how are the vaccination rates for individual vaccines related? In other words, if students are missing one vaccine are they missing all of the others?

By running a correlation test for the districts that have completed their vaccination reports, while also sub-setting the data by the relevant numeric variables, we can determine how each of these variables are related to one another. Looking at the PctBelieveExempt variable, we can tell that the percentage of students not vaccinated due to "believes" has a strong inverse correlation of -0.74 to the PctUpToDate variable. This is to be expected, as we can infer that these students have not received vaccinations across the board, and therefore move in the opposite direction of the up-to-date vaccination rates reported in the PctUpToDate variable. This inverse relationship can be observed in the scatterplot below, as the higher the believe exempt rates are, the lower the up-to-date vaccination rates are:

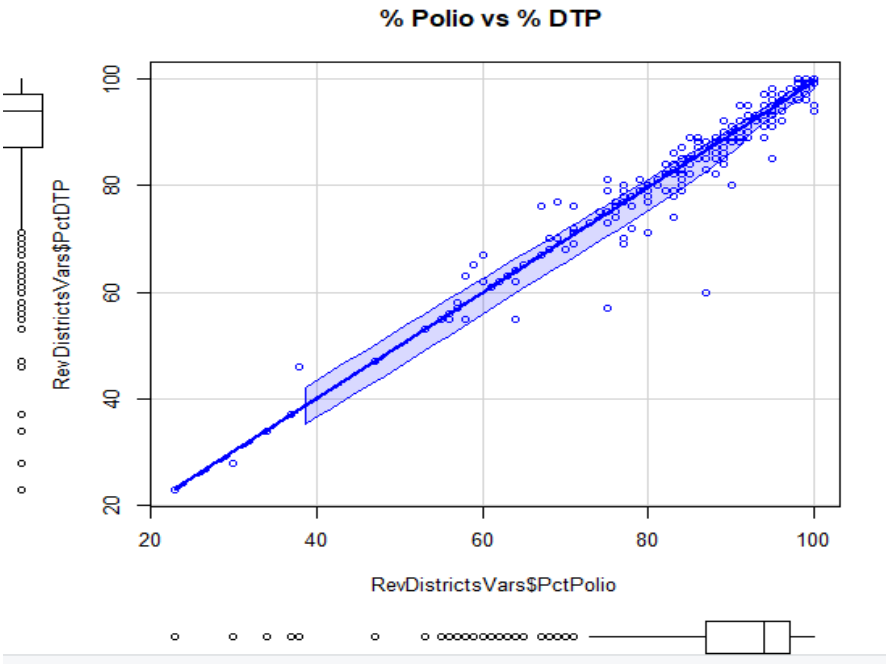


Non surprising, the same strong inverse relationship is also present for all 4 vaccines: -0.82 for DTP, -0.92 for HepB, -0.80 for MMR and -0.84 for Polio when compared to PctBelieveExempt.

The correlation analysis also shows a strong relationship of 0.75 between PctChildPoverty and PctFreeMeal and well as 0.84 between PctFamilyPoverty and PctChildPoverty. These variables have weak correlations to all 4 vaccines, ranging from 0.25 to 0.30 for PctFreeMeals, 0.20 to 0.22 for PctChildPoverty, and 0.25 to 0.27 for PctFamily Poverty. The same 3 variables also have very weak correlations to PctUpToDate: 0.21, 0.25 and 0.26 respectively. The scatterplot below is showing the strong correlation between PctFreeMeals and PctChildPoverty:



The strong correlation between the PctChildPoverty, PctFamilyPoverty, and PctFreeMeals variables, combined with their weak correlation with PctUptoDate, suggest that if a student in these districts receives free meals, there is good likelihood that the student is also below the poverty level, and that he or she is highly likely to not be up to date with vaccinations. Although correlation cannot tell us with absolute certainty if a student is missing one vaccine, he or she is missing all of them, from the low correlation rates between PctChildPoverty, PctFamilyPoverty, and PctFreeMeals to all four vaccines, we can infer that that these same students likely have low vaccination rates across the board. The opposite is true for the relationships between the vaccines themselves as they all display strong correlations among all 4 vaccines. We could perhaps infer that if a student received 1 vaccine, they likely received all 4. The scatterplot below is showing the strong correlation between PctDTP and PctPolio, as the move in the same direction and form a clear “cigar-shaped” trend pattern.




```
> cor(RevDistrictsVars)
                 PctUpToDate PctBeliefExempt PctChildPoverty PctFamilyPoverty
PctUpToDate        1.0000000      -0.7431689       0.2120931        0.2512115
PctBeliefExempt   -0.7431689       1.0000000      -0.1865198       -0.2505417
PctChildPoverty    0.2120931      -0.1865198       1.0000000        0.8477126
PctFamilyPoverty   0.2512115      -0.2505417       0.8477126        1.0000000
PctFreeMeal        0.2554940      -0.2808055       0.7549974        0.7267263
PctDTP             0.9655277      -0.8152436       0.2073955        0.2541683
PctHepB            0.8679261      -0.9180909       0.2235828        0.2787973
PctMMR             0.9753687      -0.8001370       0.2049241        0.2501737
PctPolio           0.9453674      -0.8367047       0.2059190        0.2573989
                 PctFreeMeal     PctDTP    PctHepB     PctMMR   PctPolio
PctUpToDate        0.2554940  0.9655277  0.8679261  0.9753687  0.9453674
PctBeliefExempt   -0.2808055 -0.8152436 -0.9180909 -0.8001370 -0.8367047
PctChildPoverty    0.7549974  0.2073955  0.2235828  0.2049241  0.2059190
PctFamilyPoverty   0.7267263  0.2541683  0.2787973  0.2501737  0.2573989
PctFreeMeal        1.0000000  0.2518026  0.2996478  0.2554863  0.2568728
PctDTP             0.2518026  1.0000000  0.9101956  0.9754551  0.9819677
PctHepB            0.2996478  0.9101956  1.0000000  0.9081746  0.9241032
PctMMR             0.2554863  0.9754551  0.9081746  1.0000000  0.9641282
PctPolio           0.2568728  0.9819677  0.9241032  0.9641282  1.0000000
```

Predictive Analyses:

(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.)

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

To determine if the reporting was complete based on the available variables, a logit linear regression was performed with DistrictComplete as the predicted variable, and PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools as the predictor variables.

Out of all the predictor variables, only the intercept, Enrolled and TotalSchools were statistically significant with p-values below the threshold alpha of 0.05. The other variables had p-values above 0.05, and therefore could not be used to predict if the districts' reporting was complete. They have failed to reject the null hypothesis that the probability or log-odds that the reporting was complete is equal to zero. Enrolled and Total schools have p-values of .00417 and 0.00128 respectively, and therefore provide support for the alternative hypothesis that the probability that the reporting was complete is not equal to zero.

The MCMClogit test, provides further support for the alternative hypothesis as in all instances it does not overlap with zero, and has all p-values below the alpha of 0.05. It is important to note that the predictor variable Enrolled gets close to crossing over zero with a range from- 0.003386 to -0.0008836. However, most of the values remained away from zero, as shown in the chart below. The density chart for each of the variables used in the MCMClogit test shows the HDI, or high density interval, which is the equivalent of where the majority of results falls in each of the cases.

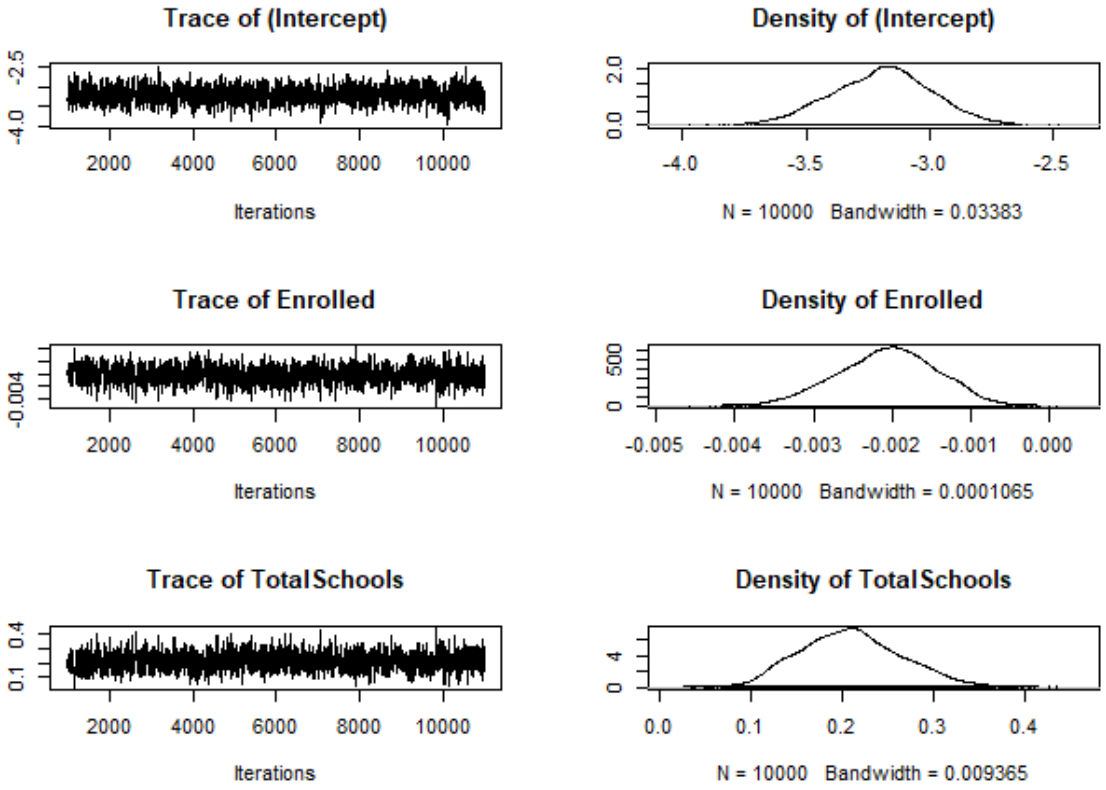
Logistic Analysis:

```
> #Output of Logistic Regression
> summary(glm_DistrictComplete)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2474  -0.3489  -0.2886  -0.2340   2.7763

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.9191908  0.4865102  -8.056  7.9e-16 ***
PctChildPoverty  -0.0290118  0.0306150  -0.948 0.343317
PctFreeMeal       0.0166113  0.0112181   1.481 0.138670
PctFamilyPoverty  0.0473507  0.0386546   1.225 0.220587
Enrolled         -0.0021936  0.0006216  -3.529 0.000417 ***
TotalSchools      0.2141159  0.0558969   3.831 0.000128 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 312.23  on 699  degrees of freedom
Residual deviance: 273.35  on 694  degrees of freedom
AIC: 285.35

Number of Fisher Scoring iterations: 6
```

Bayes Logistic Analysis:

```
> summary(glm_DistrictCompleteBayes)

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)  -3.192691 0.2013425 2.013e-03      6.633e-03
Enrolled     -0.002065 0.0006449 6.449e-06      2.184e-05
TotalSchools  0.210536 0.0564284 5.643e-04      1.864e-03

2. Quantiles for each variable:

                 2.5%       25%       50%       75%      97.5%
(Intercept)  -3.599242 -3.328871 -3.184430 -3.058699 -2.8117432
Enrolled     -0.003386 -0.002486 -0.002041 -0.001636 -0.0008836
TotalSchools  0.111810  0.171496  0.208693  0.246191  0.3271097
```

Plot of Bayes Logistic Analysis:



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

A multivariable linear regression was run to predict what percentage of all enrolled students have up-to-date vaccines. The test had PctUpToDate as the predicted variable and PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled, and TotalSchools as the predictor variables.
From the resulting coefficients, the only ones that were statistically significant (p-values < 0.05), were the ones for PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and Total Schools. The linear regression also produced an adjusted p-value of 7.9e-16 for the intercept. After performing a

vif() test to check for multicollinearity, only 3 predictor variables remained: PctChildPoverty, PctFreeMeal, and PctFamilyPoverty. Enrolled and TotalSchools scored high (above 5) on the vif() test which means they are highly correlated and could not be used.

A second run of the linear regression was performed with just PctChildPoverty, PctFreeMeal and PctFamilyPoverty. The p-value for PctChildPoverty was 0.140649 this time, which meant it was not statistically significant. The p-value for PctFreeMeal was of 0.000885 and PctFamilyPoverty was 0.006127, both statistically significant, and can therefore be used as a predictor for up-to-date vaccines. The resulting adjusted R-squared of 0.06839 can be interpreted as about 6% of the up-to-date vaccination rates can be explained by the changes in rates for PctFreeMeals and PctFamilyPoverty. The Bayes factor linear model produced odds of 6175399 to 1 in favor of the alternative hypothesis that states that the relationship between the predicted variable and predictor variables are not equal to zero.

1st Linear Regression:

```
> # Output of Linear Regression
> summary(PctUpToDate_lm)

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

Residuals:
    Min      1Q  Median      3Q     Max
-67.966  -3.130   3.189   7.188  18.336

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      82.558349   1.092769  75.550  < 2e-16 ***
PctChildPoverty  -0.109325   0.078904  -1.386  0.16633
PctFreeMeal       0.093253   0.028946   3.222  0.00133 **
PctFamilyPoverty  0.293322   0.111424   2.632  0.00867 **
Enrolled          0.006023   0.001900   3.170  0.00159 **
TotalSchools     -0.536448   0.175638  -3.054  0.00234 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.08 on 694 degrees of freedom
Multiple R-squared:  0.08668,   Adjusted R-squared:  0.0801
F-statistic: 13.17 on 5 and 694 DF,  p-value: 2.831e-12

> # Test for Multicollinearity
> vif(PctUpToDate_lm)
 PctChildPoverty      PctFreeMeal PctFamilyPoverty         Enrolled
        4.277070         2.439704         3.876561        85.959980
    TotalSchools
       85.886669
```

2nd Linear Regression:

```
> summary(PctUpToDate_lm2)

Call:
lm(formula = PctUpToDate ~ PctChildPoverty + PctFreeMeal + PctFamilyPoverty,
    data = newDist)

Residuals:
    Min      1Q  Median      3Q     Max
-68.525  -3.378   3.592   7.295  18.233

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      82.26739    1.08748  75.650  < 2e-16 ***
PctChildPoverty  -0.11701    0.07932  -1.475 0.140649
PctFreeMeal       0.09711    0.02908   3.339 0.000885 ***
PctFamilyPoverty  0.30800    0.11203   2.749 0.006127 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.16 on 696 degrees of freedom
Multiple R-squared:  0.07239,   Adjusted R-squared:  0.06839
F-statistic:  18.1 on 3 and 696 DF,  p-value: 2.539e-11
```

Bayes Factor analysis:

```
> summary(PctUpToDate_BF)
Bayes factor analysis
--------------
[1] PctChildPoverty + PctFamilyPoverty : 6175399 ±0%

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

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

Once again, a multivariable linear regression was used to find the best predictor variables. With PctBelieveException as the predicted and PctChildPoverty, PctFreeMeal, PctFamilyPoverty, Enrolled and TotalSchools as the predictors. The first results are shown below with the summary output. A vif() test for multicollinearity revealed that only PctChildPoverty, PctFreeMeal,and PctFamilyPoverty could be used as predictor variables.

A second linear regression was run, and the resulting p-values were all below 0.05 and statistically significant. They are as follows: 2e-16 for the intercept, 0.000598 for PctChildPoverty, 2.6e-8 for PctFreeMeal, and 0.001772 for pctFamilyPoverty. The coefficients for PctFreeMeal and PctFamilyPoverty were negative and thus inversely related to our predicted variable: PctBelieveExempt. This suggests that the children in the PctBelieveExempt group, are less likely to be from an impoverished background.

The resulting adjusted R-squared of 0.09353 tells us that about 9% of the rates for PctBelieveExempt can be predicted by the PctChildPoverty, PctFreeMeal, PctFamilyPoverty variables. The Bayes Factor analysis also gave us very good odds of 4083819 to 1, that the individual relationships between each of the predictor variables and the predicted variable are not equal to zero.

1st Linear Regression:

```
> # Output of Linear Regression
> summary(PctBeliefExempt_lm)

Call:
lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
    PctFamilyPoverty + Enrolled + TotalSchools, data = newDist)

Residuals:
    Min      1Q  Median      3Q     Max
-12.383  -3.960  -2.079   0.556  65.794

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.669266   0.763737  12.660  < 2e-16 ***
PctChildPoverty   0.184758   0.055146   3.350 0.000851 ***
PctFreeMeal      -0.111274   0.020231  -5.500 5.34e-08 ***
PctFamilyPoverty -0.236442   0.077875  -3.036 0.002486 **
Enrolled         -0.002552   0.001328  -1.922 0.055054 .
TotalSchools      0.211806   0.122754   1.725 0.084890 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.445 on 694 degrees of freedom
Multiple R-squared:  0.1059,    Adjusted R-squared:  0.0995
F-statistic: 16.45 on 5 and 694 DF,  p-value: 2.325e-15

> # Test for Multicollinearity
> vif(PctBeliefExempt_lm)
 PctChildPoverty      PctFreeMeal PctFamilyPoverty
        4.277070         2.439704         3.876561
        Enrolled     TotalSchools
       85.959980        85.886669
>
```

2nd Linear regression:

```
> summary(PctBeliefExempt_lm_2)

Call:
lm(formula = PctBeliefExempt ~ PctChildPoverty + PctFreeMeal +
    PctFamilyPoverty, data = newDist)

Residuals:
    Min      1Q  Median      3Q     Max
-12.340  -3.921  -2.084   0.574  65.879

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.70532    0.75775  12.808  < 2e-16 ***
PctChildPoverty   0.19061    0.05527   3.449 0.000598 ***
PctFreeMeal      -0.11411    0.02026  -5.631  2.6e-08 ***
PctFamilyPoverty -0.24496    0.07806  -3.138 0.001772 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.473 on 696 degrees of freedom
Multiple R-squared:  0.09742,   Adjusted R-squared:  0.09353
F-statistic: 25.04 on 3 and 696 DF,  p-value: 2.155e-15
```

Bayes Factor analysis:

```
> summary(BFOut)
Bayes factor analysis
--------------
[1] PctChildPoverty + PctFamilyPoverty + Enrolled + TotalSchools : 4083819
±0%

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?

Based upon the foregoing analyses, we've learned that the overall vaccination rates for California's reported districts are comparable to the US overall rates as reported by the World Health Organization. The only significant difference in rates were for HebB with California's districts reporting and average of 92% compared to 74% for the WHO. We were also able to determine through the use of descriptive statistic methods, that the proportion of public schools reporting their vaccinations rates are much higher at 80%, than private schools at only 20%. Our frequentist t.test provided support to the conviction that these results are credible in the long run, and that the means of these two groups are not equal to zero.

After running a correlation model for all the numerical variables present in the California Districts data set, we observed that there is a strong correlation between PctFreeMeals, PctChildPoverty, and PctFamilyPoverty, all of which also have strong inverse correlation to the PctUptodate variable. They also have an inverse, though weak correlation, to the PctBelieveExempt variable. This suggests that children of low income are more likely to not have up-to-date vaccination rates. It also suggests that children that are part of the PctBelieveExempt demographic are less likely to be from a lower income background. The PctFreeMeals, PctChildPoverty, and PctFamilyPoverty variables all had low correlations to all vaccination rates, thus further supporting the inference that children of lower income families, or bellow poverty lines have the lowest vaccination rates across all four vaccines.

While performing various predictive analysis, it was determined through a logistic analysis that the logs-odds or probability that the reporting is complete is not equal to zero. The Bayesian logistic approach confirmed these findings as none of the predictor variables overlapped with zero during the MCMClogite test. The multivariable linear regression analysis of the predicted PctUpToDate variable, confirmed that the PctFreeMeals, PctFamilyPoverty are the best variables available to predict the rate of up-to-date vaccinations. The linear regression for the PctBelieveExempt dependent variable, also came back with PctFreeMeals, PctChildPoverty, and PctFamilyPoverty as the best predictors. The Bayes factor analysis for both linear regressions returned excellent odds in favor of the alternative hypothesis, that the relationships between predictor and predicted variables are not equal to zero, and can therefore be used to predict the rates of PctUpToDate and PctBelieveExempt.

As next steps I would suggest that the staff at the state legislator's office allocates resources to increase research to better understand the relationships between PctBelieveExempt and PctUpToDate and how they relate to different income levels of the population. According to the

findings above mentioned, there is a clear association with impoverished demographics and low vaccination rates. A strong inverse relationship to vaccination rates and PctBelieveExempt was also detected. PctBelieveExempt was weakly correlated to low-income indicators, so the approach to this demographic should in all likelihood, be different than the one applied to the low-income ones.

I would recommend a thorough analysis of the districts that have reported the highest rates of PctFreeMeals, PctFamilyPoverty, and PctChildPoverty as well as the ones that reported highest rates for PctBelieveExampt, to determine how to better allocate resources and improve vaccination rates.