IST 772 Final Examination
Name : Akash kandarkar

Instructions: This exam is open book and open notes, but you may not receive assistance, help, coaching, guidance, or support from anyone.

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. To accomplish this, you should provide sufficient numeric and graphical detail that the staff member can use to 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.

You will work with two data sets that pertain to vaccinations; both are posted in Blackboard. These first dataset is the same for everyone and mainly included to provide context for interpretation of the results. Most of the substantive analyses occur in reference to the second dataset. I will provide code so that everyone will have a different sample of data from this data set.

Here is a description of each dataset:

US Vaccine Data: USAvaccinesTS.csv – Time series data from the World Health Organization reporting vaccination rates in the U.S. for four common vaccines from 1980 to 2017. Data include:

DPT = First dose of Diphtheria/Pertussis/Tetanus vaccine
Polio = Polio third dose
MMR = Measles first dose
HepB = Hepatitis B, Birth Dose

California School District Data: DistrictData.csv – Once this data set is in R, you will run the provided code to obtain your sample. This contains California public school districts' data from the 2013 data collection, along with specific numbers and percentages for each district. Variables are:

$ SchoolDistrict:	Name of school district
$ ToDate:	Percentage of all enrolled students with completely up-to-date vaccines
$WODPT:	Percentage of students without the DPT vaccine
$ WOPolio:	Percentage of students without the Polio vaccine
$ WOMMR:	Percentage of students without the MMR vaccine
$ WOHepB:	Percentage of students without the Hepatitis B vaccine
$ ChPov:	Percentage of children in the district living below the poverty line
$ Meal:	Percentage of children in the district eligible for free student meals
$ FamPov:	Percentage of families in the district living below the poverty line
$ Enrolled:	Total number of enrolled students in the district
$ Complete:	Boolean indicating whether or not the district's reporting was complete, 0 = No and 1 = Yes
$ TotSch:	Total number of different schools in the district

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

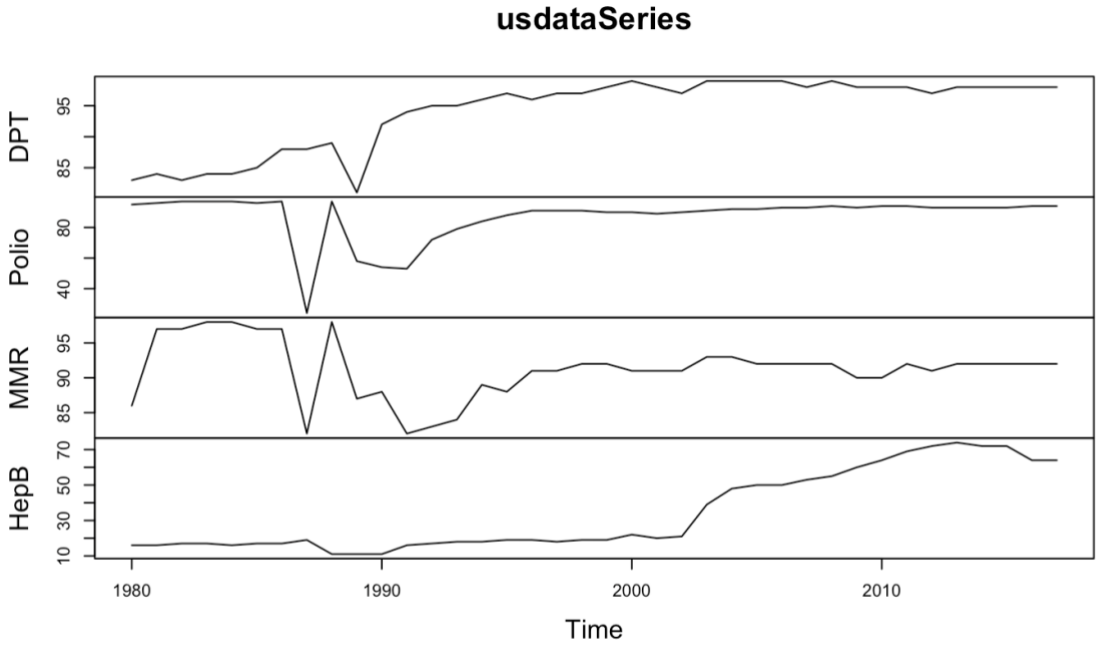
Introductory/Descriptive Reports:

1. US Vaccine Data:

a. Are US vaccination rates increasing or decreasing over this time period?
Ans ->

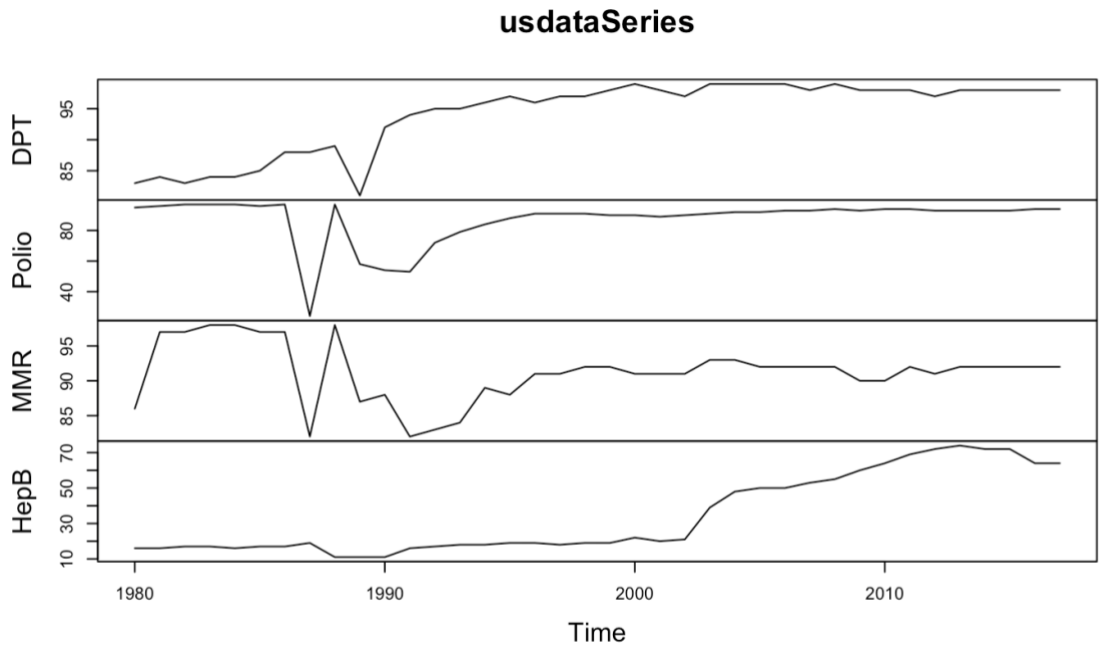
Looking at the rates of vaccines given for HepB, MMR, Polio and DPT we can see that overall there is growth in vaccination rates from 1980 to 2017. Even thought there was drop in vaccination rates in 1988-89 from MMR, Polio and DPT but HepB vaccination stayed stable till 2003-04.

Also, if we look at the data for HepB we can see there is drop in vaccination rates after 2015.
But we can say that there is an increase in vaccination rates over the period of time for all.



b. Which vaccination has the highest rate at the conclusion of the time series?

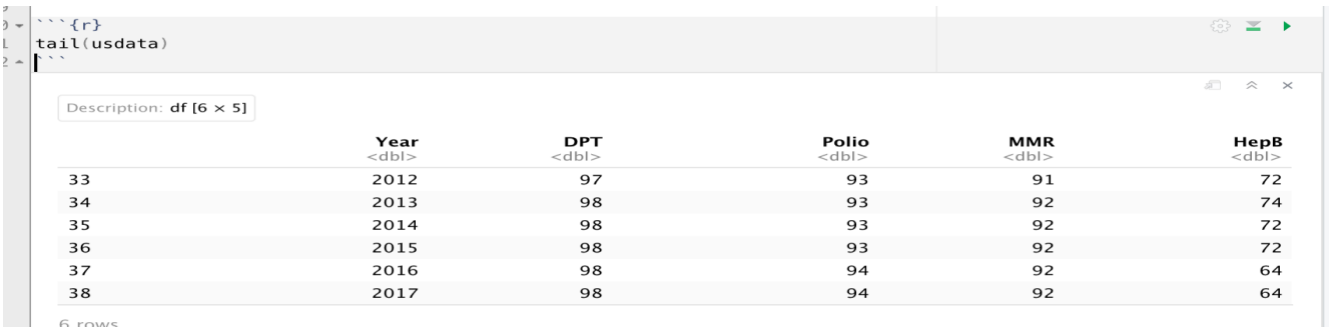
Ans ->



Looking at the summary of the data we can see except DPT and MMR , Polio and HepB rates in 1980 were very low.

```
      Year           DPT             Polio            MMR             HepB       
 Min.   :1980   Min.   :81.00   Min.   :24.00   Min.   :82.00   Min.   :11.00  
 1st Qu.:1989   1st Qu.:89.75   1st Qu.:90.00   1st Qu.:90.00   1st Qu.:17.00  
 Median :1998   Median :97.00   Median :93.00   Median :92.00   Median :19.00  
 Mean   :1998   Mean   :94.05   Mean   :87.16   Mean   :91.24   Mean   :34.21  
 3rd Qu.:2008   3rd Qu.:98.00   3rd Qu.:94.00   3rd Qu.:92.00   3rd Qu.:54.50  
 Max.   :2017   Max.   :99.00   Max.   :97.00   Max.   :98.00   Max.   :74.00  
```

Even though Polio and HepB vaccination rates have increase substantially, DPT have highest rate at the conclusion of the time series(2017) which is 98 as compare to Polio with 94,MMR with 92 and HepB with 64



```{r}
tail(usdata)
```

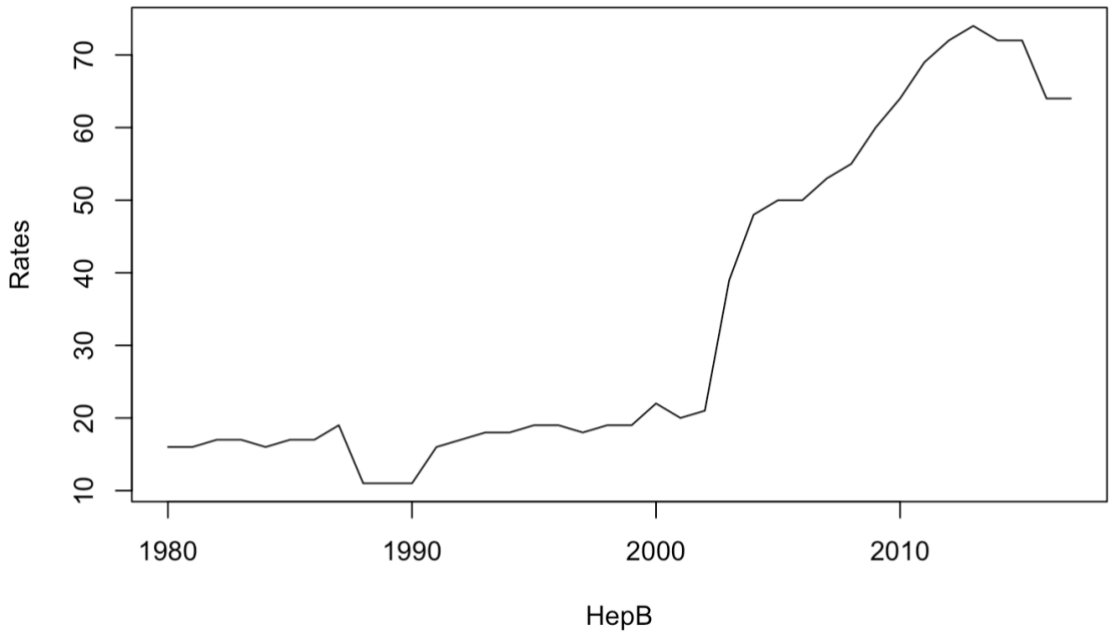
Description: df [6 × 5]

	Year <dbl>	DPT <dbl>	Polio <dbl>	MMR <dbl>	HepB <dbl>
33	2012	97	93	91	72
34	2013	98	93	92	74
35	2014	98	93	92	72
36	2015	98	93	92	72
37	2016	98	94	92	64
38	2017	98	94	92	64

6 rows

c. Which vaccination has the lowest rate at the conclusion of the time series?

Ans->



Even thought there was increase in vaccination rates for HepB, from 2002 to 2015 to highest 74, there is certain drop in vaccination rates as we can see from the data in 2017, HepB have lowest rate of vaccination,

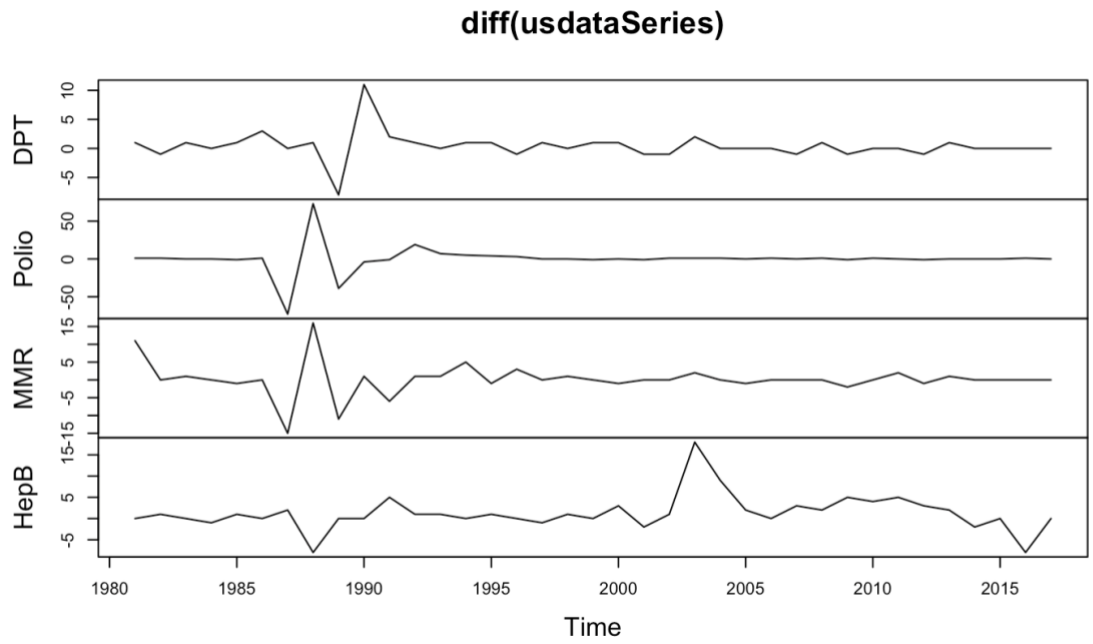
```{r}
tail(usdata)
```

Description: df [6 × 5]

	Year <dbl>	DPT <dbl>	Polio <dbl>	MMR <dbl>	HepB <dbl>
33	2012	97	93	91	72
34	2013	98	93	92	74
35	2014	98	93	92	72
36	2015	98	93	92	72
37	2016	98	94	92	64
38	2017	98	94	92	64

6 rows

d. Which vaccine has the greatest volatility?



Standard deviation in Hepatitis B, Birth Dose is 22.53, which shows it have highest volatility among the all vaccines.

```{r}
sd(usdataSeries[,"HepB"])
#Standard deviation in Hepatitis B, Birth Dose is 22.53, which shows it have highest volatility among the all vaccines.
```

```
[1] 22.53904
```

```{r}
sd(usdataSeries[,"Polio"])
#Standard deviation in Polio third dose is 15.35266, which shows it have volatility less than Hepatitis B, but greater than rest.
```

```
[1] 15.35266
```

```{r}
sd(usdataSeries[,"DPT"])
#Standard deviation in First dose of Diphtheria/Pertussis/Tetanus vaccine is 5.867673, which shows it have less volatility

```

```
[1] 5.867673
```

```{r}
sd(usdataSeries[,"MMR"])
#Standard deviation in Measles first dose vaccine is 4.187718, which shows it have lowest volatility
```

```
[1] 4.187718
```

2. California School District Data:

a. What are 2013 vaccination rates for individual vaccines (i.e., DPT, Polio, MMR, and HepB) in California public schools?

Ans->

The WODPT , WOPolio, WOMMR and WOHepB column provides data regarding individuals without DPT,Polio, MMR and HepB vaccines so this mean of these columns must be subtracted from 1 to get the data with DPT,Polio, MMR and HepB vaccines.

Vaccination rates for individual vaccines are given below :

Description: df [1 × 4]

WODPT_mean <dbl>	WOPolio_mean <dbl>	WOMMR_mean <dbl>	WOHepB_mean <dbl>
89.98857	90.40857	89.93857	92.42143

1 row

89.99% individuals in California public schools are DPT vaccinated.
90.40% individuals in California public schools are Polio vaccinated.
89.93% individuals in California public schools are MMR vaccinated.
92.42% individuals in California public schools are HepB vaccinated.

b. 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)?

Ans->

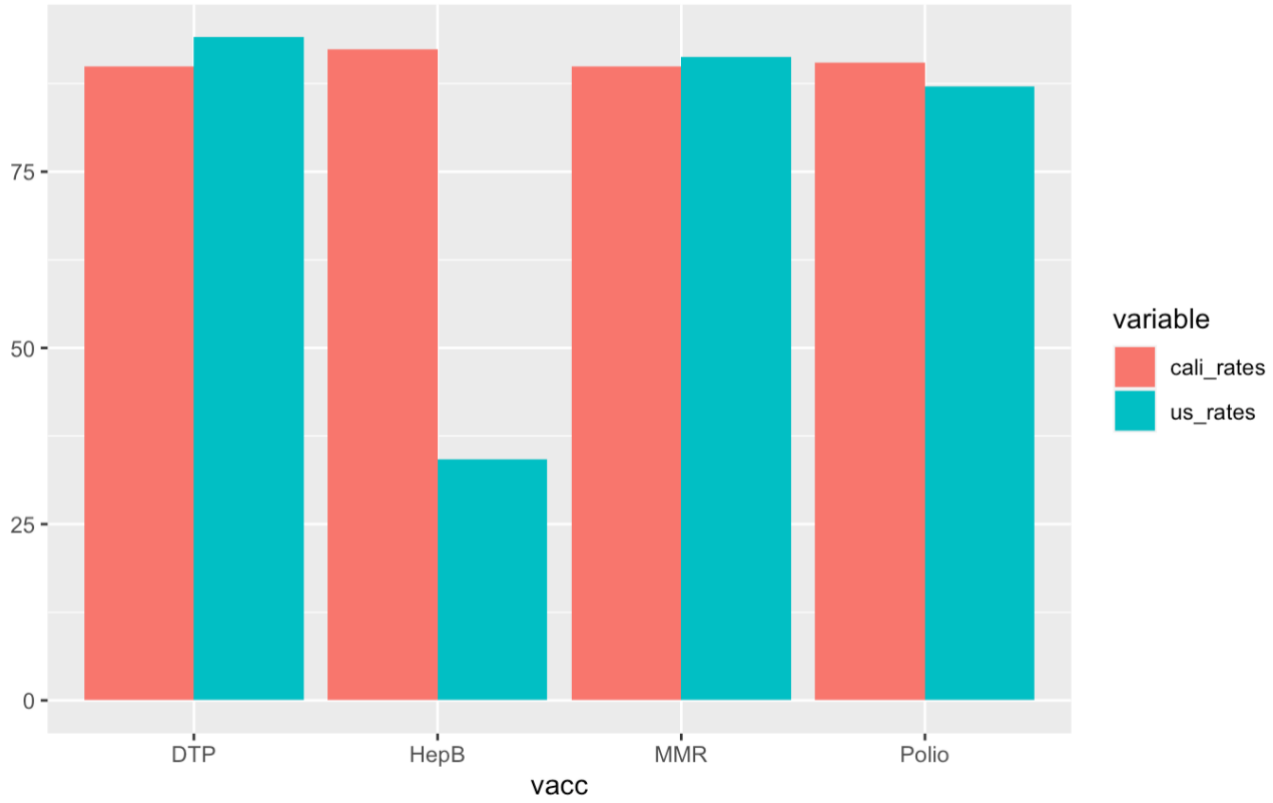
We have vaccination rates for California districts :

Description: df [1 × 4]

WODPT_mean <dbl>	WOPolio_mean <dbl>	WOMMR_mean <dbl>	WOHepB_mean <dbl>
89.98857	90.40857	89.93857	92.42143

1 row

We will compare these vaccination rates with usdata vaccination rates.



#For DPT vaccine, the overall US vaccination rates stand at 94.05% while that in California districts are 89.99%.
#There is quite a difference between these two percentages.

#For HepB vaccine, the overall US vaccination rates stand at 34.21% while that in California districts are 92.42%.
#In case of HepB vaccine the California district vaccinations rates are significantly higher.

#For MMR vaccine, the overall US vaccination rates stand at 91.23% while that in California districts are 89.94%.
#In case of MMR vaccine the percentage values are closer between USA and California.

#For Polio vaccine, the overall US vaccination rates stand at 87.15% while that in California districts are 90.40%.
#In case of Polio vaccine, the percentage values are closer between USA and California.

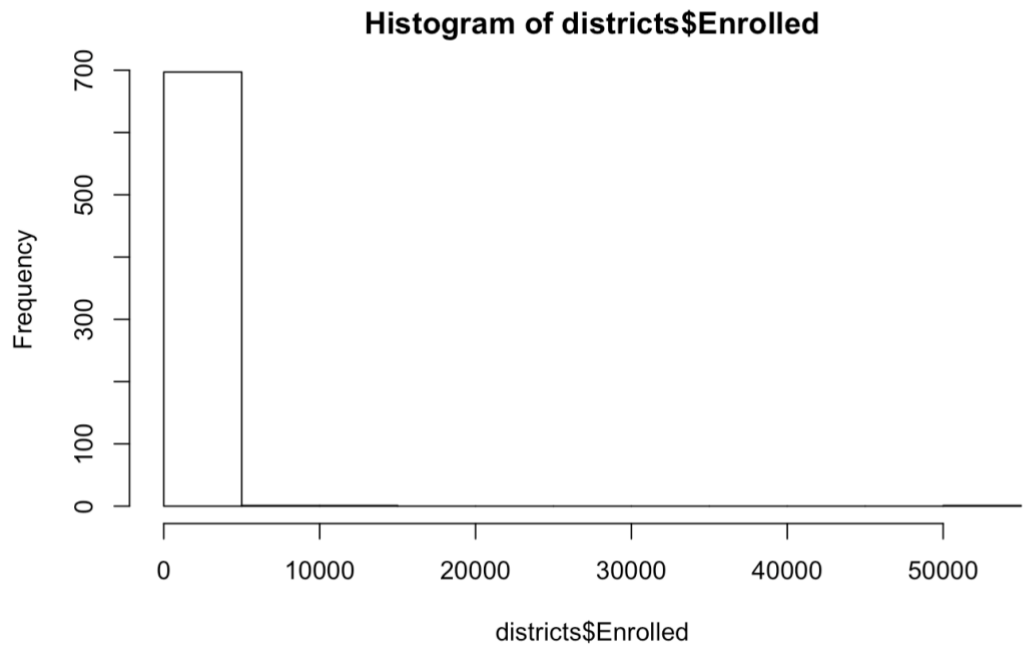
Predictive Analyses:
(For all of these analyses, use ChPov, Meal, FamPov, Enrolled, and TotSch as predictors. Transform variables as necessary to improve prediction and/or interpretability. In general, if there is a Bayesian version of an analysis available, you are expected to run that analysis in addition to the frequentist

version of the analysis.)

Ans →

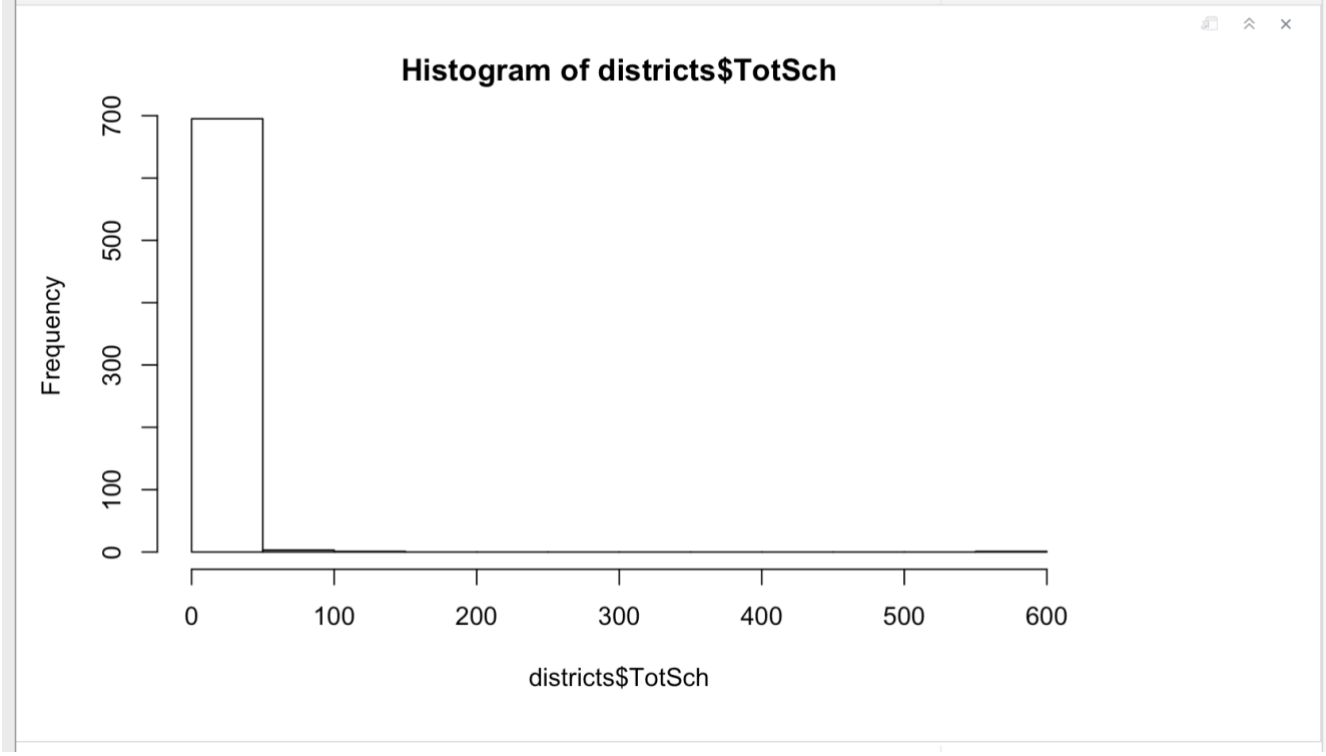
Plotting data to observe skewness:

```r
hist(districts$Enrolled)
#The histogram for TotSch shows that data is skewed to the right.
```



```r
hist(districts$TotSch)

#The histogram for TotSch shows that data is skewed to the right.
```



Correlation matrix:

	ChPov	Meal	FamPov	Enrolled	TotSch
ChPov	1.000000000	0.75226740	0.85834877	0.01346971	0.009320205
Meal	0.752267395	1.00000000	0.72412131	0.05541598	0.050871994
FamPov	0.858348771	0.72412131	1.00000000	0.02958267	0.023371620
Enrolled	0.013469707	0.05541598	0.02958267	1.00000000	0.994355922
TotSch	0.009320205	0.05087199	0.02337162	0.99435592	1.000000000

We can see from the correlation matrix that TotSch and Enrolled have a poor correlation with all the other variables, but the other variables have a strong link.

To lessen the skewness of data, we will need to modify the variables using the suitable transformation function.

Let's, Verify the skewness before transformation

```r
skewness(districts$Enrolled)#20.77401
skewness(districts$TotSch)#20.30251
```

Positive skewness indicates that the data is skewed correctly. Both the Enrolled and TotSch data have a positive skewness value, which makes reasonable given that they were both right skewed in the histogram.

Lets try different transformation to see which one reduces the skewness :

- Square-root for transformation :

```
skewness(sqrt(districts$Enrolled)) #4.186683
skewness(sqrt(districts$TotSch)) #4.826241
```

- Log for transformation

```
#Using log for transformation
skewness(log(districts$Enrolled)) #-0.05603821
skewness(log(districts$TotSch)) #0.6057657
```

- atan for transformation

```
#Using atan for transformation
skewness(atan(districts$Enrolled)) #-1.981354
skewness(atan(districts$TotSch)) #-0.1581528
```

- Reciprocal for transformation

```
#Using reciprocal for transformation
skewness(1/districts$Enrolled) #1.985216
skewness(1/districts$TotSch)# 0.2635157
```

#When we look at the skewness values after transformation, we can observe that the TotSch variable has the largest skewness decrease (-0.10) after using the atan approach. After using the log approach, the skewness of Enrolled variable is reduced the highest (-0.0020).

Creating correlation matrix again :

```
                ChPov        Meal      FamPov logEnrolled  atanTotSch
ChPov       1.00000000  0.75226740  0.85834877 -0.08710329 -0.13478888
Meal        0.75226740  1.00000000  0.72412131  0.03351441 -0.01424682
FamPov      0.85834877  0.72412131  1.00000000  0.01556182 -0.04784430
logEnrolled -0.08710329  0.03351441  0.01556182  1.00000000  0.89044042
atanTotSch  -0.13478888 -0.01424682 -0.04784430  0.89044042  1.00000000
```

3. What variables predict whether a district's reporting was complete?

   Ans->
   The Enrolled variable after log transformation, the TotSch variable after atan transformation variable are significant in predicting whether a district's reporting was complete.

4. What variables predict the percentage of all enrolled students with up-to-date vaccines?
   Ans →

   logEnrolled and Meal are significant in predicting ToDate variable.

5. 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?

   Ans ->

   The first analysis performed focused on predicting the reporting completeness in districts. Based on different predictor variables from the data supplied. We found using logistic regression analysis that variables Enrolled (Total number of enrolled students in the district), TotSch(Total number of different schools in the district) are significant in making the prediction.
   The logodds coefficient of Enrolled is 1.7735 and TotSch is -15.6983.

   However, our initial descriptive analysis had shown that the data in Enrolled variable and TotSch variable were highly skewed and hence mathematical transformations on the data were required to reduce the skewness and do the correct analysis.

   The same analysis was performed using the Bayesian methodology which also resulted in the same results.

The second analysis, focused on predicting the Percentage of all enrolled students with completely up-to-date vaccines. This analysis was conducted using Linear regression and the Bayesian equivalent of it as well. Both the analysis stated that, the significant variables in the prediction are Enrolled (Total number of enrolled students in the district) and Meal(Percentage of children in the district eligible for free student meals).

Coming to conclusion, number of increase in students enrolled should be increased to increase Vaccination. As we can see there is meal plans are not working great with vaccination reports, but showing benefits of free meal and investing in free meal service will invite new students to school which will increase vaccination rates.