BIG DATA ANALYTICS FINAL PROJECT
Immunization Data by School

Introduction

Immunization Data by school is the dataset that I choose for my Big Data Analysis Project. The dataset includes vaccination details of students from kindergartner through 12th grade in both public and private schools for the year 2016-2017. The dataset is collected from the parent's report about the immunization records of their children that was given at the schools. This data is collected from school in Washington districts and county wise. So, it covers all the schools in Washington in the year 2016-2017.The dataset has 2,595 number of observations and 35 attributes. In this project with help of visualizations, hypothesis testing, classification we will be analyzing the dataset.[1]

Who

• Who they are, what do they do?

The data maintained by Data.WA.gov and the owner for the dataset is Eavey,Joanna(DOH). They have collected the data from Office of Immunization and Child Profile, Department of Health.

Their main job is to collect this data to help the government make better decisions and bring about improvements that will makes changes in the required vaccinations. They will also release the data through an open source to the users. And people can have a clear idea about the vaccinations causes that are necessary for their children. [1]

• What is their role/purpose?

Data.WA.gov is the open data platform of the federal government, with the goal of making government more transparent and accountable. Data.gov is aimed at improving public access to high-value, machine-readable datasets created by the Federal Government's Executive Branch. The site is a repository for information provided to the public by federal, state, local, and tribal government. This offers Federal datasets (metadata) definitions, information on how to access datasets, and resources that exploit government datasets. As data sets are added, the data catalogs will continue to grow.[1]

Need

• Why did they collect this data?

The data is needed to be collected so that that the government can ensures that parents are aware about the vaccinations for their children. The open existence of publicly available data shows an organization's side that is very frequently kept under wraps. Open data helps to build people's awareness. The availability of consolidated information in a single and easily accessible location is advantageous for both the use of current and historical data collected over time. This data storage method ensures that all information will appear where and how it should be, and that it will remain for future reference at that location.[1]

• What potential questions could be answered by studying this data?

a) What are the top 5 schools that have highest Percent that have completed all immunization?
b) What is the top 5 percentage of students that having exception for hepatis B
c) How many students have medical exemption?
d) How many schools have reported the immunization details?
e) Top five county's with highest records?

• Is there any privacy, quality, or other issues with this data?

Privacy:

The main aim of this dataset is for the public to access it for their benefits. Data.WA.gov is the open data platform of the federal government with the goal of improving public access to high-value, machine-readable data sets.

Quality:

The dataset contains many blank fields (i.e NA's) which are to be removed from the dataset for preparing the dataset suitable for extracting the results.

Other issues:

The missing data and null values in the data set may cause the results to be error / inaccurate.

Requirements and Resources Needed:

Software:

a) RStudio is used for data cleaning, transformation, and visualization.
b) Tableau is used for visualization.

Hardware:

a) Processor Name: Intel Core i5
b) Processor Speed: 2.9 GHz
c) RAM – 8GB
d) System Type: 64-Bit operating system

Data Cleaning :

Data cleaning must be done to get better visualizations. The dataset that I have chosen has many N/A values . I have used RStudio and a command na.omit() clean the data that has N/A values.

Summary of the Dataset

```
> omitteddata <-na.omit(Schooldata)
> summary(omitteddata)
                      School_Name    School_year   Reported K_12_enrollment
 ROOSEVELT ELEMENTARY SCHOOL :   7   2016-17:2478   N:   0   Min.   :   0.0
 CASCADE MIDDLE SCHOOL       :   5                  Y:2478   1st Qu.: 171.0
 JEFFERSON ELEMENTARY        :   5                           Median : 418.5
 LINCOLN ELEMENTARY SCHOOL   :   5                           Mean   : 455.9
 WASHINGTON ELEMENTARY SCHOOL:   5                           3rd Qu.: 591.8
 CHINOOK MIDDLE SCHOOL       :   4                           Max.   :2577.0
 (Other)                     :2447
 Percent_complete_for_all_immunizations Percent_with_any_exemption Percent_with_medical_exemption
 Min.   :  0.00                         Min.   : 0.000            Min.   : 0.000
 1st Qu.: 82.90                         1st Qu.: 2.525            1st Qu.: 0.000
 Median : 90.30                         Median : 4.700            Median : 0.700
 Mean   : 85.35                         Mean   : 6.414            Mean   : 1.112
 3rd Qu.: 94.30                         3rd Qu.: 7.400            3rd Qu.: 1.400
 Max.   :100.00                         Max.   :78.300            Max.   :21.600

 Percent_with_personal_exemption Percent_with_religious_exemption Percent_with_religious_membership_exemption
 Min.   : 0.000                  Min.   : 0.0000                  Min.   : 0.0000
 1st Qu.: 1.600                  1st Qu.: 0.0000                  1st Qu.: 0.0000
 Median : 3.400                  Median : 0.0000                  Median : 0.0000
 Mean   : 4.988                  Mean   : 0.3657                  Mean   : 0.1046
 3rd Qu.: 5.700                  3rd Qu.: 0.4000                  3rd Qu.: 0.0000
 Max.   :75.000                  Max.   :28.6000                  Max.   :14.3000

 Percent_exempt_for_diphtheria_tetanus Percent_exempt_for_pertussis Percent_exempt_for_measles_mumps_rubella
 Min.   :  0.000                       Min.   :  0.000              Min.   :  0.000
 1st Qu.:  1.300                       1st Qu.:  1.200              1st Qu.:  1.300
 Median :  2.700                       Median :  2.500              Median :  2.700
 Mean   :  4.255                       Mean   :  4.044              Mean   :  4.303
 3rd Qu.:  4.500                       3rd Qu.:  4.300              3rd Qu.:  4.500
 Max.   :100.000                       Max.   :100.000              Max.   :100.000

 Percent_exempt_for_polio Percent_exempt_for_HepatitisB Percent_exempt_for_varicella
 Min.   :  0.000          Min.   :  0.000               Min.   : 0.000
 1st Qu.:  1.300          1st Qu.:  1.300               1st Qu.: 1.800
 Median :  2.600          Median :  2.700               Median : 3.600
 Mean   :  4.221          Mean   :  4.434               Mean   : 5.273
 3rd Qu.:  4.500          3rd Qu.:  4.700               3rd Qu.: 5.900
 Max.   :100.000          Max.   :100.000               Max.   :99.500

 Number_complete_for_all_immunizations Number_with_any_exemption Number_with_medical_exemption
 Min.   :   0.0                        Min.   :  0.00            Min.   :  0.000
 1st Qu.: 132.0                        1st Qu.:  6.00            1st Qu.:  0.000
 Median : 370.0                        Median : 17.00            Median :  3.000
 Mean   : 398.0                        Mean   : 24.06            Mean   :  5.301
 3rd Qu.: 532.8                        3rd Qu.: 32.00            3rd Qu.:  6.000
 Max.   :2532.0                        Max.   :332.00            Max.   :207.000

 Number_with_personal_exemption Number_with_religious_exemption Number_with_religious_membership_exemption
 Min.   :  0.0                  Min.   : 0.000                  Min.   : 0.0000
 1st Qu.:  4.0                  1st Qu.: 0.000                  1st Qu.: 0.0000
 Median : 12.0                  Median : 0.000                  Median : 0.0000
 Mean   : 17.8                  Mean   : 1.354                  Mean   : 0.4023
 3rd Qu.: 24.0                  3rd Qu.: 2.000                  3rd Qu.: 0.0000
 Max.   :232.0                  Max.   :36.000                  Max.   :45.0000

 Number_exempt_for_diphtheria_tetanus Number_exempt_for_pertussis Number_exempt_for_measles_mumps_rubella
 Min.   :  0.0                        Min.   :  0.00              Min.   :  0.00
 1st Qu.:  4.0                        1st Qu.:  3.00              1st Qu.:  4.00
 Median : 10.0                        Median : 10.00              Median : 10.00
 Mean   : 14.2                        Mean   : 13.31              Mean   : 13.96
 3rd Qu.: 19.0                        3rd Qu.: 18.00              3rd Qu.: 19.00
 Max.   :250.0                        Max.   :245.00              Max.   :247.00

 Number_exempt_for_polio Number_exempt_for_HepatitisB Number_exempt_for_varicella
 Min.   :  0.00          Min.   :  0.00               Min.   :  0.00
 1st Qu.:  3.25          1st Qu.:  4.00               1st Qu.:  5.00
```

```
Number_exempt_for_polio Number_exempt_for_HepatitisB Number_exempt_for_varicella
Min.   :  0.00          Min.   :  0.00               Min.   :  0.00
1st Qu.:  3.25          1st Qu.:  4.00               1st Qu.:  5.00
Median : 10.00          Median : 10.00               Median : 13.00
Mean   : 13.72          Mean   : 14.35               Mean   : 18.62
3rd Qu.: 19.00          3rd Qu.: 20.00               3rd Qu.: 25.00
Max.   :253.00          Max.   :249.00               Max.   :627.00

                      School_District       County                                                ESD
SEATTLE PUBLIC SCHOOLS         : 169   KING     :652   PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121:920
SPOKANE SCHOOL DISTRICT        :  71   PIERCE   :259   NORTHWEST EDUCATIONAL SERVICE DISTRICT 189  :358
LAKE WASHINGTON SCHOOL DISTRICT:  70   SNOHOMISH:202   EDUCATIONAL SERVICE DISTRICT 101            :275
TACOMA SCHOOL DISTRICT         :  70   SPOKANE  :177   EDUCATIONAL SERVICE DISTRICT 112            :209
BELLEVUE SCHOOL DISTRICT       :  58   CLARK    :138   EDUCATIONAL SERVICE DISTRICT 113            :202
NORTHSHORE SCHOOL DISTRICT     :  49   YAKIMA   :101   EDUCATIONAL SERVICE DISTRICT 123            :133
(Other)                        :1991   (Other)  :949   (Other)                                     :381
 Grade_Levels Has_kindergarten Has_6thGrade                                                Location.1
K -5   :388   N: 945           N:1372       PO BOX 476\nYELM                                    :   8
09-Dec :362   Y:1533           Y:1106       PO BOX 200\nBATTLE GROUND                           :   7
K -6   :235                                 1110 S. 6TH STREET\nSUNNYSIDE\n(46.317414, -120.012983):   5
06-Aug :234                                 P O BOX 833\nOMAK                                   :   5
PK-5   :226                                 P.O. BOX 907\nMATTAWA                               :   5
P-8    :123                                 8301 84TH STREET NE\nMARYSVILLE\n(48.07201, -122.11792):   4
(Other):910                                 (Other)                                             :2444
```

Dataset Description

The dataset with a size of 629 KB has over 35 attributes, 2595 observation in total collected by the Government and stored at Data.Gov. The 35 attributes of the dataset are described below

School Name- It contains the names of the schools that are recorded in Washington. It's a text datatype and the attribute is a nominal data. Example- Rock Creek Hutterite

Schoolyear- It contains information about the years in which the data was collected. The datatype is text and the attribute is an interval data. Example- 2016-2017

Reported- It contains information about the schools that have reported and that have not reported . The datatype is text and the attribute is an ordinal data. Example- Y, N

K_12_enrollment- It contains information about the students who have enrolled with their immunizations. The datatype is number and the attribute is a ratio data. Example- 114, 23, 549,16.

Percent_complete_for_all_immunizations- It contains data about the percentage of students who completed all the immunizations. The datatype is a number and the attribute is a interval data. Example- 81, 99.5

Percent_with_any_exemption- It contains the data about the percentage of students with any type of exemptions. The datatype is number and the attribute is ratio data. Example- 2.2, 0, 5.4.

Percent_with_medical_exemption- It contains data of the percentage of students with medical exemptions. The datatype is a number and the attribute is a ratio data. Example- 0.2, 0, 0.7.

Percent_with_personal_exemption- It contains data of the percentage of students with personal exemption. The datatype is number and the attribute is a ratio. Example- 0.2, 0.5,2.4,0.

Percent_with_religious_exemption- It contains data of the percentage of students with religious exemptions. This datatype is ratio. Example- 0, 0.1, 0.3

Percent_with_religious_membership_exemption- It contains data of the percentage of students with religious exemptions. The datatype is number and attribute is ratio data. Example- 0.5, 0.7,0.

Percent_exempt_for_diphtheria_tetanus- It contains data of the percentage of students exempted from diphtheria & tetanus. The datatype is number and the attribute is ratio. Example- 0.4, 0.5,0.2.

Percent_exempt_for_pertussis- It consists data of the percentage of students exempted from pertussis. The datatype is number and the attribute is ratio data. Example- 0.6, 0.9,0.

Percent_exempt_for_measles_mumps_rubella- It consists data of the percentage of students exempted from measles mumps and rubella. The datatype is number and the attribute is ratio. Example- 2.9, 1.1.

Percent_exempt_for_polio- It consists data of the percentage of students exempted from polio. The datatype is number and the attribute is ratio. Example- 15.9, 0.2.

Percent_exempt_for_HepatitisB- It consists data of the percentage of students exempted from Hepatitis B. The datatype is number and the attribute is ratio. Example- 1.6, 0.3.

Percent_exempt_for_varicella- It consists data of the percentage of students exempted from varicella. The datatype is number and the attribute is ratio. Example- 7.7, 0.9.

Number_complete_for_all_immunizations- It contains the data of the number of students who completed all the immunizations. The datatype is number and the attribute is interval data. Example- 476, 120.

Number_with_any_exemption- It contains the data about the number of students with any type of exemptions. The datatype is number and the attribute is ratio. Example- 4,3,0.

Number_with_medical_exemption- It contains data of the number of students with medical exemptions. The datatype is number and the attribute is ratio. Example- 1,2,3,0.

Number_with_personal_exemption- It contains data of the number of students with personal exemption. The datatype is number and the attribute is ratio. Example- 2,3,1.

Number_with_religious_exemption- It contains data of the number of students with religious exemptions. The datatype is number and the attribute is ratio. Example- 3,2,1,0.

Number_with_religious_membership_exemption- It contains data of the number of students with religious exemptions. The datatype is number and the attribute is ratio. Example- 2,3,0.

Number_exempt_for_diphtheria_tetanus- It contains data of the number of students exempted from diphtheria & tetanus. The datatype is number and the attribute is ratio. Example- 3,0,4.

Number_exempt_for_pertussis- it consists data of the number of students exempted from pertussis. The datatype is number and the attribute is ratio. Example- 0,1,3.

Number_exempt_for_measles_mumps_rubella- It consists data of the number of students exempted from measles mumps and rubella. The datatype is number and the attribute is ratio. Example-2,0,1.

Number_exempt_for_polio- It consists data of the number of students exempted from polio. The datatype is number and the attribute is ratio. Example- 2,0,1.

Number_exempt_for_HepatitisB- It consists data of the number of students exempted from Hepatitis B. The datatype is number and the attribute is ratio. Example- 3,2,1.

Number_exempt_for_varicella- It consists data of the number of students exempted from varicella. The datatype is number and the attribute is ratio. Example- 2,3,1,4.

School_District- It consists data about the the district that the school belongs to . The datatype text and attribute is nominal data. Example- Ephrata school district.

County- It consists of the data about the county that the school belongs to . The datatype is a text and attribute is nominal. Example- Asotin, grant.

ESD- It consists of the data about the educational service district that the school belongs to. The datatype is text and attribute is nominal data . Example- Educational service district 101

Grade_Levels- It consists of the data about the grade level of the students in the schools. The datatype is text and attribute is ordinal data. Example- P-1, PK-12, P-6.

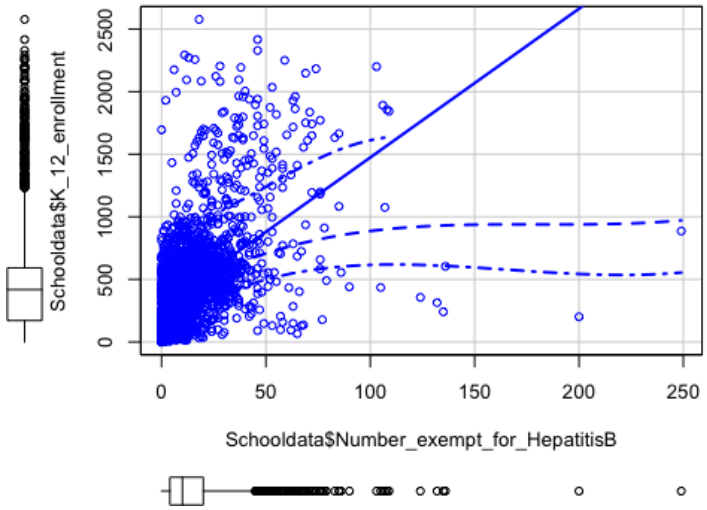
Has_kindergarten- It consists of the data if the student is a kindergarten student or not. The datatype is text and attribute is ordinal data. Example- Y,N.

Has_6thGrade- It consists of the data if the student has crossed 6th grade. The datatype is text and attribute is ordinal data. Example- Y, N.

Area- It consists data about the geographical location of the school. The datatype is text and attribute is ordinal data. Example-701 E, 3242.

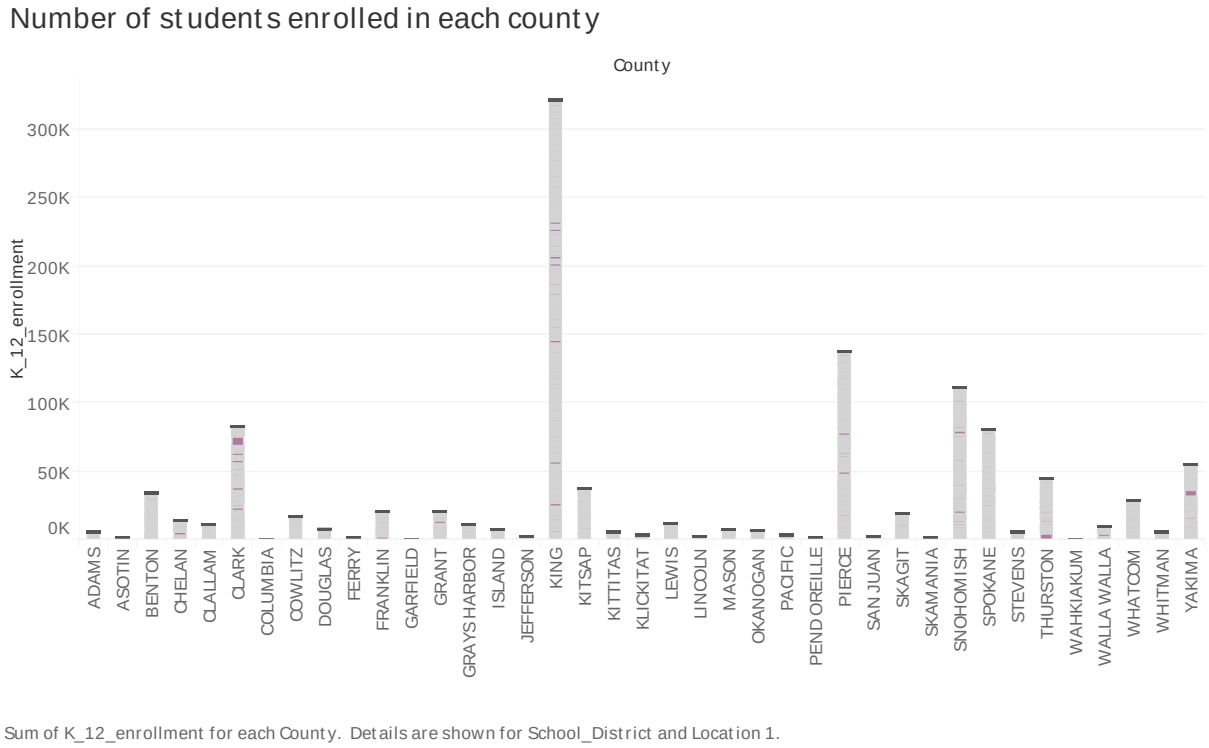
Results and Findings

a) Scatter plot



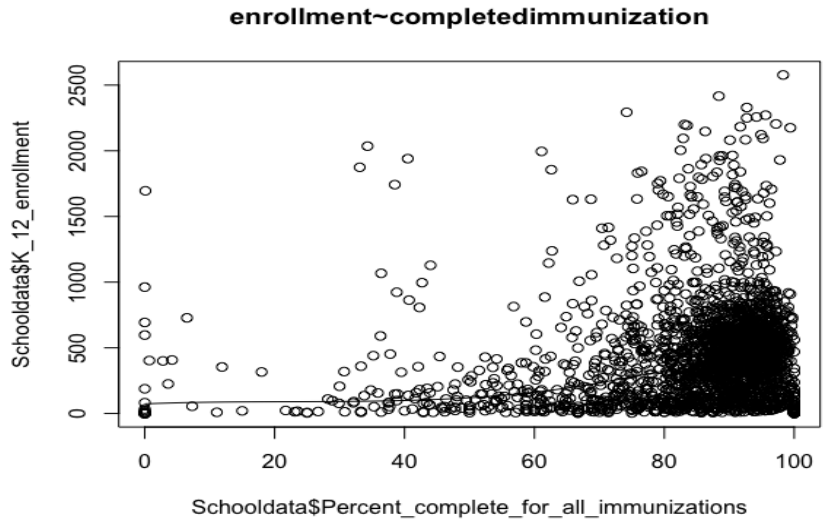
The above scatter plot is drawn between Number of students that is K_12_enrollment and Number of students that have exception from Hepatitis B. We can see the highest number of people who have exception for Hepatitis B are above 2500 and the lowest number of people that have exception for hepatitis B is zero. This visualization answers question B.

b) Boxplot



The above boxplot is plotted between k_12_enrollment and County. We can see that the highest number of enrolment's are from King county and the least from Columbia and wahkiakum. This visualization answers question e. It shows county's that have highest enrolment's.

c) Regression analysis



The scatter plot along with the smoothing line above suggests a linearly constant and then gradually increasing relationship between the

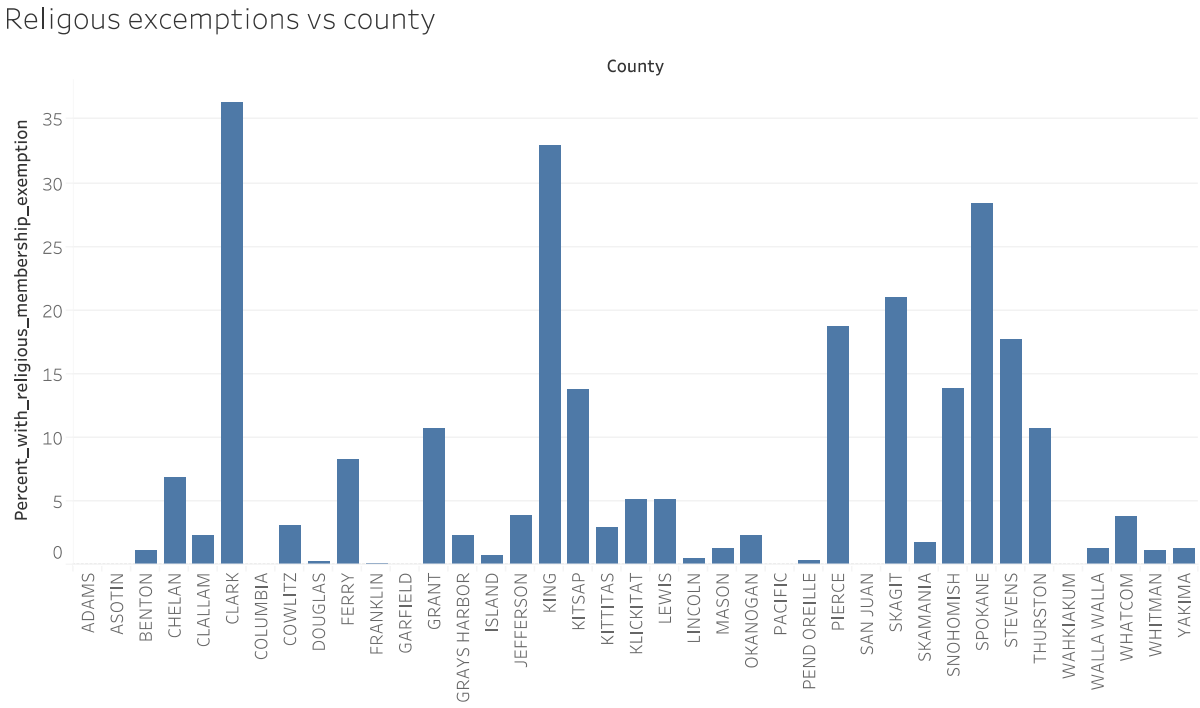
enrolment and completed immunization indicates that enrolment is constant with the complete immunization.

d) Correlation

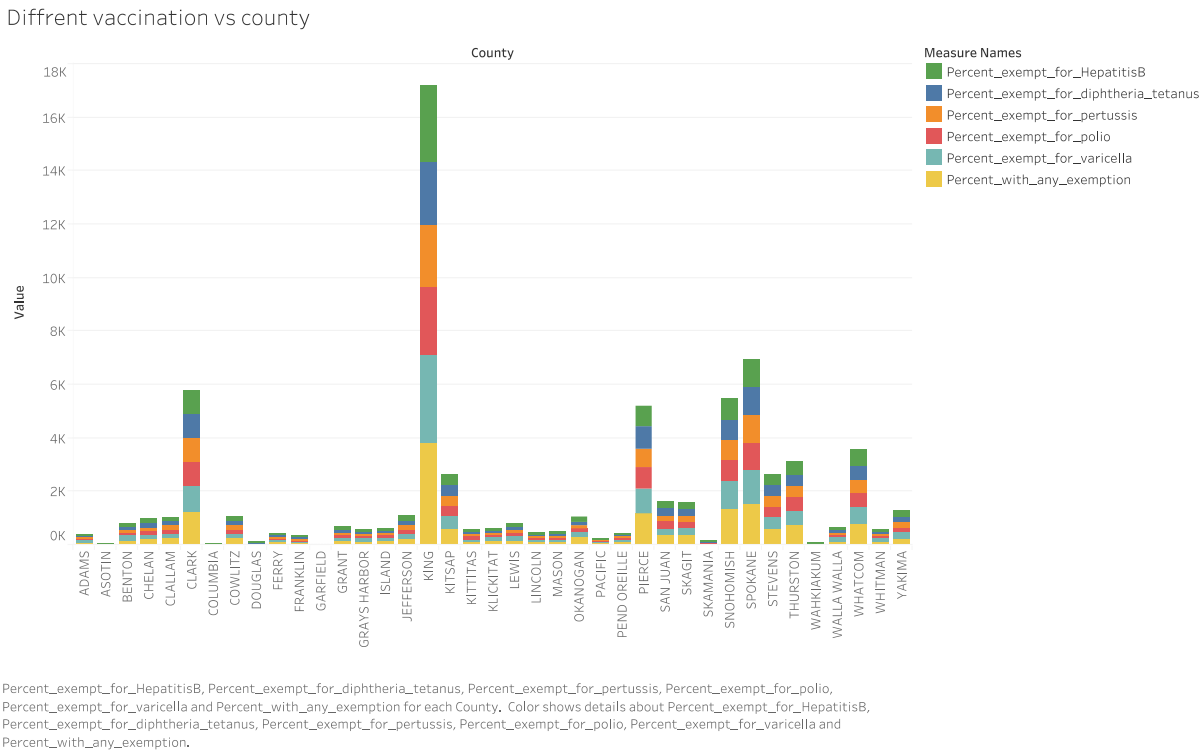
```
> #corelation
> cor(Schooldata$Percent_with_medical_exemption,Schooldata$Percent_with_personal_exemption)
[1] NA
>
```

We can see that there is no correlation between the attributes.

e) Bar graph



The highest religious exceptions is made from clark county and lowest from Lincoln district.



The bar graphs shows different vaccination exemptions in different county's

Explain/define terms:

Regression: Linear regression is used to predict the value of an outcome variable *Y* based on one or more input predictor variables *X*.

Correlation: C**orrelation analysis** is used to investigate the association between two or more variables

Works Cited

[1]All students, kindergarten through 12th grade, immunization data by school, 2016-2017,May 2017, https://data.wa.gov/Health/All-students-kindergarten-through-12th-grade-immun/9zru-c2kz