Big Data Case Study On – Student Immunization

Sai Shashank Vinnakota

G01221348

svinnak@gmu.edu

Submission Date: 12-16-2019

Subject: AIT-580 Big Data to Information

Professor: Zeehasham Rasheed

Contents

Introduction:

The Washington State Department of Health Works for a Safer and Healthier Washington. Their programs and services help prevent illness and injury, promote healthy places to live and work, provide education to help people make good health decisions and ensure our state is prepared for emergencies, helping prevent illness is a cornerstone of public health. They work to improve health through disease and injury prevention, immunization, and newborn screening for prenatal disease programs.
They work with many partners to provide educational and training programs as well as health and safety information to help people make healthy choices. **It was founded in 1989**. (Washington health facebook, n.d.) (washington state Report, n.d.)

Need:

The necessity of collecting this data was to identify the efficiency and effectiveness of the efforts taken by the Washington state department for immunization of children. Also, by using the information generated from this data, a clearer picture can be drawn which can help the organization reach the desired goal i.e. immunization of all children along with improving health of all residents. (Washington health facebook, n.d.)

List of potential questions that could be answered

- How many schools have got maximum immunizations done?
- Is religious exemption and Religious membership exemption same?
- Which county has lease exemption rate.
- How many have enrolled and how many have submitted?
- What can be the relationship between number of any exemptions made and total immunizations done
- How can we make analyses on few of the common and dangerous diseases?
- How each disease is related to each other? How can we analyze?

Requirements and Resources needed:

This data requires computing resources and software such as Microsoft Excel, Python (or any third-party application for python like PyCharm, Rodeo, etc.,), R-Studio, Tableau 10.5 and SQL Workbench. Hardware resources include compatible Operating System (such as Windows 10), at least 8gb Random Access Memory. Also, a significant amount of preprocessing is required in order to eliminate all the null values and blank cells.

Data set description: The dataset "School Immunization Record" consists of over his page is your guide to school immunization data in the state for school years 2016-2017, 2017-2018, and 2018-2019. Most schools and school districts are represented in the data. However, schools and school districts with less than 10 students for the 2018-2019 school year have been suppressed from public data for privacy reasons.

The list of columns in the csv file and their description is as follows:

- **School name**: Name of the school
- **School_year**: Records that have been registered in which year
- **Reported:** Parents have reported or not
- **K_12_enrollment**: How many students have got enrolled for Immunization
- **Percent complete for all immunization:** percentage of students who completed all immunizations
- **Percent with any exemptions**: Percentage of students with any exceptions
- **Percent with medical exemptions**: Percentage of students with medical exemptions
- **Percent with personal exemptions**: Percentage of students with personal exemptions
- **Percent with religious exemptions**: Percentage of students with religious exemptions
- **Percent with religious membership exemptions**: Percentage of students with religious membership exemptions
- **Percent exempt for diphtheria and tetanus**: Percentage of students with diphtheria and tetanus exemptions
- **Percent exempt for pertussis**: Percentage of students with pertuis exemption
- **Percent exempt for mmr**: Percentage of students with mmr exemption
- **Percent exempt for polio**: Percentage of students with polio exemption
- **Percent for hepatitis**: Percentage of students with Hepatitis exemption
- **Percent exempt for varicella**: Percentage of students with varicella exemptions
- **Number complete for all immunizations**: Number of students who completed all immunizations

- **Number with any exemptions**: Number of students with any exceptions
- **Number with medical exemption**: Number of students with medical exemptions
- **Number with personal exemption**: Number of students with personal exemptions
- **Number with religious exemption**: Number of students with religious exemptions
- **Number with religious membership exemption**: Number of students with religious membership exemptions
- **Number exempt for diphtheria tetanus**: Number of students with diphtheria and tetanus exemptions
- **Number exempt for pertussis**: Number of students with pertuiss exemption
- **Number exempt for mmr**: Number of students with mmr exemption
- **Number exempt for polio**: Number of students with polio exemption
- **Number exempt for hepatitis B**: Number of students with Hepatitis B exemption
- **Number exempt for varicella**: Number of students with Varicella
- **School _district**: District to which the school is located
- **County**: County to which the school is located
- **ESD**: To which EDUCATIONAL SERVICE DISTRICT the school belongs to
- **Grade level**: The grade level to which the school offers
- **Has Kindergarten**: Does the school have kindergarten (Yes/No)
- **Has 6th Grade**: Does the school have 6th Grade
- **Location:** location where the school has been constructed.

Issues:

Missed opportunities for administration of vaccines: A missed opportunity is defined as any healthcare provider office visit at which a child does not receive all needed vaccinations that could have been safely and appropriately administered. Most missed opportunities occur during acute care visits or health supervision and follow-up visits. Several studies have shown that vaccination coverage could be increased up to 20 percent by eliminating missed opportunities. (washington state Report, n.d.)

Challenges to receiving vaccinations on time: Many children do not receive all vaccinations within the recommended time frame. A recent study has shown that up to half of babies in the United States miss or delay at least some of the recommended vaccinations. Other studies have shown that over one-third of children were under vaccinated for over six months and that children in western states have lower rates of on time vaccinations than children in eastern states. (washington state Report, n.d.)

Barriers:
Barriers to immunizations include physical ones such as distance to travel, transportation problems, and inconvenient clinic hours and waiting times, as well as psychological barriers such as unpleasant experiences, complexity and continual change of the immunization

schedule, and concerns and fears over vaccine safety.15 Also of importance are family financial barriers to accessing primary care coupled with the situation that some providers are stopping providing immunizations due to diminishing reimbursements for those services. (washington state Report, n.d.)

Protective factors: Data indicates that children 19–35 months old who are the firstborn child are more likely to be fully immunized with age appropriate recommended vaccines than children who have older siblings. These findings suggest it may be helpful for immunization providers to pay attention to the vaccination status of children who are not firstborn. NIS data also suggest that children whose mothers are married or who have never been married are more likely to be fully vaccinated than children whose mothers are widowed, divorced, separated or deceased. (washington state Report, n.d.)

Results/Findings

```
> str(Da)
Classes 'tbl_df', 'tbl' and 'data.frame':       2476 obs. of  35 variables:
 $ School_Name                                : chr  "GRANGER MIDDLE SCHOOL" "MUKILTEO ACADEMY" "SACRED HEART SCHOOL" "ST. CHARLES BORROMEO
 SCHOOL" ...
 $ School_year                                : chr  "2016-17" "2016-17" "2016-17" "2016-17" ...
 $ Reported                                   : chr  "Y" "Y" "Y" "Y" ...
 $ K_12_enrollment                            : num  480 23 370 519 526 ...
 $ Percent_complete_for_all_immunizations     : num  99.2 100 97.8 99.4 94.3 95.6 70.2 94.2 100 61.6 ...
 $ Percent_with_any_exemption                 : num  0.2 0 2.2 0.6 4.4 2.1 12.8 2 0 0.6 ...
 $ Percent_with_medical_exemption             : num  0 0 0 0.2 1.7 1.1 2.1 0.4 0 0 ...
 $ Percent_with_personal_exemption            : num  0 0 2.2 0.4 1.7 0.9 6.4 1.5 0 0.6 ...
 $ Percent_with_religious_exemption           : num  0.2 0 0 0 1 0.1 4.3 0 0 0 ...
 $ Percent_with_religious_membership_exemption: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Percent_exempt_for_diphtheria_tetanus      : num  0 0 1.1 0.4 3.4 0.7 6.4 1 0 0.3 ...
 $ Percent_exempt_for_pertussis               : num  0 0 1.1 0.4 3.4 0.6 6.4 0.7 0 0 ...
 $ Percent_exempt_for_measles_mumps_rubella   : num  0 0 1.4 0.4 3.6 0.9 6.4 0.7 0 0.3 ...
 $ Percent_exempt_for_polio                   : num  0 0 1.4 0.4 3.4 0.6 4.3 0.9 0 0.3 ...
 $ Percent_exempt_for_HepatitisB              : num  0 0 2.2 0.6 2.9 0.6 4.3 0.7 0 0.3 ...
 $ Percent_exempt_for_varicella               : num  0.2 0 1.4 0.4 4 1.5 10.6 0.9 0 0.3 ...
 $ Number_complete_for_all_immunizations      : num  476 23 362 516 496 ...
 $ Number_with_any_exemption                  : num  1 0 8 3 23 47 6 46 0 2 ...
 $ Number_with_medical_exemption              : num  0 0 0 1 9 24 1 10 0 0 ...
 $ Number_with_personal_exemption             : num  0 0 8 2 9 20 3 34 0 2 ...
 $ Number_with_religious_exemption            : num  1 0 0 0 5 3 2 1 0 0 ...
 $ Number_with_religious_membership_exemption : num  0 0 0 0 0 1 0 1 0 0 ...
 $ Number_exempt_for_diphtheria_tetanus       : num  0 0 4 2 18 16 3 23 0 1 ...
 $ Number_exempt_for_pertussis                : num  0 0 4 2 18 14 3 15 0 0 ...
 $ Number_exempt_for_measles_mumps_rubella    : num  0 0 5 2 19 20 3 16 0 1 ...
 $ Number_exempt_for_polio                    : num  0 0 5 2 18 14 2 20 0 1 ...
 $ Number_exempt_for_HepatitisB               : num  0 0 8 3 15 13 2 16 0 1 ...
 $ Number_exempt_for_varicella                : num  1 0 5 2 21 33 5 21 0 1 ...
 $ School_District                            : chr  "GRANGER SCHOOL DISTRICT" "MUKILTEO SCHOOL DISTRICT" "BELLEVUE SCHOOL DISTRICT" "TACOMA
 SCHOOL DISTRICT" ...
 $ County                                     : chr  "YAKIMA" "SNOHOMISH" "KING" "PIERCE" ...
 $ ESD                                        : chr  "EDUCATIONAL SERVICE DISTRICT 105" "NORTHWEST EDUCATIONAL SERVICE DISTRICT 189" "PUGET
 SOUND EDUCATIONAL SERVICE DISTRICT 121" "PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121" ...
 $ Grade_Levels                               : chr  "5-8" "P-1" "P-8" "P-8" ...
 $ Has_kindergarten                           : chr  "N" "Y" "Y" "Y" ...
 $ Has_6thGrade                               : chr  "Y" "N" "Y" "Y" ...
 $ Location 1                                 : chr  "701 E AVENUE\nGRANGER\n(46.347133, -120.18824)" "13000 BEVERLY PARK RD\nMUKILTEO\n(47.
```

SQL Schema

To Create Table:

Worksheet Query Builder

```
drop table All_students__kindergarten_through_12th_grade__immunization_data_by_school__2016_2017;
create table All_students__kindergarten_through_12th_grade__immunization_data_by_school__2016_2017
(
School_Name varchar(200),
School_year varchar(7),
Reported char(1),
K_12_enrollment number(4),
Percent_complete_for_all_immunizations  decimal(4,1),
Percent_with_any_exemption  decimal(3,1),
Percent_with_medical_exemption  decimal(3,1),
Percent_with_personal_exemption decimal(3,1),
Percent_with_religious_exemption decimal(3,1),
Percent_with_religious_membership_exemption decimal(3,1),
Percent_exempt_for_diphtheria_tetanus decimal(4,1),
Percent_exempt_for_pertussis    decimal(4,1),
Percent_exempt_for_measles_mumps_rubella    decimal(4,1),
Percent_exempt_for_polio    decimal(4,1),
Percent_exempt_for_HepatitisB   decimal(4,1),
Percent_exempt_for_varicella    decimal(3,1),
Number_complete_for_all_immunizations   number(4),
Number_with_any_exemption   number(3),
Number_with_medical_exemption   number(3),
Number_with_personal_exemption  number(3),
Number_with_religious_exemption     number(2),
Number_with_religious_membership_exemption  number(2),
Number_exempt_for_diphtheria_tetanus    number(3),
Number_exempt_for_pertussis     number(3),
Number_exempt_for_measles_mumps_rubella     number(3),
Number_exempt_for_polio     number(3),
Number_exempt_for_HepatitisB        number(3),
Number_exempt_for_varicella     number(3),
School_District varchar(150),
County  varchar(25),
ESD varchar(150),
Grade_Levels    varchar(20),
Has_kindergarten char(1),
Has_6thGrade    char(1),
Location1 varchar(300)
);
```

QUERY 1:

```
select School_Name, K_12_enrollment, Number_complete_for_all_immunizations
from All_students__kindergarten_through_12th_grade__immunization_data_by_school__2016_2017
order by K_12_enrollment;
```

Script Output | Query Result

SQL | Fetched 50 rows in 0.048 seconds

	SCHOOL_NAME	K_12_ENROLLMENT	NUMBER_COMPLETE_FOR_ALL_IMMUNIZATIONS
1	RE-ENTRY MIDDLE SCHOOL	0	0
2	THE CHILDREN'S INN ACADEMY	0	0
3	SKILSKIN HIGH SCHOOL	0	0
4	BEST NIGHT SR HIGH SCHOOL	0	0
5	MONTESANO LEARNING ACADEMY	0	0
6	NORTHERN LIGHTS MONTESSORI SCHOOL	0	0
7	DARTMOOR SCHOOL--SEATTLE	1	1
8	NORTH COUNTRY CHRISTIAN SCHOOL	1	1
9	TROJAN ALTERNATIVE SCHOOL	1	1
10	BK PLAY ACADEMY FOR GIFTED CHILDREN	2	2
11	UNIVERSITY PLACE SPECIAL EDUC	2	0
12	RE-ENTRY HIGH SCHOOL	2	2
13	ALCUIN SCHOOL	2	2
14	DARTMOOR SCHOOL--WOODINVILLE	2	2
15	LITTLE STAR MONTESSORI SCHOOL	3	2
16	SPECIAL EDUCATION SCHOOL	3	3
17	TOUCHSTONE SCHOOL	3	0
18	DISCOVERY DEPOT MONTESSORI SCHOOLHOUSE	3	2
19	THE MOOSE PROJECT	3	3
20	DECATUR ELEMENTARY	4	3
21	GOOD SHEPHERD MONTESSORI	4	1
22	SEATTLE MINI MEDICAL SCHOOL	5	5
23	EVERGREEN ACADEMY OF ARTS & SCIENCES	6	6
24	DOLAN ACADEMY & LEARNING CENTER	6	6
25	DARTMOOR SCHOOL--ISSAQUAH	6	6
26	CHEWELAH ALTERNATIVE	6	3
27	IMAGINATION SCHOOL OF EDUCATION	6	6
28	SPRING ACADEMY	7	6
29	LES LILAS FRENCH BILINGUAL COMMUNITY SCHOOL	7	7
30	COMMUNITY MONTESSORI	7	7
31	BRIGHTMONT ACADEMY--SEATTLE CAMPUS	7	7
32	HOLDEN VILLAGE COMMUNITY SCHOOL	7	4
33	LAKESHORE MONTESSORI SCHOOL	7	3
34	GRAYS HARBOR ADVENTIST CHRISTIAN SCHOOL	7	6
35	SOLVE FOR X SCHOOL	7	7

From the above query we can see that the number of students that registered and the number that have completed all the immunizations. In few cases we can observe that few people have neglected their immunizations as the number of enrollments is greater than number of students who have completed their vaccinations completed.

Query 2

```
select count(Reported) as Reported_Yes, count(Has_kindergarten) as Kindergarden_Yes from All_students_kindergarten_through_12th_grade_immunization_data_by_school_2016_2017
where Reported = 'Y' and Has_kindergarten = 'Y';
```

Output:

Script Output | Query Result

SQL | All Rows Fetched: 1 in 0.581 seconds

	REPORTED_YES	KINDERGARDEN_YES
1	1533	1533

From the above query we can conclude that the total number of people who have reported have kindergarten in their school.

Visualization1: Scatter plot

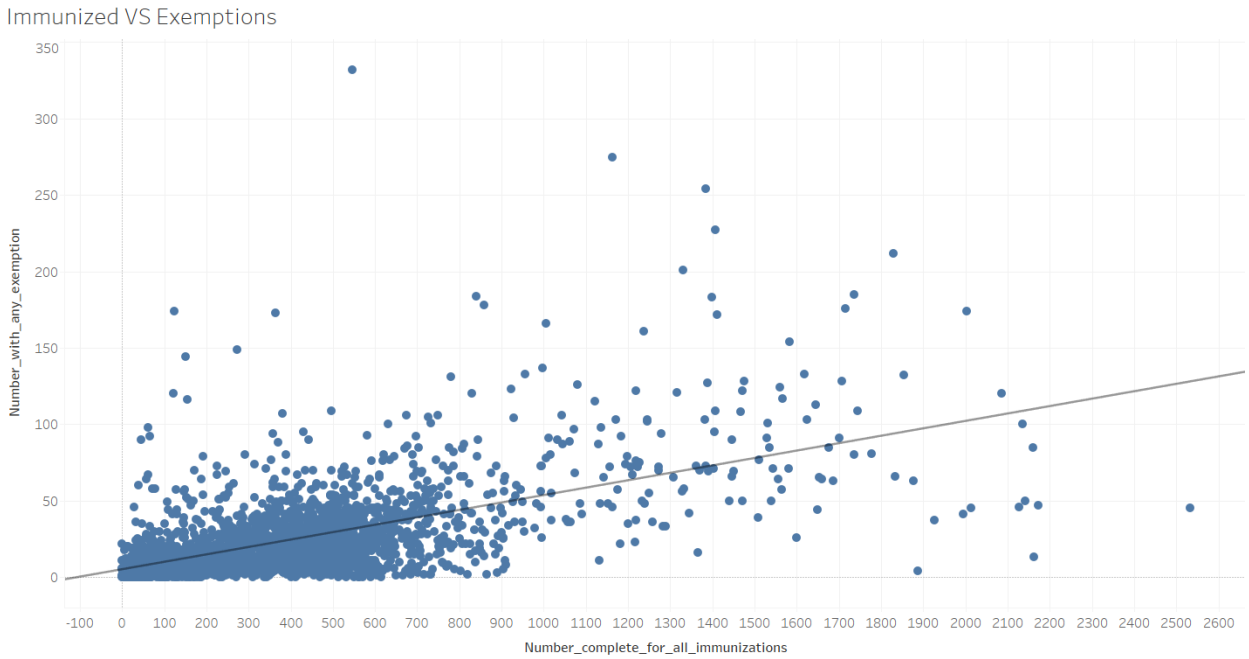


Fig 1

The above graph (fig 1) shows the scatter plot between total number of completed immunizations for all the student's vs total number of any exemptions for vaccinations that are given to students.

We can observe that the highest number of students who were exempted for any reason is around 330 and the lowest falls to 45, whereas on the flip side least number of students who have completed the all vaccinations reaches up to 546 and highest makes it to 2532 students.

The dense cloud indicates the students with all vaccinations completed range between 0-800 and students with any exemptions range between 0-25.

This shows that the total number of students who are vaccinated are more than the students that haven been exempted for any reason possible.

Visualization2: Boxplot

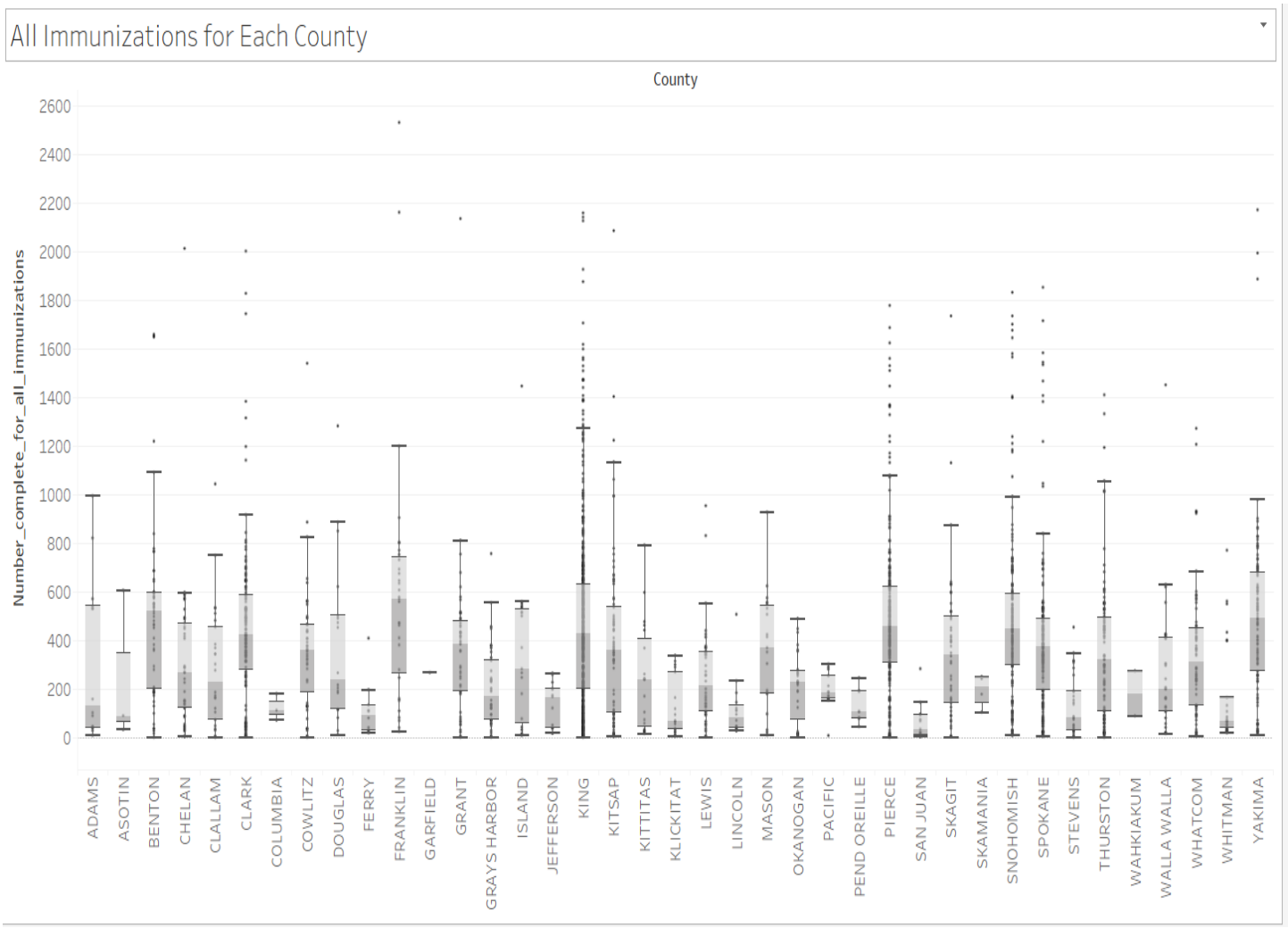


Fig 2

From fig 2 we can get an inkling as to which county has the best immunization record. For which I have taken a boxplot graph into consideration which can easily tell that which county the highest number of vaccinations have been completed, with the numbers range that are defined in the given quartiles.

The outliers help in finding out the maximum and minimum count in each county. It can be found that KING County has the most vaccinations registered for all the students that way ahead of all other counties and County FERRY has the least number of students who have got vaccinated.

Visualization 3: Bar Graph

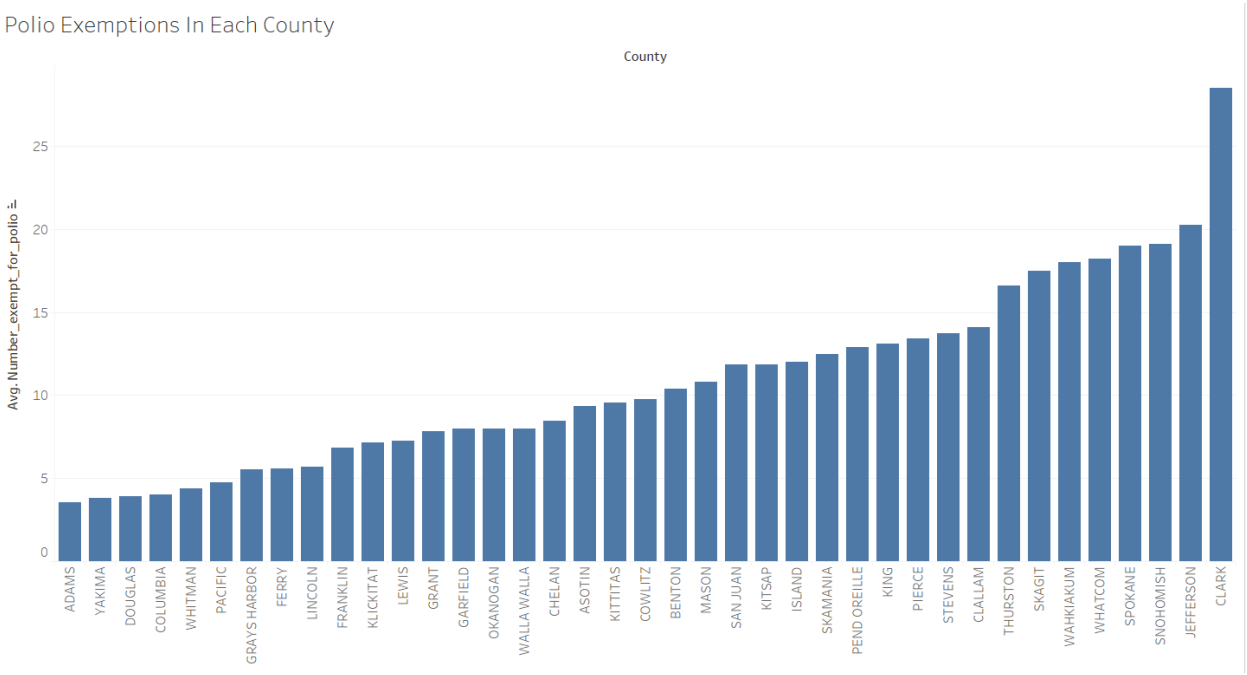


Fig 3

Fig 3 illustrates one of the most important Vaccine that has affected most of the children due to neglecting the dosage i.e. polio. In order to understand which county, the best and worst immunized records belong to, I have plotted a bar graph which clearly shows that Clark county has the highest exemptions and Adams country has the least exempted.

As observed in the graph the counties to the right have a high number of exemptions for polio, hence the graph is left skewed.

Analysis:

<u>Linear Regression</u>

I have tried out linear regression in order to get the relationship between the predictor variable Enrollment and all independent variables in the data set.

On running the linear Regression, I have found out the following:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2809.75  -151.32   -40.26    92.46  1748.34

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               224.0315    36.5584   6.128 1.03e-09 ***
Da$Percent_complete_for_all_immunizations   1.2850     0.3921   3.277  0.00106 **
Da$Percent_with_any_exemption              15.1280    11.0281   1.372  0.17026
Da$Percent_with_medical_exemption         -31.5469    10.7467  -2.935  0.00336 **
Da$Percent_with_personal_exemption        -20.7126    10.8446  -1.910  0.05626 .
Da$Percent_with_religious_exemption       -29.3766    11.2692  -2.607  0.00919 **
Da$Percent_with_religious_membership_exemption -65.6333    15.1147  -4.342 1.47e-05 ***
Da$Number_exempt_for_diphtheria_tetanus     9.1573     1.3728   6.671 3.14e-11 ***
Da$Number_exempt_for_pertussis              2.2375     1.5611   1.433  0.15191
Da$Percent_exempt_for_pertussis             3.2534     3.0059   1.082  0.27920
Da$Percent_exempt_for_polio               -19.0524     2.8464  -6.694 2.69e-11 ***
Da$Number_exempt_for_HepatitisB             0.5277     1.1268   0.468  0.63958
Da$Number_exempt_for_varicella              4.3561     0.3633  11.991  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 269.6 on 2463 degrees of freedom
Multiple R-squared:  0.5114,    Adjusted R-squared:  0.509
F-statistic: 214.8 on 12 and 2463 DF,  p-value: < 2.2e-16
```

The best predictors for enrollment are the ones with p-value less that 0.05 as we have taken confidence interval as 5%, and more than 0.05 are bad predictors.

Over all model is significant as the p value overall is 2.2e-16

Bad Predictors

Percentage with any exemption

Percentage with personal exemption

Percent exempt for pertussis

Number of exempts for pertussis

Adjusted R2 is 50 that means the model is an average model. This can be improved by taking into consideration of transforming or using log function.

Random Forest:

I have tried out Random forest in order to get to know better results for the predictors and getting a better Adjusted R square value or Variance value.

```
Call:
 randomForest(formula = Da$K_12_enrollment ~ Da$Percent_complete_for_all_immunizations +      Da$Percent_with_any_exemption + Da$Per
cent_with_medical_exemption +      Da$Percent_with_personal_exemption + Da$Percent_with_religious_exemption +      Da$Percent_with_r
eligious_membership_exemption + Da$Number_exempt_for_diphtheria_tetanus +      Da$Number_exempt_for_pertussis + Da$Percent_exempt_fo
r_pertussis +      Da$Percent_exempt_for_polio + Da$Number_exempt_for_HepatitisB +      Da$Number_exempt_for_varicella, data = Da, m
try = 4, ntree = 5000,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 5000
No. of variables tried at each split: 4

          Mean of squared residuals: 10749.78
                    % Var explained: 92.74
```

We can clearly see that the model has a very good % Variance i.e. 92.47 which far better in when compared to the linear regression output.

K-Means

MMR being one of the commonly repeating disease and serious one, I have tried to get the plot out this analysis using cluster analysis (K=3).

The output shows that total number of Exceptions for MMR doesn't go high with the increase in number of any exemptions, except for few cases here and there, hence we can say that the negligence for MMR is less and not on a high scale.

Hypothesis Testing:

```
        Two Sample t-test

data:  Da$Number_with_religious_exemption and Da$Number_with_religious_membership_exemption
t = 17.444, df = 4950, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.844953 1.058924
sample estimates:
mean of x mean of y
1.3546042 0.4026656
```

I have tried out two variables that have almost the same meaning so hence I was interested in knowing if they were by any chance equal.

H0(null hypothesis): Number with religious exemption and Number with religious membership exemption are equal.

H1(alternative hypothesis): Number with religious exemption and Number with religious membership exemption are not equal.

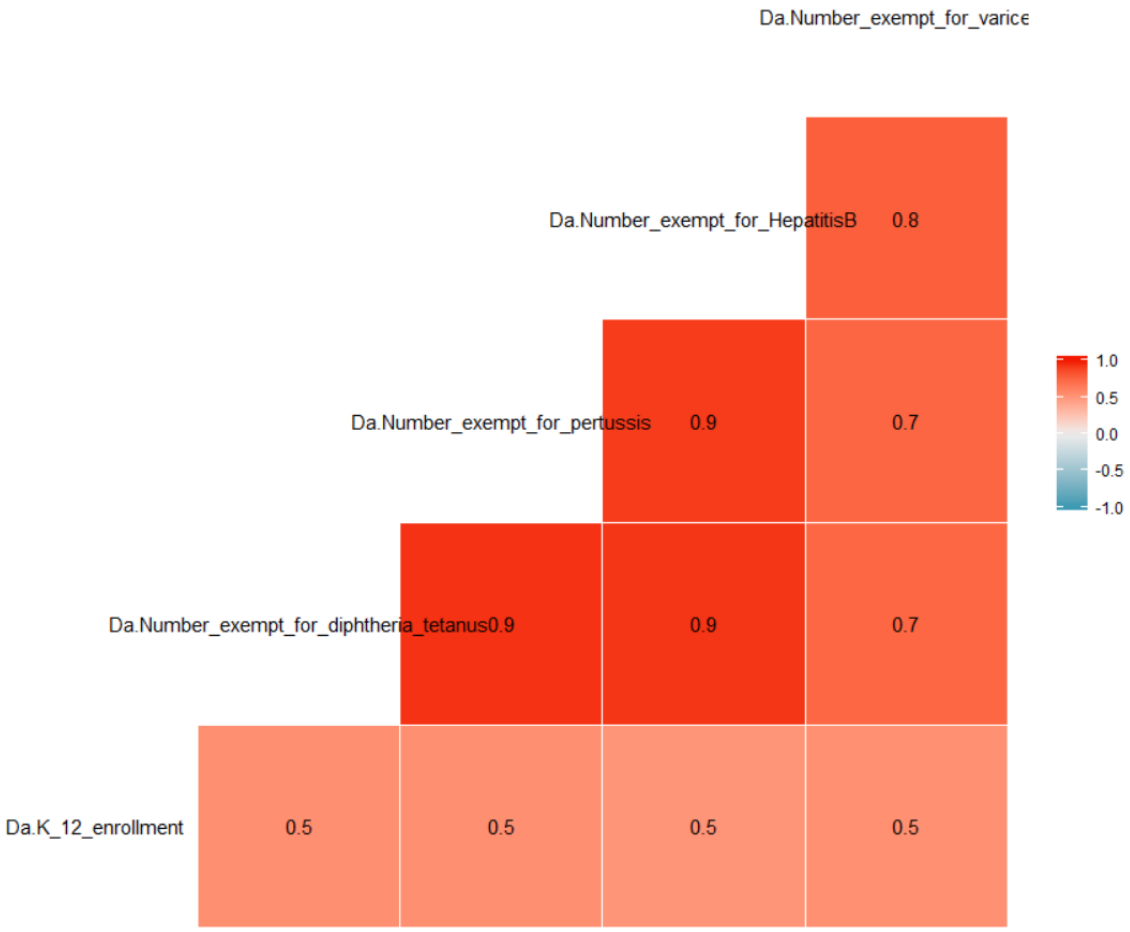
For that I have tried out the hypothesis testing that clearly shows that P-Value is less than 0.05, that mean null hypothesis is not rejected.

Correlation Matrix:

I have found out the correlation matrix with the predicted variable and few other dependent variables.

From the below graph we can clearly make out that all the independent variables have good correlation between each other but have the same correlation with the predictor variable with 0.5 that is a good enough relation between them of 0.5 where highest being 1 and lowest going -1.

`



Glossary:

1. **Linear regression:** it is way of predicting the future using linear model of regression where you have 1 predictor variable and all other variable fall into independent category, these independent variables help in predicting the dependent (predictor) variable. The relation between them tells the significance of the model.

   Regression formula: Y=Mx1+Mx2+……+C

   Y= predictor

   X= Independent variables

   M= slope

   C= Constant

2. **K means Clustering**: It is one of the techniques to group the similar kind of observations together and made into clusters in order.
   This makes it easy for understanding about what type of observations are similar and analysis becomes easy. We need to define the number of clusters to be used by K.

3. **Hypothesis Testing:** Hypothesis testing is used to get the ambiguity of a situation to a solution. Null hypothesis is where we are considering the present scenario and alternative hypothesis is what has been questioned on the null hypothesis. We prove that either null hypothesis is accepted, or null hypothesis is rejected.

4. **Correlation Matrix:** correlation matrix helps in getting to know how the variables are related to each other, by getting to know the relation between the variables, we can have a clear idea on what variables can we have analysis done.

5. **Random forest:** Random forest is an ensemble learning method classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

References

Washington health facebook. (n.d.). Retrieved from https://www.facebook.com/pg/WADeptHealth/about/?ref=page_internal

washington state Report. (n.d.). Retrieved from https://www.doh.wa.gov/DataandStatisticalReports/HealthBehaviors/HealthofWashingtonStateReport/MostRecentReport