# AIT 580
# FINAL PROJECT

By

DEEPTHI TAMMA

G01241465

**All students, kindergarten through 12th grade, immunization data by school, 2016-2017**

**Introduction:**
This dataset contains the information regarding the immunization records through all grades i.e., from Kindergarten to 12th standard for various public and private schools of the Washington state. It contains the percentage as well as number of the total number of different kinds of vaccines given to the school as well as those who have negotiated to take the vaccinations due to the various reasons and all these are reported by the Parent reports to the school. A study of the Immunization records can provide the assess to know the scope of practice and quality of care that paves a path to identify opportunities for the development of models of collaboration between the public and the private health sector. [1]

**Who:**
Eavey Joanna (DOH) is the one who has collected the data and reported the data to the Office of Immunization and Child Profile, Department of Health by 12/31/2015.
Need:
The primary purpose of collecting this data is to which find the percent of vaccinations taken by the students of each school which could also help protecting children and communities from the contagious diseases. This could help in reducing the risk from the vaccine preventable diseases. Some of the questions that can be answered for this dataset can be done with the help of visualizing and modeling this data.

- To know on an average how many of students of that schools got all the vaccinations (Complete) as well as if the school has been reported or not?
- Does there is any relationship between the number of students who have signed for the exemption for religious exemption and religious- membership exemption?
- What is the median number of schools are done with all the immunizations who has Kindergarten?
- Is there any relation between the number of members exempt for the polio for all schools and number of members exempt for Hepatitis B?

**Quality Issues & Privacy Issues:**
Data Quality issues needs to be resolved as accurately as possible because they may lead to the misleading or the invalid conclusions. Some of the data quality issues may be due to missing data values, duplicate data or the inconsistent data. This is occurred due to the possibility of occurrence of noise or the poorly defined data. These can be identified, and it can be corrected with many ways and this process is known as Data Cleaning. Some of them are

- To completely eliminate the null values
- To replace the null values with the average value of that column for the numerical data attribute

Here, in this dataset, I have removed the null values with the help of the omit function in R studio. There is no privacy issue with this dataset according to my knowledge because it is accessible to everyone.

**Requirements & Resources Needed:**
**Hardware & Software Requirements:**
**Hardware Requirements:**
The entire project is done using the laptop which are having the configurations that include

- Name of the Processor- 1.4 GHz Quad-Core Intel Core i5
- Memory- 8 GB 2133 MHz LPDDR3
- Graphics- Intel Iris Plus Graphics 645 1536 MB
- Operating System- macOS Catalina

**Software Requirements:**

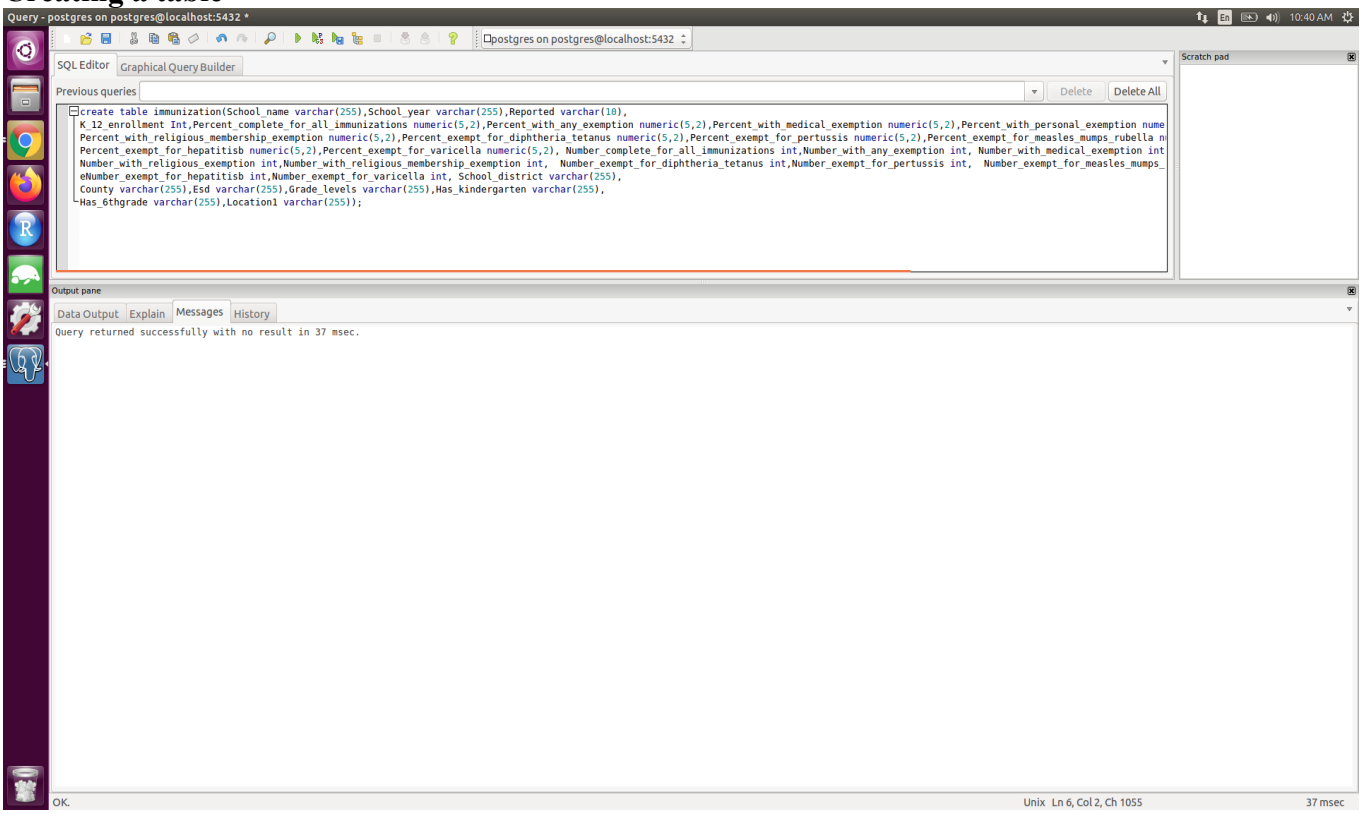
- R Studio is an open source is used for the data modelling.
- Tableau is also used for the visualization of the dataset.
- Virtual Machine is used building the dataset schema and also useful for querying in the PostgreSQL
- Anaconda (Jupyter Notebook) is used in order to find the regression models and the scatter plots.

**Description of the Dataset:**
The size of the dataset is 636kb with 2595 rows and 35 columns. This dataset describes about the total and Percentage of students who has taken the vaccinations as well as who has not taken due to the exceptions for the relevant schools (Name of the schools given). The data includes if that students of that school has been from kindergarten or not as well it checks if they have completed the $6^{th}$ grade. The dataset also has the information whether they have been reported or not. The exemptions for the medications include the religious, personal, medical, religious-membership and many more. It also gives the count of those exemptions for those schools.
Here, the complete indicates that the student met with all the requirements of the schools and the corresponding grade. For conditional, the students don't provide the appropriate documentation for the necessary immunizations. For the exempt, it means that the student has signed the document of exempt fopr the particular documentations. The datatype of the attributes are as follows:
School_name – Nominal (Eg: HEIGHTS ELEMENTARY, P C JANTZ ELEMENTARY)
School_year- Ordinal (E.g.: 2016-17)
Reported- Nominal (E.g., Y, N)
K_12_enrollment- Numerical (Discrete) (E.g., 0,480,125)
Percent_complete_for_all_immunizations -Numerical (Continuous) E.g., 90,99.2,100
Percent_with_any_exemption Numerical (Continuous) E.g.,0,0.2,3.2
Percent_with_medical_exemption – Numerical (Continuous) E.g.,0.1,12.5,0
Percent_with_personal_exemption– Numerical (Continuous) E.g., 0,0.5,0.4
Percent_with_religious_exemption – Numerical (Continuous) E.g.,3.8,0,0.4
Percent_with_religious_membership_exemption– Numerical (Continuous) E.g.,0.1,0.5,0.2
Percent_exempt_for_diphtheria_tetanus– Numerical (Continuous) E.g.,7,12.5, 2.9
Percent_exempt_for_pertussis– Numerical (Continuous) E.g.,1.1,0,5.4
Percent_exempt_for_measles_mumps_rubella– Numerical (Continuous) E.g., 6.4,0.7,0
Percent_exempt_for_polio– Numerical (Continuous) E.g.,12,13.5,0.4
Percent_exempt_for_hepatitisb – Numerical (Continuous) E.g.,4.4,0,12.5
Percent_exempt_for_varicella – Numerical (Continuous) E.g.,0,0.6,6.1
Number_complete_for_all_immunizations – Numerical (Discrete) E.g., 476,120,521
Number_with_any_exemption - Numerical (Discrete) E.g.,1,4,19

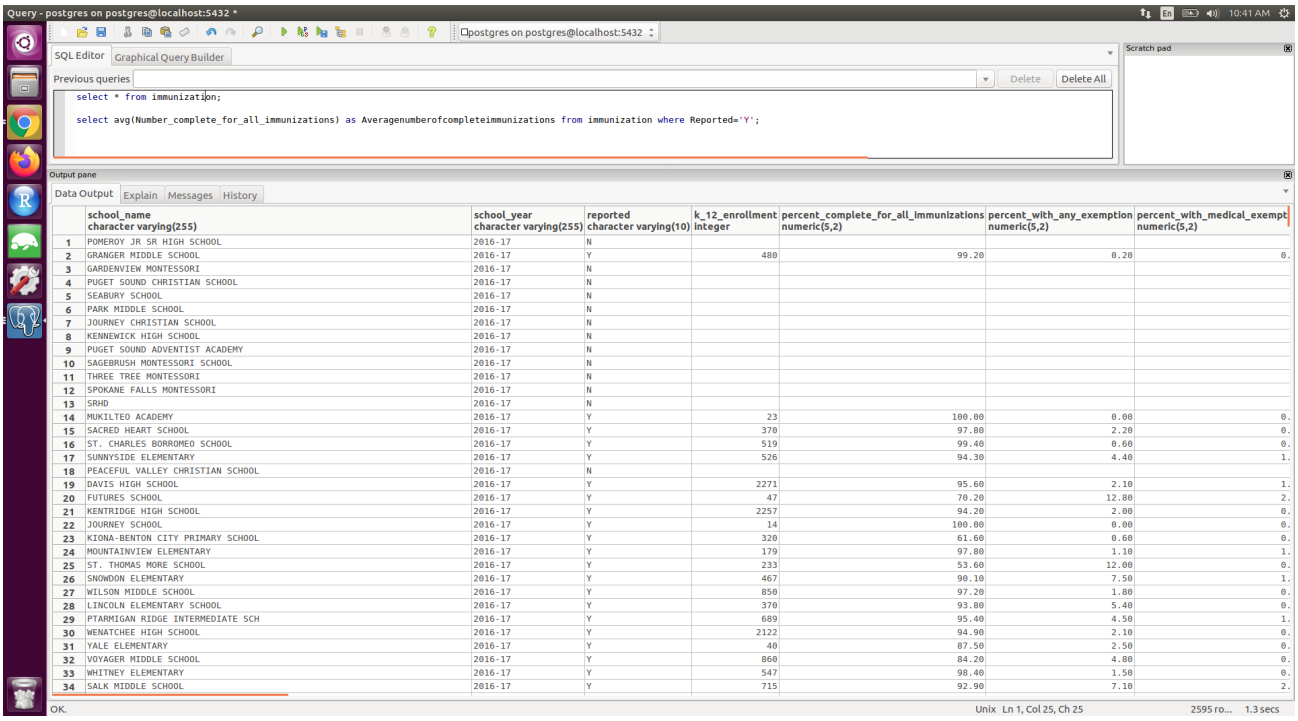
Number_with_medical_exemption – Numerical (Discrete) E.g.,10,24,1
Number_with_personal_exemption – Numerical (Discrete) E.g.,18,12,9
Number_with_religious_exemption – Numerical (Discrete) E.g.,2,5,8
Number_with_religious_membership_exemption- Numerical (Discrete) E.g.,1,2,0
Number_exempt_for_diphtheria_tetanus- Numerical (Discrete) E.g.,4,8,10
Number_exempt_for_pertussis – Numerical (Discrete) E.g.,8,11,14
Number_exempt_for_measles_mumps_rubella – Numerical (Discrete) E.g.,3,10,14
Number_exempt_for_polio – Numerical (Discrete) E.g.,2,17,12
Number_exempt_for_hepatitisb – Numerical (Discrete) E.g.,8,11,15
Number_exempt_for_varicella – Numerical (Discrete) E.g., 21,41,16
School_district – Categorical (E.g., CLARKSTON SCHOOL DISTRICT, SPOKANE SCHOOL DISTRICT)
County – Categorical (E.g., WHATCOM KING)
Esd- Nominal (E.g., NORTHWEST EDUCATIONAL SERVICE DISTRICT 189)
Grade_levels – Nominal (E.g., PK-8, K-12,PK-6)
Has_kindergarten- Categorical (E.g., Y,N)
Has_6thgrade- Nominal ( E.g.,Y,N)
Location1 – Text (E.g., 1801 NE 53RD ST TACOMA (47.304686, -122.423352) ) [2]

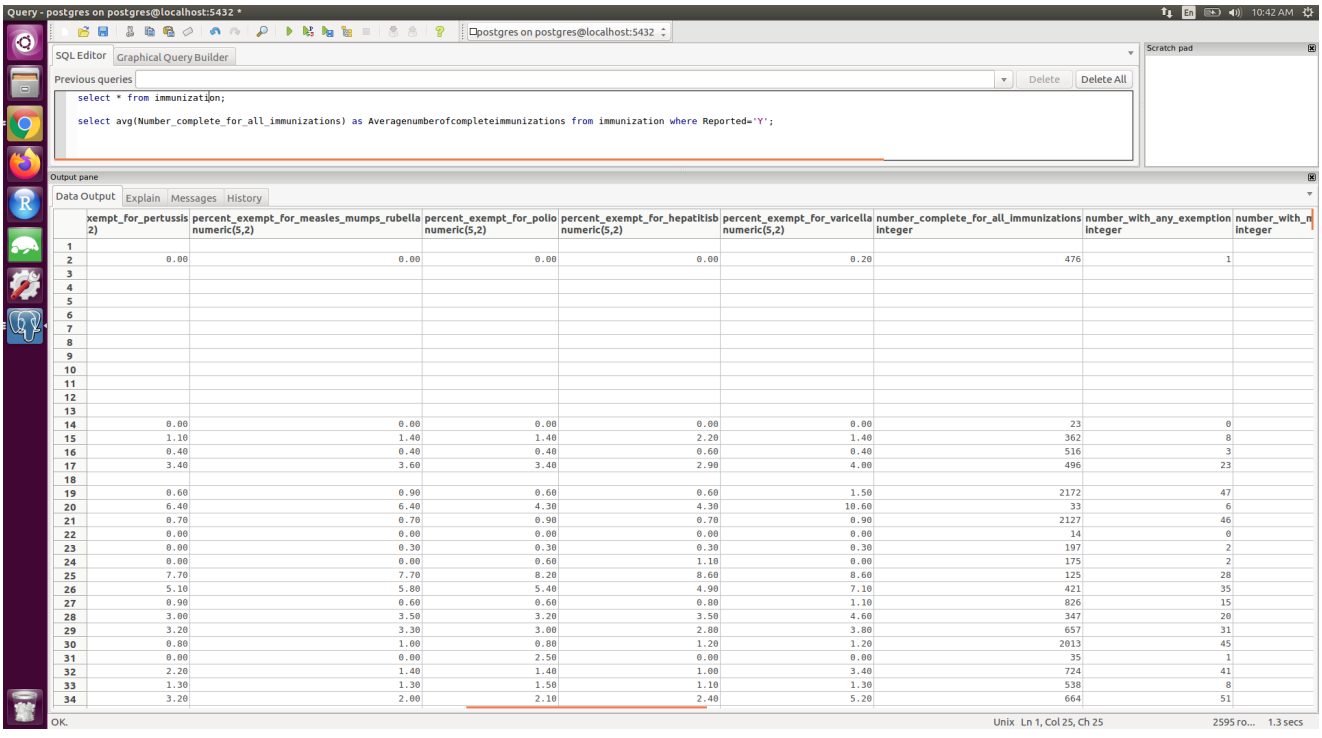
**SQL Schema:**

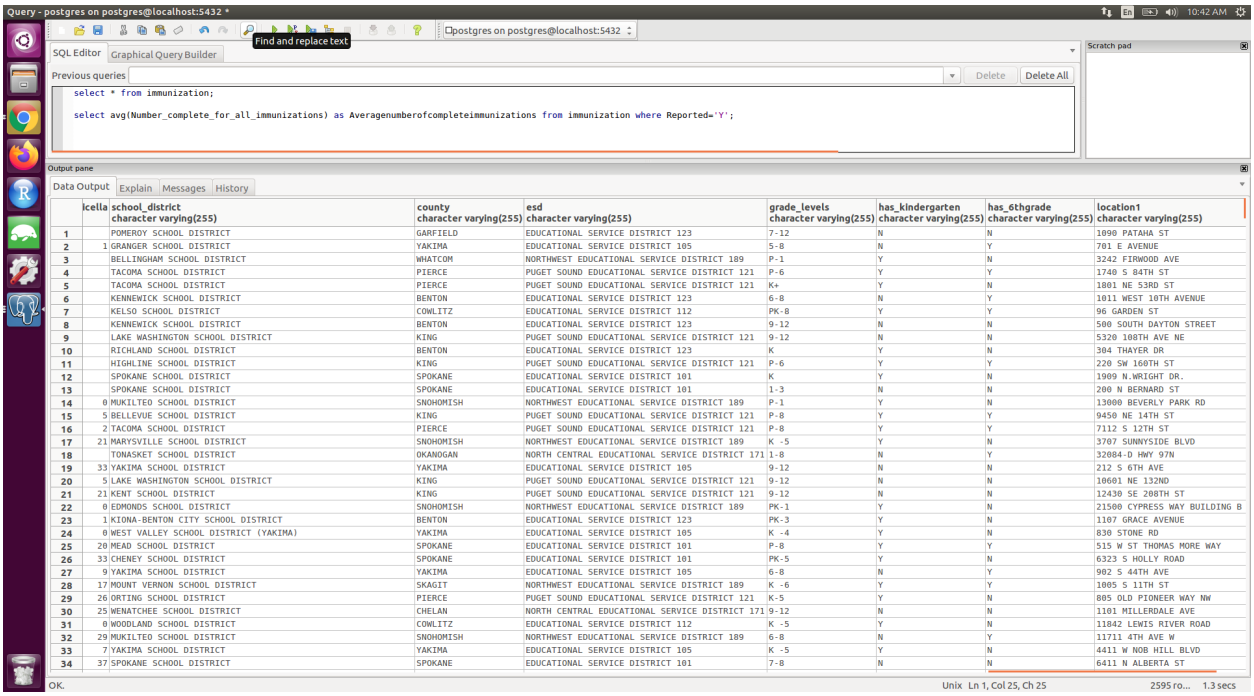
**Creating a table**




After that the file needs to be imported and then we can display table with the help of queries


Query - postgres on postgres@localhost:5432 *
10:41 AM
postgres on postgres@localhost:5432
SQL Editor
Graphical Query Builder
Previous queries
Delete
Delete All
Scratch pad
select * from immunization;
select avg(Number_complete_for_all_immunizations) as Averagenumberofcompleteimmunizations from immunization where Reported='Y';
Output pane
Data Output
Explain
Messages
History
school_name character varying(255)
school_year character varying(255)
reported character varying(10)
k_12_enrollment integer
percent_complete_for_all_immunizations numeric(5,2)
percent_with_any_exemption numeric(5,2)
percent_with_medical_exempt numeric(5,2)
1 POMEROY JR SR HIGH SCHOOL 2016-17 N
2 GRANGER MIDDLE SCHOOL 2016-17 Y 480 99.20 0.20 0.
3 GARDENVIEW MONTESSORI 2016-17 N
4 PUGET SOUND CHRISTIAN SCHOOL 2016-17 N
5 SEABURY SCHOOL 2016-17 N
6 PARK MIDDLE SCHOOL 2016-17 N
7 JOURNEY CHRISTIAN SCHOOL 2016-17 N
8 KENNEWICK HIGH SCHOOL 2016-17 N
9 PUGET SOUND ADVENTIST ACADEMY 2016-17 N
10 SAGEBRUSH MONTESSORI SCHOOL 2016-17 N
11 THREE TREE MONTESSORI 2016-17 N
12 SPOKANE FALLS MONTESSORI 2016-17 N
13 SRHD 2016-17 N
14 MUKILTEO ACADEMY 2016-17 Y 23 100.00 0.00 0.
15 SACRED HEART SCHOOL 2016-17 Y 370 97.80 2.20 0.
16 ST. CHARLES BORROMEO SCHOOL 2016-17 Y 519 99.40 0.60 0.
17 SUNNYSIDE ELEMENTARY 2016-17 Y 526 94.30 4.40 1.
18 PEACEFUL VALLEY CHRISTIAN SCHOOL 2016-17 N
19 DAVIS HIGH SCHOOL 2016-17 Y 2271 95.60 2.10 1.
20 FUTURES SCHOOL 2016-17 Y 47 70.20 12.80 2.
21 KENTRIDGE HIGH SCHOOL 2016-17 Y 2257 94.20 2.00 0.
22 JOURNEY SCHOOL 2016-17 Y 14 100.00 0.00 0.
23 KIONA-BENTON CITY PRIMARY SCHOOL 2016-17 Y 320 61.60 0.60 0.
24 MOUNTAINVIEW ELEMENTARY 2016-17 Y 179 97.80 1.10 1.
25 ST. THOMAS MORE SCHOOL 2016-17 Y 233 53.60 12.00 0.
26 SNOWDON ELEMENTARY 2016-17 Y 467 90.10 7.50 1.
27 WILSON MIDDLE SCHOOL 2016-17 Y 850 97.20 1.80 0.
28 LINCOLN ELEMENTARY SCHOOL 2016-17 Y 370 93.80 5.40 0.
29 PTARMIGAN RIDGE INTERMEDIATE SCH 2016-17 Y 689 95.40 4.50 1.
30 WENATCHEE HIGH SCHOOL 2016-17 Y 2122 94.90 2.10 0.
31 YALE ELEMENTARY 2016-17 Y 40 87.50 2.50 0.
32 VOYAGER MIDDLE SCHOOL 2016-17 Y 860 84.20 4.80 0.
33 WHITNEY ELEMENTARY 2016-17 Y 547 98.40 1.50 0.
34 SALK MIDDLE SCHOOL 2016-17 Y 715 92.90 7.10 2.
OK.
Unix Ln 1, Col 25, Ch 25
2595 ro...
1.3 secs


Query - postgres on postgres@localhost:5432 *
10:42 AM
postgres on postgres@localhost:5432
SQL Editor
Graphical Query Builder
Previous queries
Delete
Delete All
Scratch pad
select * from immunization;
select avg(Number_complete_for_all_immunizations) as Averagenumberofcompleteimmunizations from immunization where Reported='Y';
Output pane
Data Output
Explain
Messages
History
xempt_for_pertussis 2)
percent_exempt_for_measles_mumps_rubella numeric(5,2)
percent_exempt_for_polio numeric(5,2)
percent_exempt_for_hepatitisb numeric(5,2)
percent_exempt_for_varicella numeric(5,2)
number_complete_for_all_immunizations integer
number_with_any_exemption integer
number_with_n integer
1
2 0.00 0.00 0.00 0.00 0.20 476 1
3
4
5
6
7
8
9
10
11
12
13
14 0.00 0.00 0.00 0.00 0.00 23 0
15 1.10 1.40 1.40 2.20 1.40 362 8
16 0.40 0.40 0.40 0.60 0.40 516 3
17 3.40 3.60 3.40 2.90 4.00 496 23
18
19 0.60 0.90 0.60 0.60 1.50 2172 47
20 6.40 6.40 4.30 4.30 10.60 33 6
21 0.70 0.70 0.90 0.70 0.90 2127 46
22 0.00 0.00 0.00 0.00 0.00 14 0
23 0.00 0.30 0.30 0.30 0.30 197 2
24 0.00 0.00 0.60 1.10 0.00 175 2
25 7.70 7.70 8.20 8.60 8.60 125 28
26 5.10 5.80 5.40 4.90 7.10 421 35
27 0.90 0.60 0.60 0.80 1.10 826 15
28 3.00 3.50 3.20 3.50 4.60 347 20
29 3.20 3.30 3.00 2.80 3.80 657 31
30 0.80 1.00 0.80 1.20 1.20 2013 45
31 0.00 0.00 2.50 0.00 0.00 35 1
32 2.20 1.40 1.40 1.00 3.40 724 41
33 1.30 1.30 1.50 1.10 1.30 538 8
34 3.20 2.00 2.10 2.40 5.20 664 51
OK.
Unix Ln 1, Col 25, Ch 25
2595 ro...
1.3 secs



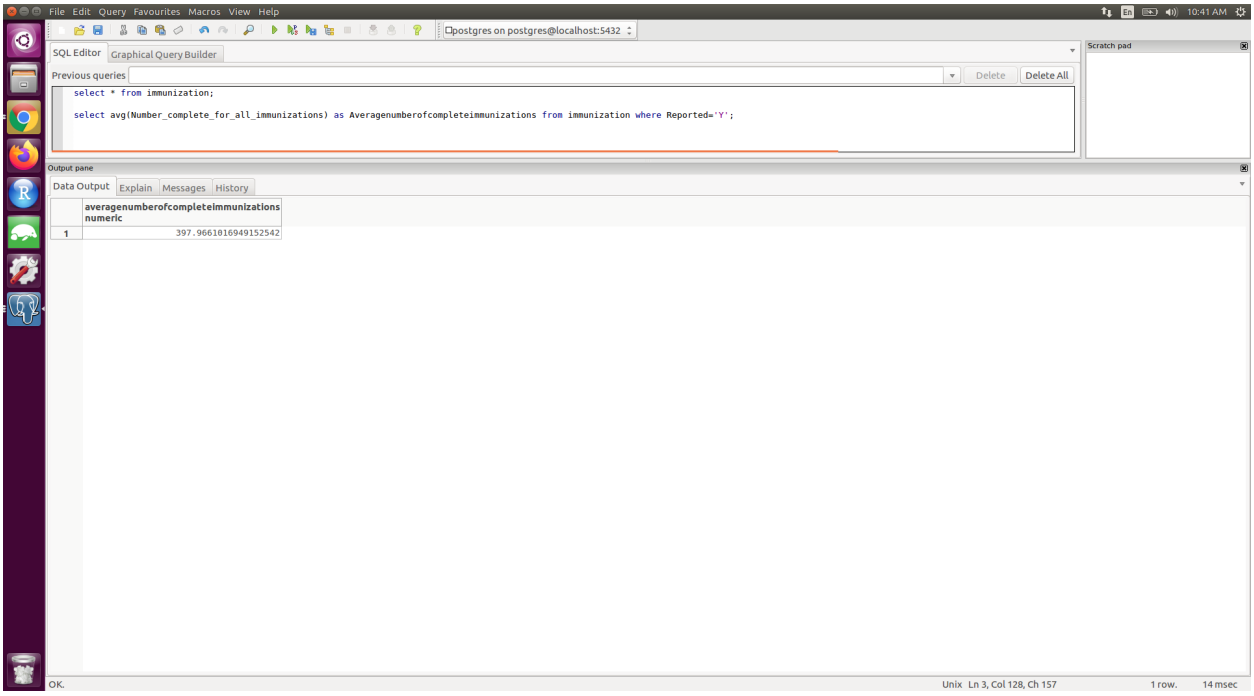

| | lcella | school_district character varying(255) | county character varying(255) | esd character varying(255) | grade_levels character varying(255) | has_kindergarten character varying(255) | has_6thgrade character varying(255) | location1 character varying(255) |
|---|---|---|---|---|---|---|---|---|
| 1 | | POMEROY SCHOOL DISTRICT | GARFIELD | EDUCATIONAL SERVICE DISTRICT 123 | 7-12 | N | N | 1090 PATAHA ST |
| 2 | 1 | GRANGER SCHOOL DISTRICT | YAKIMA | EDUCATIONAL SERVICE DISTRICT 105 | 5-8 | N | Y | 701 E AVENUE |
| 3 | | BELLINGHAM SCHOOL DISTRICT | WHATCOM | NORTHWEST EDUCATIONAL SERVICE DISTRICT 189 | P-1 | Y | N | 3242 FIRWOOD AVE |
| 4 | | TACOMA SCHOOL DISTRICT | PIERCE | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | P-6 | Y | Y | 1740 S 84TH ST |
| 5 | | TACOMA SCHOOL DISTRICT | PIERCE | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | K+ | Y | N | 1801 NE 53RD ST |
| 6 | | KENNEWICK SCHOOL DISTRICT | BENTON | EDUCATIONAL SERVICE DISTRICT 123 | 6-8 | N | Y | 1011 WEST 10TH AVENUE |
| 7 | | KELSO SCHOOL DISTRICT | COWLITZ | EDUCATIONAL SERVICE DISTRICT 112 | PK-8 | Y | Y | 96 GARDEN ST |
| 8 | | KENNEWICK SCHOOL DISTRICT | BENTON | EDUCATIONAL SERVICE DISTRICT 123 | 9-12 | N | N | 500 SOUTH DAYTON STREET |
| 9 | | LAKE WASHINGTON SCHOOL DISTRICT | KING | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | 9-12 | N | N | 5320 108TH AVE NE |
| 10 | | RICHLAND SCHOOL DISTRICT | BENTON | EDUCATIONAL SERVICE DISTRICT 123 | K | Y | N | 304 THAYER DR |
| 11 | | HIGHLINE SCHOOL DISTRICT | KING | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | P-6 | Y | Y | 220 SW 160TH ST |
| 12 | | SPOKANE SCHOOL DISTRICT | SPOKANE | EDUCATIONAL SERVICE DISTRICT 101 | K | Y | N | 1909 N.WRIGHT DR. |
| 13 | | SPOKANE SCHOOL DISTRICT | SPOKANE | EDUCATIONAL SERVICE DISTRICT 101 | 1-3 | N | N | 200 N BERNARD ST |
| 14 | 0 | MUKILTEO SCHOOL DISTRICT | SNOHOMISH | NORTHWEST EDUCATIONAL SERVICE DISTRICT 189 | P-1 | Y | N | 13000 BEVERLY PARK RD |
| 15 | 5 | BELLEVUE SCHOOL DISTRICT | KING | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | P-8 | Y | Y | 9450 NE 14TH ST |
| 16 | 2 | TACOMA SCHOOL DISTRICT | PIERCE | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | P-8 | Y | Y | 7112 S 12TH ST |
| 17 | 21 | MARYSVILLE SCHOOL DISTRICT | SNOHOMISH | NORTHWEST EDUCATIONAL SERVICE DISTRICT 189 | K -5 | Y | N | 3707 SUNNYSIDE BLVD |
| 18 | | TONASKET SCHOOL DISTRICT | OKANOGAN | NORTH CENTRAL EDUCATIONAL SERVICE DISTRICT 171 | 1-8 | N | Y | 32084-D HWY 97N |
| 19 | 33 | YAKIMA SCHOOL DISTRICT | YAKIMA | EDUCATIONAL SERVICE DISTRICT 105 | 9-12 | N | N | 212 S 6TH AVE |
| 20 | 5 | LAKE WASHINGTON SCHOOL DISTRICT | KING | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | 9-12 | N | N | 10601 NE 132ND |
| 21 | 21 | KENT SCHOOL DISTRICT | KING | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | 9-12 | N | N | 12430 SE 208TH ST |
| 22 | 0 | EDMONDS SCHOOL DISTRICT | SNOHOMISH | NORTHWEST EDUCATIONAL SERVICE DISTRICT 189 | PK-1 | Y | N | 21500 CYPRESS WAY BUILDING B |
| 23 | 1 | KIONA-BENTON CITY SCHOOL DISTRICT | BENTON | EDUCATIONAL SERVICE DISTRICT 123 | PK-3 | Y | N | 1107 GRACE AVENUE |
| 24 | 0 | WEST VALLEY SCHOOL DISTRICT (YAKIMA) | YAKIMA | EDUCATIONAL SERVICE DISTRICT 105 | K -4 | Y | N | 830 STONE RD |
| 25 | 20 | MEAD SCHOOL DISTRICT | SPOKANE | EDUCATIONAL SERVICE DISTRICT 101 | P-8 | Y | Y | 515 W ST THOMAS MORE WAY |
| 26 | 33 | CHENEY SCHOOL DISTRICT | SPOKANE | EDUCATIONAL SERVICE DISTRICT 101 | PK-5 | Y | N | 6323 S HOLLY ROAD |
| 27 | 9 | YAKIMA SCHOOL DISTRICT | YAKIMA | EDUCATIONAL SERVICE DISTRICT 105 | 6-8 | N | Y | 902 S 44TH AVE |
| 28 | 17 | MOUNT VERNON SCHOOL DISTRICT | SKAGIT | NORTHWEST EDUCATIONAL SERVICE DISTRICT 189 | K -6 | Y | Y | 1005 S 11TH ST |
| 29 | 26 | ORTING SCHOOL DISTRICT | PIERCE | PUGET SOUND EDUCATIONAL SERVICE DISTRICT 121 | K-5 | Y | N | 805 OLD PIONEER WAY NW |
| 30 | 25 | WENATCHEE SCHOOL DISTRICT | CHELAN | NORTH CENTRAL EDUCATIONAL SERVICE DISTRICT 171 | 9-12 | N | N | 1101 MILLERDALE AVE |
| 31 | 0 | WOODLAND SCHOOL DISTRICT | COWLITZ | EDUCATIONAL SERVICE DISTRICT 112 | K -5 | Y | N | 11842 LEWIS RIVER ROAD |
| 32 | 29 | MUKILTEO SCHOOL DISTRICT | SNOHOMISH | NORTHWEST EDUCATIONAL SERVICE DISTRICT 189 | 6-8 | N | Y | 11711 4TH AVE W |
| 33 | 7 | YAKIMA SCHOOL DISTRICT | YAKIMA | EDUCATIONAL SERVICE DISTRICT 105 | K -5 | Y | N | 4411 W NOB HILL BLVD |
| 34 | 37 | SPOKANE SCHOOL DISTRICT | SPOKANE | EDUCATIONAL SERVICE DISTRICT 101 | 7-8 | N | N | 6411 N ALBERTA ST |

**Finding the Average in SQL:**

I have calculated the average number of students who has taken all the immunizations irrespective of the exemptions of the particular schools where I have applied a condition that the average value should be calculated for those schools only which are reported.

The Query is as follows:

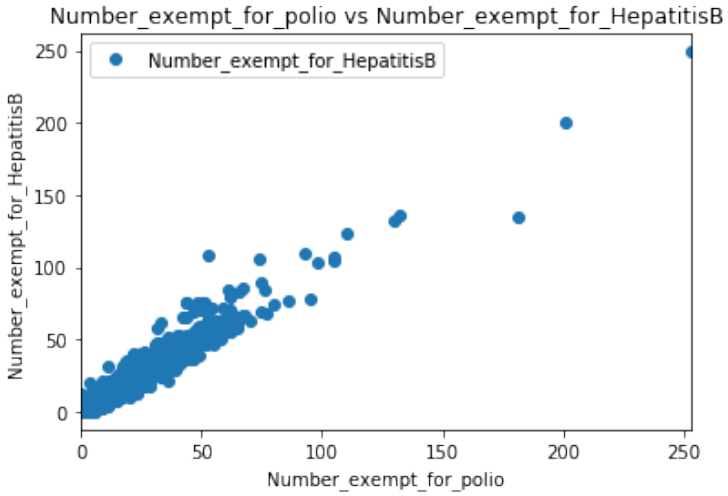
Select avg (Number_complete_for_all_immunizations) as Averagenumberofcompleteimmunizations where Reported=’Y’;




| | averagenumberofcompleteimmunizations numeric |
|---|---|
| 1 | 397.966101694915254 2 |

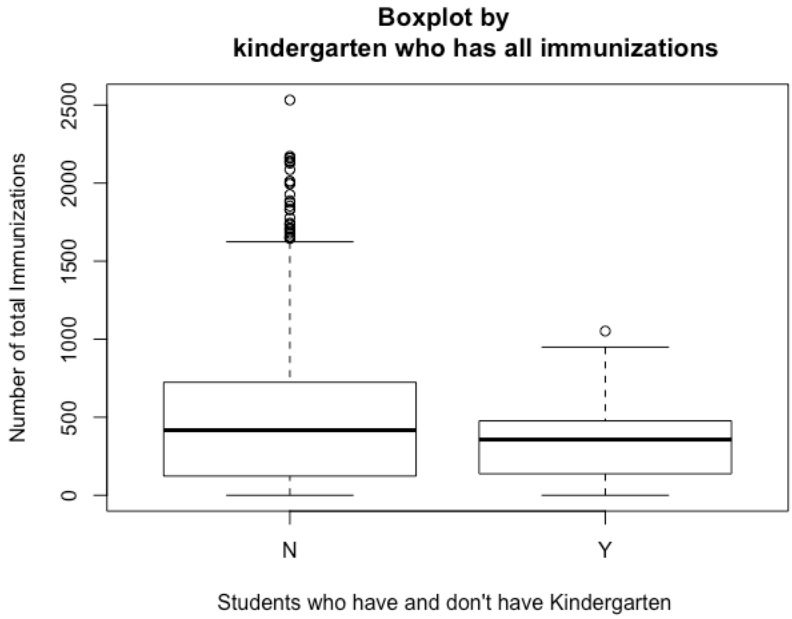
It means that there are about 397 schools which are reported, and they have taken all the immunizations

**Scatter-Plot Analysis:**




The above scatter plot shows the relationship between the Number of members exempt for the polio for the particular school to the number of members exempt for Hepatitis B. It can be observed that the values are distributed in an increasing manner. It can be said that the number of people who do not take the polio vaccine also take the hepatitis because of the linearity between them. Hence, it is concluded that the relation between them is Linear and Increasing exponentially. [3]

**Box-Plot Analysis:**




The above box-plot shows the relationship between the total number of students in the particular schools who have all the immunizations to the schools who have kindergarten or not. The plot for the immunizations who has and who doesn't have kindergarten the least value of 0 and the median

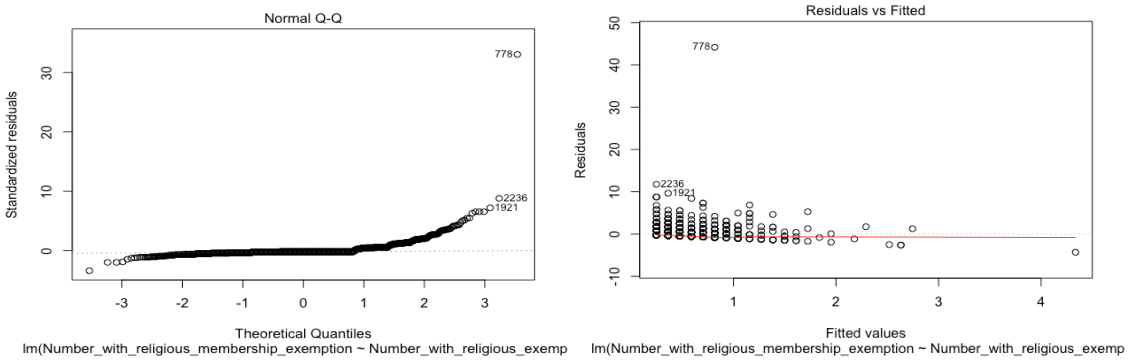
of both the plots is almost same which is around 450. From the box-plot, it can be observed that there are only very few number who has kindergarten and the total number of immunizations who doesn't have kindergarten is high. Here, there is only one outlier which means that there would not be much difference in the value.

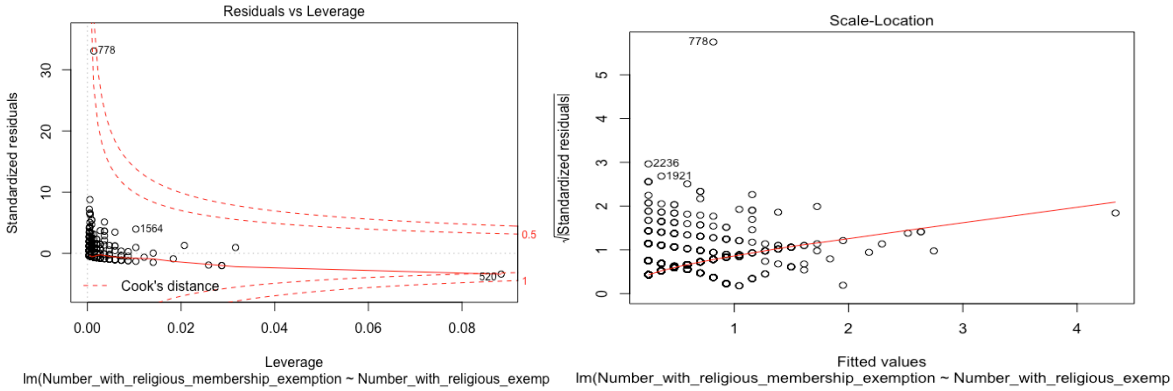
**Hypothesis Testing:**

```
> t.test(newdata$Number_with_any_exemption,mu=24,alternative = "less",conf.level = 0.95)

        One Sample t-test

data:  newdata$Number_with_any_exemption
t = 0.10201, df = 2477, p-value = 0.5406
alternative hypothesis: true mean is less than 24
95 percent confidence interval:
    -Inf 24.9678
sample estimates:
mean of x
  24.0565

>
```

Hypothesis testing uses the p-value in order to accept or reject the null hypothesis i.e., it provides the way to measure the strength of the evidence to accept or reject the hypothesis. The value of p lies between 0 and 1. Here, the value of mu is used which means that the mean value against which the sample should be tested.
From the value of p, we can state whether it accepts or rejects the Null Hypothesis. Whenever, the value of p is large, which is p>0.05 which means that it provides very less evidence to reject the null hypothesis. Hence, alternative hypothesis is rejected, and null hypothesis is accepted.
For the above hypothesis testing, it is clear that the value of p is large i.e., p=0.540 which shows that the null hypothesis can't rejected. [4]

**Regression Analysis:**





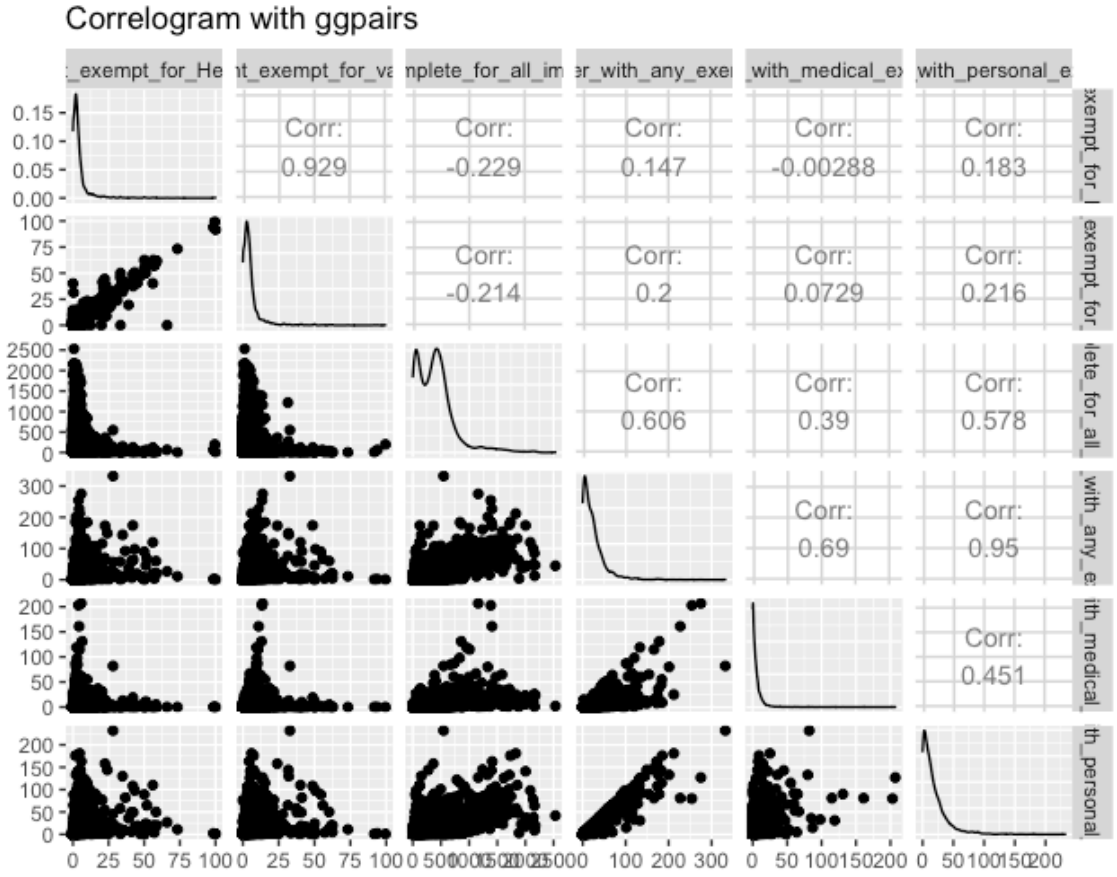

From the plots, it can be seen that the fitted line for the Normal Q-Q plot, Residuals Vs Fitted, Residuals Vs Leverage and Scale-distribution does signify any relationship between the two variables. It is neither increasing nor decreasing i.e., the variation is constant. Since, most of the points in the Q_Q plot fits the straight line, and the deviations are also minim al which means that the data is normally distributed. It concludes that he dataset variables has the compatible distributions.

```
#Cleaning the data
#Removing all the Null Values
newdata<-na.omit(mydata)
View(newdata)
#Linear Regression
model1<-lm(Number_with_religious_membership_exemption~Number_with_religious_exemption,
           data=newdata)
model1
summary(model1)
plot(model1)
```

```
> summary(model1)

Call:
lm(formula = Number_with_religious_membership_exemption ~ Number_with_religious_exemption,
    data = newdata)

Residuals:
   Min     1Q Median     3Q    Max
-4.336 -0.362 -0.249 -0.249 44.184

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    0.24867    0.03102   8.017 1.66e-15 ***
Number_with_religious_exemption  0.11353    0.01145   9.916  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.338 on 2476 degrees of freedom
Multiple R-squared:  0.0382,    Adjusted R-squared:  0.03781
F-statistic: 98.33 on 1 and 2476 DF,  p-value: < 2.2e-16
```

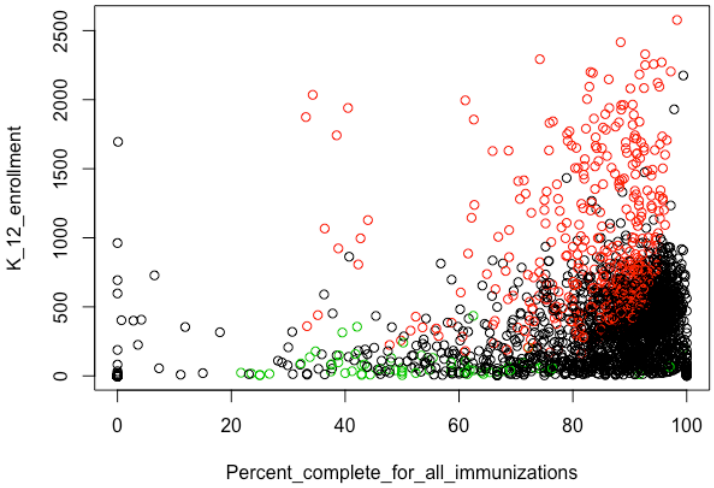
Here, for the regression analysis, I have first cleaned the data since it contains the null values due to which regression can't be done. I have used the function na.omit() which is used to remove null values and it creates a data frame.
From the summary, residuals give the basic idea about the dependent and the independent variable which means that it gives the values of minimum, maximum, quantiles and the median. R squared and the adjusted R-squared values re always between 0 and 1. Here only significant value is considered for the R squared.
F-statistic shows the relation between the predictor and the response variable where the higher value rejects the null hypothesis and it shows the importance of the model but not a specific parameter.
The value of p plays a vital role in order to know whether the model is statistically significant or not. The model can be considered statistically significant if the p value is less than the pre-determined level which is 0.05. This can be known with the number of stars beside the p-value and more number shows that model is significant.
In linear regression, Null hypothesis is given as the association between the co-efficient of the variables is 0 and for alternative hypothesis, it is not zero. Therefore, there exists no relation between the Number of people with religious membership exemption and the number with the religious exemption. [5]

**Scatter-Plot Matrix Analysis:**



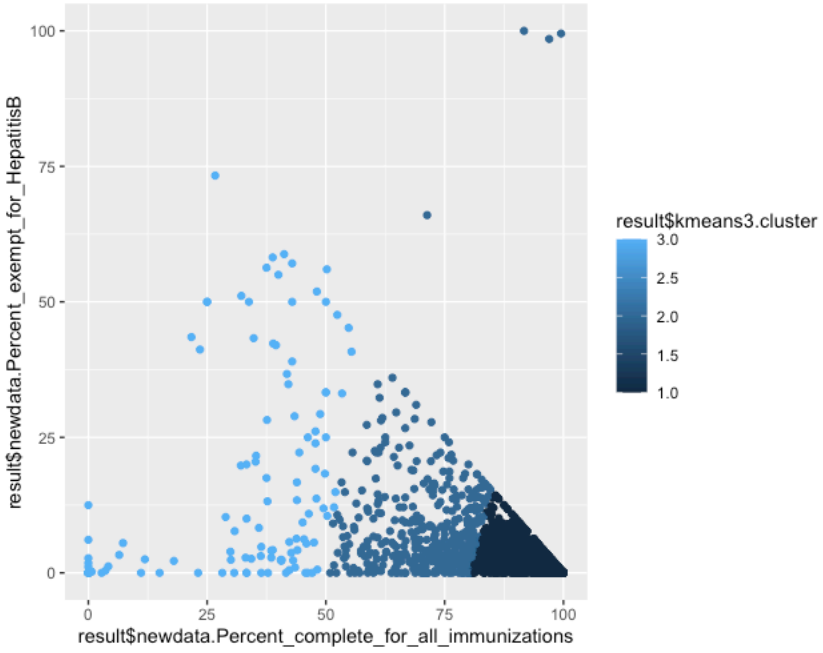
One of the best possible way to visualize and analyze observations of the possible combination of variables is Scatter matrix analysis. The above scatter plot matrix is the combination of the plots that are plotted for each of the numerical variables or the given attributes to be plotted. This can be done with the help of the ggpairs function in the R studio with the help of ggally package. It shows that scatter plot for the numerical variable at the left and it shows the correlation values to the right which indicates how much of those variables are correlated to each other.
Generally, the correlation coefficient lies between +1 and -1. The maximum correlation occurs for the variables number with any exemption to the number with the personal exemption which is 0.95which means that the correlation between them is strong and dense and there occurs the least relation between the number who are exempt for the varicella and number with all the complete immunizations because the datapoints are spread throughout the plot area and the correlation is -0.214 and hence it can be said that it has the least correlation. [6]

**K-means Clustering Analysis**




There are three clusters formed which are of different sizes i.e., 2084, 447, 64. It also provides the clustering mean which are the centers for those attributes. Clustering vector indicates that which school belongs to which cluster. For example, the 1st school belongs to the 1st cluster and similarly the schools belong to those particular clusters. It provides the information about the within cluster variability which is 16192.76 for the 1st cluster and 13452.92 for 2nd cluster and 10323.61 for 3rd cluster. The variability for the 3rd cluster is lower which indicates that they are close to each other in distance. The 1st cluster is black in color which are 2084 Clustering is good when between clustering distance is high than when distance is low but in this, we observe that the clusters are being overlapped. [7]

K-Means Clustering (k=3)




There are three clusters formed depending upon the percentage of students who have completed all the immunizations to the percentage of the students who have exempt for the Hepatitis B for the various schools. The sizes of the clusters are1899, 119, 460. The cluster variability for the first cluster is less when compared to the other two and the value is 48019.27 which means that there is good separation between the values within the cluster. Hence, we find that there is no overlapping between the clusters which means that between clustering distance is high and it means that it is a good clustering. [6]

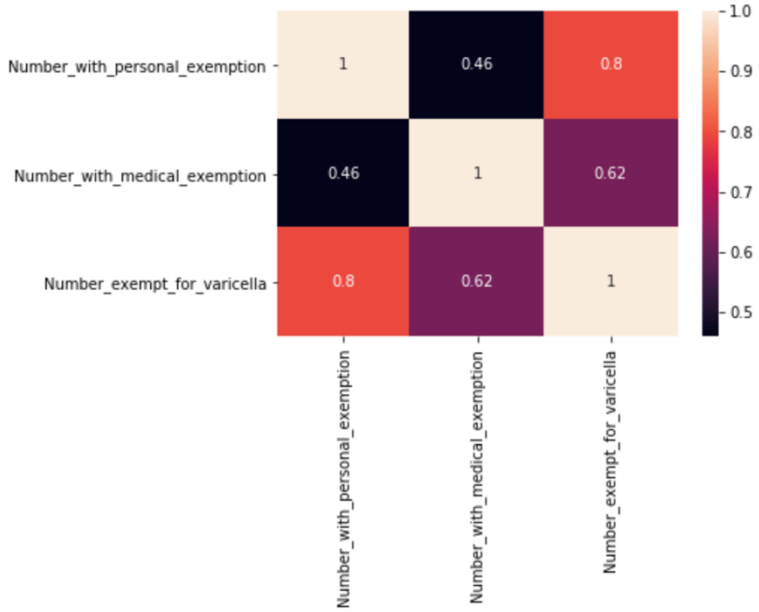
**Correlation Analysis:**

```
In [51]: df=data[['Number_with_personal_exemption','Number_with_medical_exemption','Number_exempt_for_varicella']]

In [52]: df.corr(method='pearson')
```

Out[52]:

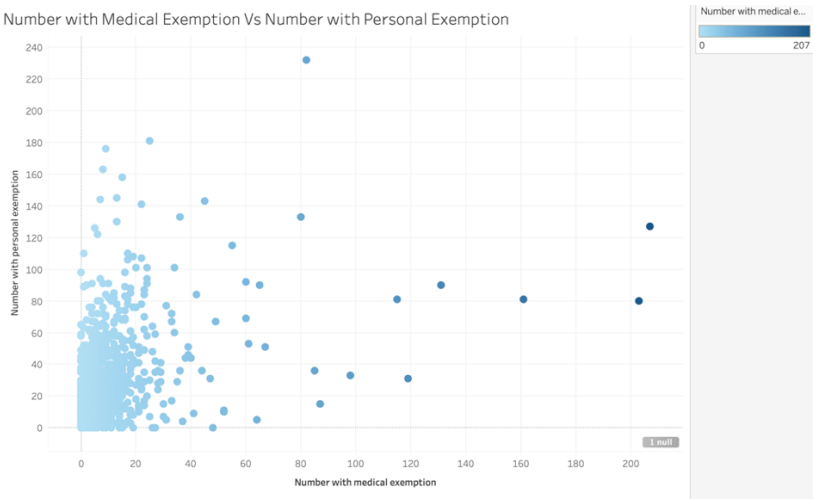
| | Number_with_personal_exemption | Number_with_medical_exemption | Number_exempt_for_varicella |
|---|---|---|---|
| **Number_with_personal_exemption** | 1.000000 | 0.459444 | 0.795049 |
| **Number_with_medical_exemption** | 0.459444 | 1.000000 | 0.623143 |
| **Number_exempt_for_varicella** | 0.795049 | 0.623143 | 1.000000 |

```
In [57]: corrMatrix=df.corr()
         sns.heatmap(corrMatrix, annot=True)
         plt.show()
```




Correlation analysis is one of the statistical methods and it is s done in order to find out how strong the different random variables are associated with each other. In other words, it can also be defined as the strength between the pair of random variables. When the value of correlation coefficient is 0, it indicates that there is no correlation between those two variables. Here I have used the Pearson method of correlation which Above analysis shows the Correlation analysis where it is done for three variables which are Number exempt for the personal, medical and varicella. The maximum correlation is between the number with the personal and the varicella exemptions which is around 0.79 and the least correlated variables are Number with medical and personal exemption which is around 0.49. [8]

**Scatter Plot Analysis**



The above scatter plot shows the relation between the number of students with the medical exemption to the number of people who have personal exemption. In the above plot, the difference between the number of medical exemptions is shown in the intensification of the color. It shows that there are a greater number of students with no medical exemption when compared to the members with the personal exemption and this graph is plotted with respect to the number of people with the personal exemptions. [9]

**Conclusion:**

- To know on an average how many of students of that schools got all the vaccinations (Complete) as well as if the school has been reported or not?

Here, I have found the average number of schools which are being reported who has taken all the immunizations is done by using the PostgreSQL of the Virtual Machine. It is found that the average number is about 397 schools.

- Does there is any relationship between the number of students who have signed for the exemption for religious exemption and religious- membership exemption?

In this analysis, it is found that there is no relation between the number of students who have exempt for the religious and the religious membership and it can be realized with the help of the regression analysis. From the plots that have been obtained, it shows that the values of those attributes lie on the fitted line itself and there is minimum deviation.

- What is the median number of schools are done with all the immunizations who has Kindergarten?

With the help of Box-plot, it is possible to find the median value where the students of those particular schools who have the kindergarten and who has taken all the immunizations is around 450 which means that it is the median value of those who have kindergarten with all vaccines taken.

- Is there any relation between the number of members exempt for the polio for all schools and number of members exempt for Hepatitis B?

From the scatter plot, it can be seen those who have not taken the polio vaccine have not taken the hepatitis vaccine as well because it shows the plot as linearly increasing which means that only if the students took the polio vaccine then hepatitis is taken by the students of their respective schools.

**Explain Terms:**

- R Studio: R Studio is an IDE which is helpful for graphics and the statistical computing. This is helpful in analyzing, visualizing the data where it is possible to code, view the data simultaneously.
- Tableau: Tableau is an open source software which is used to visualize the data more easily where it divides the attributes of the data set into measures and dimensions where measures indicate that these fields can be aggregated and the measures are those can't be aggregated and they are of categorical data type. It simply displays the plots by dragging and dropping the attributes in the rows and columns.
- PostgreSQL: It is one of the open source database system which is used for the execution of complex queries. It supports both Sql for the relational database as well as json files for the non-relational databases. It Is mainly used for the data warehousing and applications of data analysis.

- Linear Regression: Linear Regression is of the most common and basic predictive analysis that can be performed upon the numerical attributes. It specifies the relationship among the two attributes whether there is relation between the independent or the explanatory variable and the dependent variable which is response variable.
- Correlation: Correlation is also one of the basic methods which is used to find the statistical relationship between the two random variables. In this it shows the value of the correlation which indicates how strong one variable is correlated to the other variable.

**Link for dataset download:**
https://catalog.data.gov/dataset/all-students-kindergarten-through-12th-grade-immunization-data-by-school-2016-2017

# References


[1] E. Joanna, "data.wa.gov," 22 April 2019. [Online]. Available: https://data.wa.gov/Health/All-students-kindergarten-through-12th-grade-immun/9zru-c2kz.

[2] E. Joanna, "data.wa.gov," 22 April 2019. [Online]. Available: https://data.wa.gov/Health/All-students-kindergarten-through-12th-grade-immun/9zru-c2kz.

[3] D. Friedman, "dfrieds.com," Scatter Plots using Matplotlib, 2020. [Online]. Available: https://dfrieds.com/data-visualizations/scatter-plots-python-matplotlib.

[4] "Introduction to Hypothesis Testing in R," 2020. [Online]. Available: https://data-flair.training/blogs/hypothesis-testing-in-r/.

[5] S. Prabhakaran, "r-statistics.co," 2017. [Online]. Available: http://r-statistics.co/Linear-Regression.html.

[6] "Scatterplot matrix with ggpairs()," 2018. [Online]. Available: https://www.r-graph-gallery.com/199-correlation-matrix-with-ggally.html.

[7] "K-means Cluster Analysis," [Online]. Available: https://uc-r.github.io/kmeans_clustering.

[8] "Correlation Analysis," sciencedirect.com, 2020. [Online]. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/correlation-analysis.

[9] "help.tableau.com," Build a Scatter Plot, 2020. [Online]. Available: https://help.tableau.com/current/pro/desktop/en-us/buildexamples_scatter.htm.