Final Report: School Immunizations

19030017 & 20030004 & 20030058 & 20030061

1 Dataset

Dataset of immunization(vaccination) status of kindergarten students in California in schools from year 2016-2019.The annual assessment was conducted to measure the immunization coverage amongst the students entering kindergarten. There are 11 features in dataset which are: year of immunization, unique code of school, country, school sector, city name, school name, Immunizations report was conducted or not(Y /N), total number of enrollees, types of immunizations, total number of enrollees vaccinated, percentage of enrollees vaccinated.

2 Duplication of Data

No duplicate data found.

3 Attributes & Observations

- Count of Records: 263071
- Count of Attributes: 11

4 Null Values

Number of null values that are found in attributes are:

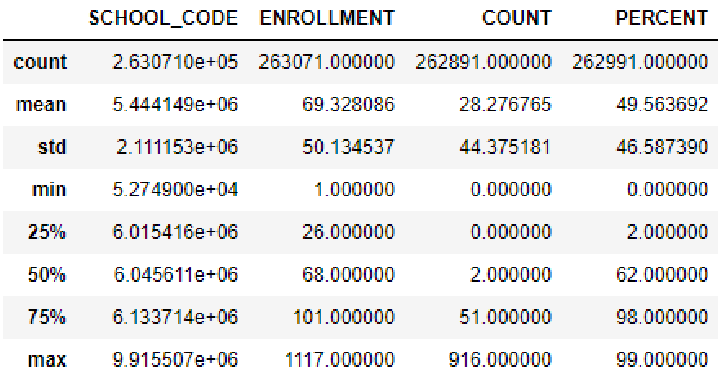
- Total number of enrollees: 9884
- Total number of enrollees vaccinated: 181137
- Percentage of enrollees vaccinated: 55249

Other features have non-null values.

5 Filling of Null Values

5.1 Backward Fill

After backward filling, "Total number of enrollees vaccinated" and "Percentage of enrollees vaccinated" have still null values as 180 and 80 respectively.The summary statistics after backward filling is given below.

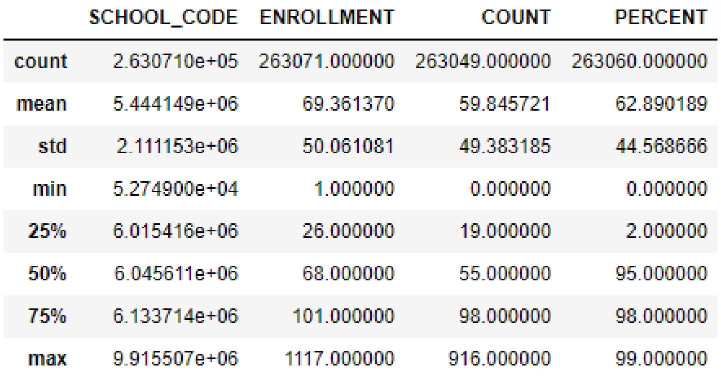


	SCHOOL_CODE	ENROLLMENT	COUNT	PERCENT
count	2.630710e+05	263071.000000	262891.000000	262991.000000
mean	5.444149e+06	69.328086	28.276765	49.563692
std	2.111153e+06	50.134537	44.375181	46.587390
min	5.274900e+04	1.000000	0.000000	0.000000
25%	6.015416e+06	26.000000	0.000000	2.000000
50%	6.045611e+06	68.000000	2.000000	62.000000
75%	6.133714e+06	101.000000	51.000000	98.000000
max	9.915507e+06	1117.000000	916.000000	99.000000

Figure 1: Summary Statistics after Backward Filling

5.2 Forward Fill

After forward filling, "Total number of enrollees vaccinated" and "Percentage of enrollees vaccinated" have still null values as 22 and 11 respectively.The summary statistics after forward filling is given below.

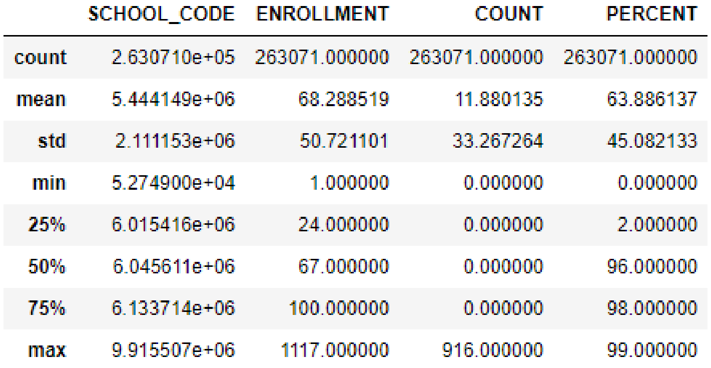


	SCHOOL_CODE	ENROLLMENT	COUNT	PERCENT
count	2.630710e+05	263071.000000	263049.000000	263060.000000
mean	5.444149e+06	69.361370	59.845721	62.890189
std	2.111153e+06	50.061081	49.383185	44.568666
min	5.274900e+04	1.000000	0.000000	0.000000
25%	6.015416e+06	26.000000	19.000000	2.000000
50%	6.045611e+06	68.000000	55.000000	95.000000
75%	6.133714e+06	101.000000	98.000000	98.000000
max	9.915507e+06	1117.000000	916.000000	99.000000

Figure 2: Summary Statistics after Forward Filling

5.3 Mode Fill

After mode filling, no null values are found. The summary statistics after mode filling is given below.

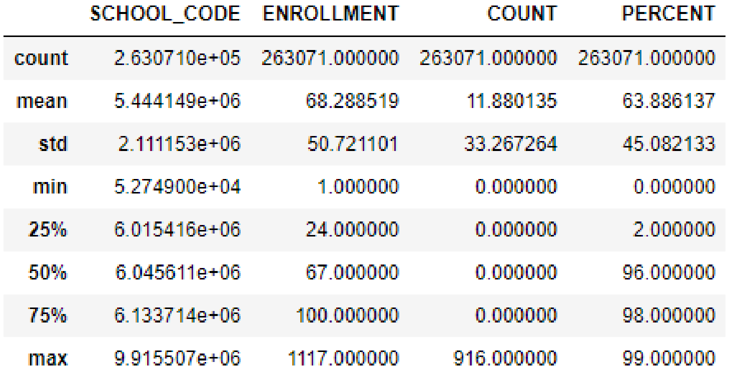


	SCHOOL_CODE	ENROLLMENT	COUNT	PERCENT
count	2.630710e+05	263071.000000	263071.000000	263071.000000
mean	5.444149e+06	68.288519	11.880135	63.886137
std	2.111153e+06	50.721101	33.267264	45.082133
min	5.274900e+04	1.000000	0.000000	0.000000
25%	6.015416e+06	24.000000	0.000000	2.000000
50%	6.045611e+06	67.000000	0.000000	96.000000
75%	6.133714e+06	100.000000	0.000000	98.000000
max	9.915507e+06	1117.000000	916.000000	99.000000

Figure 3: Summary Statistics after Mode Filling

5.4 Mean + Mode Fill

Numerical attributes are filled by mean and categorical attributes are filled by mode. There are no null values found after filling. The summary statistics after filling is given below.



	SCHOOL_CODE	ENROLLMENT	COUNT	PERCENT
count	2.630710e+05	263071.000000	263071.000000	263071.000000
mean	5.444149e+06	68.288519	11.880135	63.886137
std	2.111153e+06	50.721101	33.267264	45.082133
min	5.274900e+04	1.000000	0.000000	0.000000
25%	6.015416e+06	24.000000	0.000000	2.000000
50%	6.045611e+06	67.000000	0.000000	96.000000
75%	6.133714e+06	100.000000	0.000000	98.000000
max	9.915507e+06	1117.000000	916.000000	99.000000

Figure 4: Summary Statistics after Mean & Mode Filling

6 Summary Statistics & Visualization

- Three tenures are considered for this dataset: 2016-2017, 2017-2018, 2018-2019. in which 2016-2017 period has more count.
- 58 countries are included with different cities in them. Los Angeles has highest count.
- In those cities, public private sectors are targeted such as count of Public Sectors > Private Sectors.
- Ratio of positive conducted immunization is greater.
- PME, DTP, HEPB, MMR VARI, POLIO vaccinations have highest ratio as compared to other vaccinations.
- Highest number and percentage of the students who got vaccinated are 209589 and 98 percent respectively.

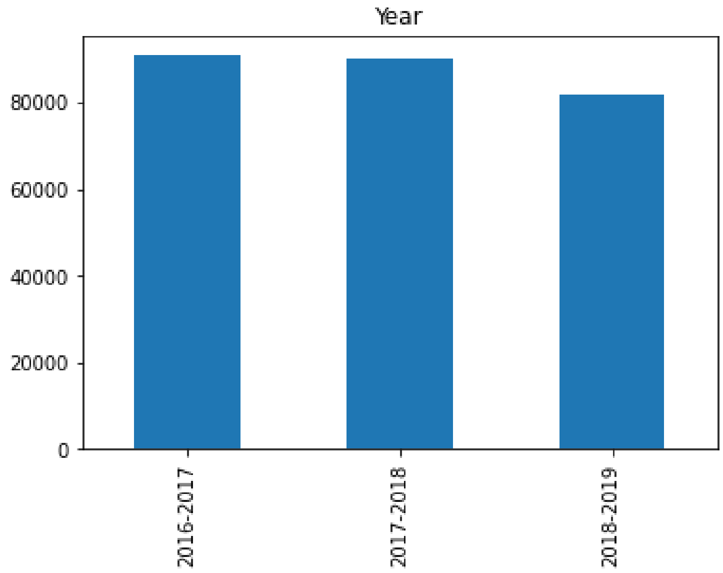


Figure 5: Years of Immunization

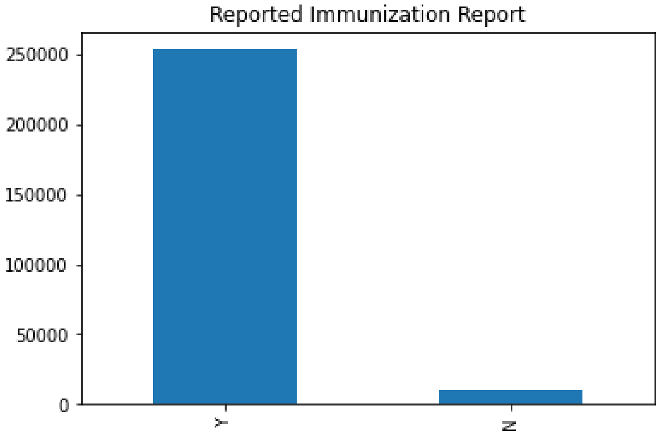


Figure 6: Reported Immunization

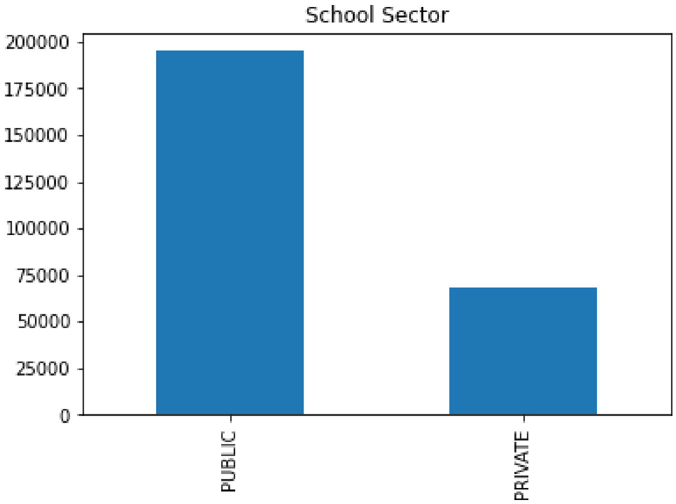


Figure 7: School Sectors

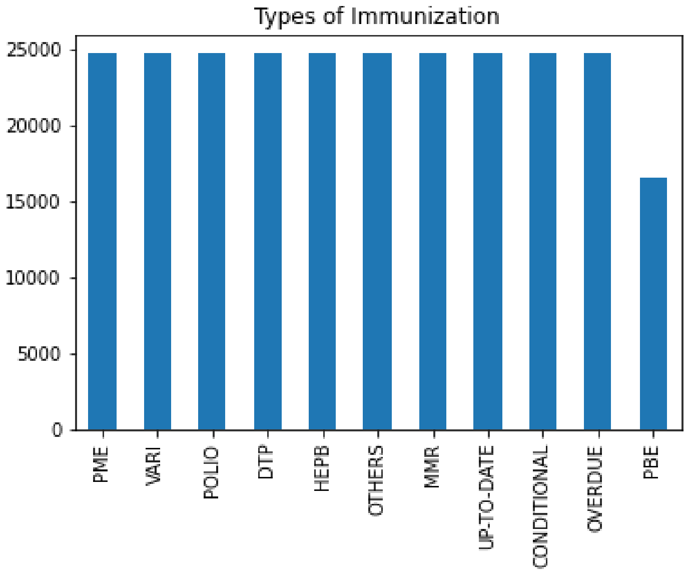


Figure 8: Types of Immunization

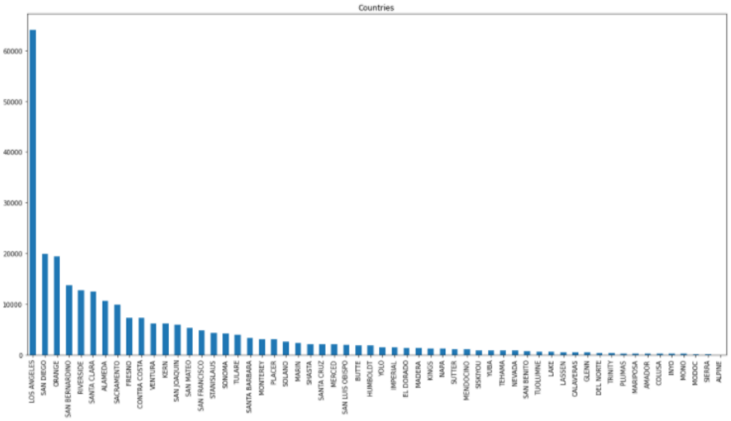


Figure 9: Countries

7 Correlation of Numerical Attributes

Correlation of numerical attributes is as:

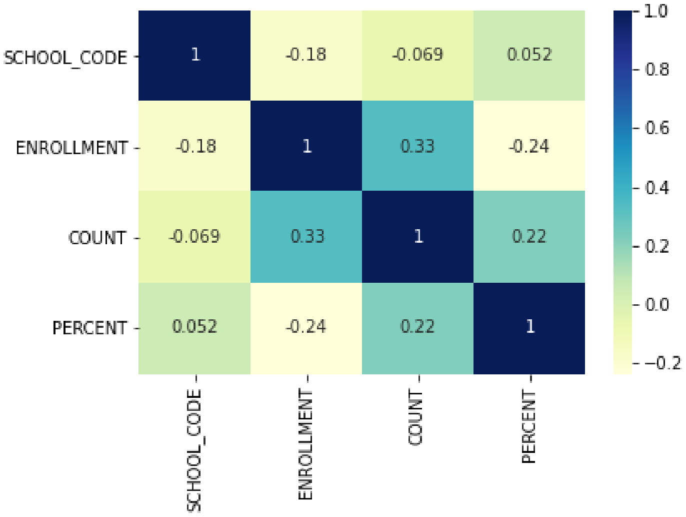


Figure 10: Correlation matrix

- Correlation score of each school and total enrolled students who are vaccinated is 0.0524, so both have slight direct relation with each other.
- Correlation score of total number of enrolled students and total number of enrooled vaccinated students is 0.33, so both have direct positive effect on each other.
- Correlation score of total number of enrollees vaccinated and percentage of enrollees vaccinated is 0.22, so both have direct positive effect on each other.

8 Covariance of Numerical Attributes

Covariance of numerical attributes is as:

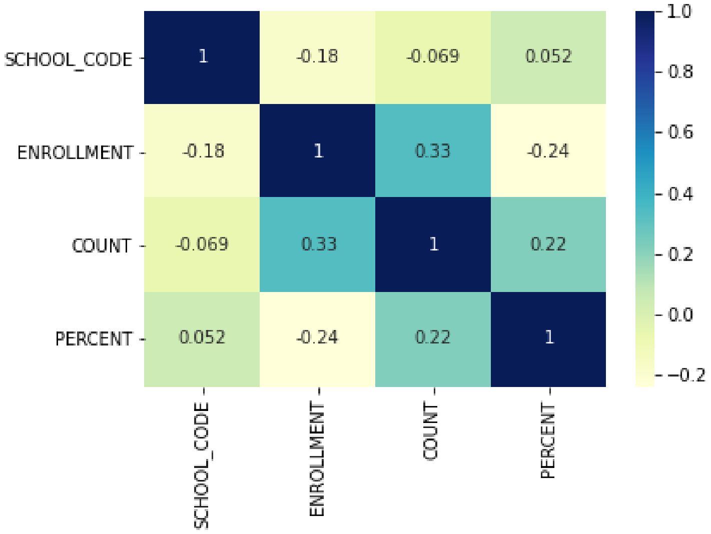


Figure 11: Covariance matrix

School, total number of enrollees, total number of enrollees vaccinated and percentage of enrollees vaccinated have high positive covariance among them and have larger contribution in dataset.

9 Correlation of All Attributes

Correlation of all attributes is as:

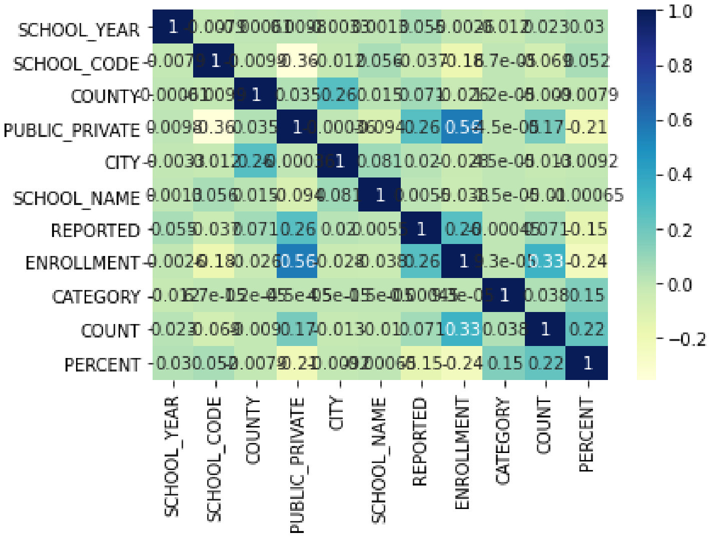


Figure 12: Correlation matrix

9.1 Covariance of All Attributes

Covariance of all attributes is as:

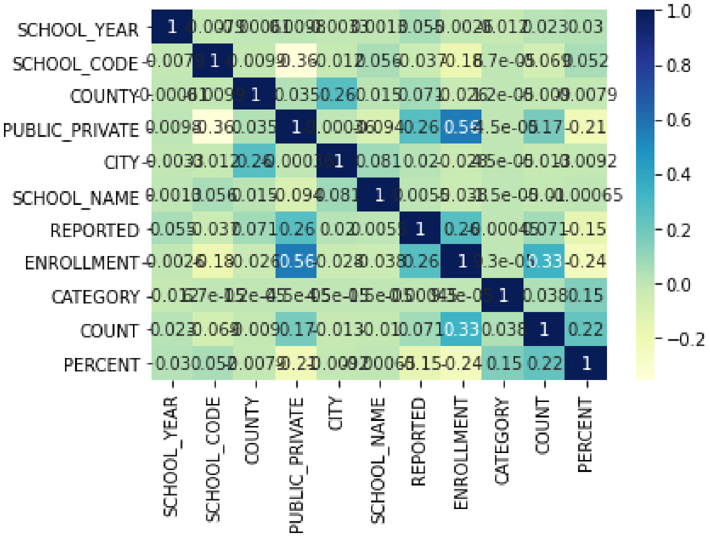


Figure 13: Covariance matrix

10 Graphs

Graphs of different features are as:

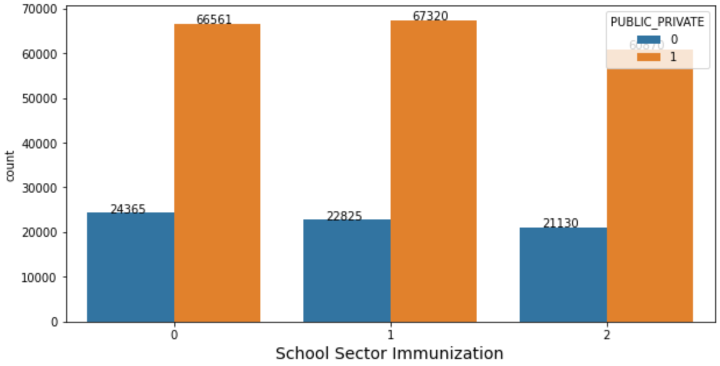


Figure 14: Count of School Sector Immunization

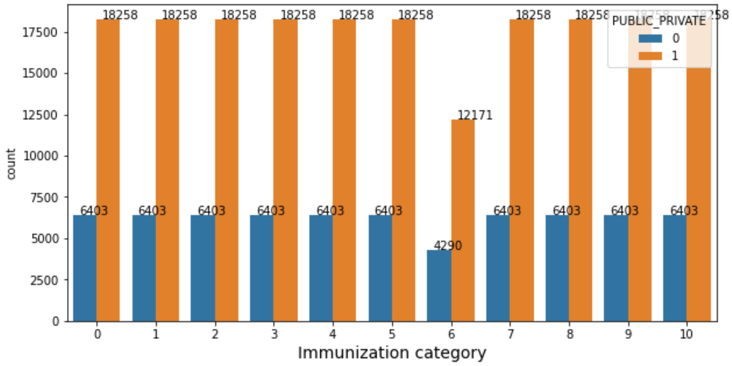


Figure 15: Immunization Category

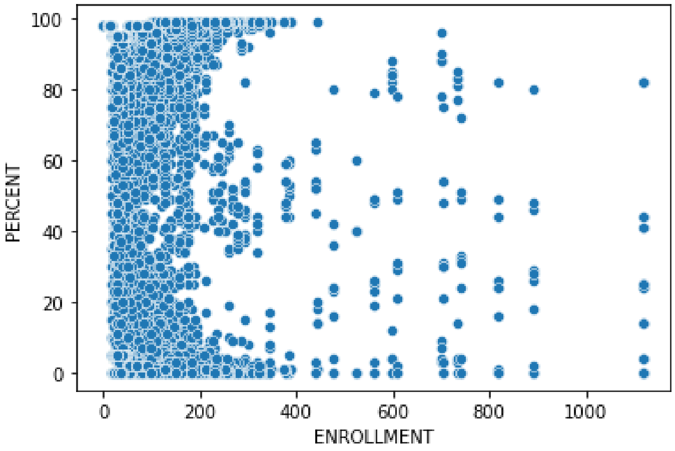


Figure 16: Total number of Enrollees Vs Percentage of Enrollees Vaccinated

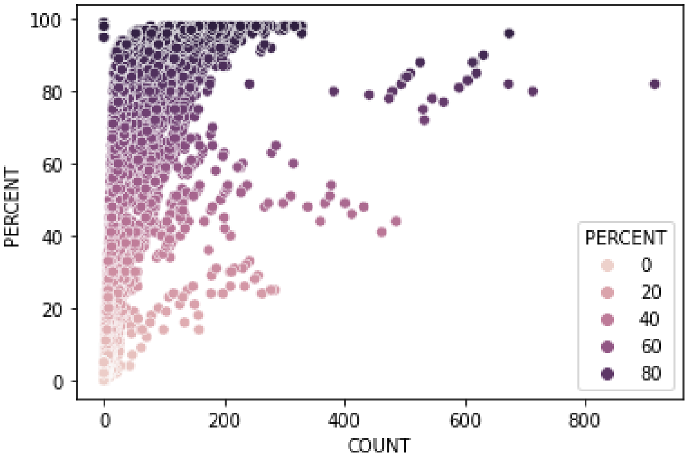


Figure 17: Count Vs Percentage of Enrollees Vaccinated

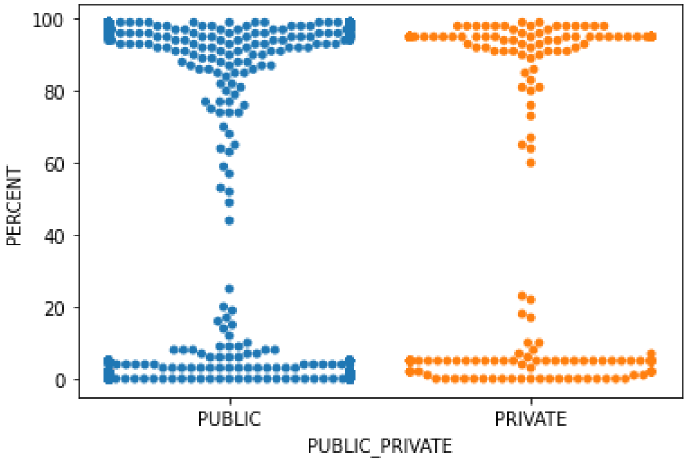


Figure 18: Percentage of Enrollees Vaccinated in Public/Private Sector

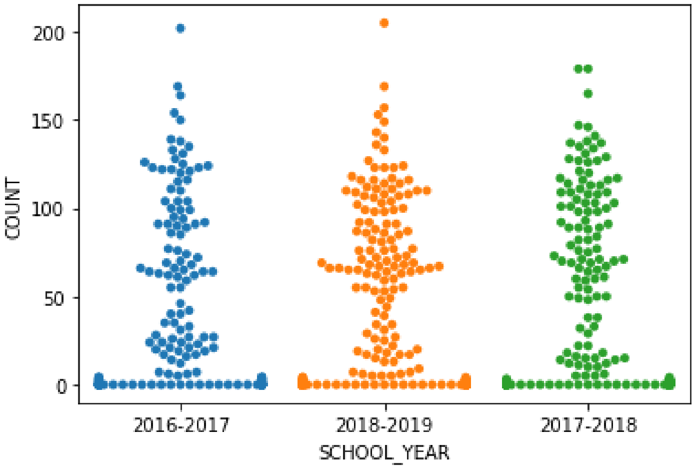


Figure 19: Total number of Enrollees vaccinated in Specific Year

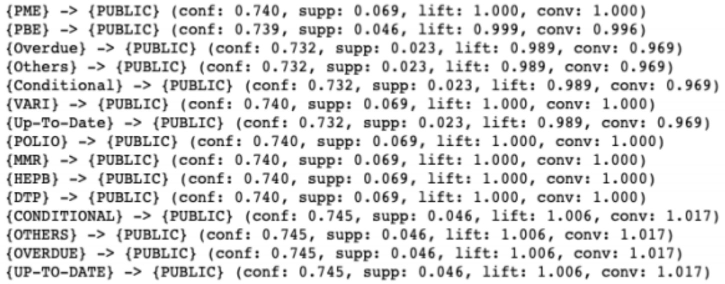
Graphs clearly shows that:

- Public Sector has greater number of Immunizations as compared to private sector.
- PME, DTP, HEPB, MMR VARI, POLIO vaccinations have highest ratio as compared to other vaccinations.
- Greater the number of enrollees vaccinated, greater is it's percentage value.

- Large portion of enrolled students got vaccinated.
- In the period of 2018-2019; count of enrollees who are vaccinated are larger in number.

11 Frequent Pattern Mining

5187 numbers of frequent item sets are found. School immunization data was modified and merged and nnumber of each row contains all the immunizations covered by a specific city. It was obsereved with confidence 1 and support 1 that in any school if there is immunization of Conditional, HEPB, MMR, OTHERS then we can say with 100 percent confidence that DTP immunization will also be covered at that school. Similarly if there is immunization of Conditional, DTP, HEPB, MMR then we can say with 100 percent confidence that others immunization will also be covered at that school. There are hundereds of such rules and are shown above



```
{PME} -> {PUBLIC} (conf: 0.740, supp: 0.069, lift: 1.000, conv: 1.000)
{PBE} -> {PUBLIC} (conf: 0.739, supp: 0.046, lift: 0.999, conv: 0.996)
{Overdue} -> {PUBLIC} (conf: 0.732, supp: 0.023, lift: 0.989, conv: 0.969)
{Others} -> {PUBLIC} (conf: 0.732, supp: 0.023, lift: 0.989, conv: 0.969)
{Conditional} -> {PUBLIC} (conf: 0.732, supp: 0.023, lift: 0.989, conv: 0.969)
{VARI} -> {PUBLIC} (conf: 0.740, supp: 0.069, lift: 1.000, conv: 1.000)
{Up-To-Date} -> {PUBLIC} (conf: 0.732, supp: 0.023, lift: 0.989, conv: 0.969)
{POLIO} -> {PUBLIC} (conf: 0.740, supp: 0.069, lift: 1.000, conv: 1.000)
{MMR} -> {PUBLIC} (conf: 0.740, supp: 0.069, lift: 1.000, conv: 1.000)
{HEPB} -> {PUBLIC} (conf: 0.740, supp: 0.069, lift: 1.000, conv: 1.000)
{DTP} -> {PUBLIC} (conf: 0.740, supp: 0.069, lift: 1.000, conv: 1.000)
{CONDITIONAL} -> {PUBLIC} (conf: 0.745, supp: 0.046, lift: 1.006, conv: 1.017)
{OTHERS} -> {PUBLIC} (conf: 0.745, supp: 0.046, lift: 1.006, conv: 1.017)
{OVERDUE} -> {PUBLIC} (conf: 0.745, supp: 0.046, lift: 1.006, conv: 1.017)
{UP-TO-DATE} -> {PUBLIC} (conf: 0.745, supp: 0.046, lift: 1.006, conv: 1.017)
```

Figure 20: Frequent ItemSets

12 Clustering

12.1 K Means Clustering

Boxplots of features using K means clutering are:

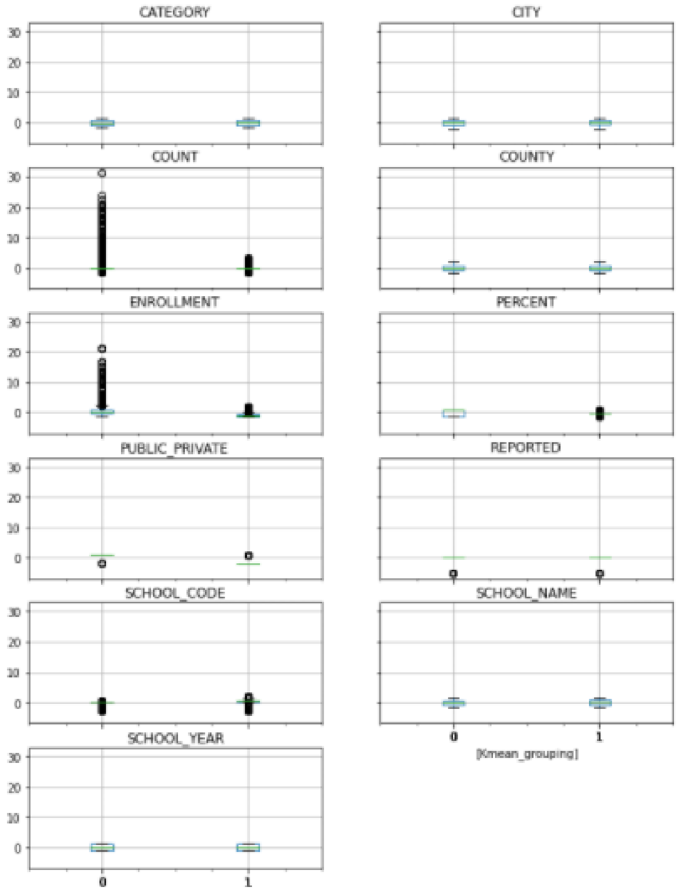


Figure 21: Boxplot using Clustering

Value of k using elbow method

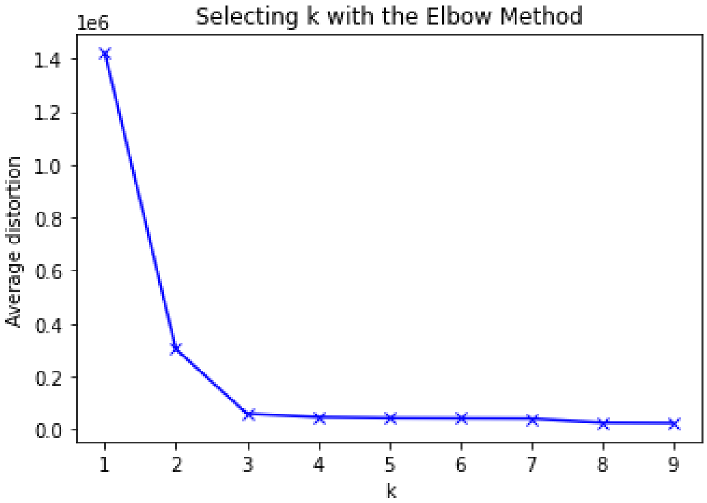


Figure 22: Elbow Method

Clusters are as follows:

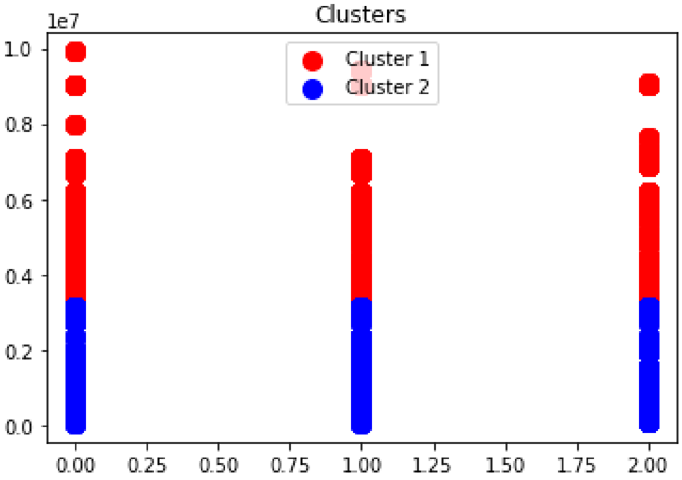


Figure 23: Clusters

12.2 Hierarchical Clustering

Boxplots of features using hierarchical clutering are:

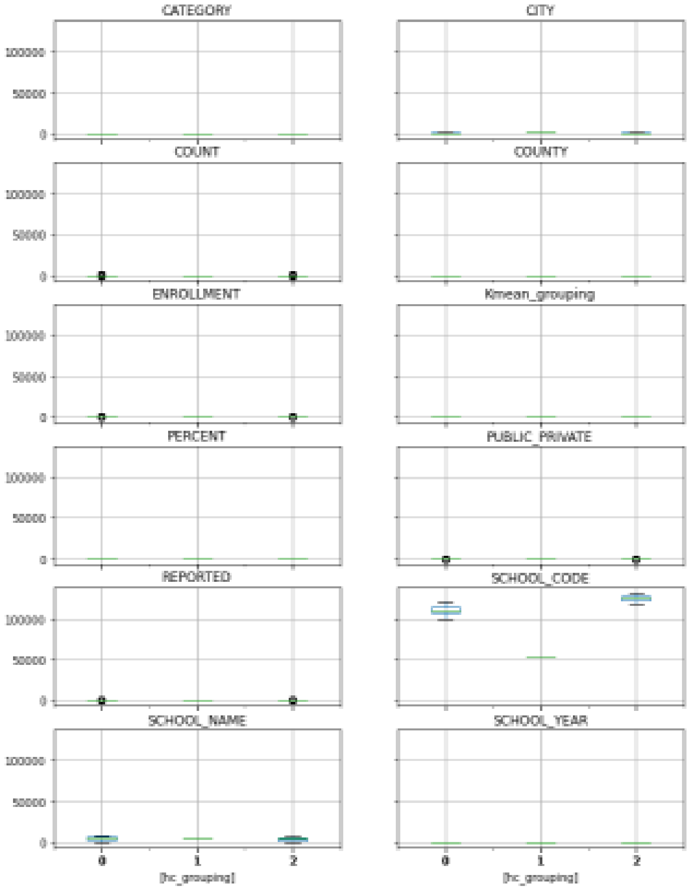


Figure 24: Boxplot using Clustering

Value of k using elbow method

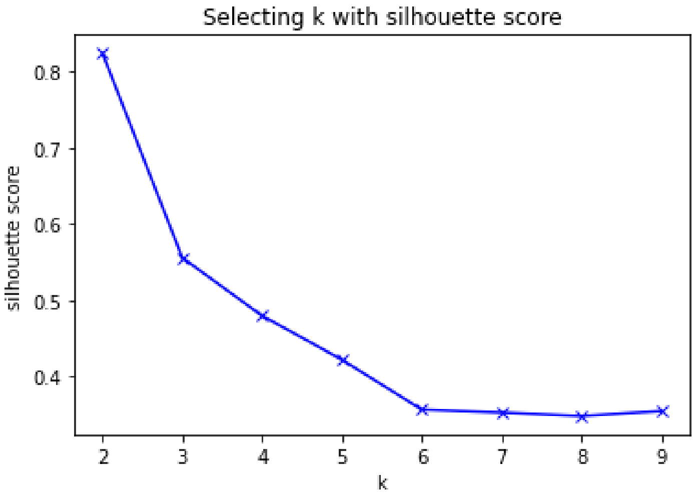


Figure 25: Elbow Method

Clusters are as follows:

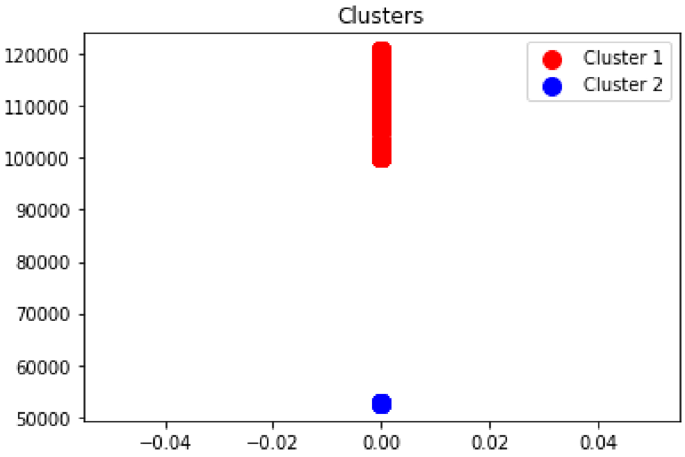


Figure 26: Clusters