PRT564 Data Analytics and Visualization

Assignment 1

Table of Contents

[Overview 3](#_Toc37368912)

[Analysis – 3](#_Toc37368913)

[Analysis and Insights for the VIC state 11](#_Toc37368914)

[Conclusion – 15](#_Toc37368915)

[References – 15](#_Toc37368916)

Table of Contents –

[Figure 1: Summary 4](#_Toc37338930)

[Figure 2:School count per State 5](#_Toc37338931)

[Figure 3:Code 1 5](#_Toc37338932)

[Figure 4:Government school per state 6](#_Toc37338933)

[Figure 5:Catholic Sector 6](#_Toc37338934)

[Figure 6:Independent School 7](#_Toc37338935)

[Figure 7:Pie chart code 7](#_Toc37338936)

[Figure 8:Sector distribution 8](#_Toc37338937)

[Figure 9:Sector code 8](#_Toc37338938)

[Figure 10:Teaching\_Staff\_Statewise 9](#_Toc37338939)

[Figure 11:Staff Code 10](#_Toc37338940)

[Figure 12:Bottom Sea Quarter % 10](#_Toc37338941)

[Figure 13:Bottom Sea Quarter 11](#_Toc37338942)

[Figure 14: VIC School\_Sector 11](#_Toc37338943)

[Figure 15:VIC sector 12](#_Toc37338944)

[Figure 16:VIC Teaching vs Non-Teaching 12](#_Toc37338945)

[Figure 17:Teaching Staff vs Non Teaching Staff 13](#_Toc37338946)

[Figure 18:VIC-Year\_Range 13](#_Toc37338947)

[Figure 19:Year range VIC state 14](#_Toc37338948)

[Figure 20:K value for cluster 14](#_Toc37338949)

[Figure 21K:means 15](#_Toc37338950)

# Overview

The Data is about the school profiles of different Schools based in Australia. There are about 9535 rows and 31 columns. The major insights that can be found in the data are subject as question here.

1. To look for Total Enrolments based on the type of School.
2. To check the Teaching Staff based on School Type so to know the distribution.
3. School Sector wise distribution of States so to know what kind of schools are present in each States.
4. Distribution of Teaching Staff as per the State.
5. Distribution of Teaching Staff based on Geolocation of Australia.

# Analysis

To Analyze the data first look for the summary of the data. The summary of the data will provide the important insights for the data like mean ,mode ,median of the numerical data .All the plots and analysis of the data is being done in R by using Library ggplot2 (Son and Kim, 2016).

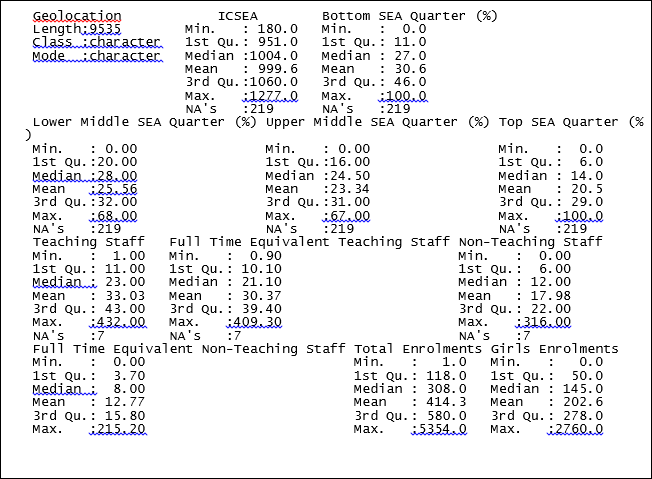


Figure : Summary

From the data the mean enrolment is the 414.3 Means that in year the mean number of enrolment in any School is about 414.3. With the mean number of Boys Enrolment are about 211.688468 and Mean number of Girls Enrolment in any given school in year 2018 is 202.6 . The data suggests that the greater number of boys are being enrolled in school as compared to girls in Australian School.

Now here are some visualization about the data –

The most important part is to know how the data is being distributed. To count the number of school per state from the dataset lets count each state as this will provide a number of schools present in the each state as per the dataset.

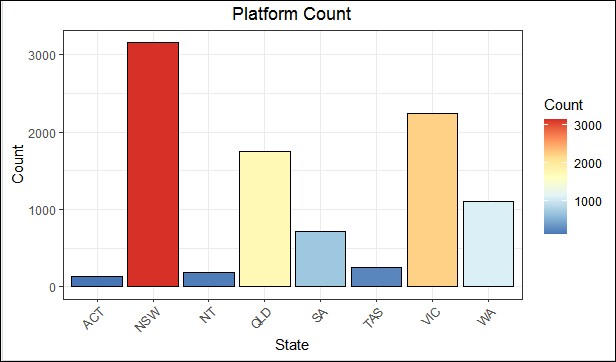


Figure 2:School count per State

Code for this Plotting –

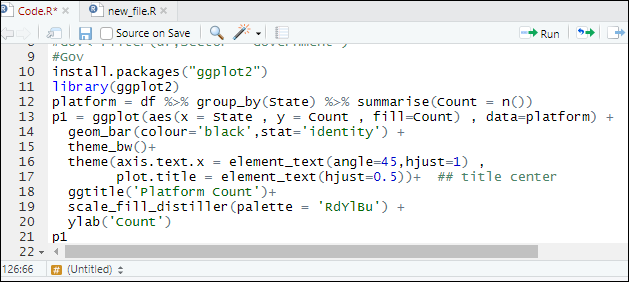


Figure 3:Code 1

As it can easily interpreted that maximum number of school are in NSW state of Australia with a count of approximately 3200 school. The second highest number of Schools are in State VIC of Australia with count of 2500 schools. This shows that there might be highly educated people in NSW and VIC. But as compared to state NSW and VIC most of the state have very low number of schools like state ACT , NT ,TAS,SA .

It’s very important to know school sector like Government, Catholic and Independent distribution as per state.

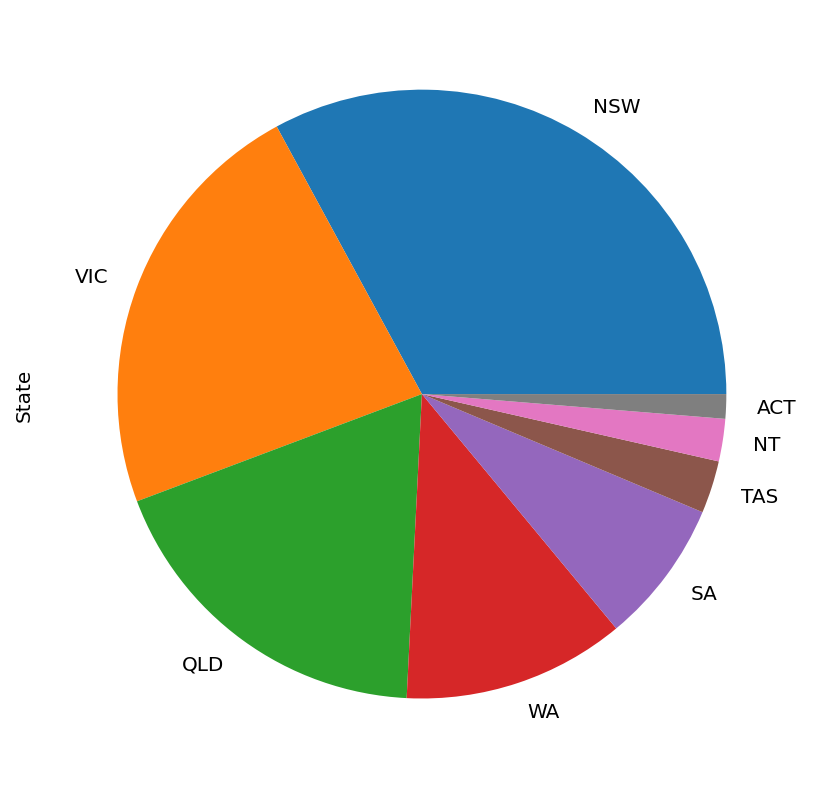


Figure 4:Government school per state

From the pie chart distribution above, we can see that most of the government school are in state NSW. As the overall number of schools distribution are biased toward the state NSW as state NSW has the most number of school in the dataset.

Distribution of Catholic school and Independent school.

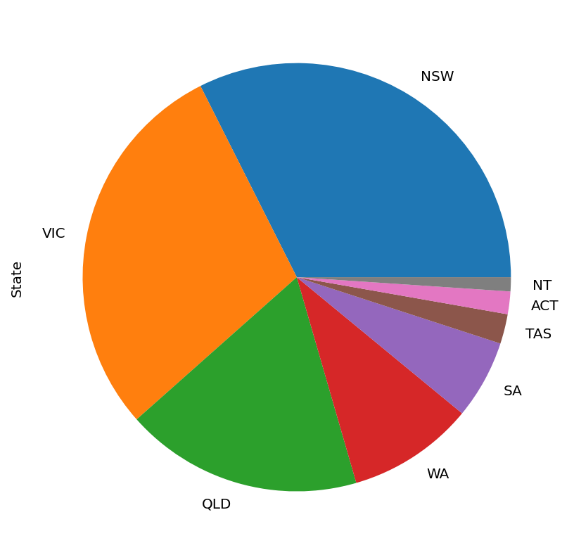


Figure 5:Catholic Sector

As from the data distribution most of the catholic school lies in state NSW and VIC similar to government sector.

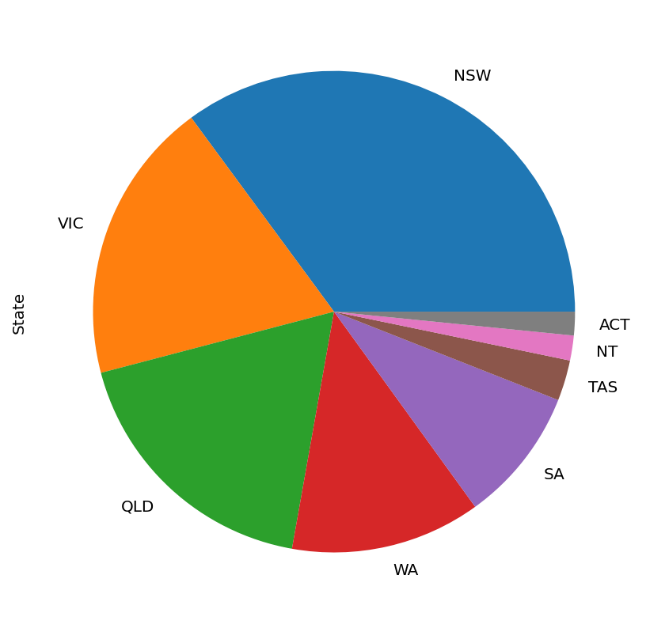


Figure 6:Independent School

The Independent school sector shows similar distribution as of catholic and government sector. But the NT state has a greater number of independent school as compared to ACT. But in catholic sector ACT has a greater number of catholic School as compared to the NT.

Code for all three pie charts –

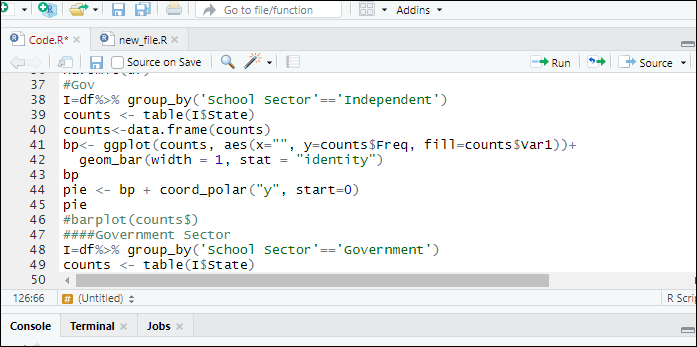


Figure 7:Pie chart code

Now let’s check the overall distribution of School Sector in the data .

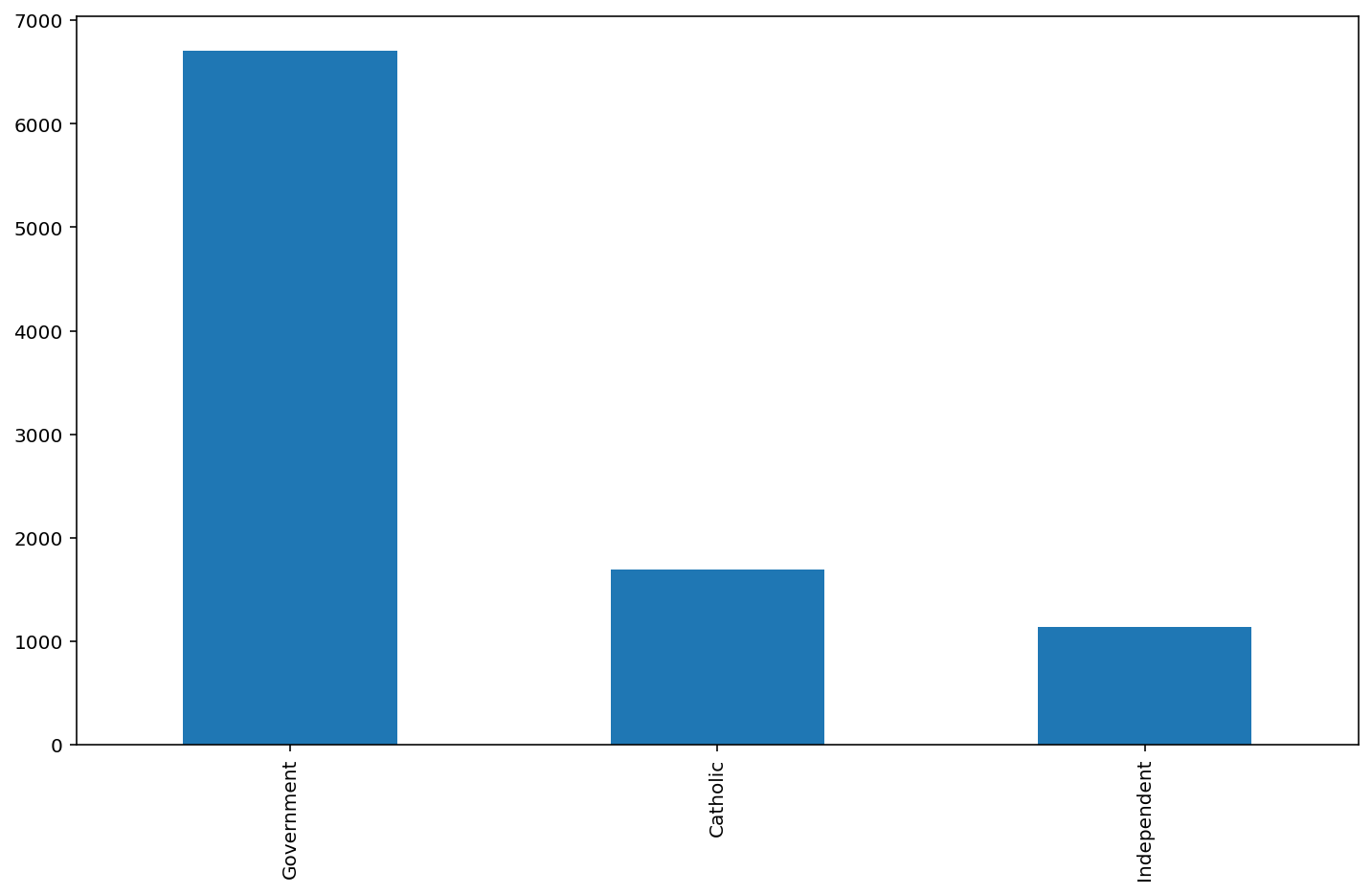


Figure 8:Sector distribution

Sector Distribution code

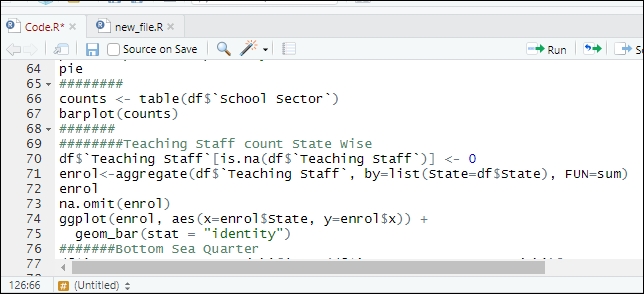


Figure 9:Sector code

From the graph we can see that in Australia the government school as more as compared to Catholic or independent schools . And from the pie chart of the government sector the most number of the government school are in NSW state.

It’s very important to know the distribution of the teaching staff . Teaching staff will help to compute the teacher to student ratio in any state. The teacher to student ration is the ration that explains the number of teacher allocated to given students.

Code for this bar plot

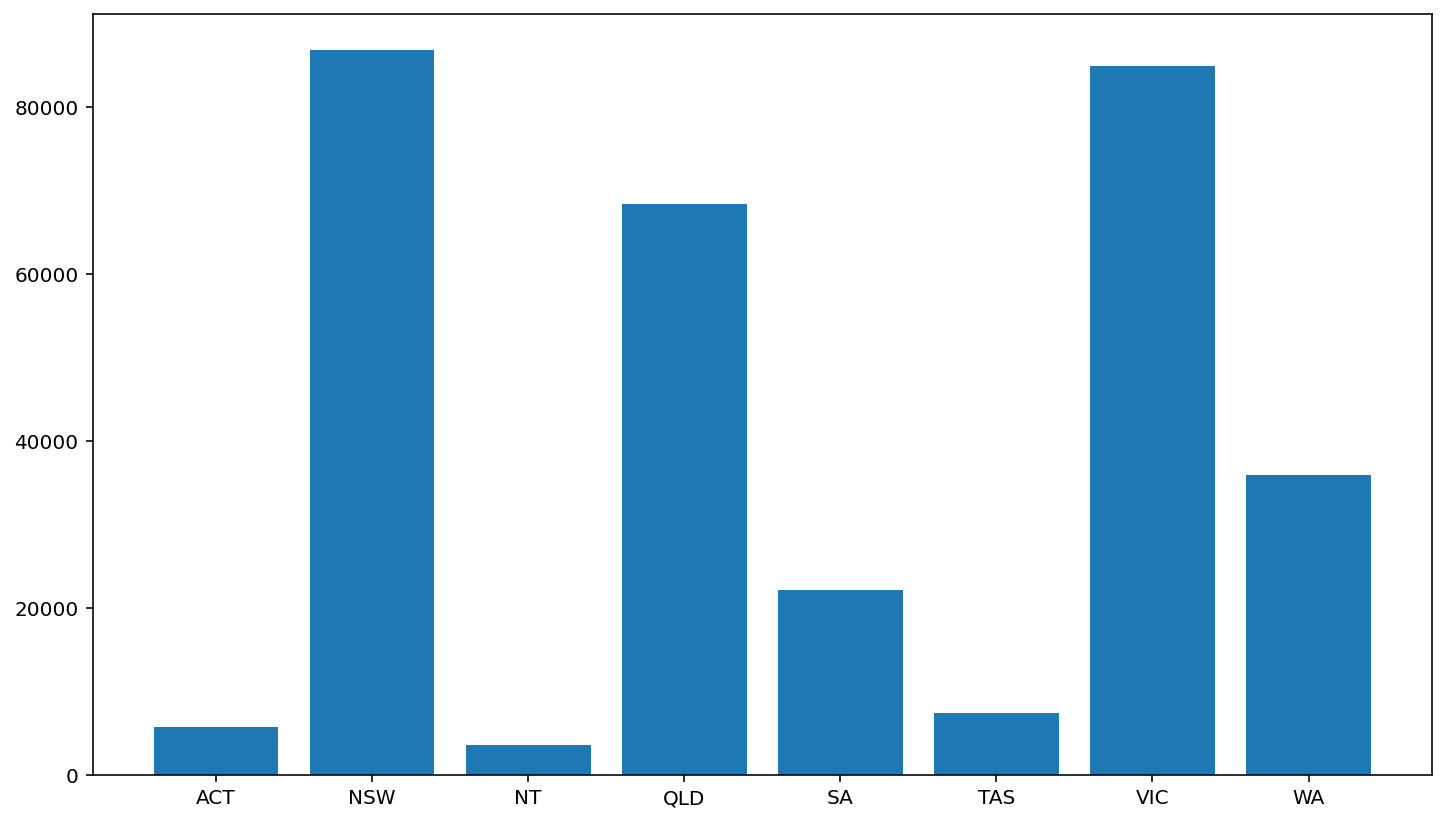


Figure 10:Teaching\_Staff\_Statewise

As from the graph we can see that the teaching staff of NSW is approximately equal to VIC . But from the above graphs we have seen that number of NSW has much more students than compared to VIC . But from the teaching Staff we have clearly seen that the number of the teaching staff of VIC state is same as that of NSW from this we can conclude that Teaching to Student ratio in state VIC is much greater than NSW.

Bottom Sea Quarter % Explains the number of students that are positioned in lowest socio-education advantage group. This data has to be read carefully state wise as this provide the information of students that belong to the low socio –education background in different States.

Staff Code-

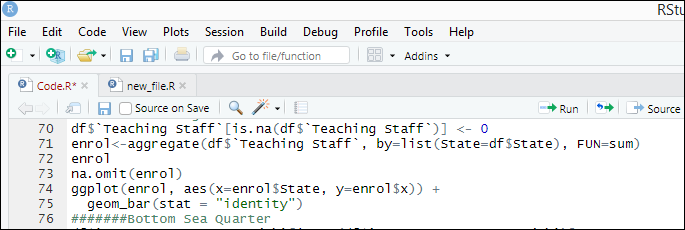


Figure 11:Staff Code

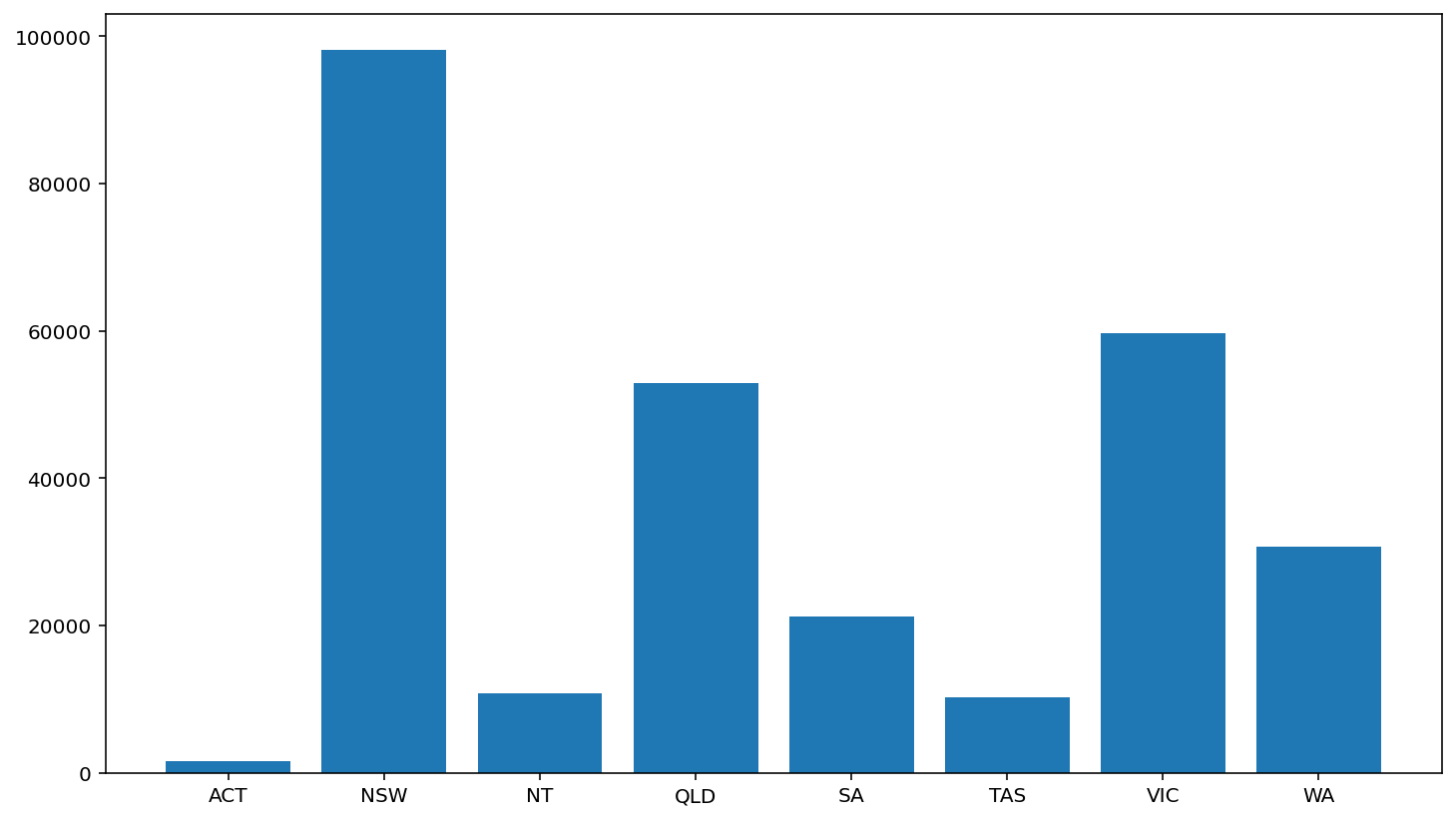


Figure 12:Bottom Sea Quarter %

Here you can see that the analysis of the of Bottom Sea Quarter group of people and as per the sum NSW has the highest value among States . This means that lower background students have a good chance of completing Higher education in NSW as compare to other States . Similarly the values for the Bottom sea quarter is approximately same for the state QLD and VIC.

Bottom Sea Quarter Code

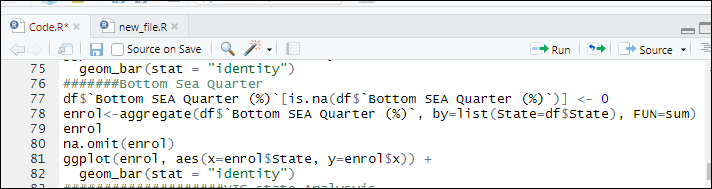


Figure 13:Bottom Sea Quarter

Analysis and Insights for the VIC state –

Lets compare the type of school sector in VIC state .

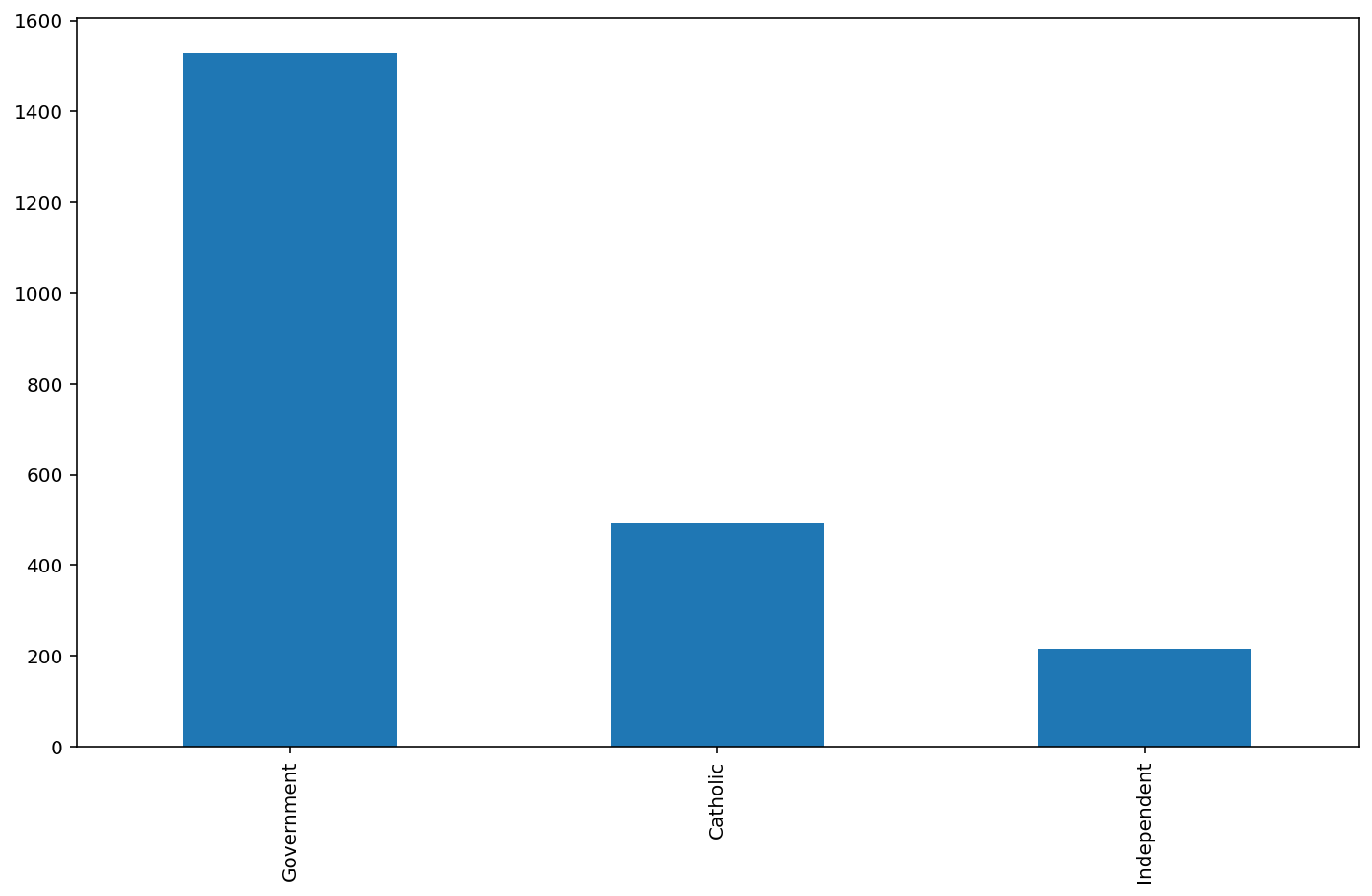


Figure 14: VIC School\_Sector

The state VIC has majority of Government school present with a count of 1400. There are more catholic schools in VIC as compared to Independent School.

VIC Sector Code-

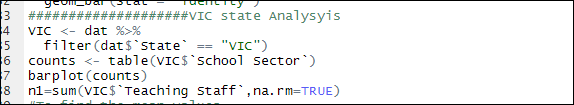


Figure 15:VIC sector

Lets figure out The difference between the Teaching Staff and non-teaching staff in state VIC .

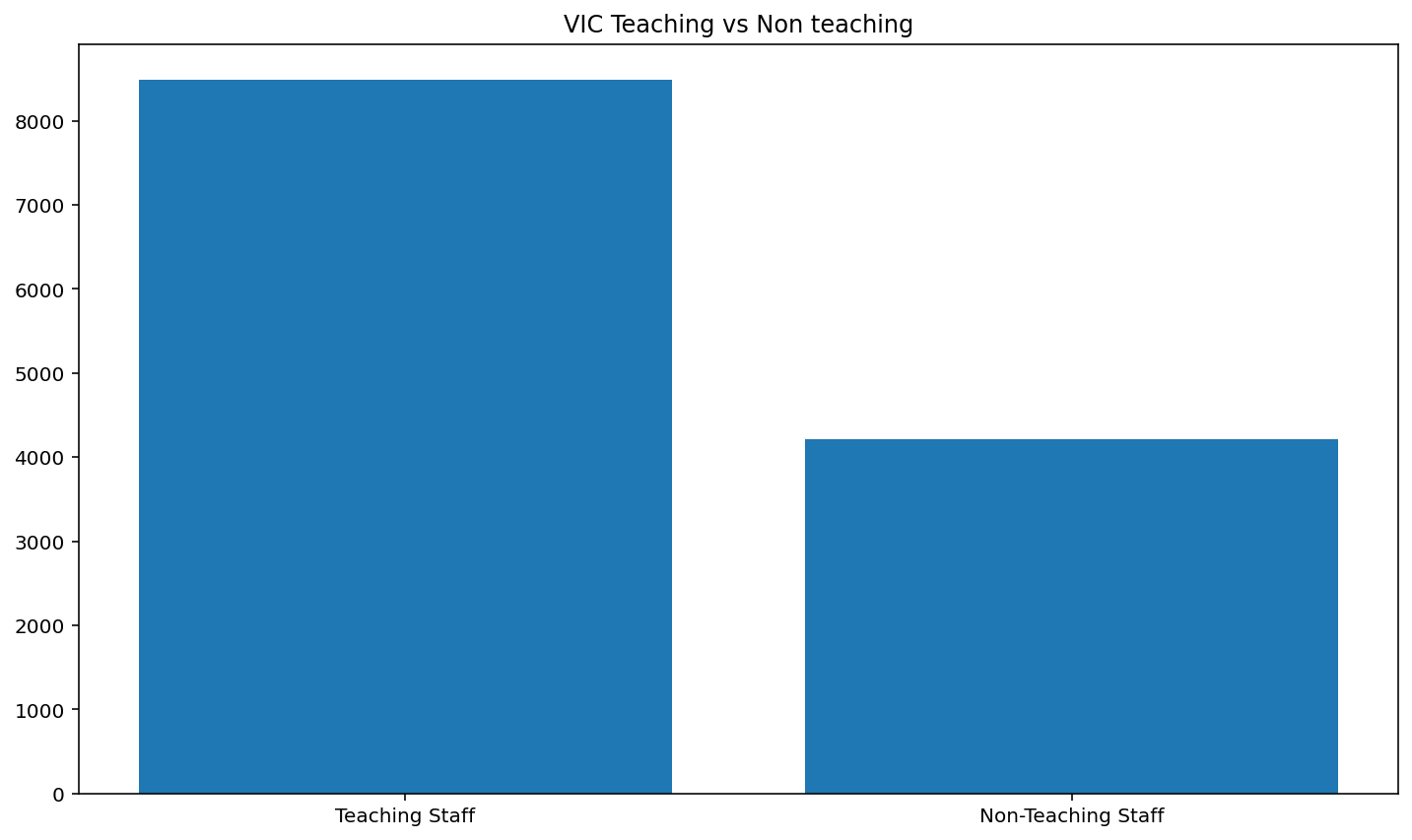


Figure 16:VIC Teaching vs Non-Teaching

It can be easily interpreted that Non-Teaching staff is approximately half to the Teaching Staff. This suggests the best ratio for teaching staff to non-teaching staff as the ratio of teaching to non-teaching does not extend to 1:3 ratio.

Teaching Staff vs Non-Teaching Staff-

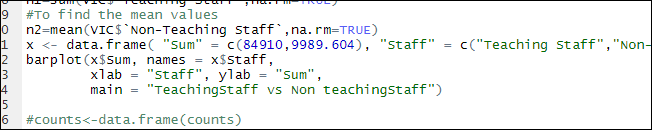


Figure 17:Teaching Staff vs Non Teaching Staff



Figure 18:VIC-Year\_Range

The year range shows the range of year level offered by school in VIC state. As from the data we can see that the maximum count of the year range is shown by Prep-6. Means that most of the school provide a course for the year Prep-6 and the count is about 1450. Other than that year range of 7-12 is the second highest count and is about 300.

Year range code-

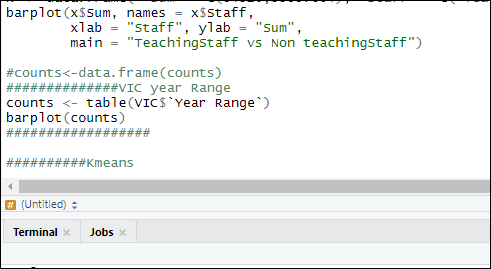


Figure 19:Year range VIC state

The state wise cluster is being formed to segment the school profile in proper clusters. K means is the technique to segment the clusters according to the attributes of the data. In R language kmeans() module is used to cluster the numerical data. To determine the value of clusters Silhouette method id used to find the optimal solution for the number of clusters. Here the attribute cost is plotted against the changing value of k(Batool and Hennig, 2019) .

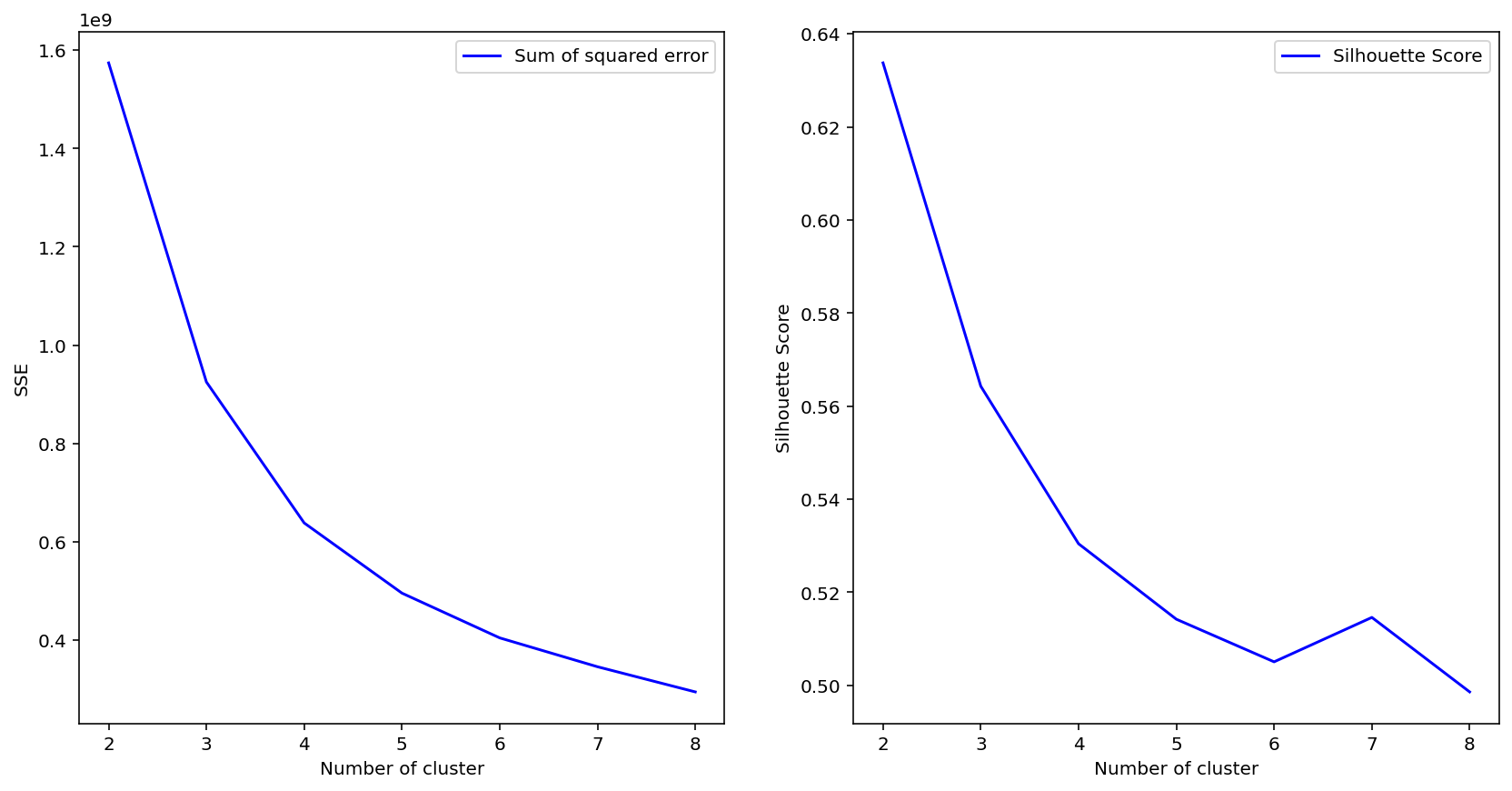


Figure 20:K value for cluster

As from the plot we can clearly see that the optimum number of clusters here is 5. As there is very less change in value of Sum Squared Error from number of clusters 5 to 6. From the Silhouette Score too the value of the optimum value of cluster is about 5 (Huang and Hu, 2018,).

K means Code-

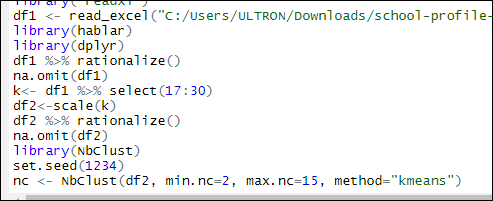


Figure 21K:means

# Conclusion –

From all the analysis it can easily concluded that in the given data the maximum number of Schools are from the state NSW. Similarly the most of the Australian school have government school. There are less number of enrollments on the colleges or schools as compared to boys. Australian government should focus on policies so to have large number of girl’s enrollment in their School. One of the major conclusion from the data is that the state VIC has good number of Student to teacher ratio as compared to other states. The number of clusters that are best fitted on data for k means clustering is 5 as described above.

# References –

Batool, F. and Hennig, C., 2019. Characterization and Development of Average Silhouette Width Clustering. *arXiv preprint arXiv:1910.11339*.

Huang, S. and Hu, P., 2018, April. K-Means Clustering Efficient Algorithm with Initial Class Center Selection. In *2018 3rd International Workshop on Materials Engineering and Computer Sciences (IWMECS 2018)*. Atlantis Press.

Son, H.S. and Kim, R.Y.C., 2016, February. A Method of Handling Metamodel Based on XML Database for SW Visulization. In *2016 International Conference on Platform Technology and Service (PlatCon)* (pp. 1-3). IEEE.