A Capstone Project Final Report On

“A COMPARATIVE STUDY ON SCHOOLS ACROSS THE STATE OF TAMIL NADU & KARNATAKA”

Submitted By

DHEERAJ

MANAS SARMAH

M ALI RAZA

Under the Guidance of

Prof. P V Subramanian

Data Scientist and Moderator

Great Learning, Chennai, TN

DATA SCIENCE AND ENGINEERING

GREAT LEARNING, BANGALORE

2018-19

Acknowledgements

We wish to place on record our deep appreciation for the guidance and help provided to us by our mentor, Prof. P. V. Subramanian. He helped us narrow down on the choice of the Project as well as the scope and focus area of the Project. He gave us valuable feedback at every stage to enhance the process and the outputs.

We would also like to place on record our appreciation for the guidance provided by all the faculty members of Great Lakes for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: October 12, 2018 Manas Sarmah

Place: Bangalore Dheeraj

M. Ali Raza

Certificate of Completion

I hereby certify that the project titled "**A COMPARATIVE STUDY ON SCHOOLS ACROSS THE STATE OF TAMIL NADU & KARNATAKA**" was undertaken and completed under my supervision by Manas Sarmah, Dheeraj and M. Ali Raza, students of the Postgraduate Program in Data Science and Engineering (PGPDSE – May 2018).

Date: October 12, 2018 (P.V. Subramanian)

Place: Bangalore Mentor

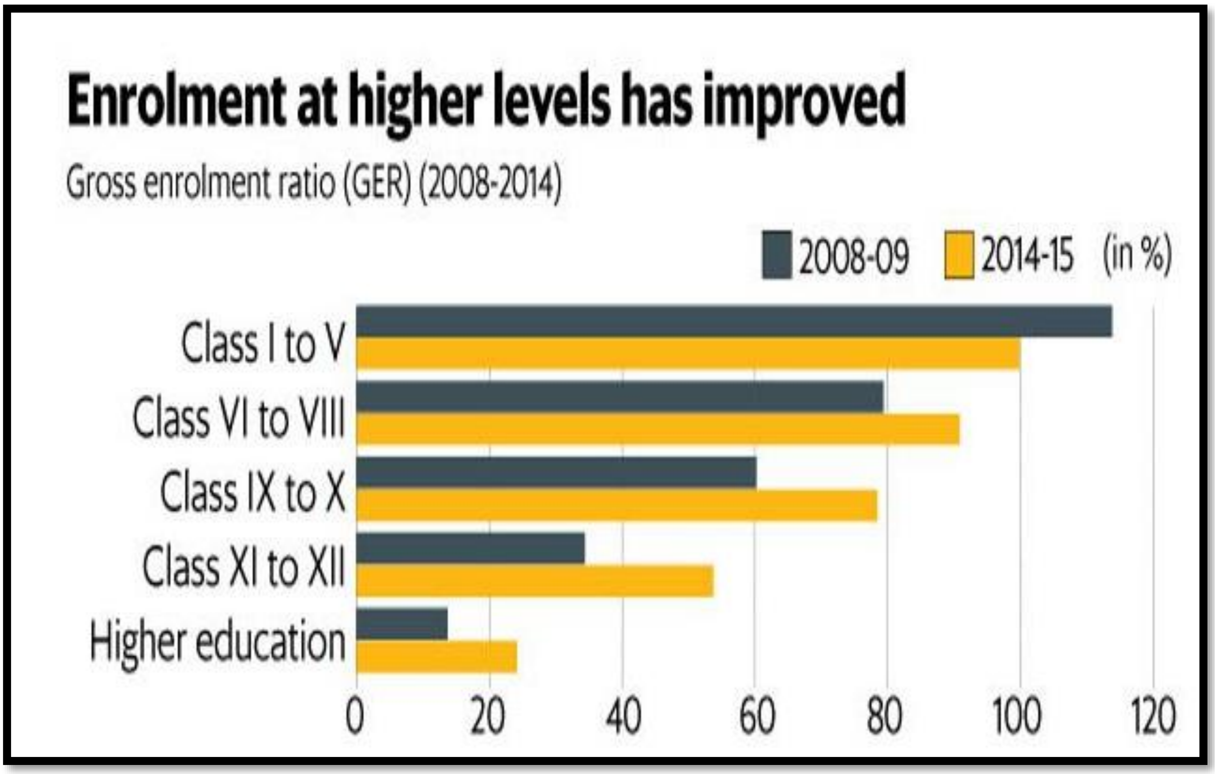
TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

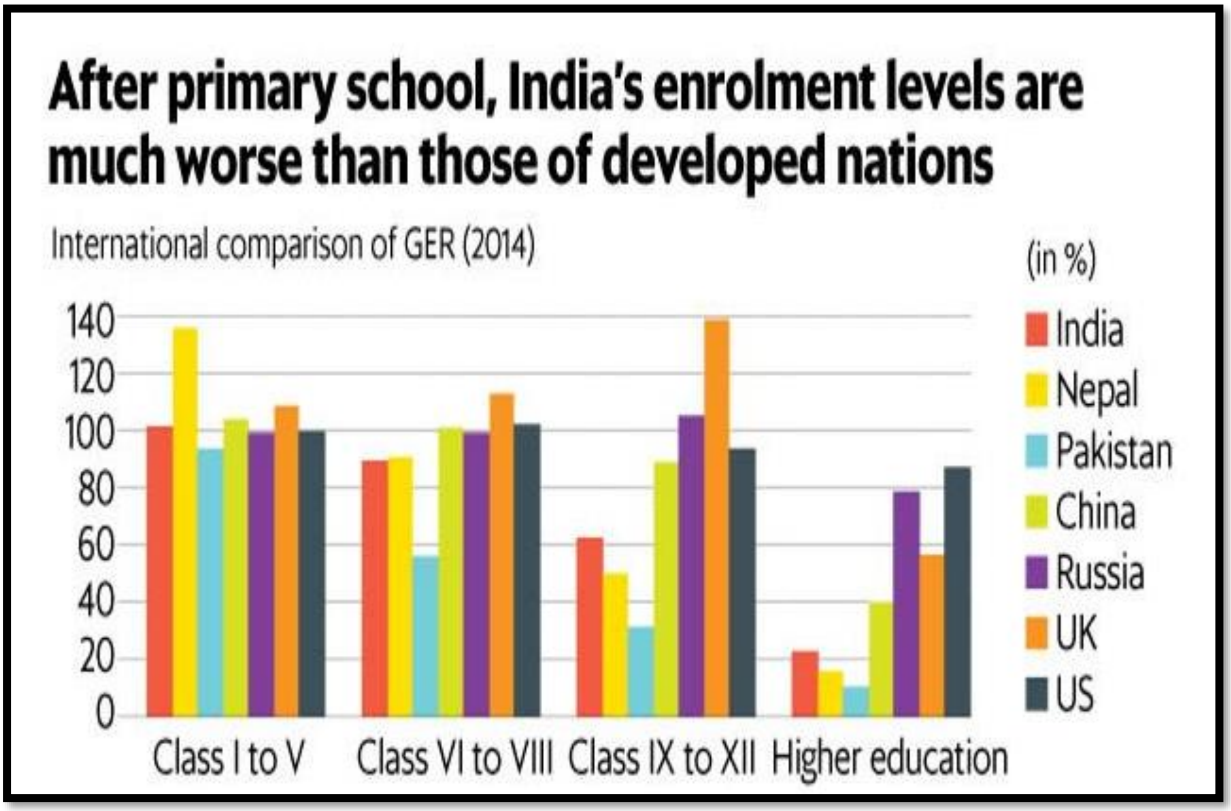
India has one of the most complex education systems with more than 1.5 million schools with approximately 315 million students. India has the second largest population in the world and so most population related numbers are bound to be big. This is not only the largest student body in the world, but Indian students would make for the fourth biggest country in the world, nearly touching the US with a population of 318 million. For comparison, China's student population is about 252 million as per UNESCO statistics.

The Union budget speech recognized the need to measure learning outcomes in schools. An amendment has been introduced to the Right to Education (RTE) Act to permit detentions after class V and class VIII after a test, remedial education and a retest. In light of this, the following data captures some trends with respect to enrolment, drop-out, and transition rates across various levels of education.



GER in class I-V reduced from 114% in 2008-2009 to 100% in 2014-2015

While universal enrolment has been achieved at the elementary level (class I-VIII), the enrolment consistently falls with successive levels of education. Gross enrolment ratio (GER) is the student enrolment as a proportion of the corresponding eligible age group in a given year. GER in class I-V reduced from 114% in 2008-09 to 100% in 2014-15. The above-100% enrolment rate in 2008-09 indicates that students enrolled in class I-V included those younger than six or older than 10 years. In 2014-15, enrolment in class I-V was about 100%, which signals a more age-appropriate class composition.



In higher education, India's enrollment rate stands at 23%, as against about 87% in the US, 57% in UK and 39% in China

India's enrolment rate in primary education (class I-V) is comparable to that of developed countries. However, it falls behind these countries after class VI. In higher education, India's enrolment rate stands at 23%, as against about 87% in the US, 57% in the UK and 39% in China.

1.1 PROBLEM STATEMENT:

In the above context, we feel, if we closely study the enrollment of students we should be able to identify the patterns, identify the correlation factors on key levels of enrollment of students across different states.

It is important to understand that there might be underlying internal and external factors like poverty, lack of interest on education, high teacher absenteeism, low student teacher ratio and other family problems influencing the enrollment of students in schools.

There were various articles and reasons mentioned that affected enrollment of students in various ways. Through this study (with the particular data available to us), we hope to uncover some insights on factors that influence enrollment of students which can prove extremely beneficial to school managements.

1.2 OBJECTIVE AND SCOPE OF THE PROJECT:

1.2.1 Objective:

The primary objectives of this study are:

- Study the enrolment of students at schools across Tamil Nadu and Karnataka.
- Identify the key factors that influences and dominates the enrolment of students
- Explore the possibility of developing a Predictive Model(s) for predicting the enrolment of students.

1.2.2 Scope:

- The scope of the study covers the conditions of the various schools, basic amenities provided, different types of funds granted, teacher student ratios and the qualification of teachers and their impact on enrolment.
- The study covers only the Indian states of Tamil Nadu and Karnataka.
- The data is collected over 2 years from educational years 2014-15 to 2015-16.

1.2.2 Out of Scope:

- A lot of articles online comment on high drop-out rates for Higher Secondary Education (Class XI and XII) students. This study will only explore the factors leading to enrolment, not the drop-out rates and other factors that lead to drop-outs.
- Due to data limitation, we are only exploring till the higher secondary education sector. Anything above that is out of the scope of this study.
- This study only pertains to the Indian states of Tamil Nadu and Karnataka. Factors influencing the enrolment of students in other Indian states are not explored and the models depicted in this report would not be justified on predicting the enrolment of students for other Indian states and territories.

1.3 DATA SOURCE:

The two main sources of educational statistics are the educational institutions and households. The educational institutions provide the data on enrolment and number of teachers whereas information on aspects like literacy, educational level of population, private expenditure on education etc., is available only from households.

We collected datasets for the year of 2014-2015 and 2015-2016 from National University of Educational Planning and Administration (NUEPA), 2 years of annual survey report by District Information System for Education (DISE) and Educational Management Information System Department (EMIS).

The dataset after collating has the following variables and structure:

1.	SCHOOL_CODE	:	School Code
2.	DISTNAME	:	District Name
3.	AC_YEAR	:	Academic year
4.	SCHOOL_NAME	:	School Name
5.	STATE	:	State Name
6.	BLOCK_NAME	:	Block Name
7.	CLUSTER_NAME	:	Cluster Name
8.	VILLAGE_NAME	:	Village Name
9.	BLDSTATUS	:	Status of the school Building
10.	CLROOMS	:	Total Classrooms used for instructional purposes
11.	CLGOOD	:	Pucca classroom Good condition
12.	CLMAJOR	:	Pucca classroom need major repair
13.	CLMINOR	:	Pucca classroom need minor repair
14.	TOILETB	:	No. of Toilet Seats Constructed/Available for Boys
15.	TOILET_G	:	No. of Toilet Seats Constructed/Available for Girls
16.	MEALSINSCH	:	Status of Mid-day Meal
17.	CAL_YN	:	Does the school have Computer Aided Learning (CAL)
18.	HMROOM_YN	:	Separate room for Head Teacher/ Principal available
19.	ELECTRIC_YN	:	Electricity connection available in the school
20.	BNDRYWALL	:	Type of Boundary wall
21.	LIBRARY_YN	:	Whether school has Library facility/Book
22.	PGROUND_YN	:	Playground Yes No
23.	BOOKINLIB	:	If Yes, No. of Books in school library
24.	WATER	:	Main source of drinking water facility
25.	MEDCHK_YN	:	Whether Medical check-up of students conducted last year
26.	RAMPS_YN	:	Whether ramp(s) is/are available
27.	COMPUTER	:	Total no. of computers available for teaching & learning purposes
28.	RURURB	:	School is in Rural area or Urban area
29.	MEDINSTR1	:	Medium of Instruction
30.	ESTDYEAR	:	Year of Establishment
31.	BOARDSEC	:	For Secondary sections
32.	BOARDHSEC	:	For Higher Secondary sections
33.	PPSEC_YN	:	Pre-primary Section available or not (yes or no)
34.	SCHRES_YN	:	Is the School Residential or not (yes or no)
35.	SCHMGT	:	Managed by (School Management)
36.	SCHCAT	:	School category
37.	SCHTYPE	:	Type of school
38.	SCHSHI_YN	:	Is the school a shift school?
39.	WORKDAYS	:	Number of instructional days (previous academic year)
40.	RESITYPE	:	If ‘Yes’, Type of residential school
41.	CONTI_R	:	Amount of School Development Grant Receipt
42.	CONTI_E	:	Amount of School Development Grant Expenditure
43.	SCHMNTCGRANT_R	:	Amount of Teacher Learning Material Receipts
44.	SCHMNTCGRANT_E	:	Amount of Teacher Learning Material Expenditure
45.	FUNDS_R	:	Funds from other sources Receipts

46. FUNDS_E : Funds from other sources Expenditure
47. APPROACHBYROAD : Whether school is approachable by all-weather roads?
48. CCE_YN : Is CCE being implemented in school at elementary level?
49. PCR_MAINTAINED : Are pupil cumulative records being maintained?
50. PCR_SHARED : Are pupil cumulative records shared with parents?
51. WSEC25P_ENROLLED : Number of student Continuing Who got admission under 25% . quota in previous year
52. SMCSDP_YN Whether SMC : Prepare the School Development Plan
53. SPLTRG_CY_ENROLLED_B : No. of children enrolled Special Training in current year– Boys
54. SPLTRG_CY_ENROLLED_G : No. of children enrolled Special Training in current year– Girls
55. SPLTRG_CY_PROVIDED_B : No. of children provided Special Training in current year– Boys
56. SPLTRG_CY_PROVIDED_G : No. of children provided Special Training in current year– Girls
57. SPLTRG_PY_ENROLLED_B : No. of children enrolled Special Training in previous year– Boys
58. SPLTRG_PY_ENROLLED_G : No. of children enrolled Special Training in Previous year– Girls
59. SPLTRG_PY_PROVIDED_B : No. of children provided Special Training in previous year– Boys
60. SPLTRG_PY_PROVIDED_G : No. of children provided Special Training in previous year– Girls
61. SPLTRG_BY : Who conducted the Special training?
62. SPLTRG_PLACE : Where was the training conducted?
63. SPLTRG_TYPE : Type of Training conducted
64. TXTBKRECD_YN : Where any text book was received in the current academic year?
65. TCH_MALE : Total Male Teachers
66. TCH_FEMALE : Total Female Teachers
67. TCH_NR : Teachers with No Response in Gender Column
68. HEADTCH : Total Head Teachers in Schools
69. GRADABOVE : Total Teachers Graduate and above
70. TCHWITHPROF : Total Teachers with Professional Qualification
71. DAYSINVLD : Total no. of working days spent to non-teaching assignments
72. TCHINVLD : Total teachers involved in (non-teaching assignments)
73. TB_ENROLL : Total Enrolment of Boys
74. TG_ENROLL : Total Enrolment of Girls
75. HTCHNAME : Name of the Headmaster
76. MDM_MAINTAINER : Master Data Management
77. NOINSPECT : No. of inspection

1.4 TOOLS AND TECHNIQUES:

We have the used the following analytical techniques/methodology for analyzing the Data

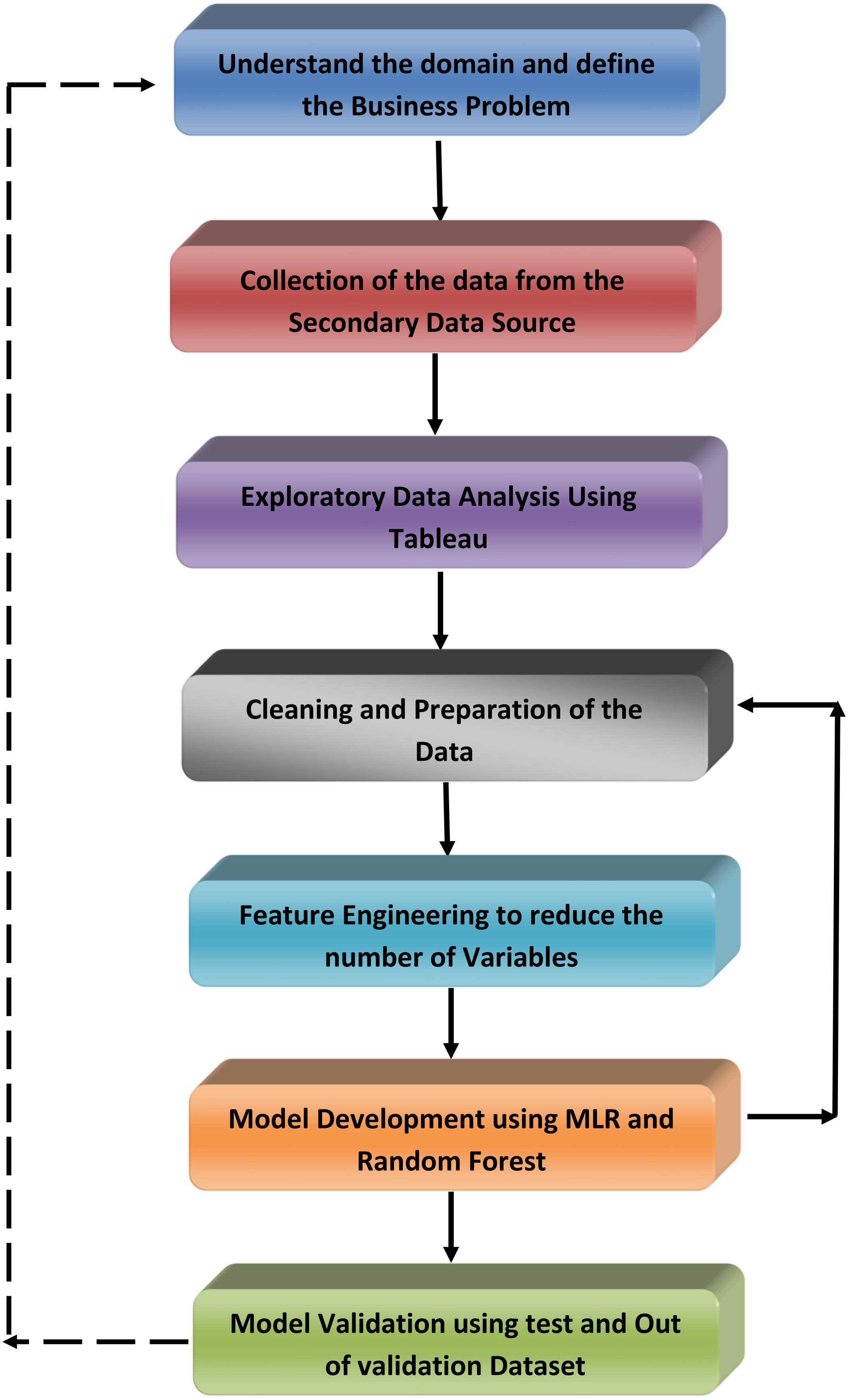
1. Summary Statistics for each variable
2. Identification of frequency of standard violation for each of the factors
3. Using Graphs and Box Plots to visually represent them
4. Identification of significant Metrological factors through correlation and regression methodology
5. Using Multiple Linear Regression, L1 and L2 regularization & Random Forest for Model Development
6. Tools used: Python, Tableau & Excel
7. Techniques: Box Plot, Histogram, Bar Chart, Line Chart, Infographics, Visual Clues, Correlation Matrix, Multiple Linear Regression, regularization and Bagging method
8. We have used Python Programming environment and Microsoft Excel for our analysis and Tableau for data visualization.

1.5 ANALYTICALAPPROACH:

The Analytical Approach will involve the following (not necessarily in the order) activities:

1. Defining the Problem Statement
    a. Based on study undertaken on the education field
2. Collection of the Data
    a. 2 years of annual survey report by District Information System for Education (DISE)
    b. datasets for the year of 2014-2015 and 2015-2016 from National University of Educational Planning and Administration (NUEPA)
    c. Educational Management Information System Department (EMIS)
3. Exploratory Data Analysis
    a. Visualization using TABLEAU
    b. Understanding the factors influencing enrolment
    c. Chart Types: Bar charts, Histograms, Boxplots, Scatterplots & Infographics
    d. Correlation Matrix
4. Cleaning and Preparation of the Data
    a. Handling the missing values
    b. Two way approaches for outlier treatments: capping and dropping
5. Feature Engineering
    a. Dropped variables with high percentage of 0's
    b. Reduced the # of variables from 77 to 53
6. Model Development
    a. Used Python Programming Language
    b. As the response variable is Enrolment of students (continuous variable), we opted for the following models:
        i. Multiple Linear Regression
        ii. Random Forest
        iii. Polynomial Regression
7. Model Validation
    a. Selection of the best fit mode
    b. Validate on test and out of validation dataset (2015-2016)

HIGH LEVEL ANALYTICAL APPROACH



Understand the domain and define the Business Problem
Collection of the data from the Secondary Data Source
Exploratory Data Analysis Using Tableau
Cleaning and Preparation of the Data
Feature Engineering to reduce the number of Variables
Model Development using MLR and Random Forest
Model Validation using test and Out of validation Dataset

1.6 LIMITATIONS:

There are few limitations that this study has w.r.t data and the methodology that can be used.

- Due to time and cost constraints we could not deploy a primary source for data collection. We have not explored how the data was collected or methods were used in collecting the data.
- Since the models are built only on data pertaining to the Indian states of Tamil Nadu and Karnataka, we cannot evaluate the model on other states of India or use it as a general purpose model for predicting enrolment of students with high accuracy.
- The data was limited to a time period of 2 years (2014-15 to 2015-16), so previous knowledge of how the data behaved in the earlier years is available to us. If that data was available, we could have evaluated the time series components of trend and seasonality into our model.

CHAPTER 2

DATA DESCRIPTION AND PREPARATION

2.1 DATA MANAGEMENT:

Based on the scope of the project, we have collated data for 2 years (2014 to 2016) which spans across 77 variables.

We collected datasets for the year of 2014-2015 and 2015-2016 from National University of Educational Planning and Administration (NUEPA), 2 years of annual survey report by District Information System for Education (DISE) and Educational Management Information System Department (EMIS).

2.2 DATA TABLE – LIST OF VARIABLES:

After collating all the data together, we arrive with two datasets. One of them for the academic year of 2015-16 and the other for the academic year of 2016-17. The final dataset (both the academic years combined) has 77 variables and 2,65,292 rows. The list of all the variables along with their description and data types are given below:

Table: List of Variables and their Data Type

Variable Abbreviation	Variable Description	Data Type
SCHOOL_CODE	School Code	Numerical
DISTNAME	Name of the District the School is present in	Categorical
AC_YEAR	Academic year	Categorical
SCHOOL_NAME	Name of the School	Categorical
STATE	State Name	Categorical
BLOCK_NAME	Name of the Block in which the School is present	Categorical
CLUSTER_NAME	The Cluster in which the School is present	Categorical
VILLAGE_NAME	Village Name in which the School is present	Categorical
BLDSTATUS	Status of the School Building	Categorical
CLROOMS	Total Classrooms used for instructional purposes	Numerical
CLGOOD	Number of Pucca classrooms in good condition	Numerical
CLMAJOR	Number of classrooms that need major repair	Numerical
CLMINOR	Number of classrooms that need minor repair	Numerical
TOILETB	Number of Toilet Seats Constructed/Available for Boys	Numerical

TOILET_G	Number of Toilet Seats Constructed/Available for Girls	Numerical
MEALSINSCH	Status of Mid-day Meal	Categorical
CAL_YN	Does the school have Computer Aided Learning (CAL) – Yes/No	Binary
HMROOM_YN	Separate room for Head Teacher/ Principal available – Yes/No	Binary
ELECTRIC_YN	Electricity connection available in the school – Yes/No	Binary
BNDRYWALL	Type of Boundary Wall	Categorical
LIBRARY_YN	Whether school has Library facility/Books – Yes/No	Binary
PGROUND_YN	Playground – Yes/No	Binary
BOOKINLIB	If Yes for LIBRARY_YN, number of books in school library	Numerical
WATER	Main source of drinking water facility	Categorical
MEDCHK_YN	Whether Medical check-up of students were conducted last year – Yes/No	Binary
RAMPS_YN	Whether ramp(s) is/are available – Yes/No	Binary
COMPUTER	Total number of computers available for teaching and learning purposes	Numerical
RURURB	Whether the School is in Rural area or Urban area	Categorical
MEDINSTR1	Medium of Instruction	Categorical
ESTDYEAR	Year of Establishment	Date
BOARDSEC	For Secondary sections	Categorical
BOARDHSEC	For Higher Secondary sections	Categorical
PPSEC_YN	Pre-primary Section available or not – Yes/No	Binary
SCHRES_YN	Is the School Residential or not – Yes/No	Binary
SCHMGT	Managed by (School Management)	Categorical
SCHCAT	School category	Categorical
SCHTYPE	Type of school	Categorical
SCHSHI_YN	Is the school a shift school – Yes/No	Binary
WORKDAYS	Number of instructional days (for previous academic year)	Numerical
RESITYPE	If ‘Yes’, Type of residential school	Categorical
CONTI_R	Amount of School Development Grant Receipt	
CONTI_E	Amount of School Development Grant Expenditure	Numerical
SCHMNTCGRANT_R	Amount of Teacher Learning Material Receipts	Numerical
SCHMNTCGRANT_E	Amount of Teacher Learning Material Expenditure	Numerical
FUNDS_R	Funds from other sources Receipts	Numerical
FUNDS_E	Funds from other sources Expenditure	Numerical
APPROACHBYROAD	Whether school is approachable by all-weather roads?	Categorical

CCE_YN	Is CCE being implemented in school at elementary level?	Binary
PCR_MAINTAINED	Are pupil cumulative records being maintained – Yes/No	Binary
PCR_SHARED	Are pupil cumulative records shared with parents – Yes/No	Binary
WSEC25P_ENROLLED	Number of student Continuing Who got admission under 25% quota in previous year	Numerical
SMCSDP_YN	Whether SMC Prepare the School Development Plan	Binary
SPLTRG_CY_ENROLLED_B	No. of children enrolled Special Training in current year – Boys	Numerical
SPLTRG_CY_ENROLLED_G	No. of children enrolled Special Training in current year – Girls	Numerical
SPLTRG_CY_PROVIDED_B	No. of children provided Special Training in current year – Boys	Numerical
SPLTRG_CY_PROVIDED_G	No. of children provided Special Training in current year – Girls	Numerical
SPLTRG_PY_ENROLLED_B	No. of children enrolled Special Training in previous year – Boys	Numerical
SPLTRG_PY_ENROLLED_G	No. of children enrolled Special Training in Previous year – Girls	Numerical
SPLTRG_PY_PROVIDED_B	No. of children provided Special Training in previous year – Boys	Numerical
SPLTRG_PY_PROVIDED_G	No. of children provided Special Training in previous year – Girls	Numerical
SPLTRG_BY	Person who conducted the Special training?	Categorical
SPLTRG_PLACE	Place where was the training conducted?	Categorical
SPLTRG_TYPE	Type of Training conducted	Categorical
TXTBKRECD_YN	Were any text book(s) received in the current academic year – Yes/No	Binary
TCH_MALE	Total Male Teachers	Numerical
TCH_FEMALE	Total Female Teachers	Numerical
TCH_NR	Teachers with No Response in Gender Column	Numerical
HEADTCH	Total Head Teachers in Schools	Numerical
GRADABOVE	Total Teachers Graduate and above	Numerical
TCHWITHPROF	Total Teachers with Professional Qualification	Numerical
DAYSINVLD	Total no. of working days spent to non-teaching assignments	Numerical
TCHINVLD	Total teachers involved in (non-teaching assignments)	Numerical
TB_ENROLL	Total Enrolment of Boys	Numerical
TG_ENROLL	Total Enrolment of Girls	Numerical
HTCNAME	Name of the Headmaster	Categorical
MDM_MAINTAINER	Master Data Management	Categorical
NOINSPECT	Whether Inspection was done Y/N	Binary

Table 1: Showing list of Variables and their Data Types

2.3 DATA QUALITY:

The ‘enrolment of students’ data collected had missing values present in it. Out of its 77 features, some features had more than 30% of the data missing. While most other features had majority of its data present with only a few percentage of missing values. Also, there were a couple of features which had majority of its values as 0’s.

Also, majority of the features had outliers present in them with some having highly skewed distributions.

We will explore all the missing values with proper statistics and numbers and come up to a conclusion on how to treat or impute them. We will look into the outliers and explore them further in our study to understand their nature and the steps needed to work around or with them.

2.4 DATA PREPARATION:

2.4.1 Variable Transformation

- Our aim is to find the total number of enrollments; therefore, we have added the two variables TB_ENROLL and TG_ENROLL which are the number of boys and girls enrolled into a single column called “Total_Enroll”.
- The feature HTCNAME had blank spaces in the front which didn’t show up as missing values. We replaced all the values in HTCNAME where a “whitespace” was present as NULL values.
- Since most of our features are categorical in nature which cannot be used in Linear Models, we have transformed them into numerical values with the help of Label Encoders.
- For our Random Forest models, we have not done any transformations.

2.4.2 Missing Value Treatment

Initially, the data collected contained a total of 77 variables with 2,65,292 rows. The features TB_ENROLL and TG_ENROLL were combined into a single feature called “Total_Enroll”.

After this we are left with a final dataset with 76 features and 2,65,292 rows.

We explored the missing values present in the data in two ways:

- Primarily by checking for the amount of missing values that a particular feature has in comparison to the whole dataset. If the certain feature has more than 40% of its data missing, we have removed that particular feature itself as it would give to added advantage to the model as almost half its data is missing. However, features have not been removed just on the above mentioned rule. We have also checked to see what that feature is and if it would still actually have an impact in the model from a business point of view.
- And secondarily, by checking for the remaining amount of missing values present in the data and check if dropping all of them would result in losing a considerable amount of data.

The variables with their corresponding number of missing values present in them are shown below:

```
BLDSTATUS             8           SCHCAT                4
MEALSINSCH         1935           SCHTYPE              12
CAL_YN              185           SCHSHI_YN           688
HMROOM_YN           221           NOINSPECT        149596
ELECTRIC_YN          68           RESITYPE          15741
BNDRYWALL           435           APPROACHBYROAD       51
LIBRARY_YN           75           CCE_YN             3589
PGROUND_YN           79           PCR_MAINTAINED     4699
WATER                76           PCR_SHARED         4725
MEDCHK_YN           327           SMCSDP_YN           119
RAMPS_YN            533           MDM_MAINTAINER   159731
RURURB                7           SPLTRG_BY           492
BOARDSEC          23070           SPLTRG_PLACE       2642
BOARDHSEC         22881           SPLTRG_TYPE        1654
PPSEC_YN           2834           TXTBKRECD_YN      12878
SCHMGT                5           HTCHNAME         208417
```

Here, MDM_MAINTAINER, HTCHNAME and NOINSPECT have more than 40% of missing values. The below table shows the exact percentage of missing values present in these columns in comparison to the total size of the whole dataset.

```
HTCHNAME          78.561359
MDM_MAINTAINER    60.209505
NOINSPECT         56.389186
```

- HTCNAME is the name of the head-teacher that the school has. 78% of the data is already missing and since the name of a head-teacher cannot provide us with any insight for the model, we have decided to drop this feature.
- MDM_MAINTAINER is the master data management maintainer. 60% of the data is missing for it. For the 40% of data that is available we hardly get any insights as the categories are divided into 6 groups with ‘others’ being the most prominent group. So, we have decided to drop this variable also as it adds no value to our model.
- NOINSPECT is whether an inspection was done or not on the school. No further information of when is detailed. Moreover, 56% of the data is missing for this feature. So, we have decided to drop this variable as well as it provides no value to our model also.

After dropping these 3 columns, we are left with 74 features and 2,65,292 attributes.

Looking at the other missing values, each feature has less than 10% of missing values.

These missing values constitute 16 percent of the whole data available. Now, if we decide to drop these missing values row-wise, we will lose just 16% of the original data. We are still left with 84% of the data for modelling. Considering that imputing these missing values using MICE package in R, we might change the structure of the data; we have better decided to drop these missing values which is only 16 percent of the original data and it still leaves us with more than 2 lakh rows for modelling.

After dropping missing values, we are left with 2,21,242 rows and 74 columns.

We also saw that a few of the variables were sparse in nature (i.e. most of the values were 0’s).

These variables were:

```
CLMAJOR                  84.171631
CLMINOR                  79.539599
WSEC25P_ENROLLED         93.609260
SPLTRG_CY_ENROLLED_B     97.810994
SPLTRG_CY_ENROLLED_G     97.994956
SPLTRG_CY_PROVIDED_B     98.421638
SPLTRG_CY_PROVIDED_G     98.577576
SPLTRG_PY_ENROLLED_B     99.117256
SPLTRG_PY_ENROLLED_G     99.189123
SPLTRG_PY_PROVIDED_B     99.004710
SPLTRG_PY_PROVIDED_G     99.073413
TCH_NR                   99.800671
DAYSINVLD                97.894161
TCHINVLD                 97.894161
```

This table represents the amount of 0’s present in that columns. We queried to find columns which had more than 70% of the data as 0’s. So, we have decided to drop those variables which have over 70% values as 0’s.

These variables are CLMAJOR, CLMINOR, WSEC25P_ENROLLED, SPLTRG_CY_ENROLLED_B, SPLTRG_CY_ENROLLED_G, SPLTRG_CY_PROVIDED_B, SPLTRG_CY_PROVIDED_G, SPLTRG_PY_ENROLLED_B, SPLTRG_PY_ENROLLED_G, SPLTRG_PY_PROVIDED_B, SPLTRG_PY_PROVIDED_G, TCH_NR, DAYSINVLD, and TCHINVLD.

Since, most of the data present in these columns are sparse, we have decided to remove these columns as they add no value to our models.

After this, we are left with 2,21,242 rows and 60 columns.

We also remove 'SCHOOL_CODE', 'DISTNAME', 'SCHOOL_NAME', 'BLOCK_NAME', 'CLUSTER_NAME', 'VILLAGE_NAME', which are nothing but names and codes of cities and schools.

After this, we are left with 2,21,242 rows and 53 columns.

2.4.3 Outlier Treatment:

For each of the variables, we have done boxplots and histograms to understand the distribution and the presence of outliers in a better way.

The approach followed in this study for outliers are done in two ways:

1. The outliers are capped to their whisker values i.e., +/- 1.5 multiplied by the Interquartile Range:

$$+/-\ (1.5\ *\ IQR)$$

   This results in 2,21,242 rows and 53 columns.

2. The outliers are dropped for those particular rows. This results in 1,61,234 rows and 53 columns.

All the predictive models explored in this study has been done for both these two approaches: Once, when the outliers have been capped to their respective whisker values and the other, when the outliers have been dropped.

CHAPTER 3

EXPLORATORY DATA ANALYSIS

3.1 EXPLORATORY DATA ANALYSIS APPROACH:

EDA is an approach to analyze data sets to summarize their main characteristics, often with visual effects. Exploratory Data Analysis involves both graphical displays of data and numerical summaries of data. A data set is often represented as a matrix. There is a row for each unit. There is a column for each variable.

We are using box plots for the exploratory data analysis using Tableau.

We are doing EDA so that we get insights about the dataset before we start the modelling or perform any kind of operations on them. Generally, EDA is a compulsory step in the problem solving process. It helps to understands how variables are distributed and what they signify in real life. We can also identify outliers which may be a major problem in model building. We have done the following plots:

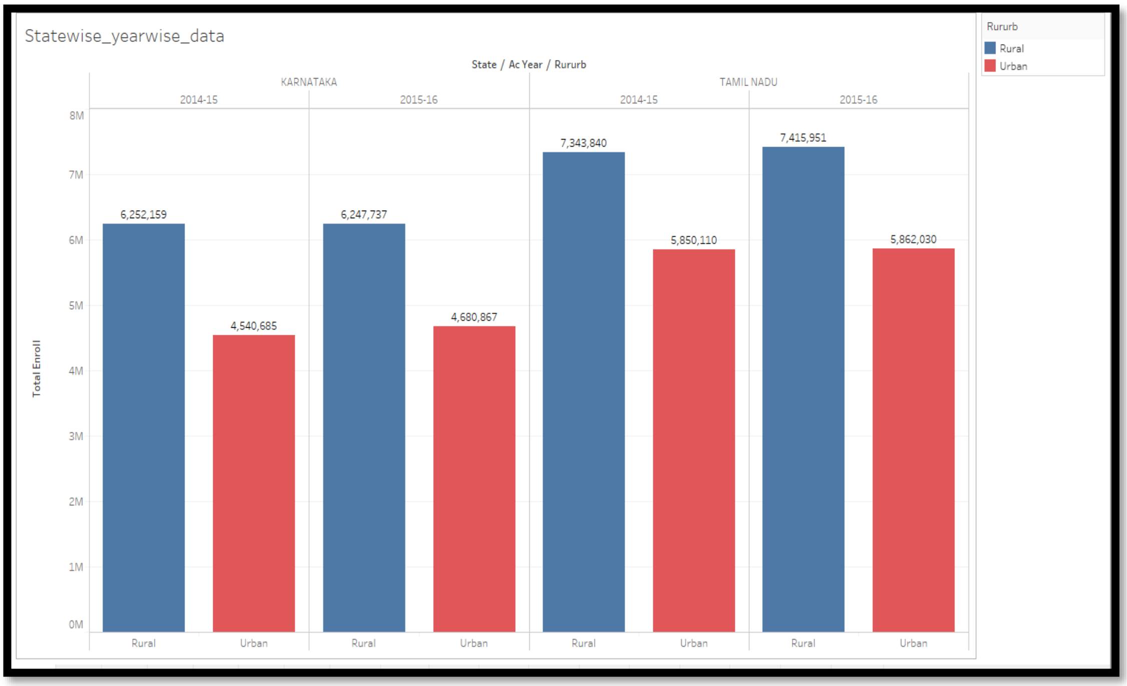
- For each of the variables, we have done boxplots and histograms for before and after outlier treatment to understand them better
- We have not done visualization for ESTDYEAR as it is the year of establishment and we will be converting this variable to categorical.
- We will be looking at the histograms to see the distribution of the variables along with the boxplots to see the outliers for each variable. Also, afterwards we will be replacing the outliers with the whisker values.

We are doing EDA so that we can have a deep insights about the dataset before we start the modelling or perform any kind of machine learning algorithms. Generally Eda is a compulsory step in the problem solving process. It helps to understands how variables behave. We can identify outliers which may be a major problem in model building.

We can perform EDA on the variables as follows:

1. **State-wise – Year wise – Total Enrollment:**

- Since we have 2 years of separate data, we can combine all the boys and girls enrolled as total enrollment and plot it state wise and year wise.
- We have taken total enrollment on y axis and Rural and Urban areas in the Primary X axis and in the secondary X axis we have taken state wise and year wise categories.

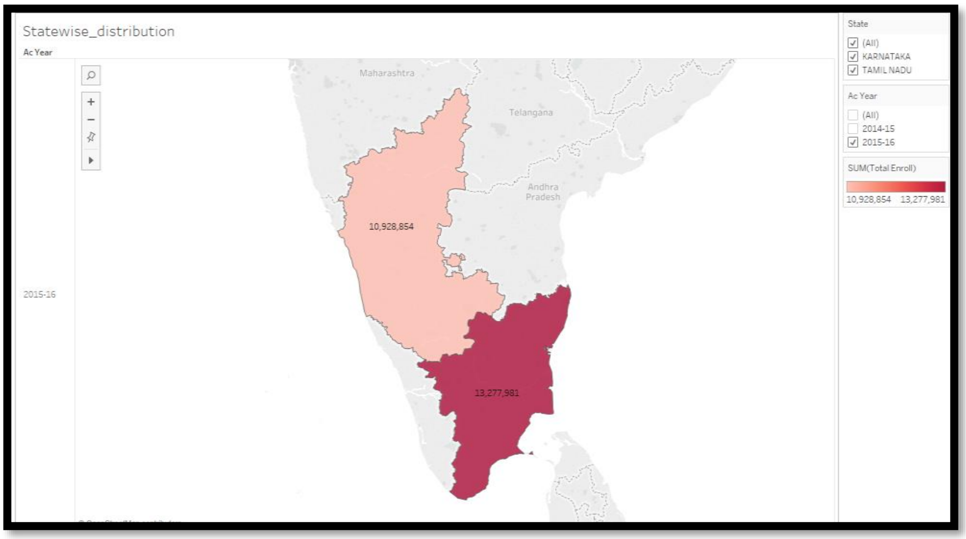


Insights from the above graph:

- When we see the state-wise enrolment, Tamil Nadu has highest enrolment of students for both the academic Years 2014-2015 and 2015-16 when compared to the Karnataka state.
- From the graph we can see that the rural region has high enrolment of students when compared to the urban region in both the states.

2. State wise Enrolment (Geographic View)

- It is easy for to understand any data if we have a location by using geographical plots. We can see how the distribution of data is in Karnataka and Tamil Nadu in our case.
- The dark red color represents the Tamil Nadu state.
- The light red color represents the Karnataka state.


State
☑ (All)
☑ KARNATAKA
☑ TAMIL NADU

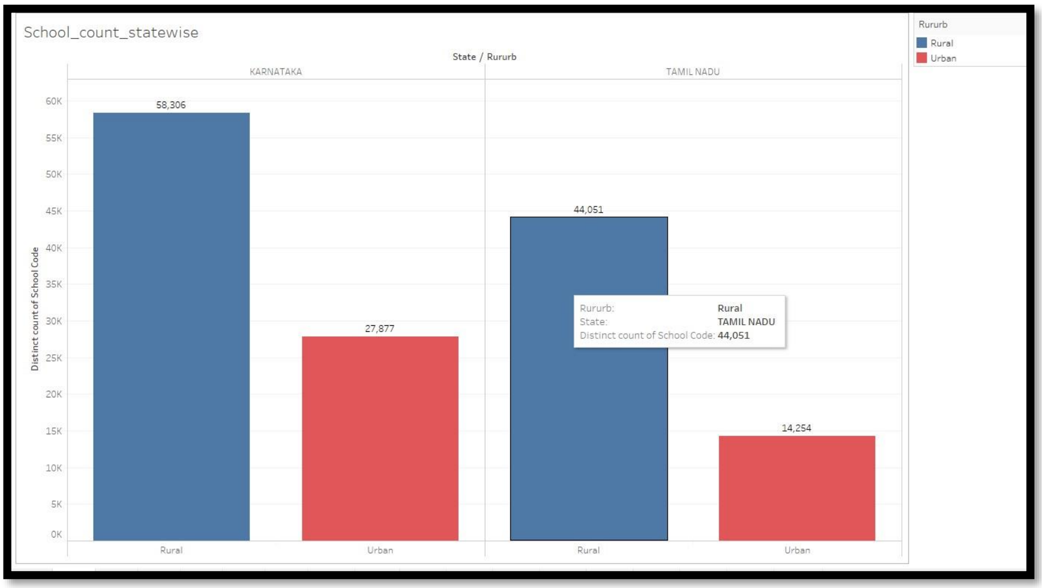
Ac Year
□ (All)
□ 2014-15
☑ 2015-16

Insights:

- Tamil Nadu has the highest number of enrollments with enrollment of 13,277,981 students.
- Karnataka has enrollments of 10,928,854 students.

3. Count of schools – State wise

- We have done distinct count on the school code, in order to get the school count in both the states.

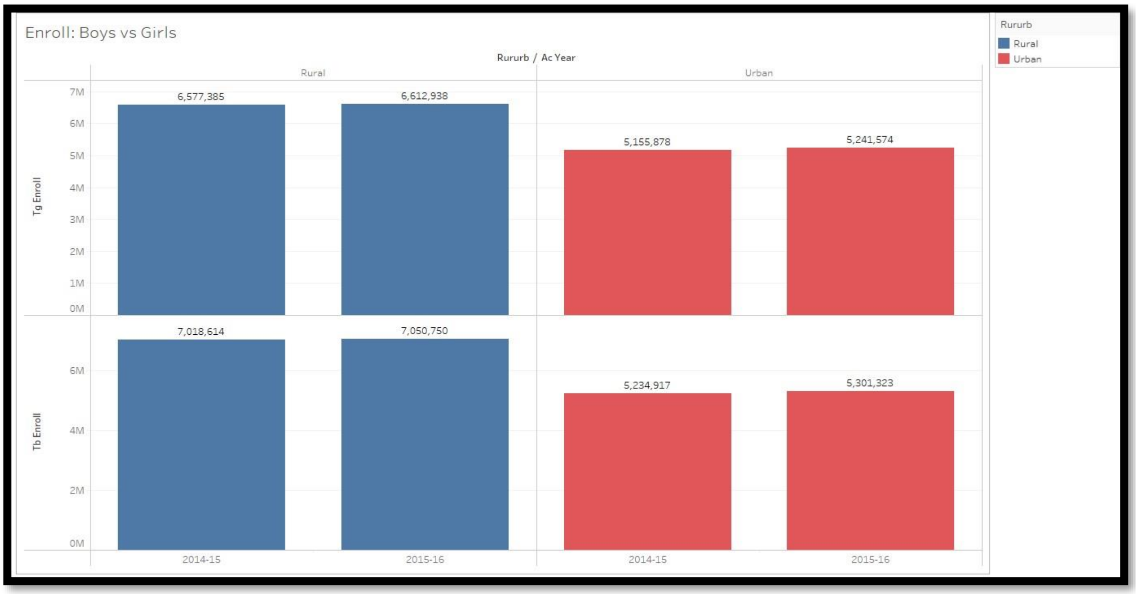


Insights:

- From the above plot we can see that Karnataka state has the highest number of schools when compared to the Tamil Nadu state. Karnataka has 86,183 number of schools and Tamil Nadu has 58,315 schools.
- Region wise, in both the states, there are more number of schools in the rural region than in the urban region.

4. Enrollment of Boys and Girls:

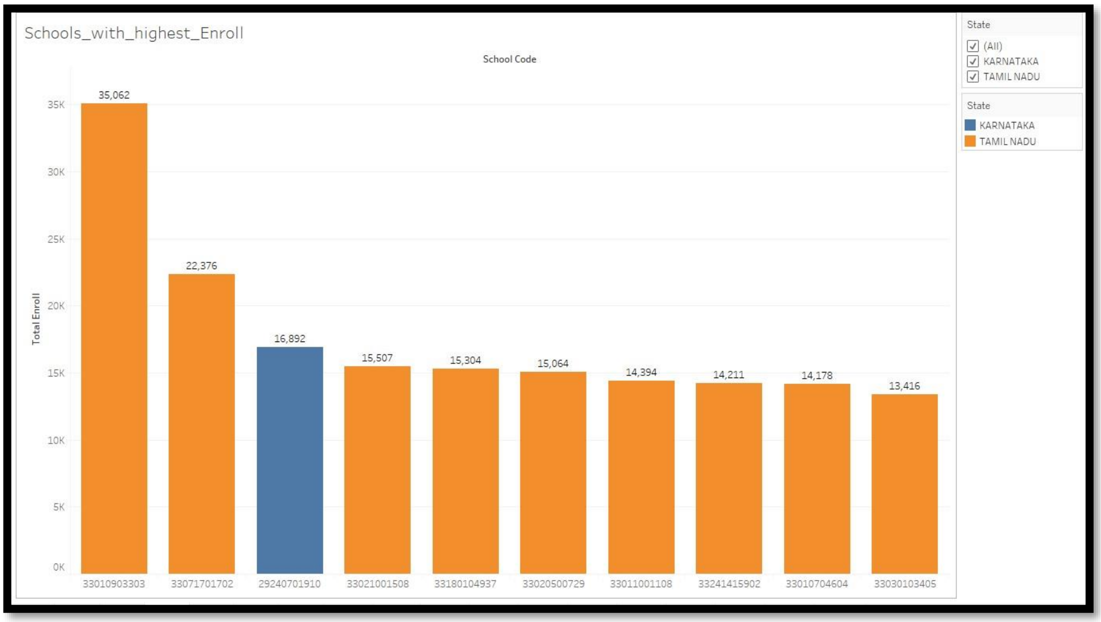
- This gives us the idea that how number of boys and girls are enrolled in the schools in both the states.
- We have considered both the academic years in the plot.



Insights:

- Overall more number of boys are enrolled when compared to the girls. We can also see that the enrollment of both boys and girls has increased from academic year 2014-15 to 2015-2016.
- Enrollment in Rural areas is high when compared to the urban regions.

5. **Schools with Highest Enrollment:**
   - This gives us the individual schools enrollment and which school has the highest enrollment.

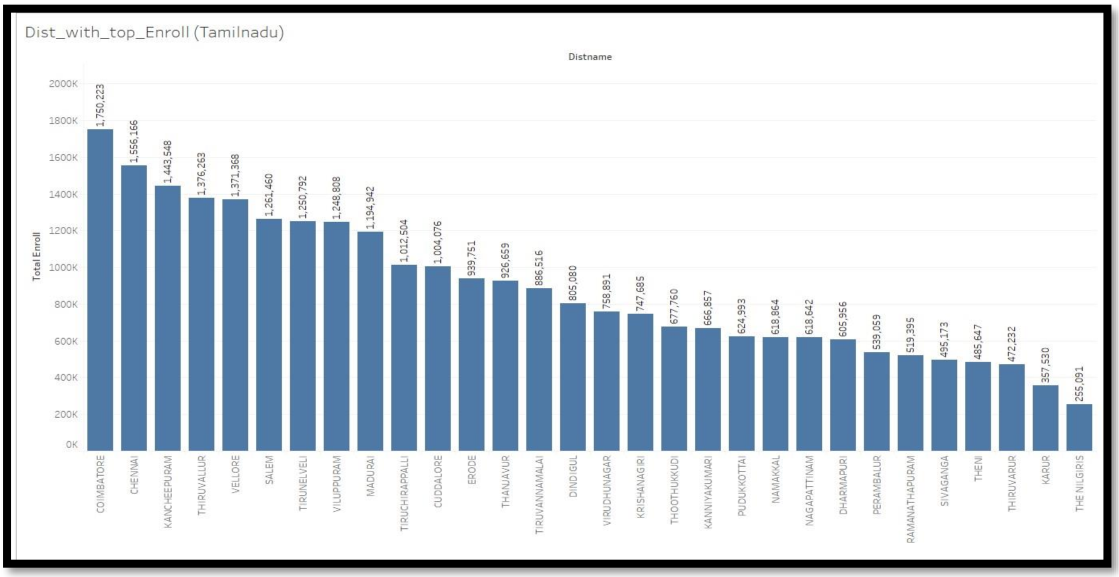


Insights:

- Over all the schools in Tamil Nadu have the high enrollment of students when compared to the Karnataka schools.
- In Tamil Nadu, the school with Id 33010903303 has the highest number of enrollment with over 35k students.
- In Karnataka, the school with Id 29240701910 has the highest number of enrollment with the enrollment of 16892 students.

6. District wise Enrollment for Tamil Nadu(Top)

- This plot gives us the idea that which district in Tamil Nadu has the highest enrollment number.

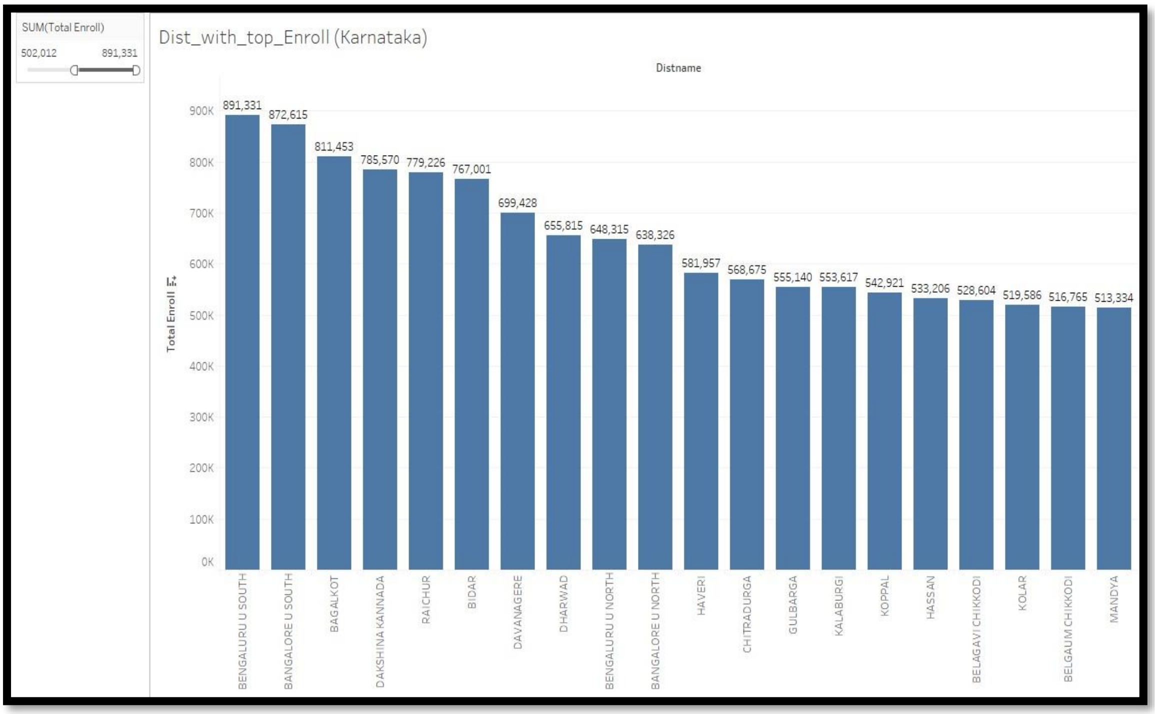


Insights:

- Coimbatore District has the highest enrollment of the students than any other districts. The enrollment in Coimbatore is over 1.7 million.
- Chennai is the second highest in enrollment of the students. The enrollment in Chennai is over 1.5 million.

7. District wise Enrollment for Karnataka (Top)

- This plot gives us the idea that which district in Karnataka has the highest enrollment number.

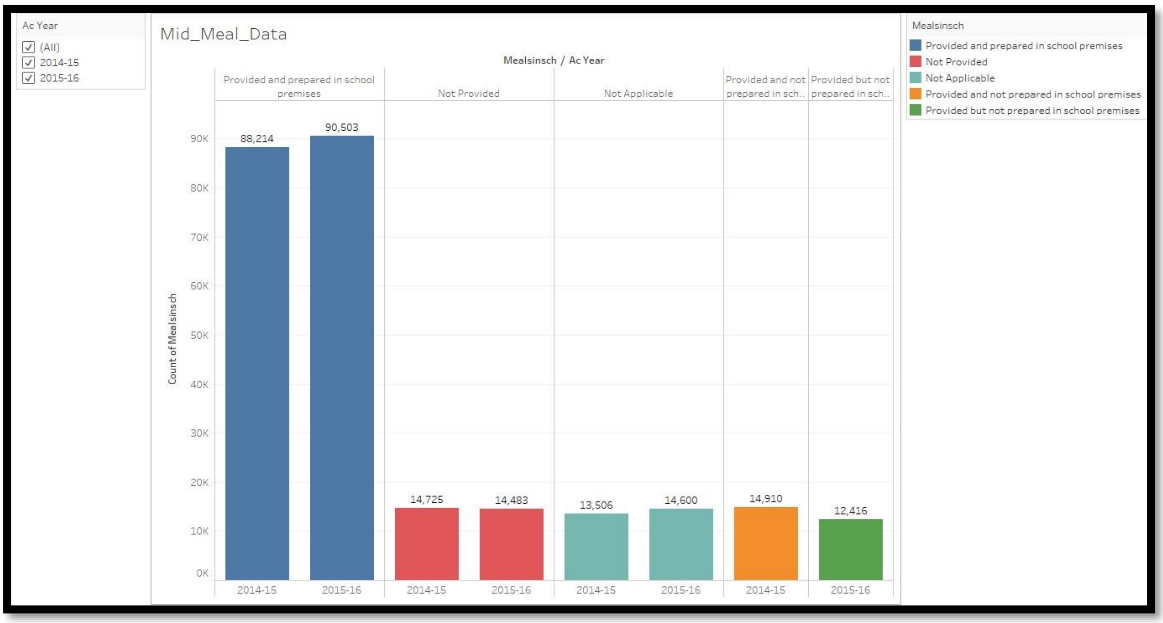


Insights:

- Bengaluru south has the highest number of enrollments in Karnataka than any other districts. The enrollment in Bengaluru South is approximately 90K students.
- Bagalkot has the second highest number of enrollments. The enrollment in Bagalkot is over 81K students.

8. Mid-Day Meals(Lunch in school)

- This plot shows which schools provide Lunch in the school premises**.**
- There are many schools which do not provide Lunch also, hence in order to see that we plot it.

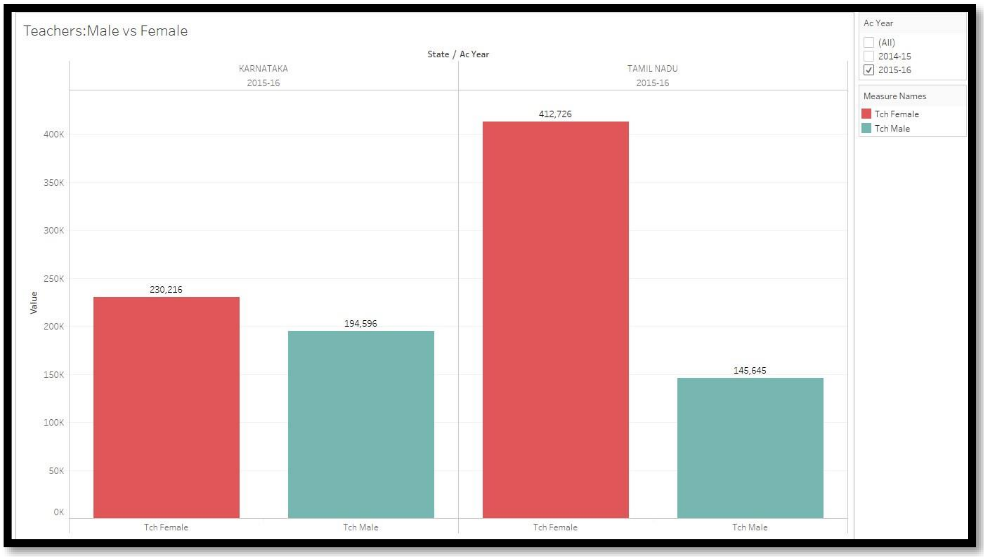


Insights:

- IN most of the schools, the lunch is provided and prepared in the school premises itself.
- Approximately 90k schools are provided with lunch in schools.
- We can see that the number of schools providing with lunch has increased from the year 2014-15 to 2015-16, whereas the schools which do not provide lunch has decreased from 14,725 to 14,483 when compared to the academic years.
- There are 14k schools where the lunch is schools in not applicable.
- There are around 15k schools where the lunch is provided, but it is not prepared in the schools. However, the count of the schools has been reduced to 12.5k from 2014-15 to 205-16.

9. Male Vs Female Teachers (State-wise)

- This plot is to how the Male and Female are involved in teaching in both the states.
- We can also the distribution of Male and Female teachers.

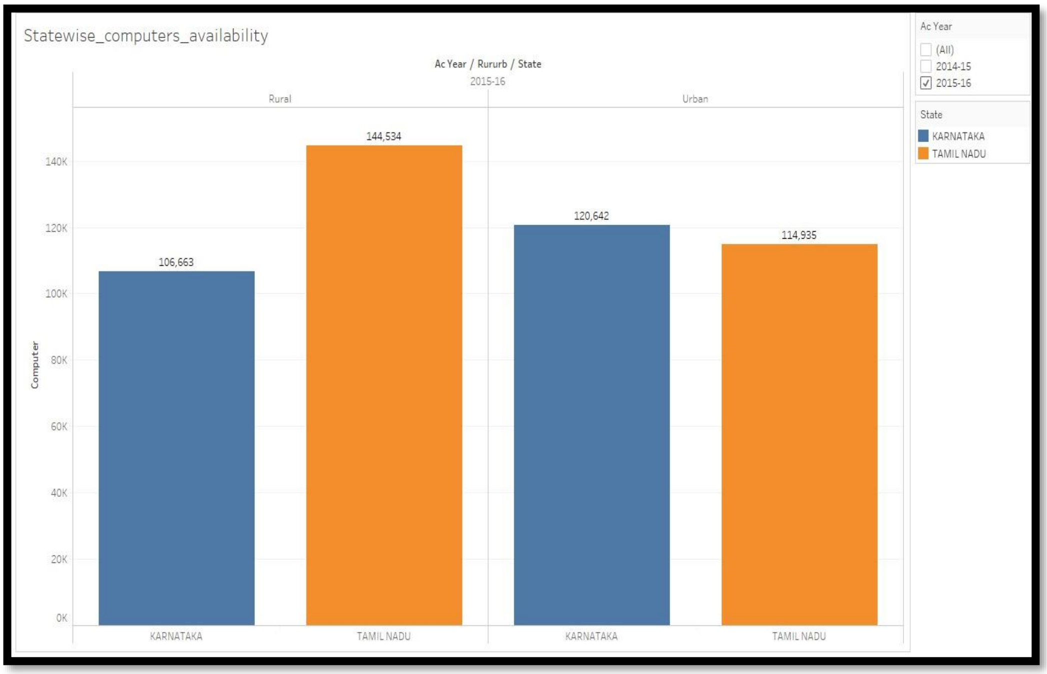


Insights:

- Over all Tamil Nadu has the highest number of teachers around 5.5Laksh.
- In Karnataka there are around 4.25 lakh teachers.
- In Tamil Nadu and in Karnataka there are more Female teachers than the Male teachers.

10. Computer availability (State wise)

- In this modern technology world, the necessity of the computer in the schools is mandatory.
- This plot shows how the schools in both the states have provided its students with the facility of computer.

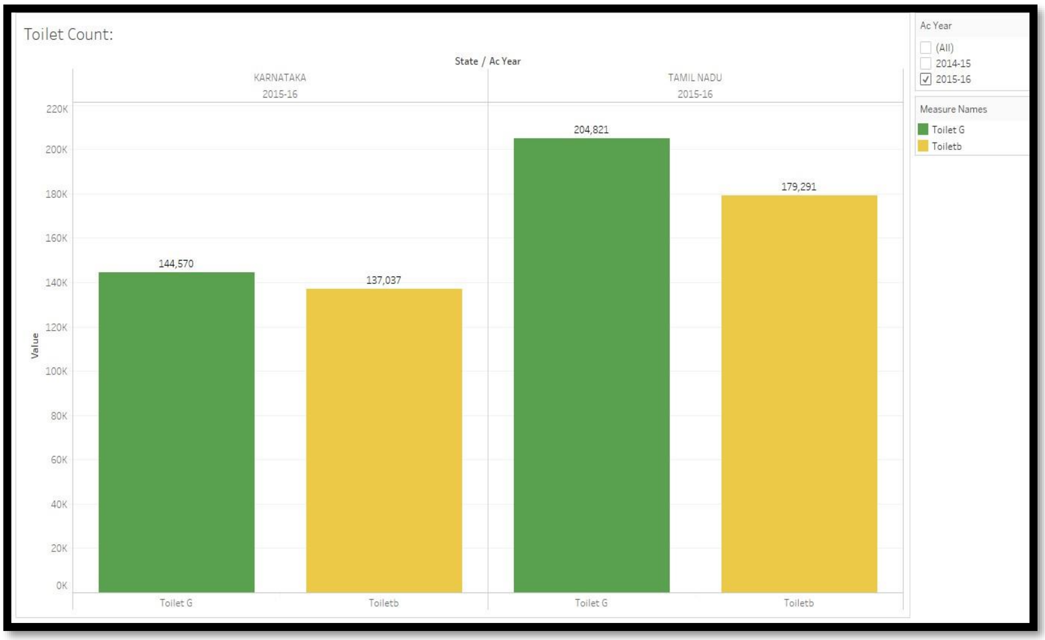


Insights:

- Tamil Nadu has the highest number of computers. There are over 2.5 Lakh computers.
- In Karnataka there are over 2.2 Lakh computers.
- Even though there are more schools in Karnataka, most of the schools do not have the facility of the computers when compared to the schools in Tamil Nadu.

11. Toilet Count in Schools (State wise)

- This gives the idea of total number of toilets in both Karnataka and Tamil Nadu.

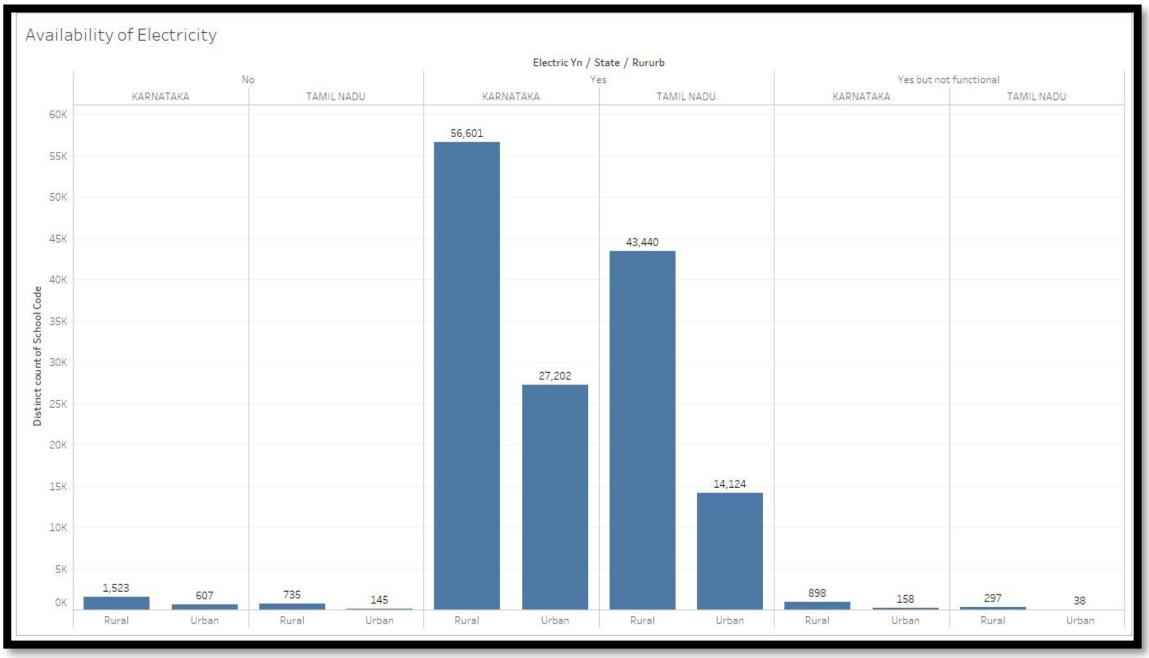


Insights:

- Tamil Nadu has the highest number of toilets in the schools. There are about 3.84 Lakh toilets in schools.
- Karnataka has over 2.81 Lakh toilets in schools.
- Overall there are more Girl toilets than the boy toilets.

12. Availability of Electricity in schools (State-wise)

- This plot is to show the availability of the electricity in schools in both Karnataka and Tamil Nadu.

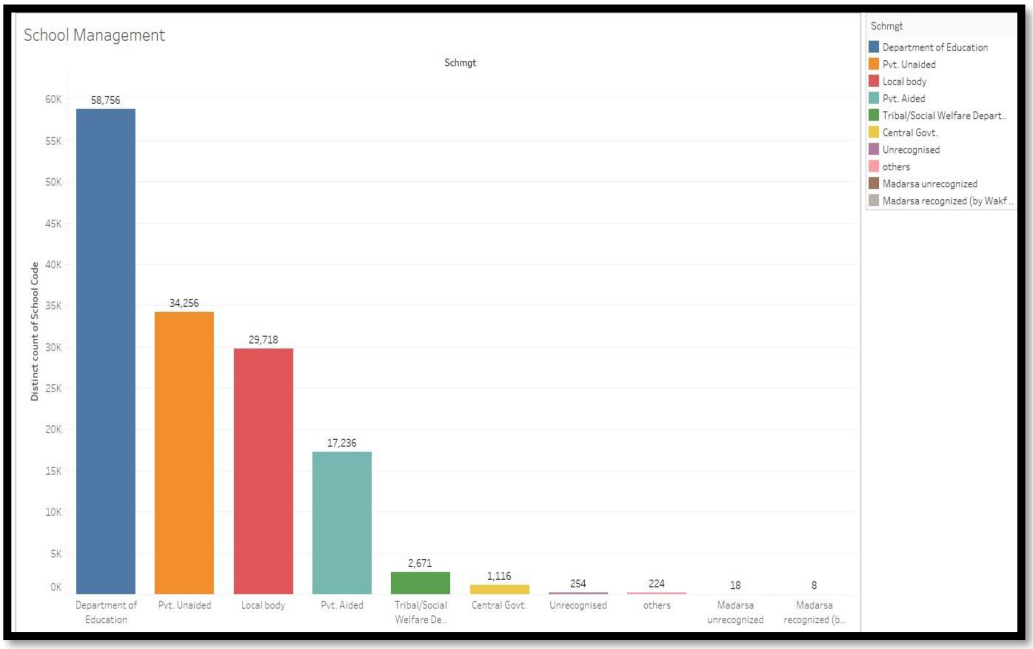


Insights:

- Karnataka has the highest availability of electricity to the schools when compared to Tamil Nadu.
- In Karnataka rural areas has more number of schools with no electricity when compared with the urban areas.
- In Tamil Nadu has less number of schools where "there is electricity but not functional" when compared to Karnataka.

13. School Management:

- This plot is to show how the schools are managed in both the states.
- There are various managements like Department of Education, Private Unaided, local body (Local state board), private aided, central government, unrecognized boards.

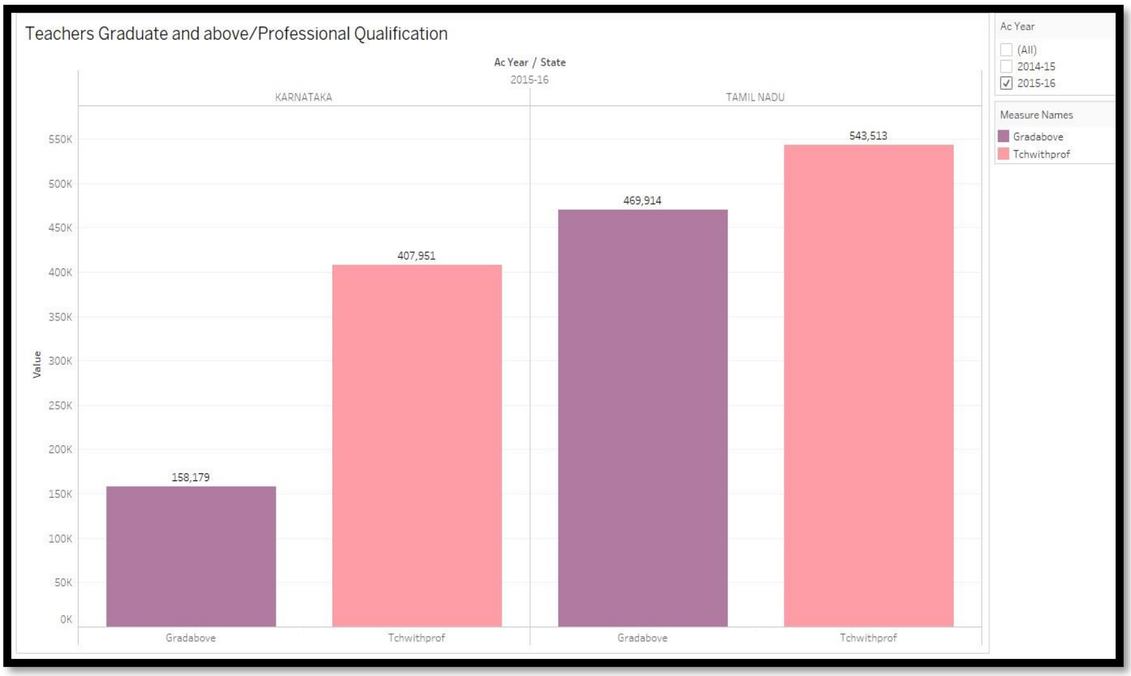


Insights:

- There are more number of schools where the schools are State board (Department of education).
- In second there private and unaided schools.
- There are very less schools which follow the Central Government Board.

14. Professional Qualification of Teachers:

- This is to show how the teacher are qualified in both Karnataka and Tamil Nadu.
- There are two categories: Graduate teachers and teacher with professional degree.

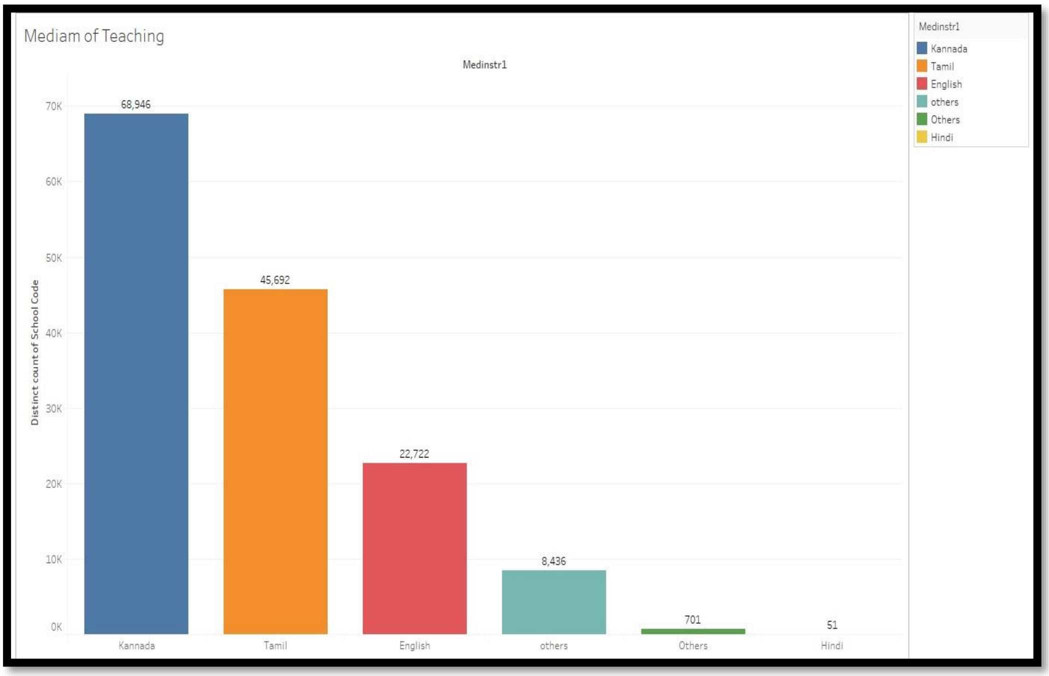


Insights:

- Tamil Nadu has a higher number of teachers as compared to Karnataka.
- There are more number of teachers who are professionally qualified to become teachers in both Tamil Nadu and Karnataka. There are less number of teacher who are graduate and above.
- So, there are teachers who have had previous experience in teaching and are teachers now. However, if you want to compare the quality of the teachers, there are much less number of teachers who have even graduated themselves. This is especially true in Karnataka.

15. Medium of Instruction:

- This is to show how the schools in Karnataka and Tamil Nadu are using different medium of teaching.
- The different categories are Kannada, Tamil, Hindi, English and other languages.

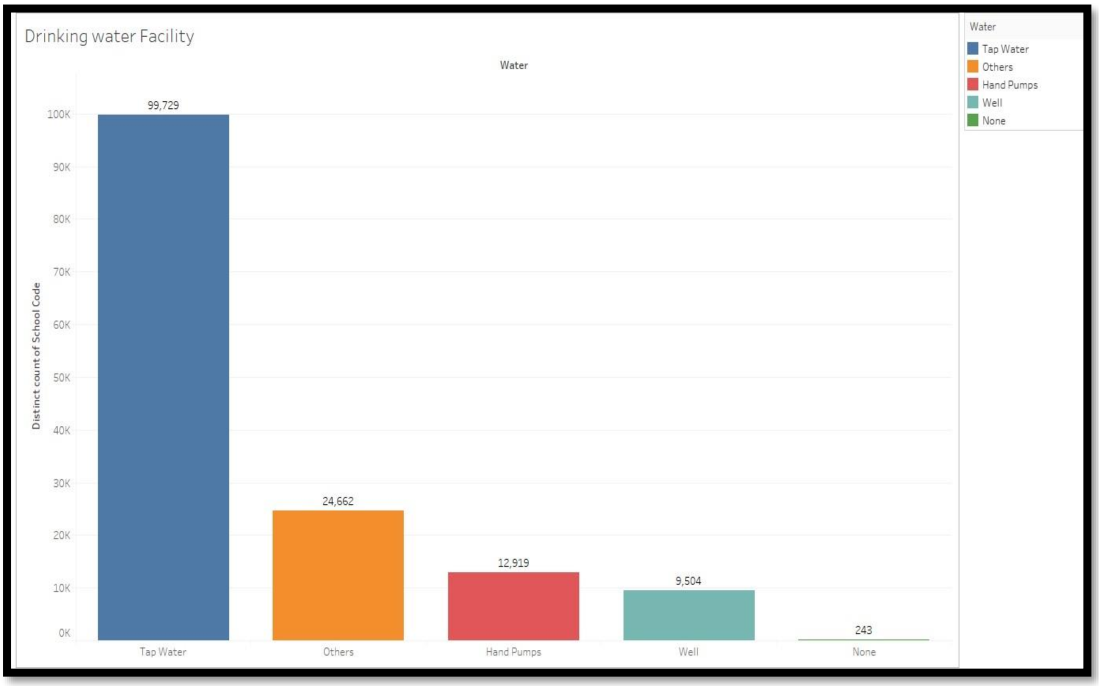


Insights:

- On comparing both Karnataka and Tamil Nadu we can see that the schools with "Kannada" as the Medium of Teaching are highest with a number around 70k.
- Second comes the "Tamil" as the medium of instruction.
- Tough English is the universal language, it comes in third in Karnataka and Tamil Nadu.

16. Drinking Water Facility:

- This is to show how schools provide the drinking water in both Karnataka and Tamil Nadu.

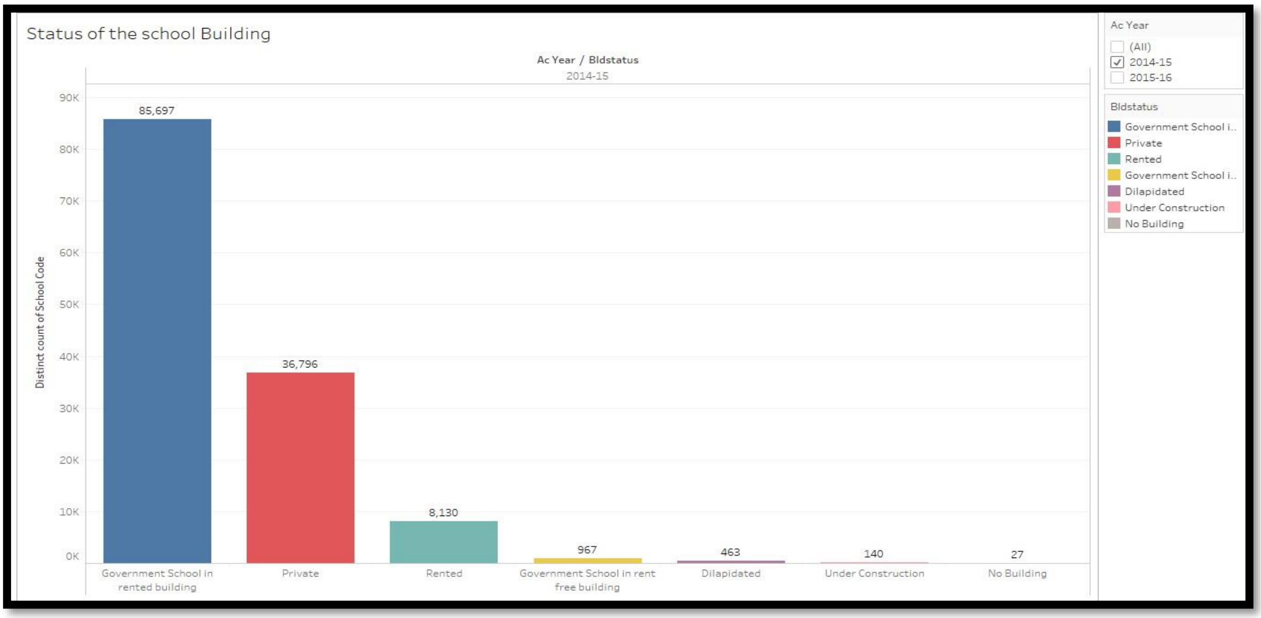


Insights:

- In most of the schools, both in Karnataka and in Tamil Nadu, the source of drinking water is "Tap water".
- There are other resources of drinking water also. But there are still schools where the drinking water is collected from Hand Pumps and wells.

17. School Building Status:

- This plot shows how the schools building are owned in the both the states.
- The various categories are: Government school, private, rented, under construction, no building.

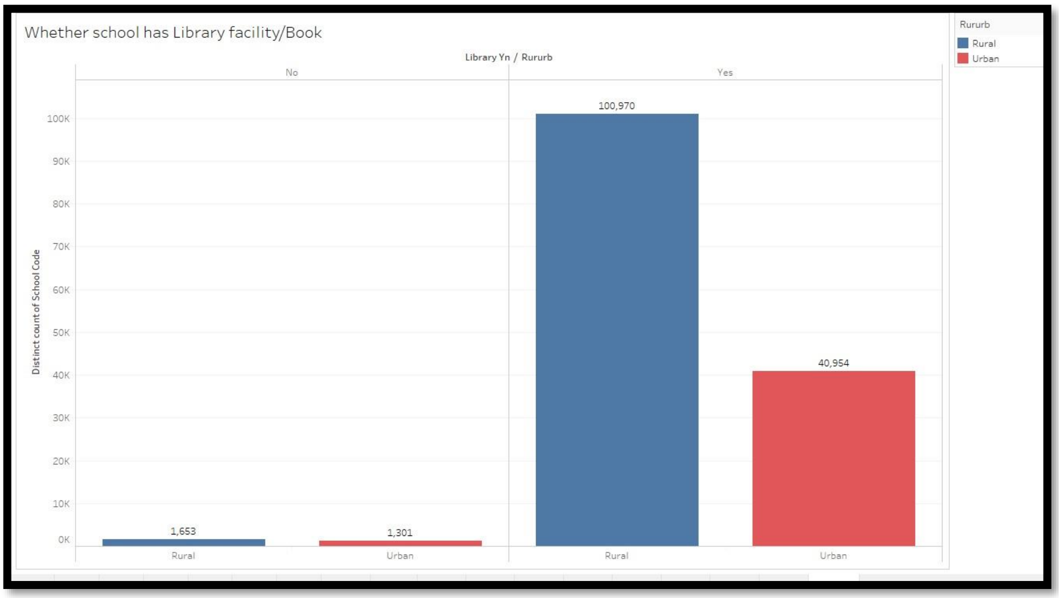


Insights:

- Most of the schools in Karnataka and Tamil Nadu are Government schools which are in rented buildings.
- In second number comes the private schools. There are schools in Rented buildings also.
- There are schools in Rent free buildings and under construction also. In remote areas there are no buildings for schools.

18. Availability of Library/Book in schools:

- This plot is to show how many schools have provided the facility of the library in the schools in Karnataka and Tamil Nadu.

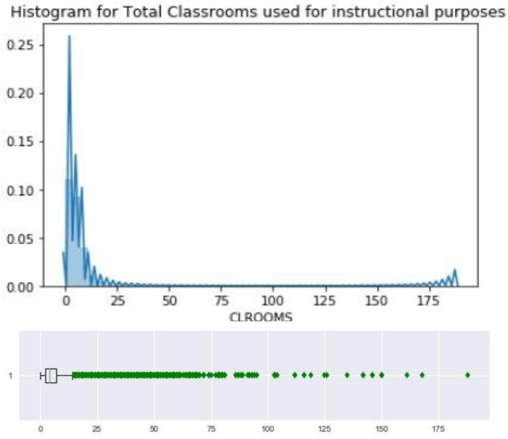


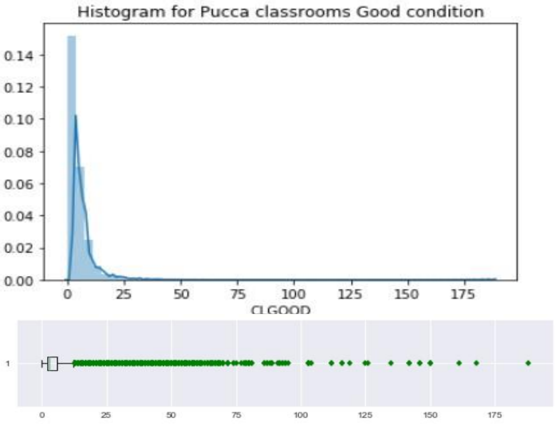
Insights:

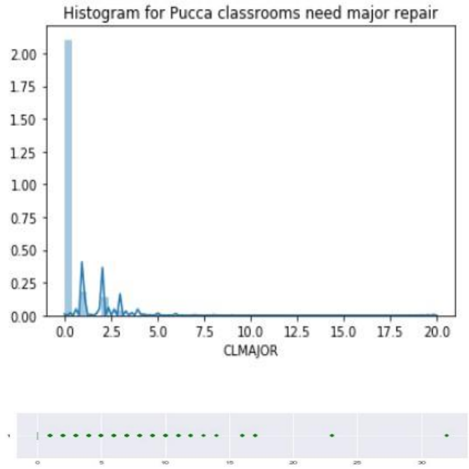
- Most of the schools have the facility of library and the books are available in the library.
- Most of the schools which are located in the rural areas have library, where as in urban areas the number of libraries are less.

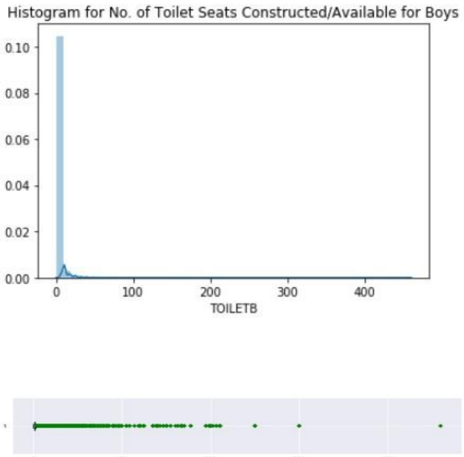
3.2: Boxplots and Histograms:

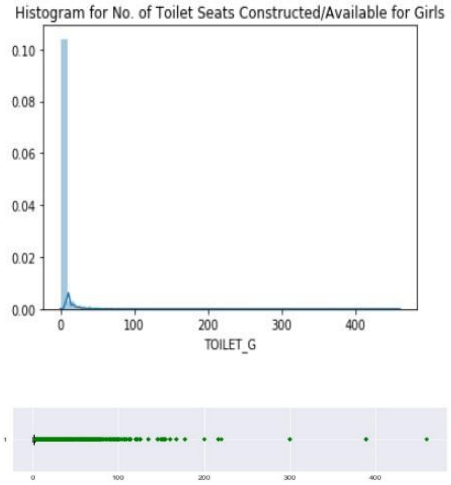
Before Outlier Treatment:

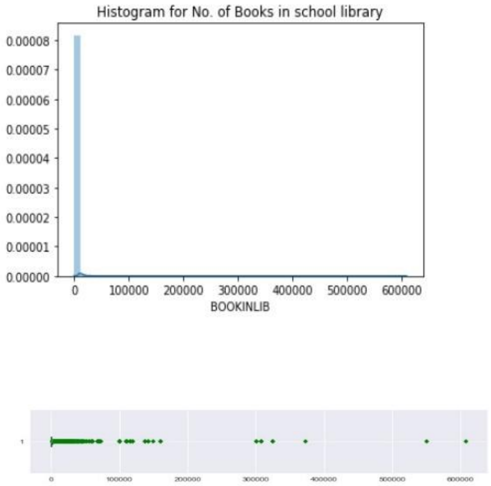


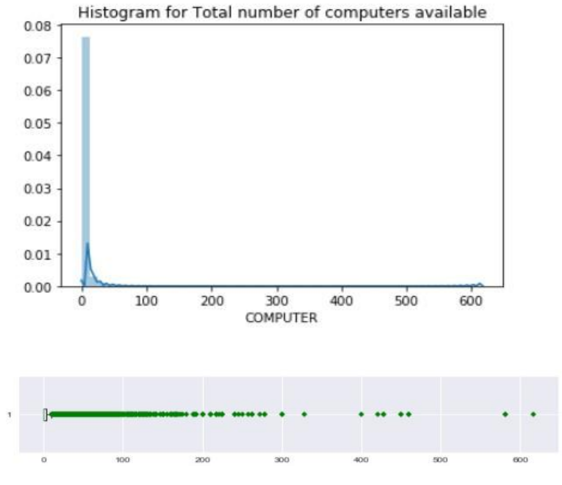


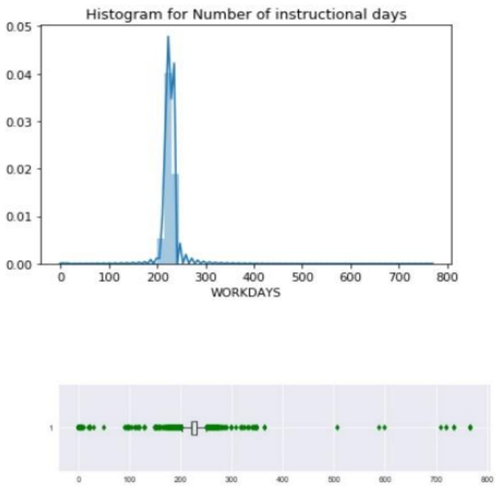

Histogram for Pucca classrooms need major repair
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CLMAJOR
1
0
5
10
15
20
25
30

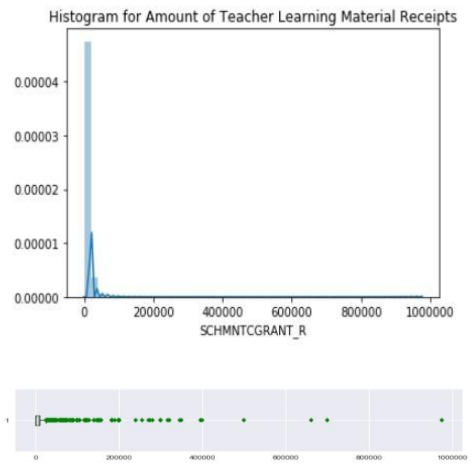

Histogram for No. of Toilet Seats Constructed/Available for Boys
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
TOILETB
1
0
100
200
300
400

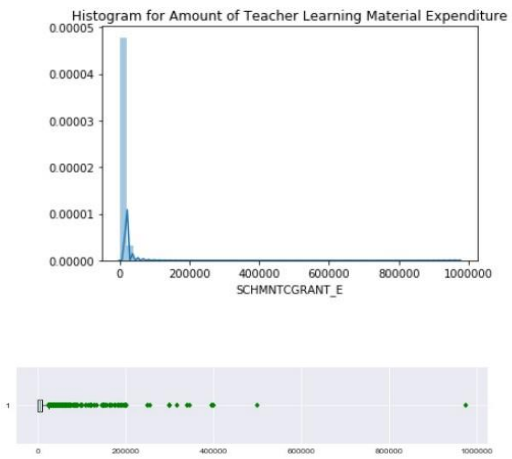

Histogram for No. of Toilet Seats Constructed/Available for Girls
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
TOILET_G
1
0
100
200
300
400

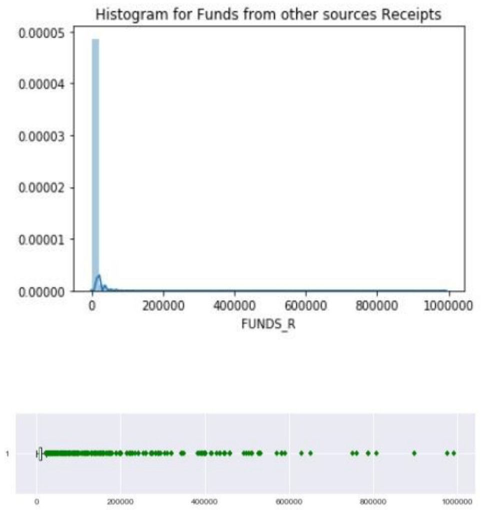

Histogram for No. of Books in school library
0.00008
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
100000
200000
300000
400000
500000
600000
BOOKINLIB
1
0
100000
200000
300000
400000
500000
600000

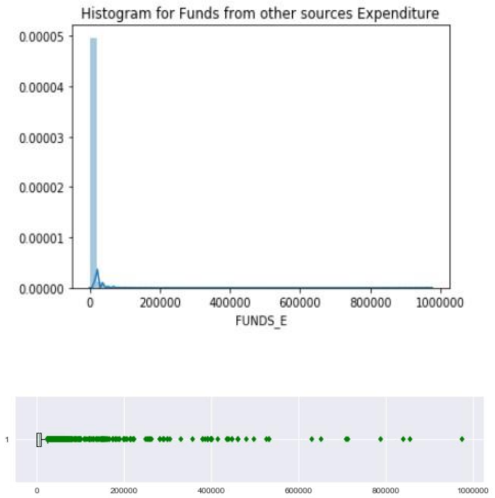

Histogram for Total number of computers available
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
600
COMPUTER
1
0
100
200
300
400
500
600

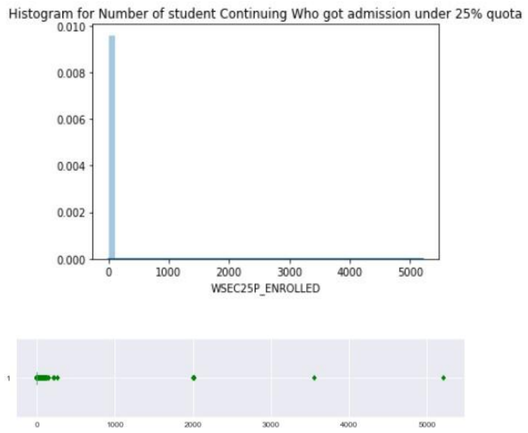

Histogram for Number of instructional days
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
600
700
800
WORKDAYS
1
0
100
200
300
400
500
600
700
800

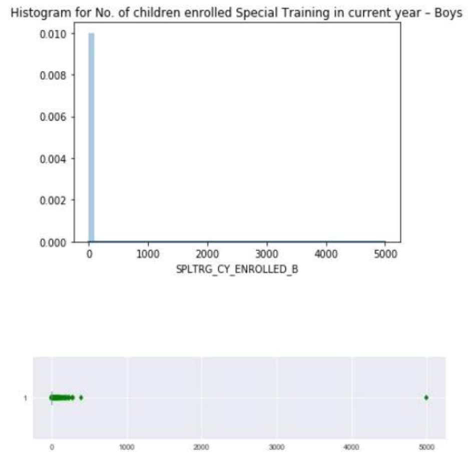

Histogram for Amount of Teacher Learning Material Receipts
0.00004
0.00003
0.00002
0.00001
0.00000
0
200000
400000
600000
800000
1000000
SCHMNTCGRANT_R
1
0
200000
400000
600000
800000
1000000

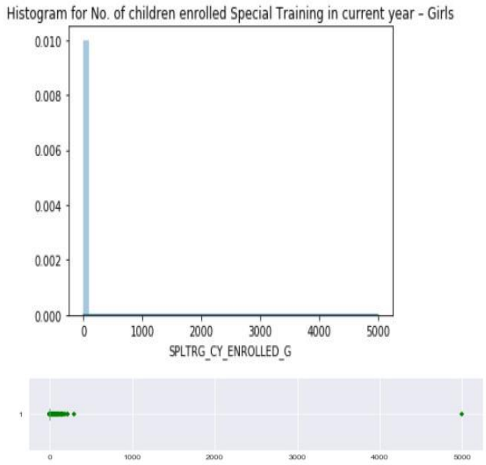

Histogram for Amount of Teacher Learning Material Expenditure
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
200000
400000
600000
800000
1000000
SCHMNTCGRANT_E
1
0
200000
400000
600000
800000
1000000

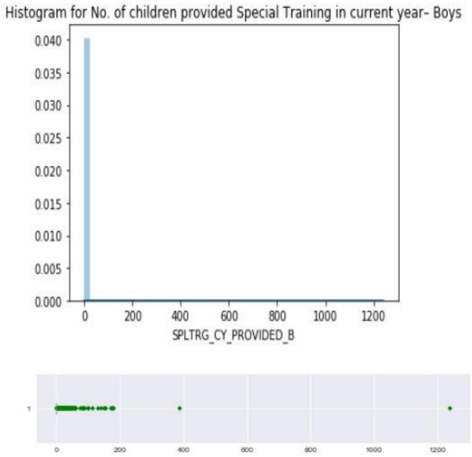

Histogram for Funds from other sources Receipts
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
200000
400000
600000
800000
1000000
FUNDS_R
1
0
200000
400000
600000
800000
1000000

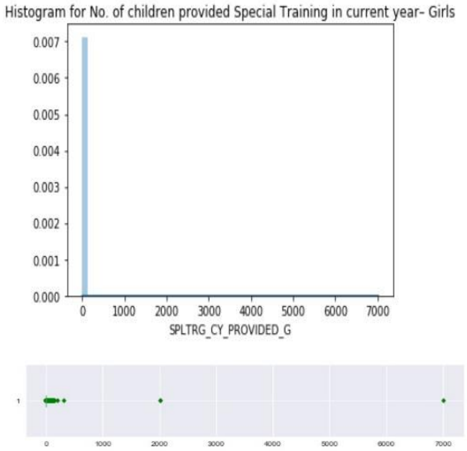

Histogram for Funds from other sources Expenditure
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
200000
400000
600000
800000
1000000
FUNDS_E
1
0
200000
400000
600000
800000
1000000

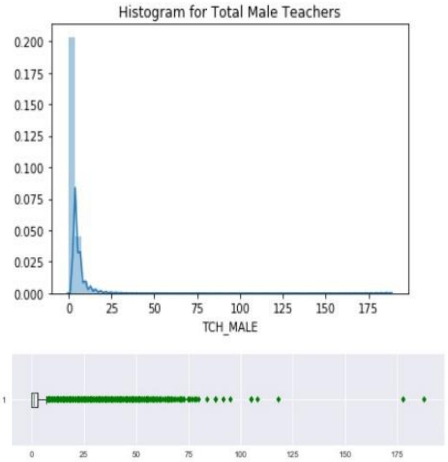

Histogram for Number of student Continuing Who got admission under 25% quota
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
WSEC25P_ENROLLED
1
0
1000
2000
3000
4000
5000

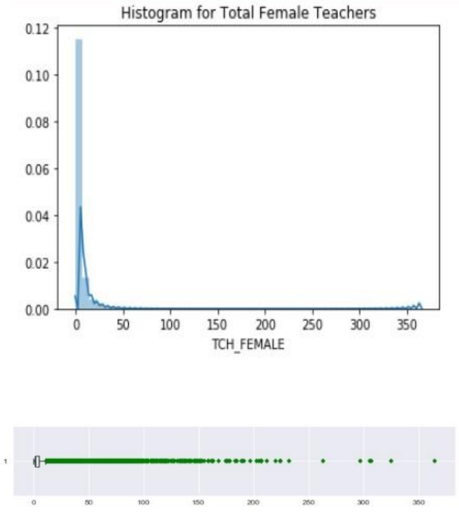

Histogram for No. of children enrolled Special Training in current year - Boys
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
SPLTRG_CY_ENROLLED_B
1
0
1000
2000
3000
4000
5000

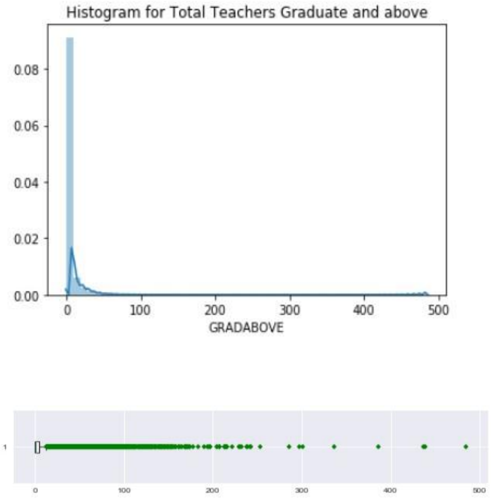

Histogram for No. of children enrolled Special Training in current year - Girls
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
SPLTRG_CY_ENROLLED_G
1

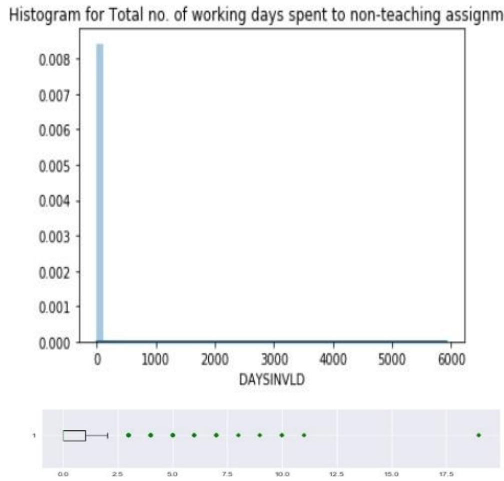

Histogram for No. of children provided Special Training in current year- Boys
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
200
400
600
800
1000
1200
SPLTRG_CY_PROVIDED_B
1

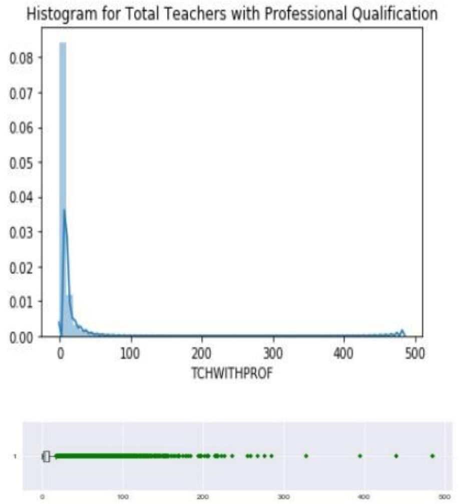

Histogram for No. of children provided Special Training in current year- Girls
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
1000
2000
3000
4000
5000
6000
7000
SPLTRG_CY_PROVIDED_G
1

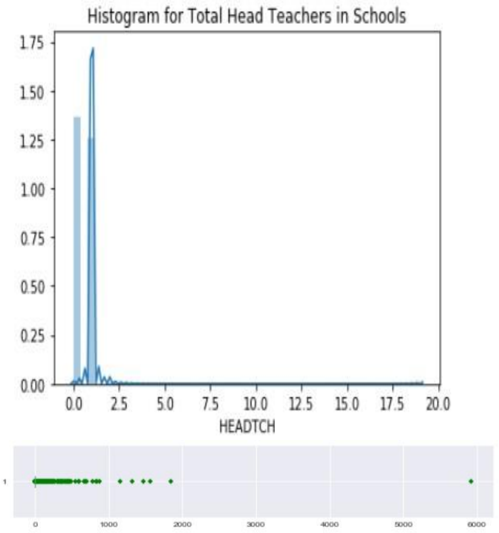

Histogram for Total Male Teachers
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
25
50
75
100
125
150
175
TCH_MALE
1

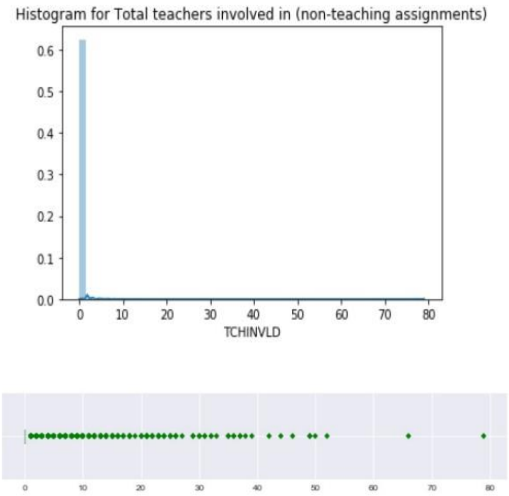

Histogram for Total Female Teachers
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
300
350
TCH_FEMALE
1

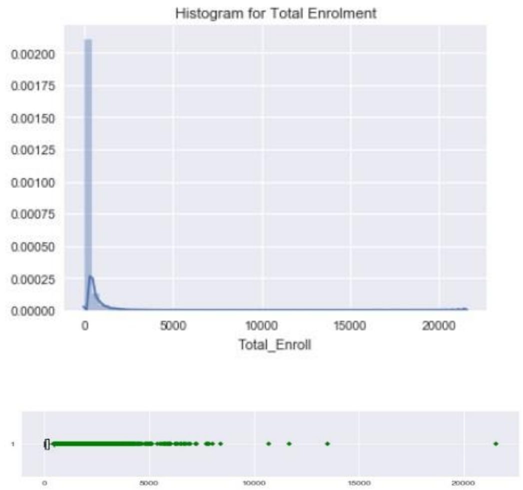

Histogram for Total Teachers Graduate and above
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
500
GRADABOVE
1


Histogram for Total no. of working days spent to non-teaching assignments
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
1000
2000
3000
4000
5000
6000
DAYSINVLD
1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

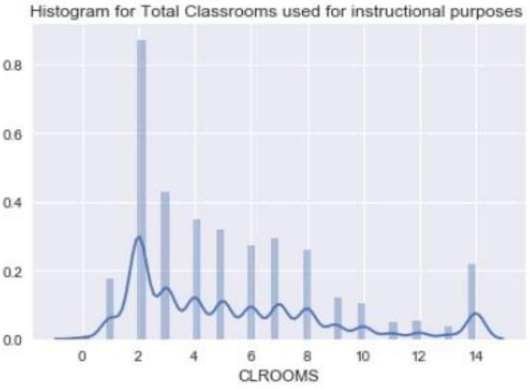

Histogram for Total Teachers with Professional Qualification
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
100
200
300
400
500
TCHWITHPROF
1
0
100
200
300
400
500

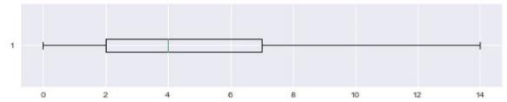

Histogram for Total Head Teachers in Schools
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
HEADTCH
1
0
1000
2000
3000
4000
5000
6000

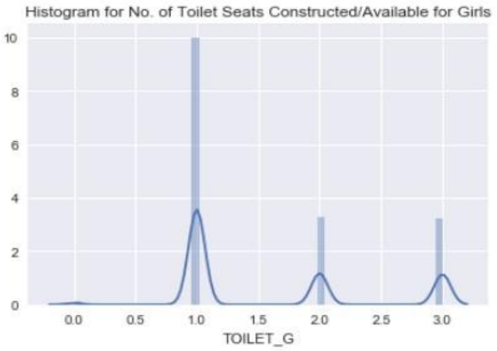

Histogram for Total teachers involved in (non-teaching assignments)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
TCHINVLD
1
0
10
20
30
40
50
60
70
80

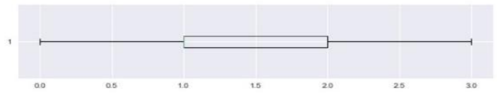

Histogram for Total Enrolment
0.00200
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
5000
10000
15000
20000
Total_Enroll
1
0
5000
10000
15000
20000

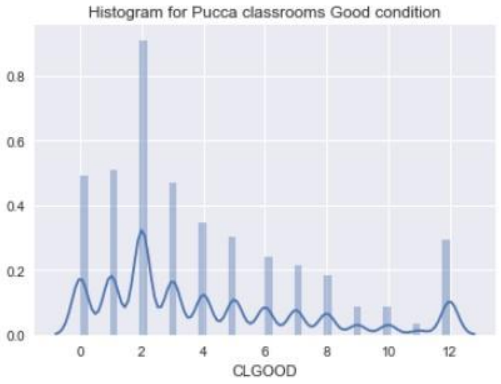
Histograms and Boxplots after Outlier Treatment:

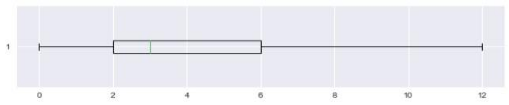


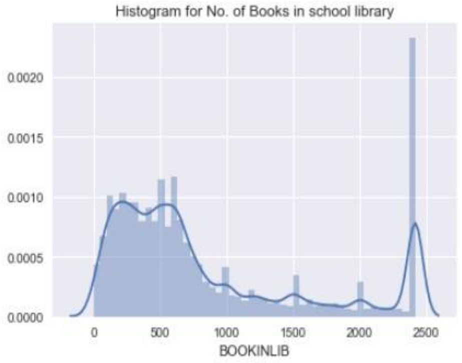


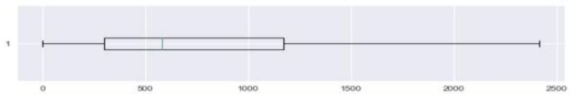


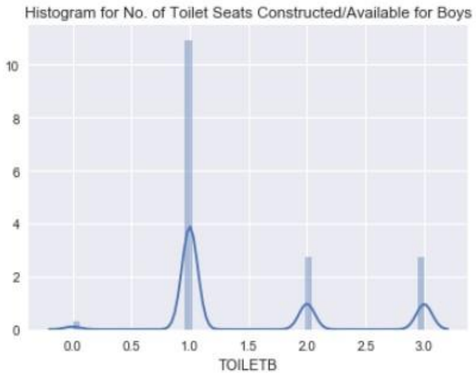


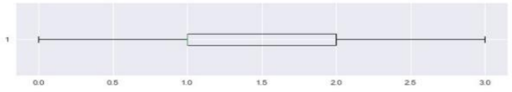


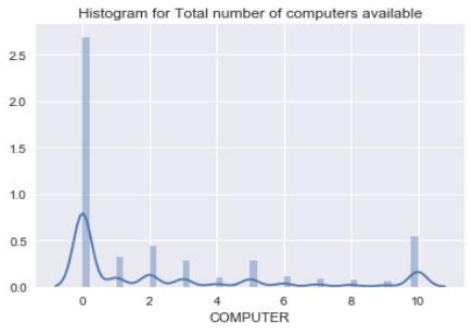


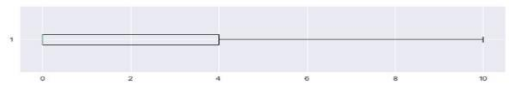


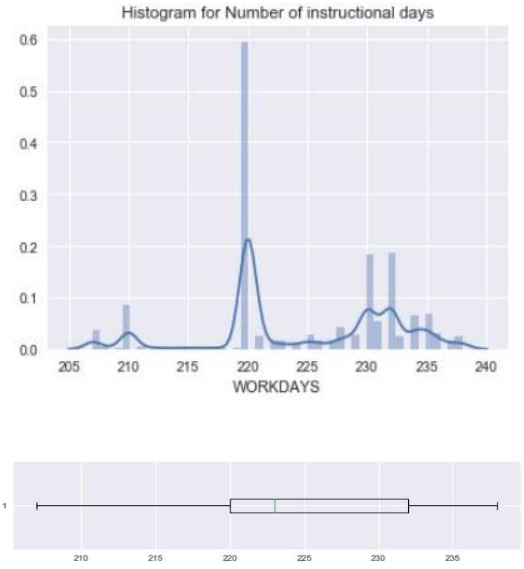


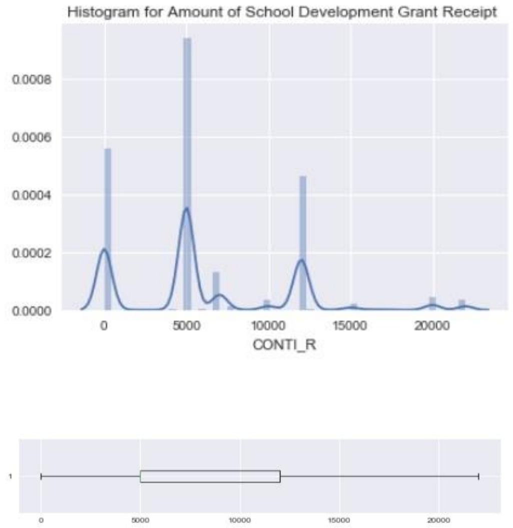


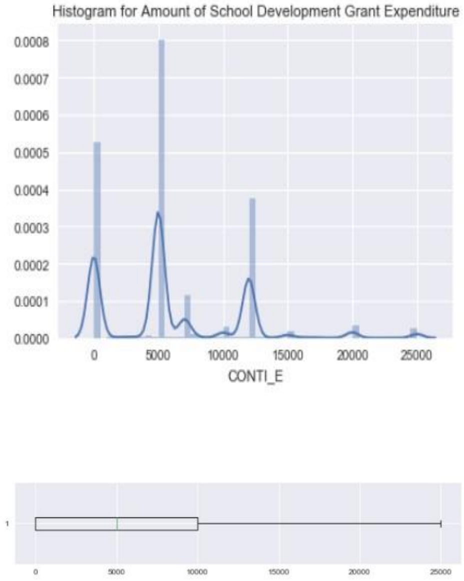


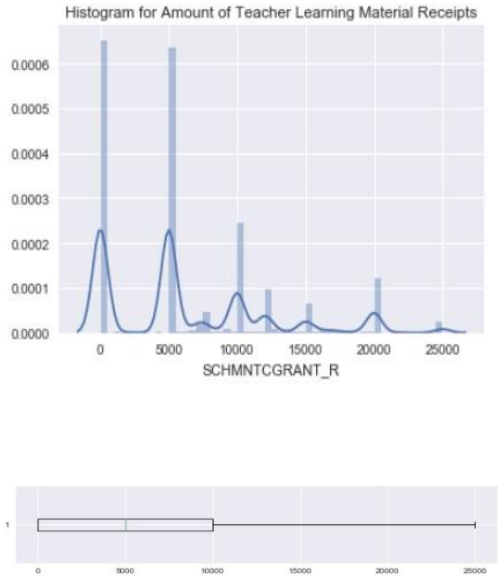


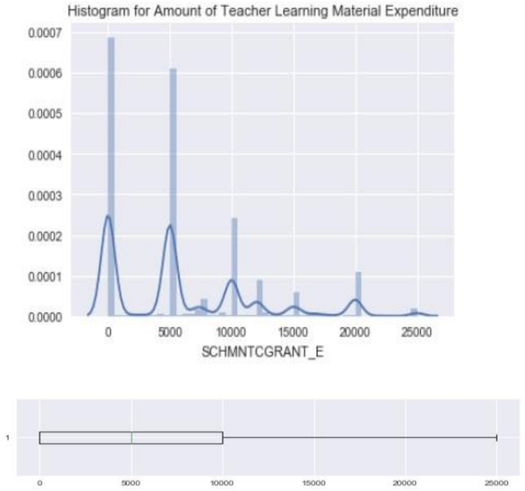


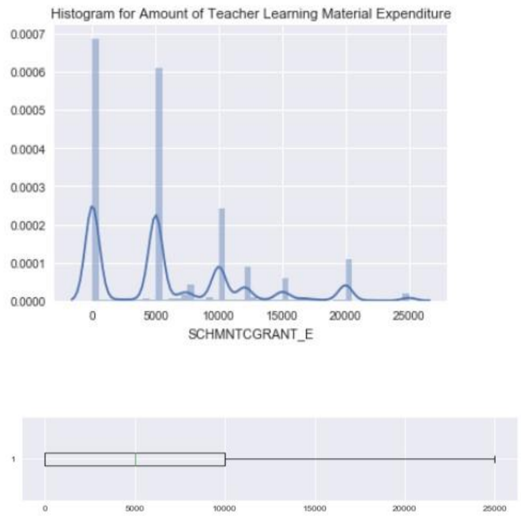

Histogram for Number of instructional days
0.6
0.5
0.4
0.3
0.2
0.1
0.0
205 210 215 220 225 230 235 240
WORKDAYS
1
210 215 220 225 230 235

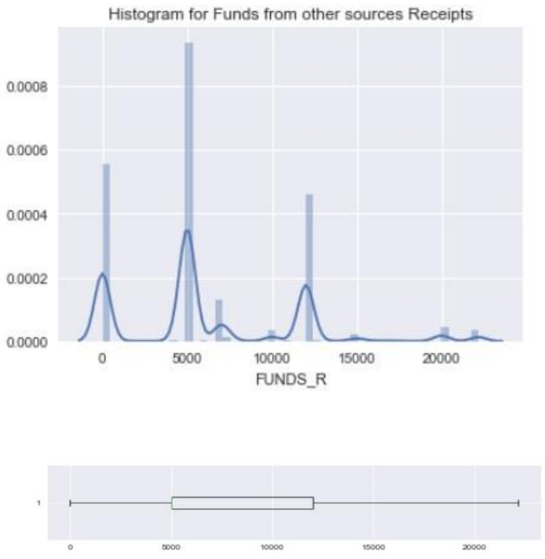

Histogram for Amount of School Development Grant Receipt
0.0008
0.0006
0.0004
0.0002
0.0000
0 5000 10000 15000 20000
CONTI_R
1
0 5000 10000 15000 20000

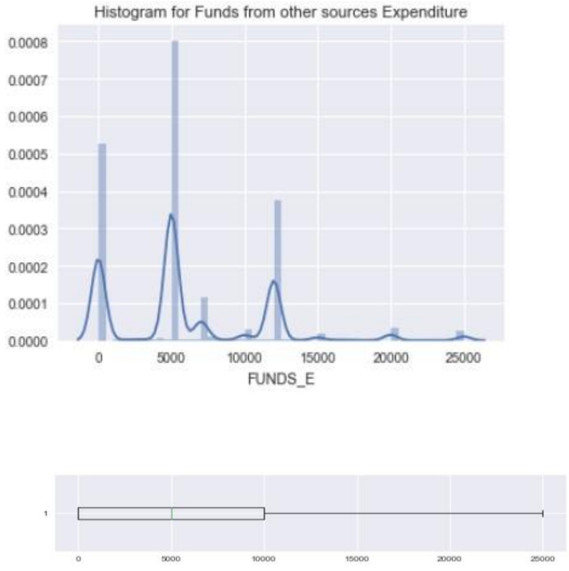

Histogram for Amount of School Development Grant Expenditure
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0 5000 10000 15000 20000 25000
CONTI_E
1
0 5000 10000 15000 20000 25000

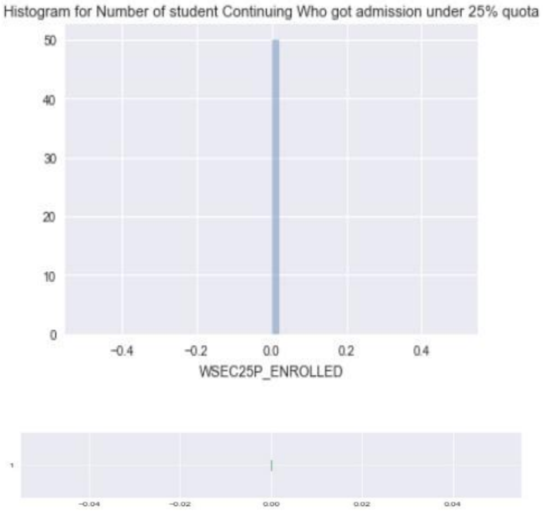

Histogram for Amount of Teacher Learning Material Receipts
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0 5000 10000 15000 20000 25000
SCHMNTCGRANT_R
1
0 5000 10000 15000 20000 25000

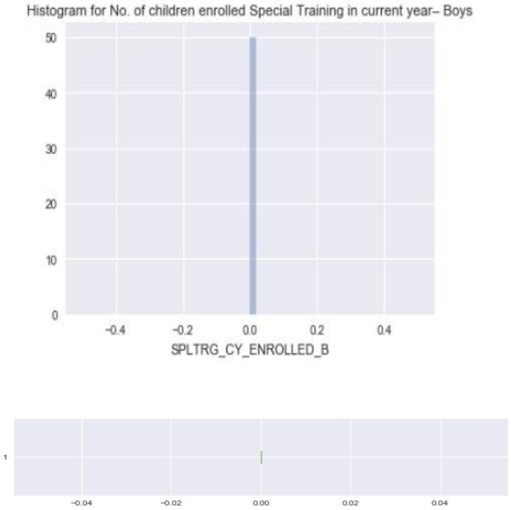

Histogram for Amount of Teacher Learning Material Expenditure
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0 5000 10000 15000 20000 25000
SCHMNTCGRANT_E
1
0 5000 10000 15000 20000 25000

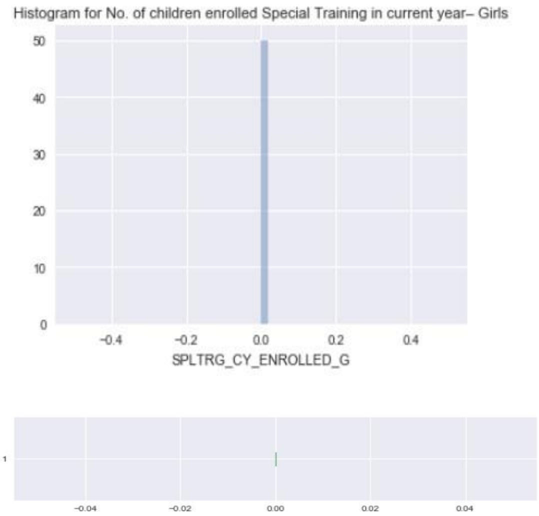

Histogram for Amount of Teacher Learning Material Expenditure
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0 5000 10000 15000 20000 25000
SCHMNTCGRANT_E
1
0 5000 10000 15000 20000 25000

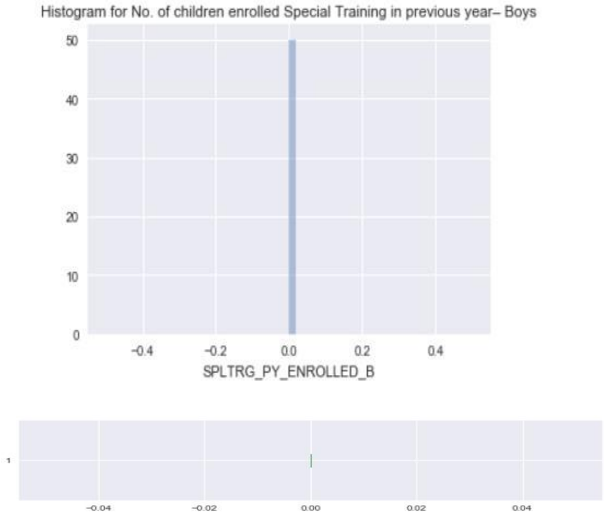

Histogram for Funds from other sources Receipts
0.0000
0.0002
0.0004
0.0006
0.0008
0
5000
10000
15000
20000
FUNDS_R
1
0
5000
10000
15000
20000

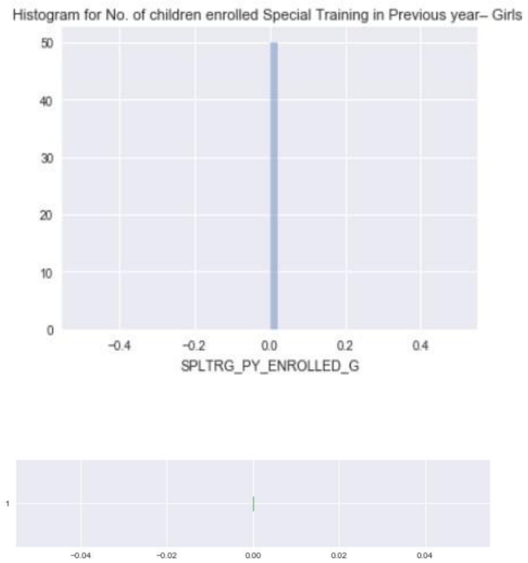

Histogram for Funds from other sources Expenditure
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
0.0008
0
5000
10000
15000
20000
25000
FUNDS_E
1
0
5000
10000
15000
20000
25000

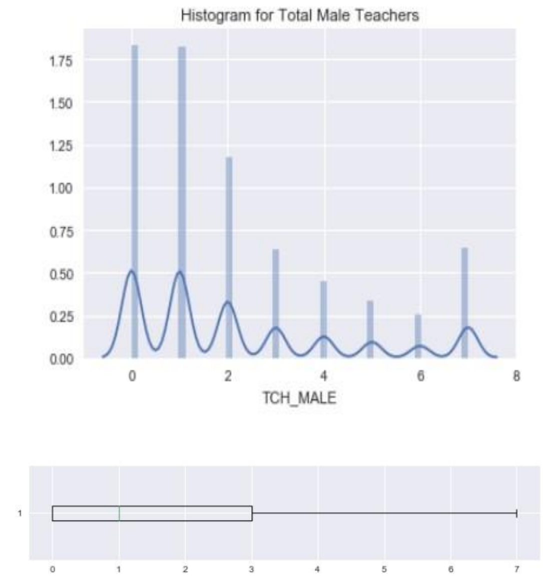

Histogram for Number of student Continuing Who got admission under 25% quota
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
WSEC25P_ENROLLED
1
−0.04
−0.02
0.00
0.02
0.04

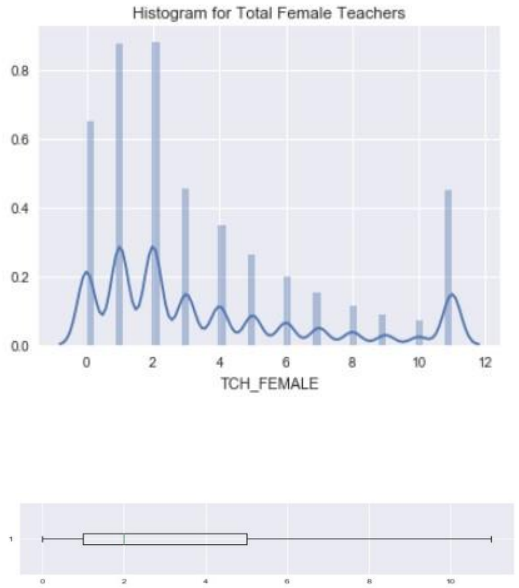

Histogram for No. of children enrolled Special Training in current year– Boys
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
SPLTRG_CY_ENROLLED_B
1
−0.04
−0.02
0.00
0.02
0.04

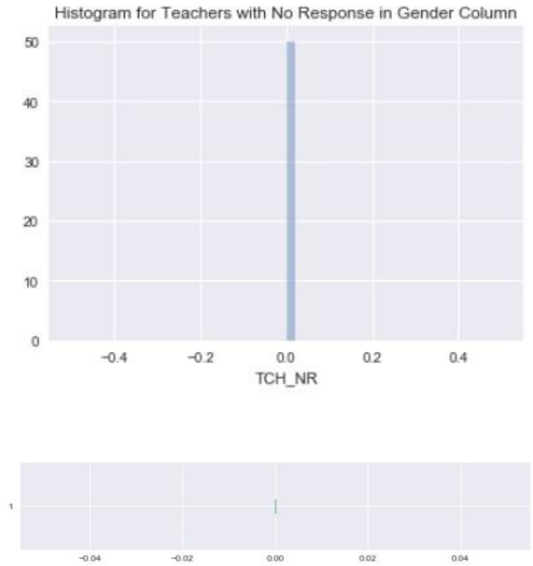

Histogram for No. of children enrolled Special Training in current year– Girls
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
SPLTRG_CY_ENROLLED_G
1
−0.04
−0.02
0.00
0.02
0.04

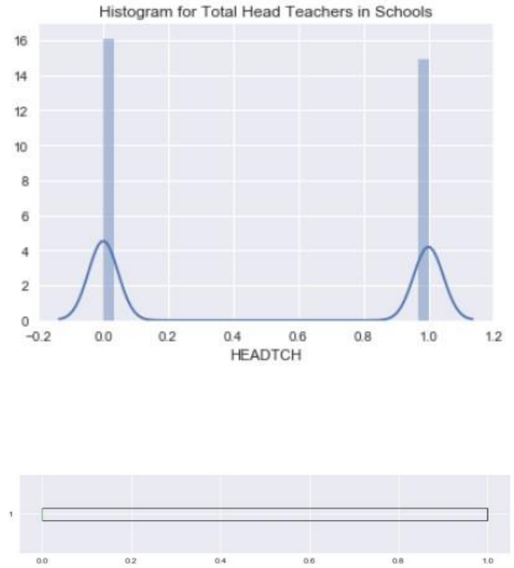

Histogram for No. of children enrolled Special Training in previous year– Boys
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
SPLTRG_PY_ENROLLED_B
1
−0.04
−0.02
0.00
0.02
0.04

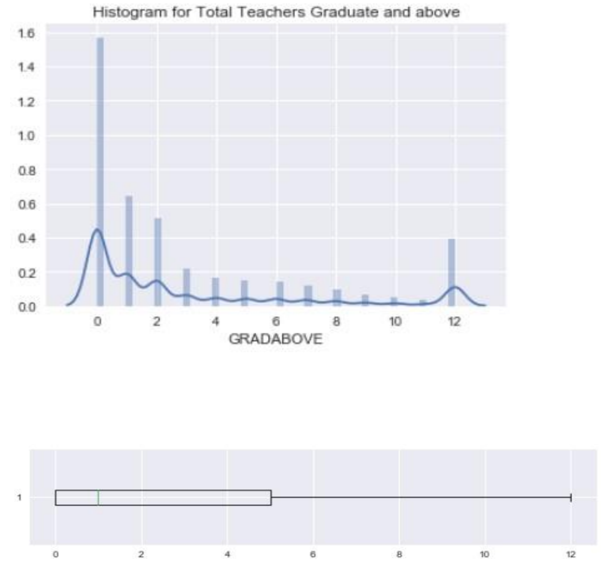

Histogram for No. of children enrolled Special Training in Previous year– Girls
50
40
30
20
10
0
−0.4
−0.2
0.0
0.2
0.4
SPLTRG_PY_ENROLLED_G
1
−0.04
−0.02
0.00
0.02
0.04

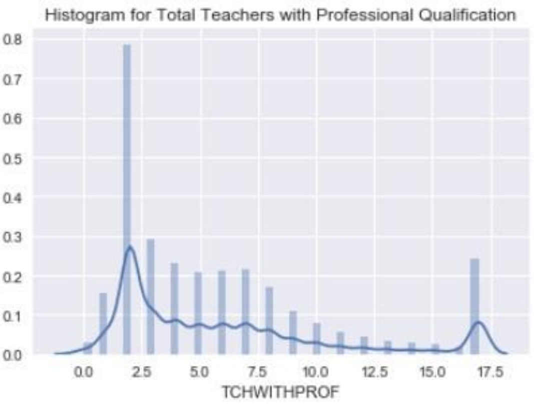

Histogram for Total Male Teachers
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
TCH_MALE
1
0
1
2
3
4
5
6
7

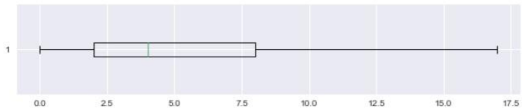

Histogram for Total Female Teachers
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
TCH_FEMALE
1
0
2
4
6
8
10

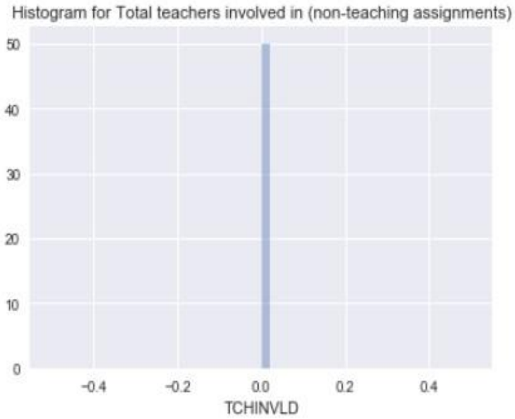

Histogram for Teachers with No Response in Gender Column
50
40
30
20
10
0
−0.4
−0.2
0.0
0.2
0.4
TCH_NR
1
−0.04
−0.02
0.00
0.02
0.04

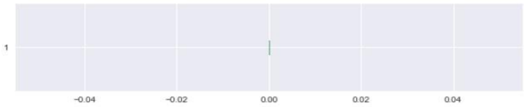

Histogram for Total Head Teachers in Schools
16
14
12
10
8
6
4
2
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
HEADTCH
1
0.0
0.2
0.4
0.6
0.8
1.0

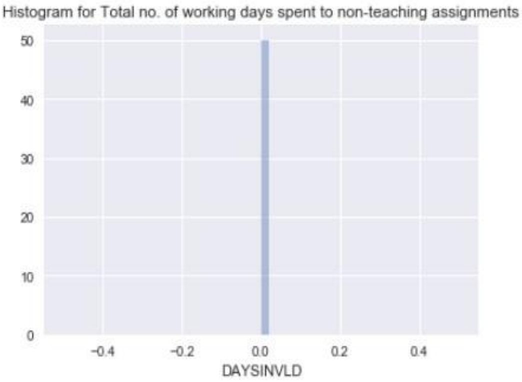

Histogram for Total Teachers Graduate and above
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
GRADABOVE
1
0
2
4
6
8
10
12

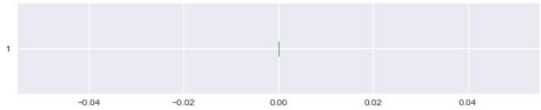

Histogram for Total Teachers with Professional Qualification
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
TCHWITHPROF


1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Histogram for Total teachers involved in (non-teaching assignments)
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
TCHINVLD


1
−0.04
−0.02
0.00
0.02
0.04


Histogram for Total no. of working days spent to non-teaching assignments
0
10
20
30
40
50
−0.4
−0.2
0.0
0.2
0.4
DAYSINVLD


1
−0.04
−0.02
0.00
0.02
0.04

CHAPTER 4

PREDICTIVE MODEL DEVELOPMENT

4.1 Multiple Linear Regression Model (MLR) & Random Forest Model (RF)

Here our objective is to build a model to predict the enrolment of children at schools in Tamil Nadu and Karnataka. Here the target variable is continuous, so this is a Regression problem.

The Model Development was done at multiple levels to arrive at a most suitable model. In general we built several MLR and RF models, once dropping the outliers and the other time we capped outlier with the Whisker maximum from the Boxplot. Here we considered 51 independent variables to predict the school enrollment. We also used L1 and L2 regularization methods to nullify the effect of multi-collinearity. We split the whole 2014-15dataset into 2 parts, here 70% of the data is being used to build the model and we test the model performance using other 30% of the data. Also, we validate our model on 2015-16 dataset.

4.2 Model Performance:

Dataset: After dropping the outliers

Total number of predictors: 51

Total number of rows in X_train: 57726

Total number of rows in X_test: 24741

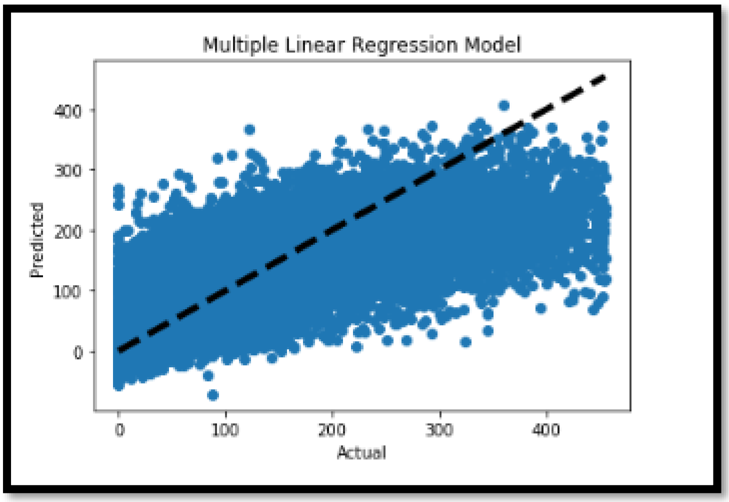
Total number of rows in X_new: 78767

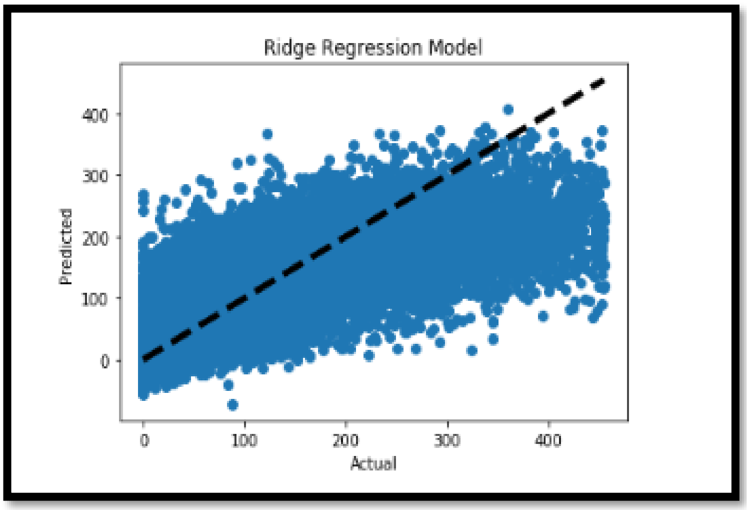
Here, 38 variables are significant out of 51 variables considered for model building

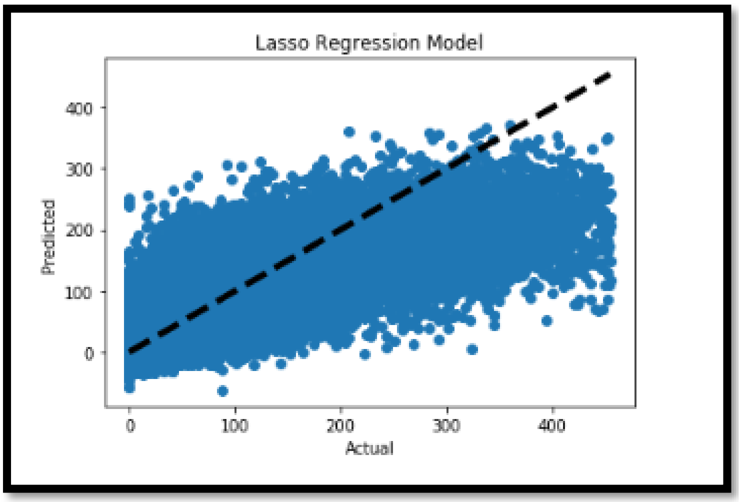
$$
\begin{aligned}
&Total\ Enrollment \\
&\quad = -63.6520 - 5.9040 * STATE + 0.0388 * BLDSTATUS + 9.9773 * CLROOMS + 0.9138 \\
&\quad * CLGOOD + 2.4032 * TOILETB + 1.4531 * TOILET\_G + 12.0706 * MEALSINSCH \\
&\quad + 7.8799 * CAL\_YN + 1.2079 * HMROOM\_YN - 0.6573 * ELECTRIC\_YN - 0.4724 \\
&\quad * BNDRYWALL + 5.7152 * LIBRARY\_YN - 0.3855 * PGROUND\_YN - 0.0031 \\
&\quad * BOOKINLIB - 1.9237 * WATER + 1.5414 * MEDCHK\_YN + 2.6104 * RAMPS\_YN \\
&\quad + 0.4042 * COMPUTER + 1.2729 * RURURB + 3.1643 * MEDINSTR1 - 0.0375 \\
&\quad * ESTDYEAR - 5.8869 * BOARDSEC - 1.5275 * BOARDHSEC - 15.5858 * PPSEC\_YN \\
&\quad + 4.5153 * SCHRES\_YN + 4.1934 * SCHMGT + 3.9423 * SCHCAT + 14.8200 * SCHTYPE \\
&\quad + 7.0701 * SCHSHI\_YN + 0.0899 * WORKDAYS - 1.3802 * RESITYPE - 0.0292 \\
&\quad * CONTI\_R + 0.0025 * CONTI\_E + 0.0055 * SCHMNTCGRANT\_R - 0.0069 \\
&\quad * SCHMNTCGRANT\_E - 0.0292 * FUNDS\_R + 0.0025 * FUNDS\_E - 3.0919 \\
&\quad * APPROACHBYROAD - 3.0037 * CCE\_YN + 2.1418 * PCR\_MAINTAINED - 2.4408 \\
&\quad * PCR\_SHARED + 0.8920 * SMCSDP\_YN + 0.0377 * SPLTRG\_BY + 10.2038 \\
&\quad * SPLTRG\_PLACE + 5.2378 * SPLTRG\_TYPE + 1.2161 * TXTBKRECD\_YN + 15.4692 \\
&\quad * TCH\_MALE + 9.5754 * TCH\_FEMALE + 0.1046 * HEADTCH - 1.8259 * GRADABOVE \\
&\quad + 5.4774 * TCHWITHPROF
\end{aligned}
$$

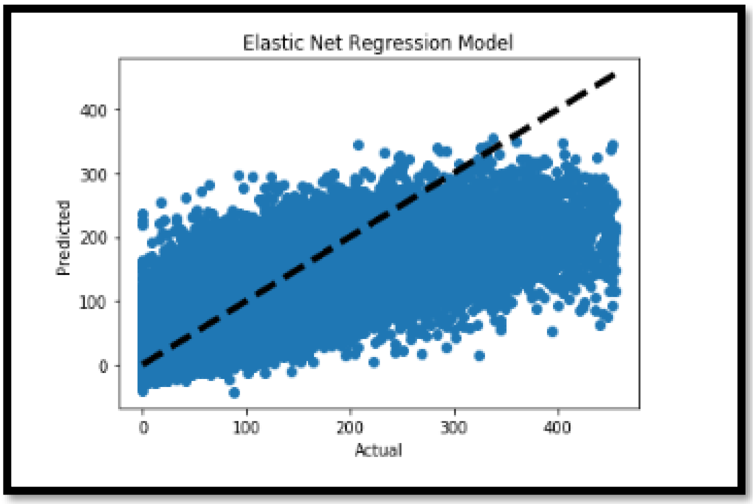
Models	R^2	Adjusted R^2	RMSE_Train	RMSE_Test	RMSE_Validation
Linear Model	66.02%	65.99%	46.58	46.24	45.75
Quadratic Model	73.11%	72.51%	41.44	43.25	79
Ridge Regression	66.02%	66.025%	46.58	46.24	45.75
Lasso Regression	65.07%	65.05%	47.23	46.96	47.34
Elastic Net	64.53%	64.53%	47.59	47.33	46.77
Random Forest	79.31%	79.30%	36.35	41.04	41.11

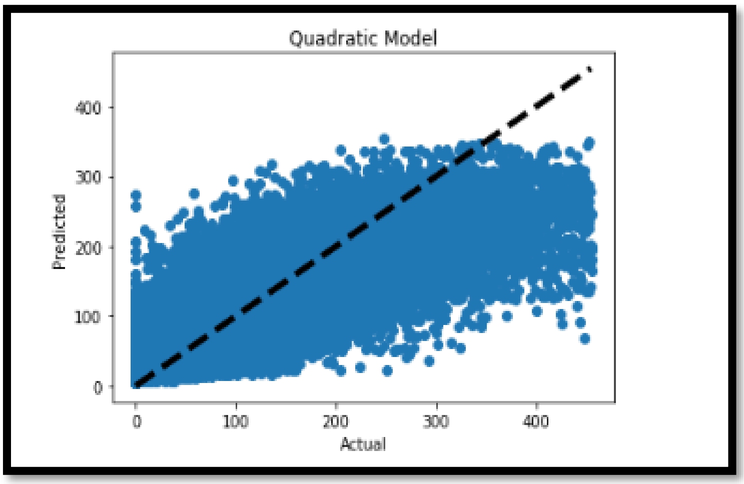
Predicted VS Actual Graphs:

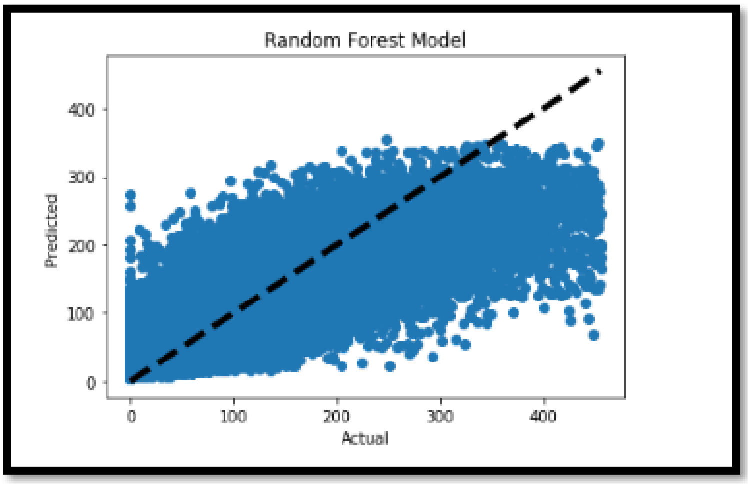












Finding the important Variable:

Top 27 variables based on their information Value

Variables	Information Value
SPLTRG_BY	1.234247e-06
BOARDSEC	1.398072e-06
MEALSINSCH	1.507714e-06
PPSEC_YN	1.654562e-06
PGROUND_YN	2.143140e-06
RURURB	2.633429e-06
SMCSDP_YN	3.003802e-06
SCHMGT	5.064561e-06
BLDSTATUS	5.170575e-06
TOILETB	6.417015e-06
BOOKINLIB	6.761313e-06
GRADABOVE	7.163585e-06
TOILET_G	7.363754e-06
SCHMNTCGRANT_E	1.077047e-05
SCHMNTCGRANT_R	1.198435e-05
COMPUTER	1.260902e-05
HMROOM_YN	1.377505e-05
CONTI_E	1.416364e-05
FUNDS_E	1.416364e-05
FUNDS_R	1.606820e-05
CONTI_R	1.606820e-05
TCH_MALE	2.129136e-05
TCH_FEMALE	2.893302e-05
CLGOOD	3.464384e-05
SCHCAT	3.712851e-05
CLROOMS	5.930342e-05
TCHWITHPROF	8.416248e-05

LMS Model	R^2	Adjusted R^2	RMSE_train	RMSE_test	RMSE_Validation
Considering Top 14 variables	63.50%	63.49%	48.28	48.01	46.65
Considering Top 27 variables	65.39%	65.38%	47.01	46.71	45.48
Only significant variables	66.01%	65.99%	46.58	46.25	46.02
Considering All the variables	66.02%	65.99%	46.58	46.24	45.75

Validation of Regression Assumptions:

1) Linear relationship between X and Y:

Most of the significant X variable has either positive or negative correlation with the Target value. So, there exists a linear relationship between predictors and the Target variable.

Predictors	Correlation
CLROOMS	0.6712
CLGOOD	0.5477
TOILETB	0.2575
TOILET_G	0.2794
BOOKINLIB	0.2469
COMPUTER	0.3848
WORKDAYS	0.0141
CONTI_R	0.2371
CONTI_E	0.2093
SCHMNTCGRANT_R	0.1944
SCHMNTCGRANT_E	0.1764
FUNDS_R	0.2371
FUNDS_E	0.2093
TCH_MALE	0.4873
TCH_FEMALE	0.5141
HEADTCH'	0.0779
GRADABOVE	0.3620
TCHWITHPROF	0.7504

2) Auto correlation:

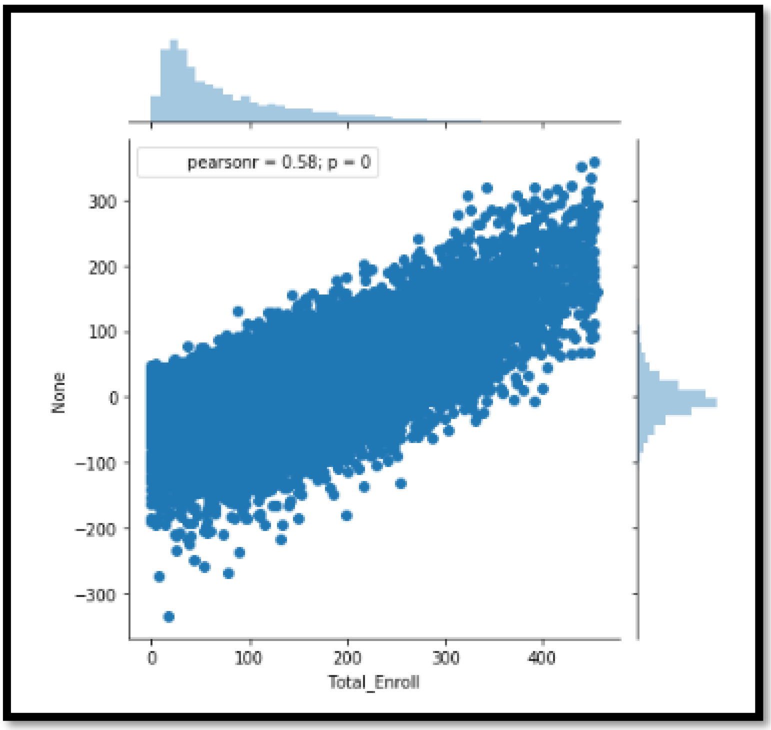
From Durbin-Watson test we found value d=2.004, which is almost equal to 2, so we can say that there is no autocorrelation present in the residuals.

3) Multi-Collinearity:

We check for multi-collinearity by fining VIF, there ae 7 variables whose VIF value is above 5, so we drop the to avoid multi-collinearity

4) Heteroscedasticity:

Residual vs fitted graph: From the graph mentioned below, we can see that the plot does not exhibit any funnel shape pattern and it is additive in nature



5) Normality of the residuals:

From the normality test, we found that p value=0.0 which is less than alpha value 0.05, so we reject the null hypothesis, which means the residuals are not normally distributed. This also says that, linear regression might not be the best method to be used for this dataset. We can achieve the normality either by transformation of the X variables or by introducing Quadratic or Cubic model.

Dataset: After treating the outliers

Total number of predictors: 51

Total number of rows in X_train: 80237

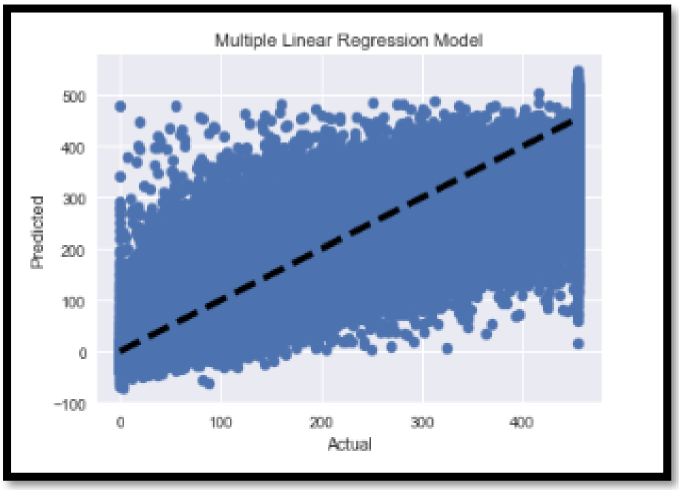
Total number of rows in X_test: 34388

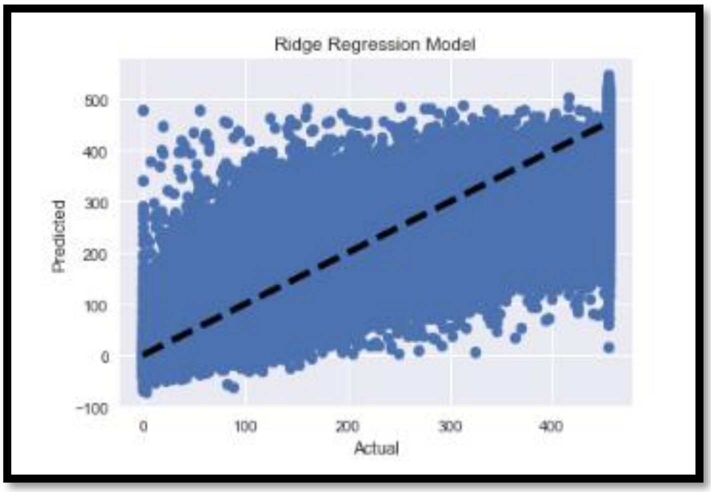
Total number of rows in X_new: 114625

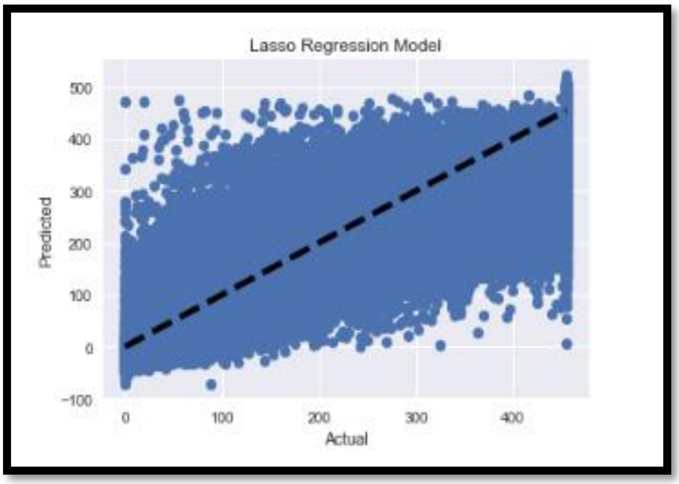
Here, 35 variables are significant out of 51 variables considered for model building

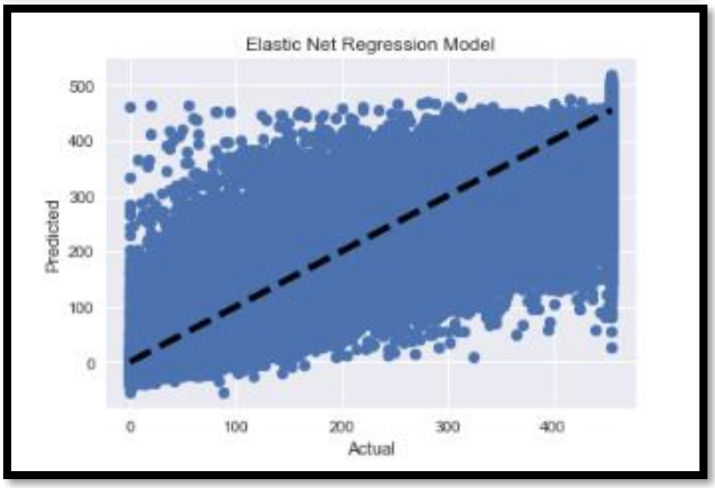
$$
\begin{aligned}
&Total\ Enrollment \\
&\quad = -134.9181 - 11.8968 * STATE + 1.4529 * BLDSTATUS + 9.6708 * CLROOMS + 1.6302 \\
&\quad * CLGOOD + 4.3229 * TOILETB + 3.0455 * TOILET_{G} + 13.0306 * MEALSINSCH - 1.3005 \\
&\quad * CAL_{YN} - 0.0080 * HMROOM_{YN} + 0.0154 * ELECTRIC_{YN} - 0.6935 * BNDRYWALL \\
&\quad + 1.2429 * LIBRARY_{YN} + 0.6065 * PGROUND_{YN} + 0.0055 * BOOKINLIB - 1.3202 \\
&\quad * WATER + 1.3612 * MEDCHK_{YN} + 3.1425 * RAMPS_{YN} + 2.5752 * COMPUTER + 2.5204 \\
&\quad * RURURB + 4.6985 * MEDINSTR1 - 0.0841 * ESTDYEAR - 4.6854 * BOARDSEC \\
&\quad + 14.5849 * BOARDHSEC - 14.5753 * PPSEC_{YN} - 2.3765 * SCHRES_{YN} + 3.8227 \\
&\quad * SCHMGT + 4.0867 * SCHCAT + 18.2671 * SCHTYPE + 21.2028 * SCHSHI_{YN} + 0.1117 \\
&\quad * WORKDAYS + 1.5803 * RESITYPE - 16.9900 * CONTI_{R} + 0.0013 * CONTI_{E} - 0.0342 \\
&\quad * SCHMNTCGRANT_{R} - 0.0010 * SCHMNTCGRANT_{E} + 16.9106 * FUNDS_{R} + 0.0013 \\
&\quad * FUNDS_{E} - 3.6049 * APPROACHBYROAD - 2.3393 * CCE_{YN} + 3.8855 * PCR_{MAINTAINED} \\
&\quad - 2.4408 * PCR_{SHARED} + 1.0464 * SMCSDP_{YN} + 0.6608 * SPLTRG_{BY} + 9.4582 \\
&\quad * SPLTRG_{PLACE} + 4.8622 * SPLTRG_{TYPE} - 0.6211 * TXTBKRECD_{YN} + 8.5023 * TCH_{MALE} \\
&\quad + 3.3537 * TCH_{FEMALE} - 1.8745 * HEADTCH - 0.8854 * GRADABOVE + 13.3303 \\
&\quad * TCHWITHPROF
\end{aligned}
$$

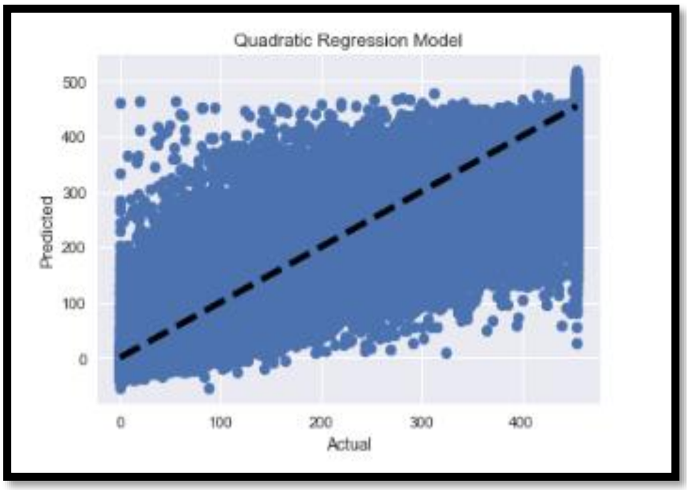
Models	R^2	Adjusted R^2	RMSE_Train	RMSE_Test	RMSE_Validation
Linear Model	79.04%	79.03%	62.74	62.17	60.81
Quadratic Model	82.98%	82.70%	56.54	57.72	57.22
Ridge Regression	79.04%	79.03%	62.74	62.17	60.81
Lasso Regression	78.65%	78.50%	63.34	62.78	60.93
Elastic Net	78.46%	78.40%	63.60	63.04	60.66
Random Forest	87.18%	87.10%	49.06	55.49	53.22

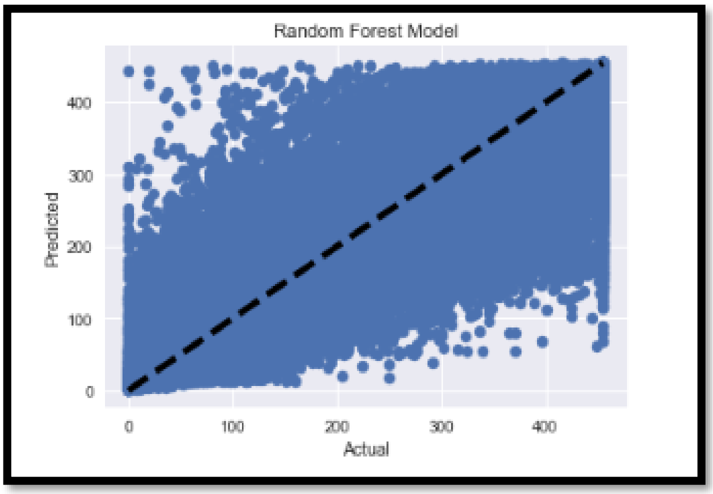

Multiple Linear Regression Model
Predicted
500
400
300
200
100
0
−100
0
100
200
300
400
Actual


Ridge Regression Model
Predicted
500
400
300
200
100
0
−100
0
100
200
300
400
Actual


Lasso Regression Model
Predicted
500
400
300
200
100
0
−100
0
100
200
300
400
Actual


Elastic Net Regression Model
Predicted
500
400
300
200
100
0
0
100
200
300
400
Actual


Quadratic Regression Model
Predicted
500
400
300
200
100
0
0
100
200
300
400
Actual


Random Forest Model
Predicted
400
300
200
100
0
0
100
200
300
400
Actual

Finding the important Variable:

Top 25 variables based on their information Value

Variables	Information Value
PPSEC_YN	2.349576e-06
MEALSINSCH	2.386834e-06
FUNDS_E	2.621026e-06
CONTI_E	2.621026e-06
FUNDS_R	2.800075e-06
CONTI_R	2.800075e-06
PGROUND_YN	3.068475e-06
SMCSDP_YN	3.242857e-06
RURURB	3.327548e-06
BOARDHSEC	3.898906e-06
SCHMGT	4.653129e-06
BLDSTATUS	5.031488e-06
BOARDSEC	5.481982e-06
GRADABOVE	7.797922e-06
BOOKINLIB	8.841864e-06
TOILETB	9.294444e-06
HMROOM_YN	1.043655e-05
TOILET_G	1.131330e-05
TCH_MALE	1.151749e-05
COMPUTER	1.241030e-05
CLGOOD	1.684294e-05
CLROOMS	2.113310e-05
TCH_FEMALE	2.129351e-05
SCHCAT	2.389661e-05
TCHWITHPROF	5.177055e-05

LMS Model	R^2	Adjusted R^2	RMSE_train	RMSE_test	RMSE_Validation
Considering Top 9 variables	77.52 %	77.52%	64.98	64.38	61.53
Considering Top 25 variables	78.65%	78.64%	63.33	62.73	60.25
Only significant variables	79.01%	79.02%	62.79	62.19	60.85
Considering All the variables	79.04%	79.03%	62.74	62.17	60.81

Validation of Regression Assumptions:

1) Linear relationship between X and Y:

Most of the significant X variable has either positive or negative correlation with the Target value. So, there exists a linear relationship between predictors and the Target variable.

Predictors	Correlation
CLROOMS	0.6713
CLGOOD	0.6343
TOILETB	0.4977
TOILET_G	0.5266
BOOKINLIB	0.4924
COMPUTER	0.6294
WORKDAYS	-0.2024
CONTI_R	-0.0073
CONTI_E	-0.0079
SCHMNTCGRANT_R	-0.0393
SCHMNTCGRANT_E	-0.0393
FUNDS_R	-0.0070
FUNDS_E	-0.0079
TCH_MALE	0.5703
TCH_FEMALE	0.7198
HEADTCH'	-0.0903
GRADABOVE	0.6270
TCHWITHPROF	0.8564

2) Auto correlation:

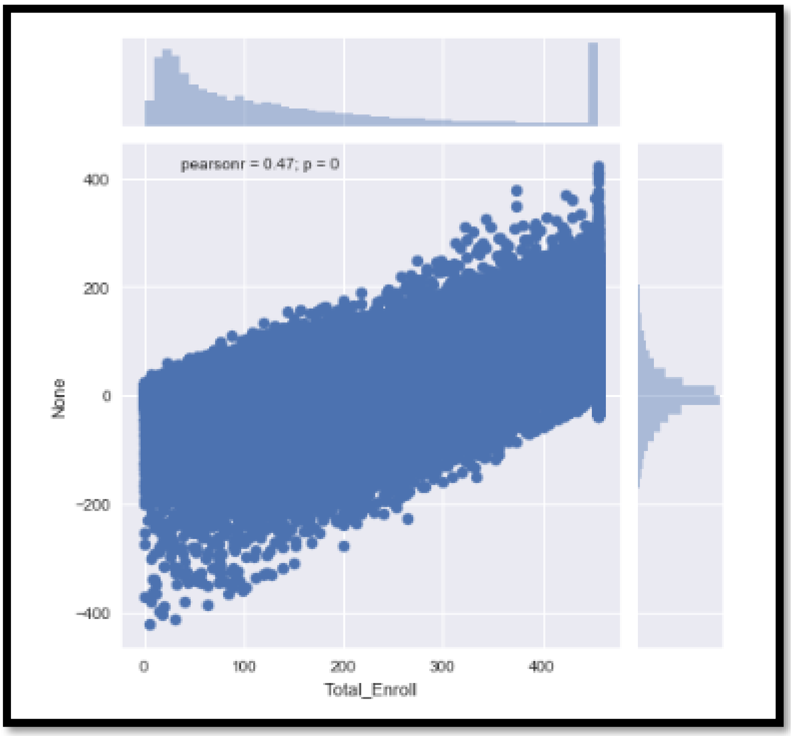
From Durbin-Watson test we found value d=1.996, which is almost equal to 2, so we can say that there is no autocorrelation present in the residuals.

3) Multi-Collinearity:

We check for multi-collinearity by finding VIF, there is only 1 variables whose VIF is above 5, so we drop it to avoid multi-collinearity

4) Homoscedasticity:

Residual vs fitted graph: From the graph mentioned below, we can see that the plot does not exhibit any funnel shape pattern and it is additive in nature



5) Normality of the residuals:

From the normality test, we found that p value=0.0 which is less than alpha value 0.05, so we reject the null hypothesis, which means the residuals are not normally distributed. This also says that, linear regression might not be the best method to be used for this dataset. We can achieve the normality either by transformation of the X variables or by introducing Quadratic or Cubic model.

4.3: Conclusion & Recommendations

- Random Forest Model gives best R^2 and lowest MSE when compare to all the models
- We use L1 and L2 regularisation methods to overcome the effect of Multi-Collinearity
- Variables which has more effect on Enrolment of Students are
  - Number of Teachers with Professional experience
  - Category of the School
  - Number of teachers who are Graduates and above
  - Amount of School Development Grant Receipt and Expenditure
  - School Building Status
  - Number of books in the Library
  - Funds from other sources receipts and Expenditures
  - Board of the Secondary and Higher Secondary sections
  - Location of the School
  - Presence of Playground
  - Status of the Mid-Day meal
  - Number of good classrooms
  - Also the number of computers, classrooms, and Number of Male and Female teachers.