Master in Computer Science (M.Sc.) – I Sem

Department of Computer Science

University of Delhi

Mathematical foundations of computer science

Exploratory Data analysis

Submitted to: Prof. Vasudha Bhatnagar

Submitted By: -

Saurav Khewal

Akash Chaudhary

Lokesh Gupta

Objective

Choose dataset(s) of your interest from data.gov.in (Open data of the Govt. of India). Explore the data using statistical tools and techniques, visualization and draw inferences. Prepare a report and submit. The report must describe the data well (what it is about), types of attributes, followed by detailed analyses. Clearly state the questions that you aim to explore, state the method you will use and the interpretation of the result.

Dataset

Name of the dataset:

Number of Schools by Availability of Infrastructure and Facilities, School Management and School Category in Maldah District of West Bengal (UDISE plus) during 2014-15

About the dataset:

The catalog contains data related to number of schools by availability of infrastructure and facilities, school management and school category (UDISE plus). U-DISE has mandate to collect infrastructure information from all recognized and unrecognized schools imparting formal education from class I to XII.

Released under:

National Data Sharing and Accessibility Policy (NDSAP)

Source:

Open Government Data (OGD) Platform India

Number of Schools by Availability of Infrastructure and Facilities, School Management and School Category in Maldah District of West Bengal (UDISE plus) during 2014-15

- The dataset contains 260 rows and 45 columns
- The dataset has the following attributes:
  - Academic_Year
  - State_Code
  - State_Name
  - District_Code
  - District_Name
  - Block_Code
  - Udise_Block_Name
  - School_Category_Id
  - School_Category_Name
  - School_Management_Id
  - School_Management_Name
  - Location School_Type_Id
  - School_Type
  - Total_Number_of_Schools
  - Building
  - Complete_Medical_Checkup
  - Computer_Available
  - Functional_Drinking_Water
  - Drinking_Water
  - Functional_Electricity
  - Functional_Boy_Toilet
  - Functional_Girl_Toilet
  - Functional_Toilet_Facility

- Functional_Toilet_and_Urinal
- Functional_Urinal
- Functional_Urinal_Boy
- Functional_Urinal_Girl
- Handwash
- Separate_Room_for_Headmaster
- Incinerator
- Internet
- Kitchen_Garden
- Land_Available
- Librarian
- Library_or_Reading_Corner_or_Book_Bank
- Medical_Checkup
- Newspaper
- Playground
- Rain_Water_Harvesting
- Ramps
- Solar_Panel
- Furniture
- Water_Purifier
- Water_Tested

Sample of the Dataset

Academic_Year	State_Code	State_Name	District_Code	District_Nam	Block_Code	Udise_Block_Name	School_Category_Id	School_Category_Name	School_Management_Id	School_Management_Name
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	3	HSS (I-XII)	7	Central Govt
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	5	HSS (VI-XII)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	5	HSS (VI-XII)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	5	HSS (VI-XII)	97	Madarsa recognized (by Wakf b
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	1	PS (I-V)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	1	PS (I-V)	4	Government Aided
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	1	PS (I-V)	5	Private Unaided (Recognized)
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	1	PS (I-V)	5	Private Unaided (Recognized)
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	7	SS (VI-X)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	7	SS (VI-X)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	7	SS (VI-X)	97	Madarsa recognized (by Wakf b
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	7	SS (VI-X)	98	Madarsa unrecognized
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	2	UPS (I-VIII)	5	Private Unaided (Recognized)
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	2	UPS (I-VIII)	4	Government Aided
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	4	UPS (VI-VIII)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190601	ENGLISH BAZAR(BLOCK)	4	UPS (VI-VIII)	5	Private Unaided (Recognized)
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	5	HSS (VI-XII)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	5	HSS (VI-XII)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	5	HSS (VI-XII)	97	Madarsa recognized (by Wakf b
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	1	PS (I-V)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	1	PS (I-V)	5	Private Unaided (Recognized)
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	6	SS (I-X)	5	Private Unaided (Recognized)
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	7	SS (VI-X)	1	Department of Education
2014-15	19	West Bengal	1906	MALDAH	190602	HARISHCHANDRAPUR-1	7	SS (VI-X)	97	Madarsa recognized (by Wakf b

Location	School_Type_Id	School_Type	Total_Number_of_Schools	Building	Complete_Medical_Checkup	Computer_Available	Functional_Drinking_Water	Drinking_Water	Functional_Electricity	Functional_Boy
Rural	3	Co-Ed	1	1	0	1	1	1	1	1
Rural	2	Girls	1	1	0	1	1	1	1	0
Rural	3	Co-Ed	11	11	0	10	11	11	11	11
Rural	3	Co-Ed	1	1	0	1	1	1	1	1
Rural	3	Co-Ed	165	161	0	3	152	165	70	155
Rural	3	Co-Ed	1	1	0	0	0	1	0	1
Rural	3	Co-Ed	45	44	0	8	37	44	32	36
Urban	3	Co-Ed	1	1	0	0	1	1	1	1
Rural	2	Girls	2	2	0	2	2	2	2	0
Rural	3	Co-Ed	9	9	0	6	9	9	9	8
Rural	3	Co-Ed	1	1	0	0	1	1	1	1
Rural	3	Co-Ed	1	1	0	0	0	1	0	1
Rural	1	Boys	1	1	0	0	0	1	1	1
Rural	3	Co-Ed	1	1	0	1	1	1	1	1
Rural	3	Co-Ed	16	16	0	0	15	16	8	15
Rural	3	Co-Ed	1	1	0	0	0	1	0	1
Rural	2	Girls	1	1	0	1	1	1	1	0
Rural	3	Co-Ed	6	6	0	5	6	6	6	6
Rural	3	Co-Ed	1	1	0	1	1	1	1	1
Rural	3	Co-Ed	144	144	0	3	96	143	49	98
Rural	3	Co-Ed	37	37	0	3	25	35	11	10
Rural	3	Co-Ed	1	1	0	0	0	1	0	0
Rural	3	Co-Ed	5	5	0	2	5	5	5	3
Rural	3	Co-Ed	1	1	0	0	1	1	1	1

Functional_Girl_Toilet	Functional_Toilet_Facil	Functional	Functional	Functional	Functional	Handwash	Separate_F	Incinerator	Internet	Kitchen_G	Land_Avail	Librarian	Library_or	Medical_C	Newspape	Playgrounc
1	1	0	0	0	0	1	1	0	1	0	0	1	1	0	1	1
1	1	0	0	0	0	0	1	0	1	0	1	0	1	1	1	1
11	11	0	0	0	0	7	6	0	5	0	10	2	11	5	8	10
1	1	0	0	0	0	0	1	0	0	0	1	0	1	1	1	1
153	161	0	0	0	0	126	37	0	0	0	109	0	150	57	4	82
0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1
38	39	0	0	0	0	32	26	0	0	0	28	0	17	4	8	23
1	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1
2	2	0	0	0	0	2	1	0	0	0	1	0	2	2	0	0
8	8	0	0	0	0	6	5	0	1	0	8	0	8	6	2	5
1	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	1
1	1	0	0	0	0	1	1	0	0	0	1	0	1	1	1	1
14	15	0	0	0	0	7	2	0	0	0	14	0	1	1	2	6
1	1	0	0	0	0	1	0	0	0	0	1	0	1	1	1	0
1	1	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1
5	6	0	0	0	0	1	5	0	3	0	6	3	6	3	5	6
1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
112	119	0	0	0	0	35	58	0	0	0	106	0	124	49	6	47
18	20	0	0	0	0	13	8	0	0	0	23	0	5	5	8	13
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	5	0	0	0	0	2	3	0	0	0	5	0	4	2	1	4
1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1

Rain_Wate	Ramps	Solar_Pane	Furniture	Water_Purifier	Water_Tested
0	1	0	1	0	0
0	1	0	1	0	0
0	10	0	1	0	0
0	1	0	1	0	0
0	130	0	33	0	0
0	0	0	1	0	0
0	1	0	18	0	0
0	0	0	1	0	0
0	1	0	1	0	0
0	7	0	3	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	1	0	1	0	0
0	7	0	4	0	0
0	1	0	0	0	0
0	1	0	1	0	0
0	5	0	1	0	0
0	1	0	0	0	0
0	89	0	59	0	0
0	2	0	14	0	0
0	0	0	0	0	0
1	5	0	3	0	0
0	1	0	1	0	0

Overview of the Dataset

Data Structures

Data Structures

Divisions	Metrics	Values
size	observations	260
size	variables	45
size	values	11,700
size	memory size (KB)	0
duplicated	duplicate observation	0
missing	complete observation	260
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

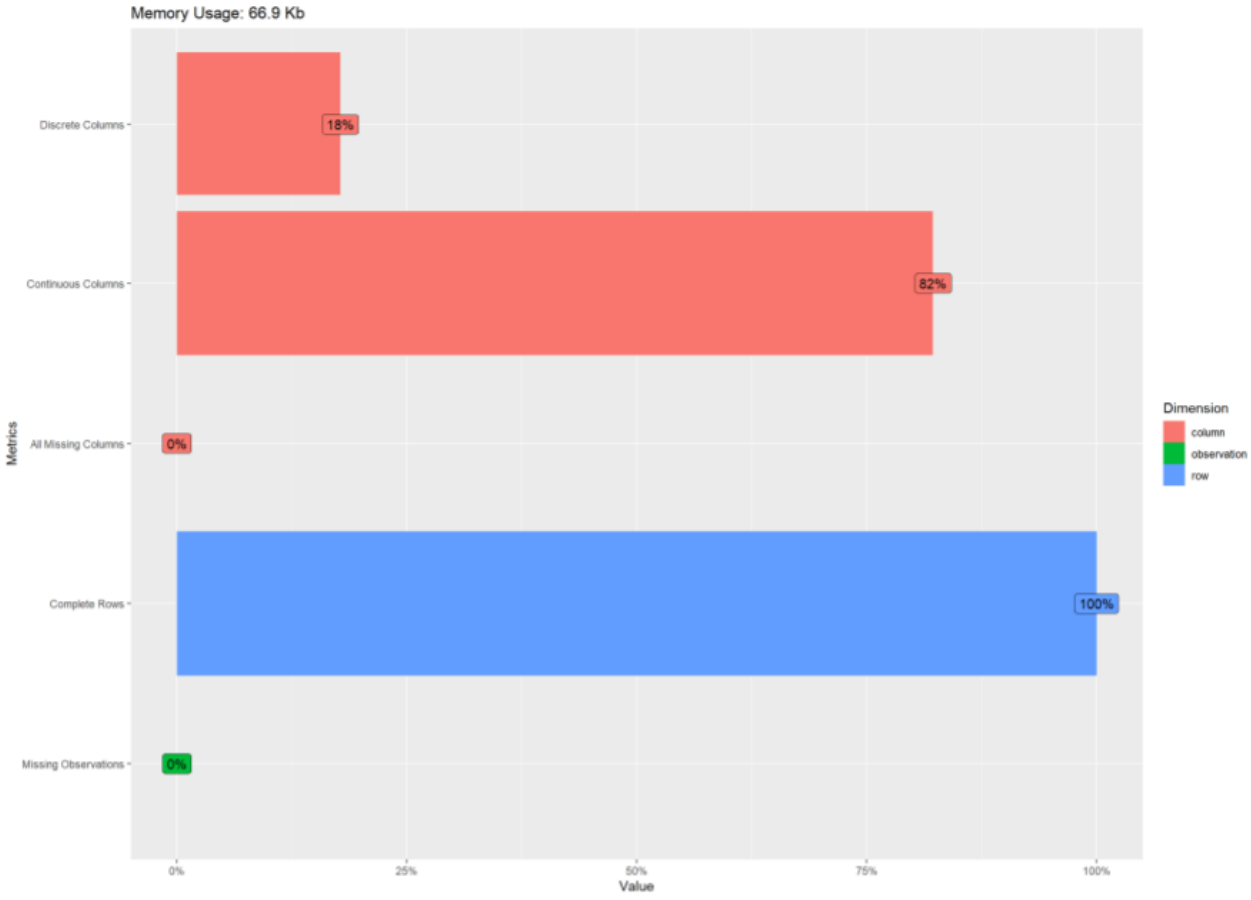
Data Types

Divisions	Metrics	Values
data type	numerics	0
data type	integers	37
data type	factors/ordered	0
data type	characters	8
data type	Dates	0
data type	POSIXcts	0
data type	others	0

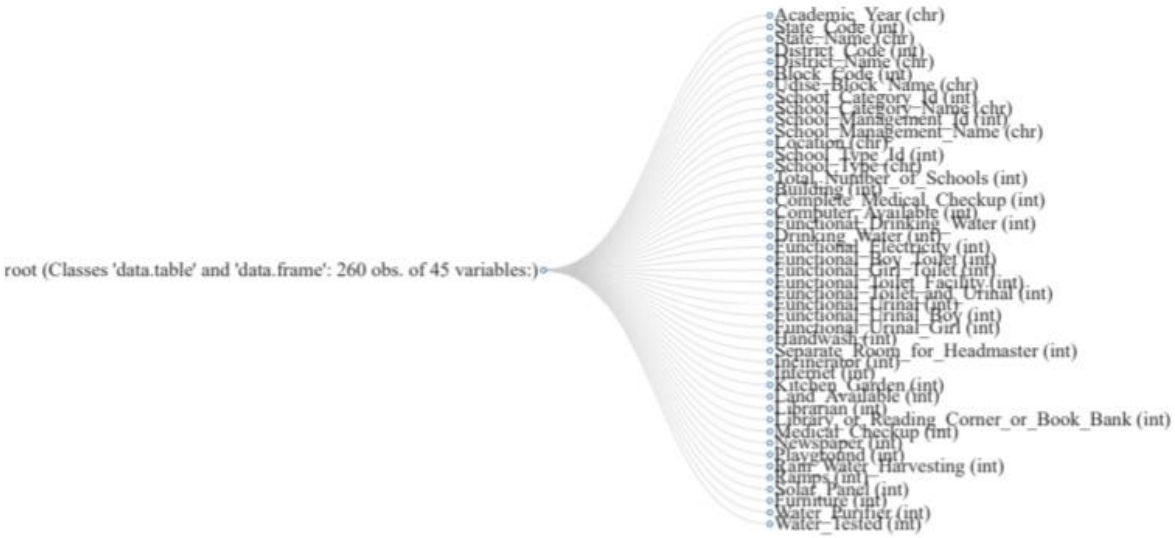
It can be observed that in the in the given dataset of the total no of values recorded are 11,700 and there are no missing observations in our datasets.

Out of the 45 features in the dataset 37 of the features are of integer type and the remaining 8 features are of character data type.

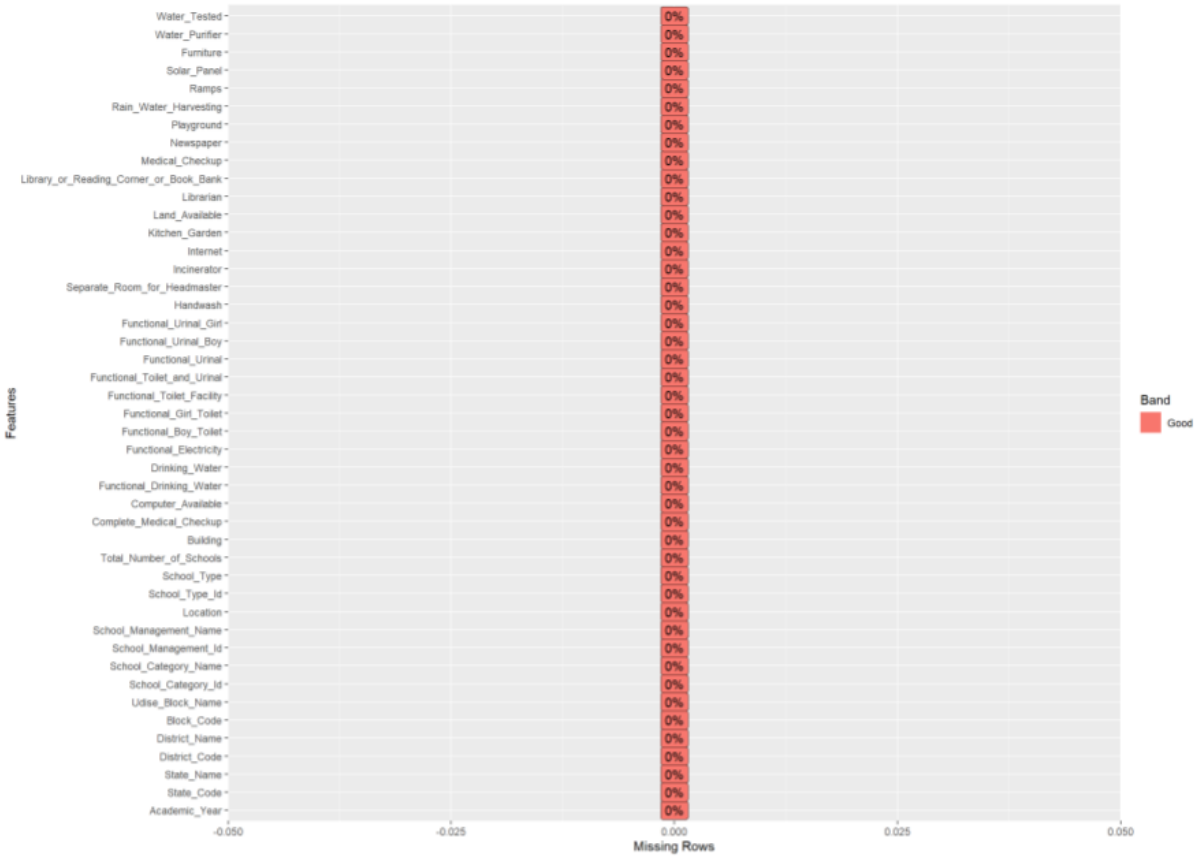
This can be seen in the following plot wherein we can observe that the 18% of the columns(features) are Discrete Columns and the rest 82% are Continuous Columns.



Data Structure



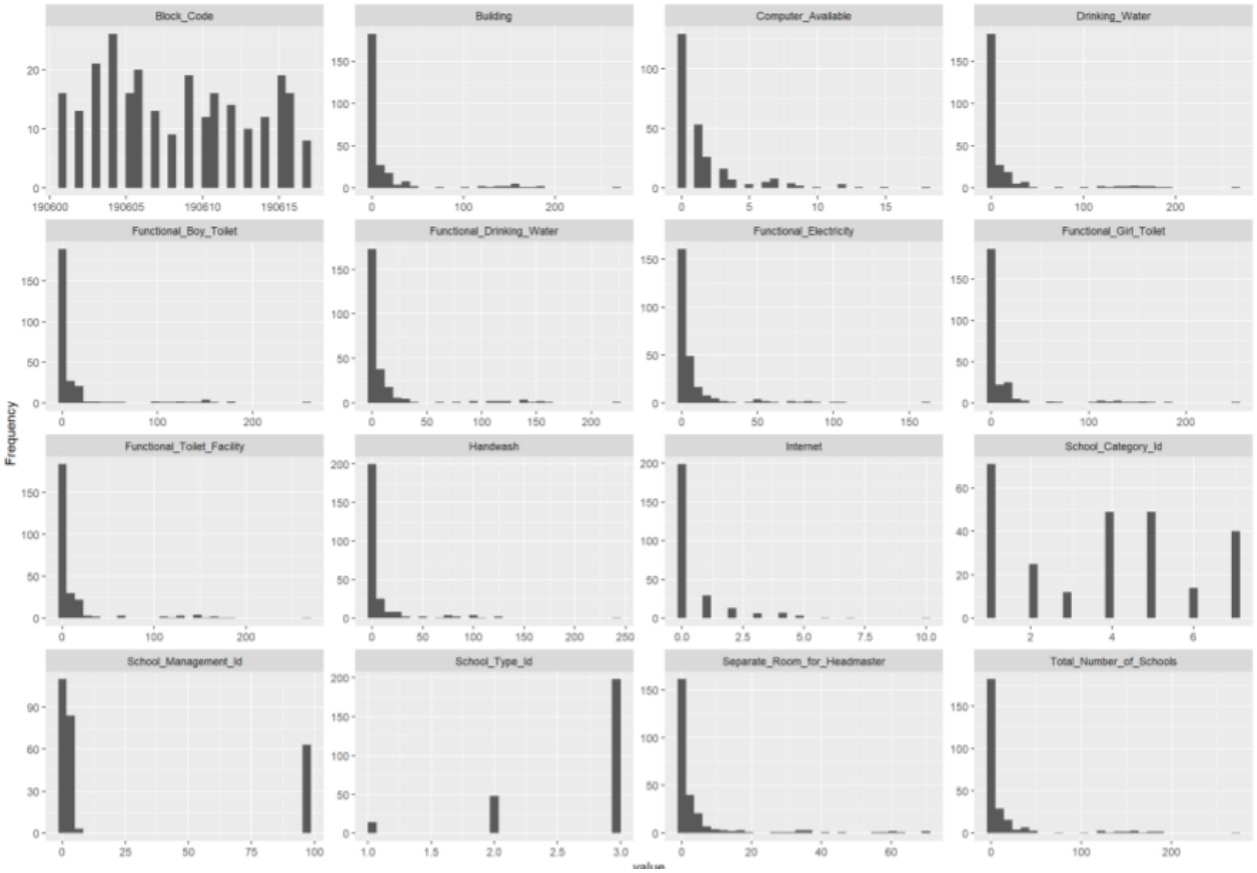
Missing values

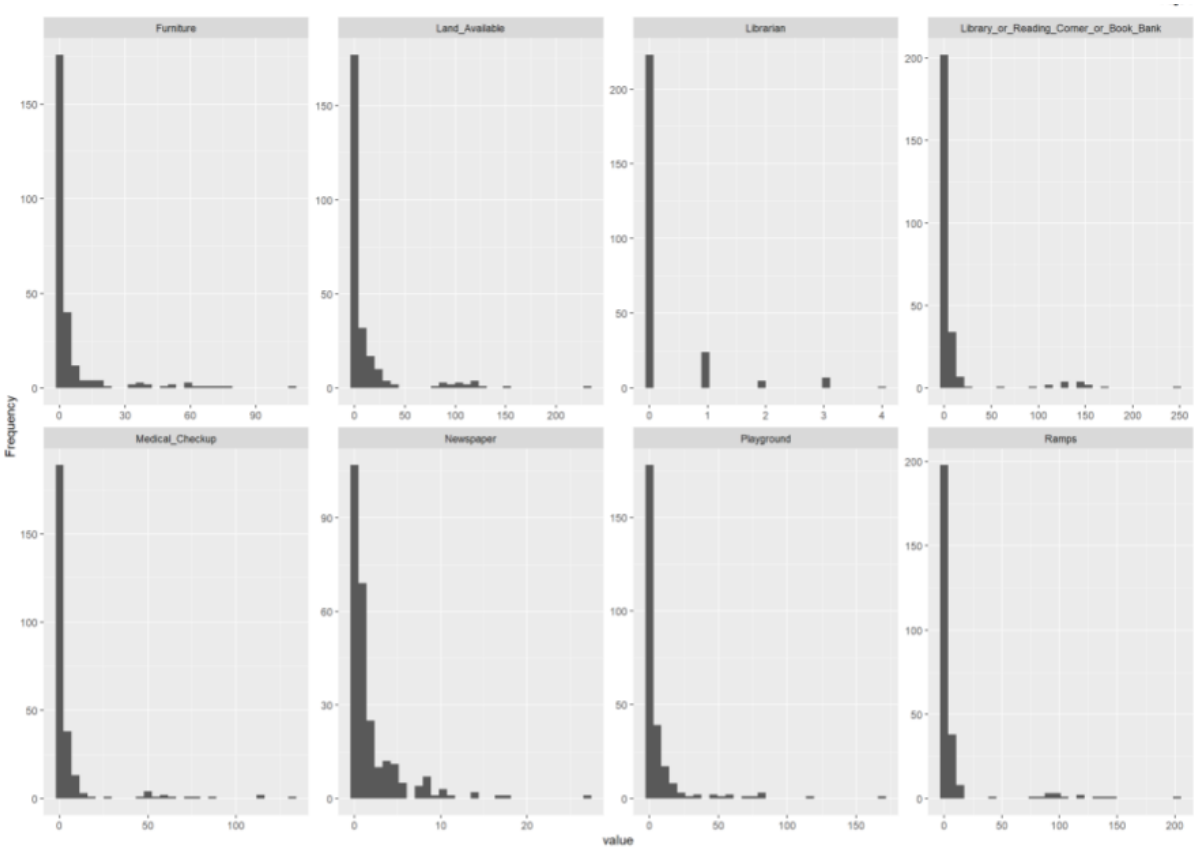


Interpretation

From the following plot one can observe that there are no missing values in our dataset

Histogram


Block_Code
Building
Computer_Available
Drinking_Water
Functional_Boy_Toilet
Functional_Drinking_Water
Functional_Electricity
Functional_Girl_Toilet
Functional_Toilet_Facility
Handwash
Internet
School_Category_Id
School_Management_Id
School_Type_Id
Separate_Room_for_Headmaster
Total_Number_of_Schools
Frequency

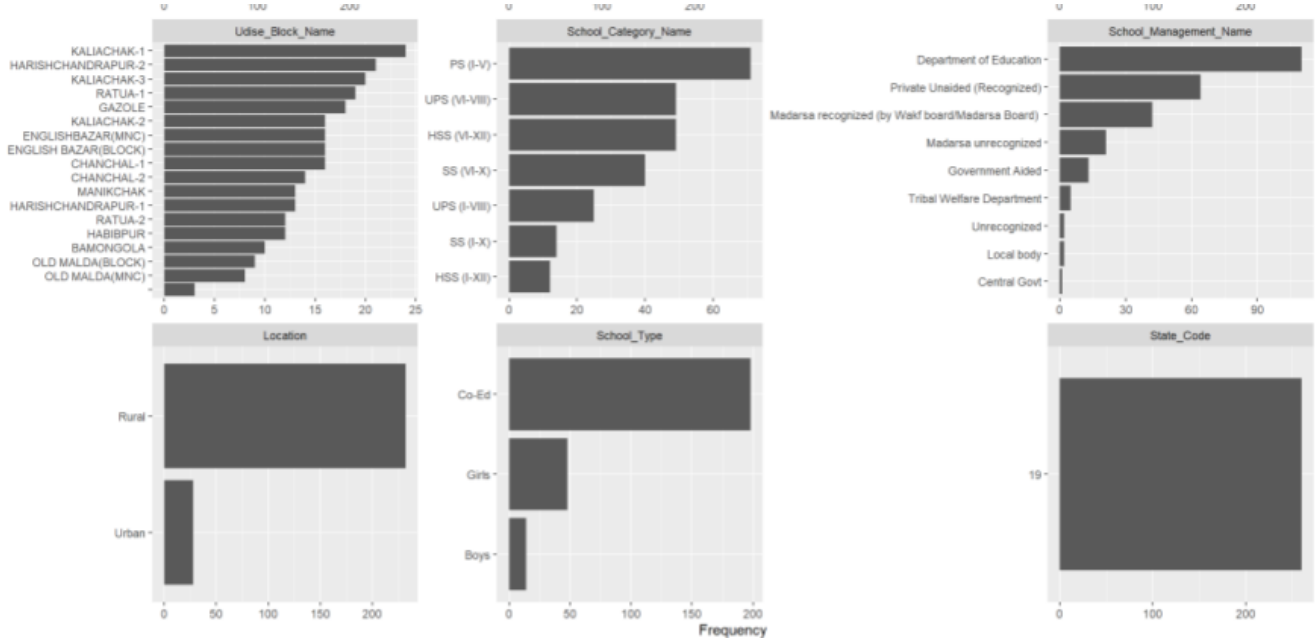


Interpretation

- The data is right-skewed

    - With right-skewed distribution (also known as "positively skewed" distribution), most data falls to the right, or positive side, of the graph's peak. Thus, the histogram skews in such a way that its right side (or "tail") is longer than its left side.

Frequency Bar Chart



Interpretation

From the following Bar Graphs the following observations can be made: -

- The Kaliachak-1 block of the Maldah district has the maximum number of schools
- Majority of the schools in the Maldah district are Primary schools whereas there are very a smaller number of Higher Secondary Schools
- Most of the schools are in Rural area.
- The majority of the schools are co-ed.
- Most of the schools are managed by the Department of Education and there are very few central government schools.

Outliers

The following tables consist of the Min, Max, Q1, Q3, Number of Outliers, Percentage of Outliers and the Position of the Outliers for all the variables.

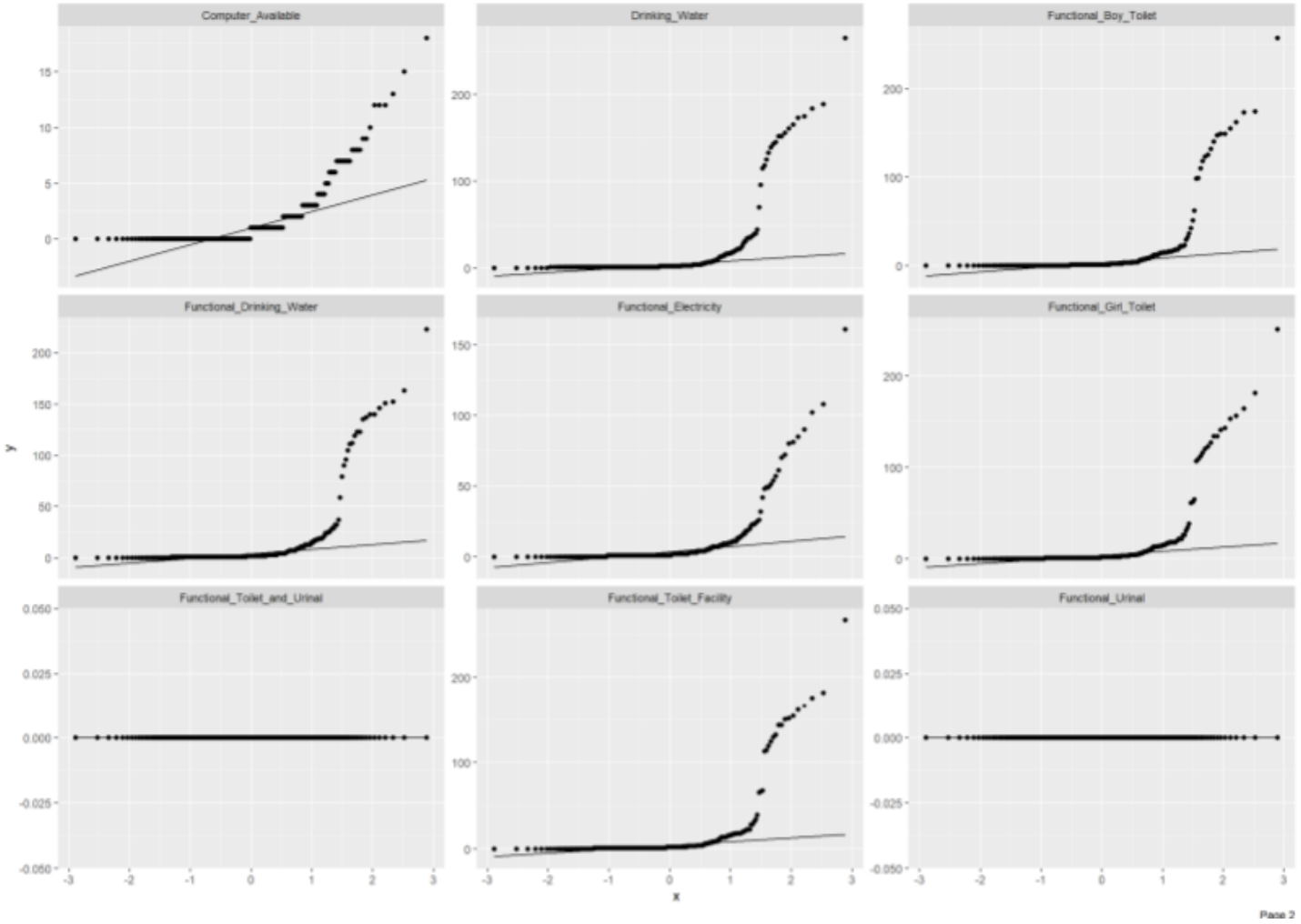
Variables	Min	Q1	Q3	Max	Outliers	Outliers (%)	Position
School_Management_Id	1	1	7.25	98	63	24.2%	Upper
School_Type_Id	1	3	3	3	62	23.8%	Lower
Total_Number_of_Schools	1	1	7.25	272	43	16.5%	Upper
Building	0	1	7.25	267	41	15.8%	Upper
Computer_Available	0	0	2	18	26	10.0%	Upper
Functional_Drinking_Water	0	1	7	223	39	15.0%	Upper
Drinking_Water	0	1	7	265	44	16.9%	Upper
Functional_Electricity	0	1	6	161	33	12.7%	Upper
Functional_Boy_Toilet	0	0	7	257	30	11.5%	Upper
Functional_Girl_Toilet	0	1	7	251	37	14.2%	Upper

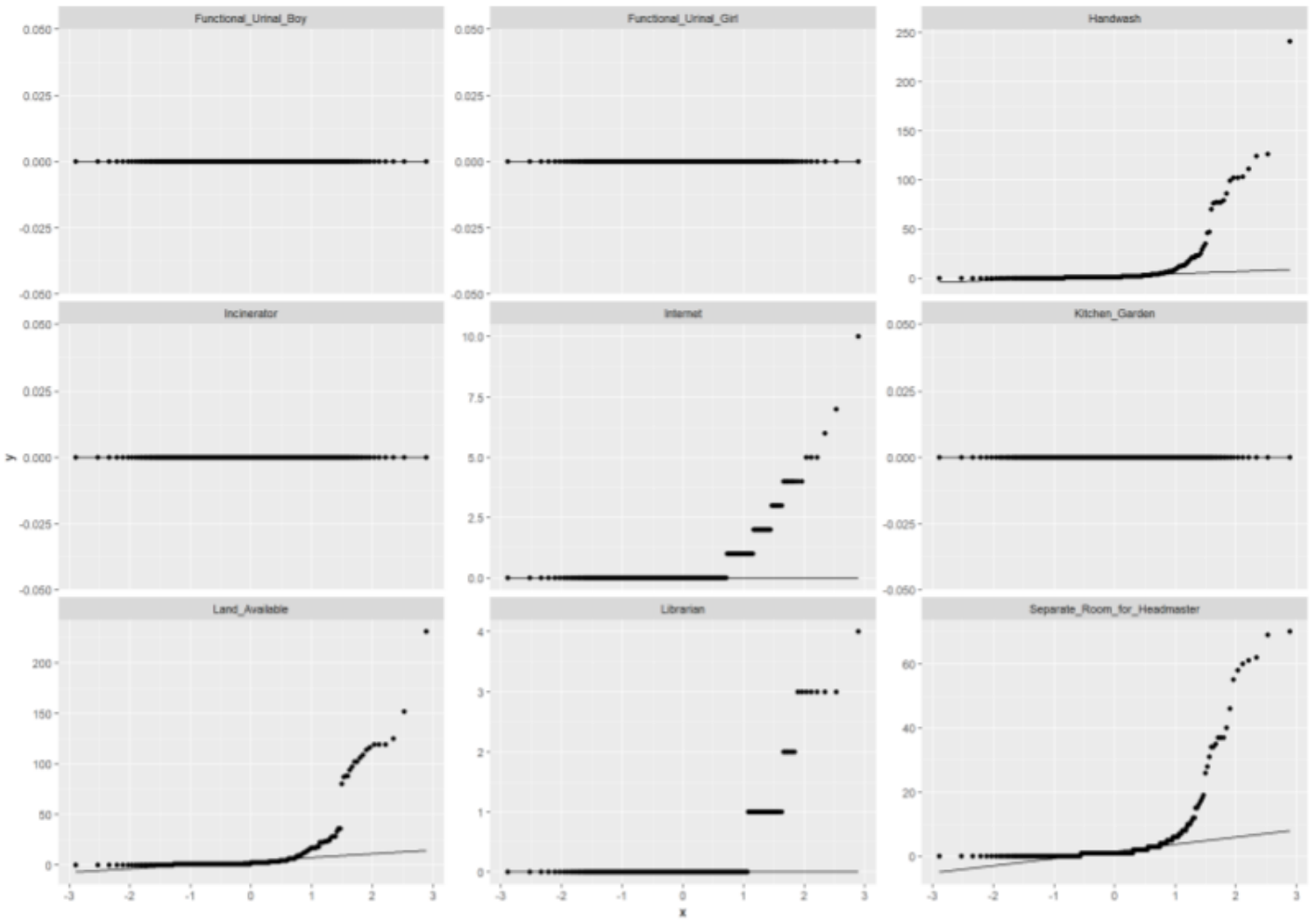
Variables	Min	Q1	Q3	Max	Outliers	Outliers (%)	Position
Rain_Water_Harvesting	0	0	0	1	18	6.9%	Upper
Ramps	0	0	3	202	32	12.3%	Upper
Furniture	0	0	3	107	39	15.0%	Upper

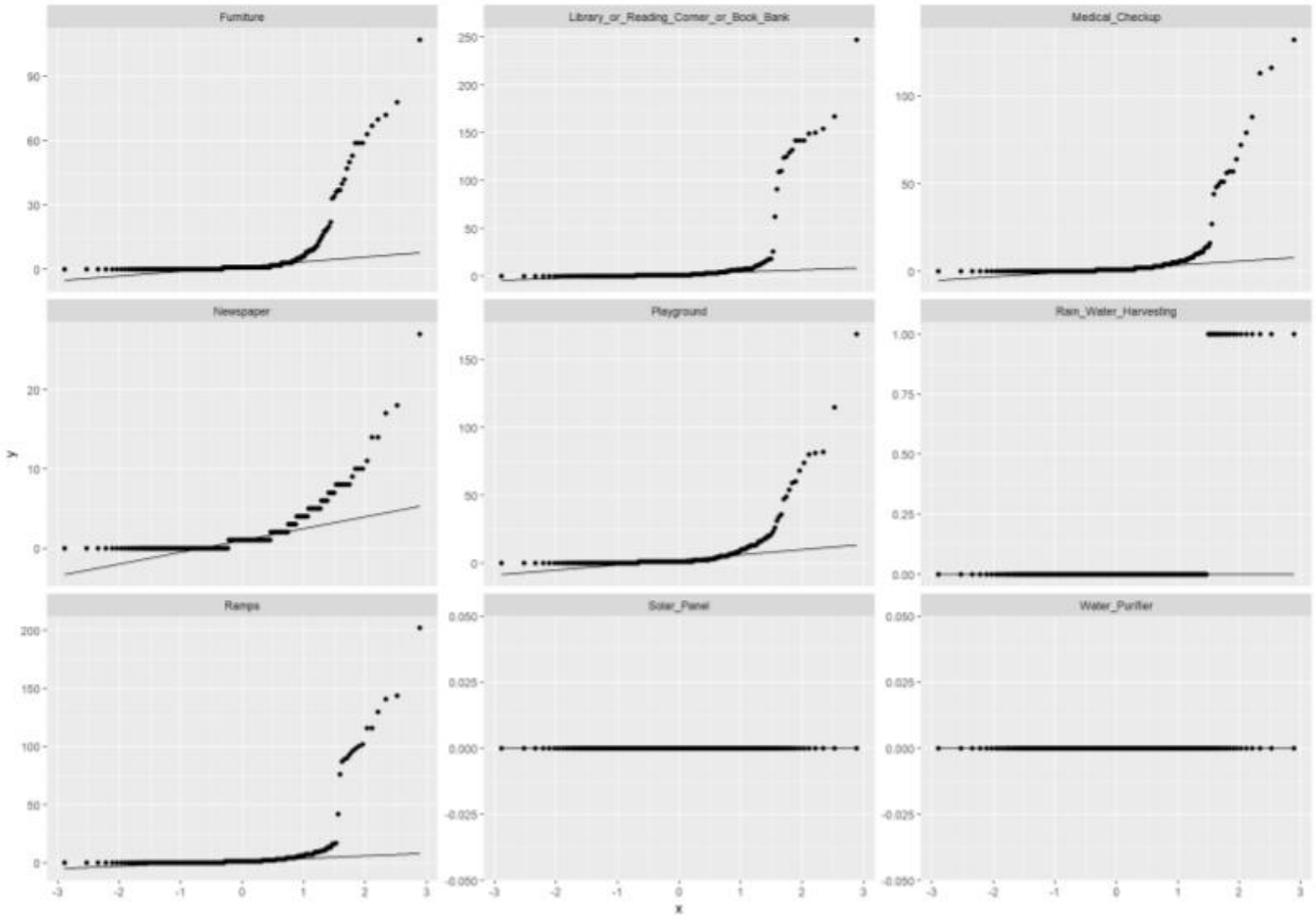
Variables	Min	Q1	Q3	Max	Outliers	Outliers (%)	Position
Functional_Toilet_Facility	0	1	7	266	39	15.0%	Upper
Handwash	0	1	4	241	42	16.2%	Upper
Separate_Room_for_Headmaster	0	0	3	70	35	13.5%	Upper
Internet	0	0	0	10	61	23.5%	Upper
Land_Available	0	1	6	231	47	18.1%	Upper
Librarian	0	0	0	4	37	14.2%	Upper
Library_or_Reading_Corner_or_Book_Bank	0	1	4	247	30	11.5%	Upper
Medical_Checkup	0	0	3	132	30	11.5%	Upper
Newspaper	0	0	2	27	26	10.0%	Upper
Playground	0	0	5	169	32	12.3%	Upper

QQ PLOT

A Q-Q plot is a plot of quantiles of two distributions against each other or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.




Functional_Urinal_Boy
Functional_Urinal_Girl
Handwash
Incinerator
Internet
Kitchen_Garden
Land_Available
Librarian
Separate_Room_for_Headmaster
y
x



Interpretation

We can observe from the Q-Q plots that our data is right-skewed (or positively skewed).