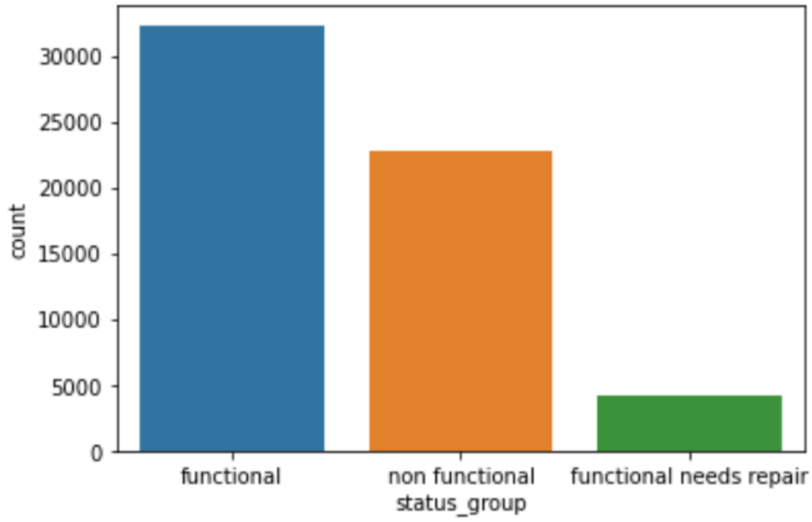
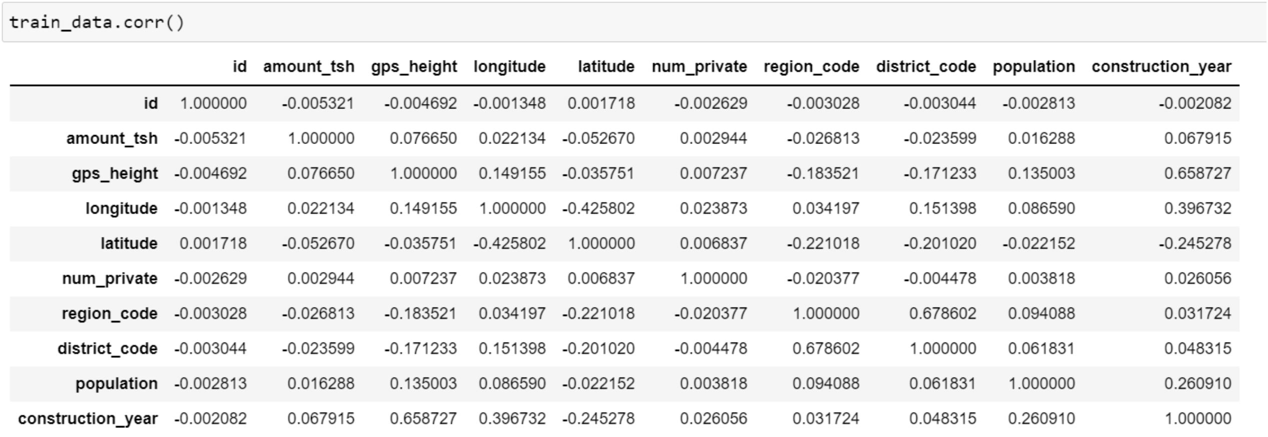
Project_82_phase_02

1. **Data-set level and output-variable analysis:**
   Given dataset have 59400 data points and 40 features. Among those features 10 are numerical features and 30 are categorical features. These data points have to classify into three categories as functional, non-functional and functional needs repair.



From the above histogram, we can conclude that given dataset is highly imbalanced as it has functional:32259, non-functional:22824, functional needs repair: 4317.

For numerical features, using correlation function formed the correlation matrix.



```
train_data.corr()
```

	id	amount_tsh	gps_height	longitude	latitude	num_private	region_code	district_code	population	construction_year
id	1.000000	-0.005321	-0.004692	-0.001348	0.001718	-0.002629	-0.003028	-0.003044	-0.002813	-0.002082
amount_tsh	-0.005321	1.000000	0.076650	0.022134	-0.052670	0.002944	-0.026813	-0.023599	0.016288	0.067915
gps_height	-0.004692	0.076650	1.000000	0.149155	-0.035751	0.007237	-0.183521	-0.171233	0.135003	0.658727
longitude	-0.001348	0.022134	0.149155	1.000000	-0.425802	0.023873	0.034197	0.151398	0.086590	0.396732
latitude	0.001718	-0.052670	-0.035751	-0.425802	1.000000	0.006837	-0.221018	-0.201020	-0.022152	-0.245278
num_private	-0.002629	0.002944	0.007237	0.023873	0.006837	1.000000	-0.020377	-0.004478	0.003818	0.026056
region_code	-0.003028	-0.026813	-0.183521	0.034197	-0.221018	-0.020377	1.000000	0.678602	0.094088	0.031724
district_code	-0.003044	-0.023599	-0.171233	0.151398	-0.201020	-0.004478	0.678602	1.000000	0.061831	0.048315
population	-0.002813	0.016288	0.135003	0.086590	-0.022152	0.003818	0.094088	0.061831	1.000000	0.260910
construction_year	-0.002082	0.067915	0.658727	0.396732	-0.245278	0.026056	0.031724	0.048315	0.260910	1.000000

For numerical features there is no strong correlation between the input variables from the above result. There is a strong correlation of 65.9% between gps_height and construiton_year. But, actually there shouldn't be relation between those features and not considering those correlations for selecting the features. Remaining categorical features has to be converting into numerical features and observe the correlations among them.

Missing values in data:

From the given dataset below are the no.of missing values are observed in Funder,Installer,SubVillage,public_meeting,Scheme_name,Scheme_management,permit columns.

funder	3635
installer	3655
subvillage	371
public_meeting	3334
scheme_management	3877
scheme_name	28166
permit	3056

We can remove the rows of missing values, but, it will cause the loss of data. For categorical features replacing the missing values with frequent label and for numerical features, replacing the missing values with median will be appropriate.

Data pre-processing:

a. Using the mode operation, it is showing clearly ID is unique for all data points. Dropping the ID feature from dataset as it is not useful for classification.

b. Removing the quantity group feature as it is same data as quantity.

c. From the output of train_data['recorded_by'].describe(). It is same for all the datapoints.so,dropping this feature.

d. In Num_private feature 98.75% of data is same and its value is 0. Removing this feature from data.

e. Date recorded feature has no correlation with the classification, as data may be recorded earlier or in delay.

f. Comparing the feature labels in both the features, below features are similar, some of them are mentioned with spelling mistakes, upper and lower case letters. Removing the below features based on the above criteria.

    1. Extraction_type_group
    2. Installer
    3. Payment_type
    4. Quality_group

5. Source_type
6. Waterpoint_type_group
7. Extraction_type
8. region
9. scheme_management
10. extraction_type

g. Region_code and district_code has a correlation of 67.8%, as both of them belong to geographical location, removing the region code feature as it has less no.of sub categories.
h. Waterpoint_type and extraction_type_class has correlation of 65% and same can be checked by comparing the labels in each feature.
i. 88% of data is sub-categorized as user-group. It is biased towards user-group. So, removing management_group feature from dataset
j. 33% of data with amount of water head is 0 and it is functional. Practically without water availability functioning is not possible. So, removing amount_tsh from data set.
k. categorcial featrures wpt_name, basin, lga,ward, sub_village all will represent about the geographical location of the water point. Among all the features basin has the less no.of sub categories. So, dropping the remaining features from the dataset

Removing Outliers:

Outliers can be detected by measuring the interquartile range and calculating the lower limit and upper limit for removing the outliers.

Data points lies in the range (Q1-1.5*IQR, Q3+1.5*IQR) will be considered for classification and remaining will be removed. From the numerical features population has only the outliers. But, in practical population will be very high in some region and less in other regions. Removing outliers will cause loss of data in other features as well. So, not removing the outliers

Univariate Analysis:

In Univariate analysis each feature is analyzed one at a time. Below are the uni-variate analysis methods.

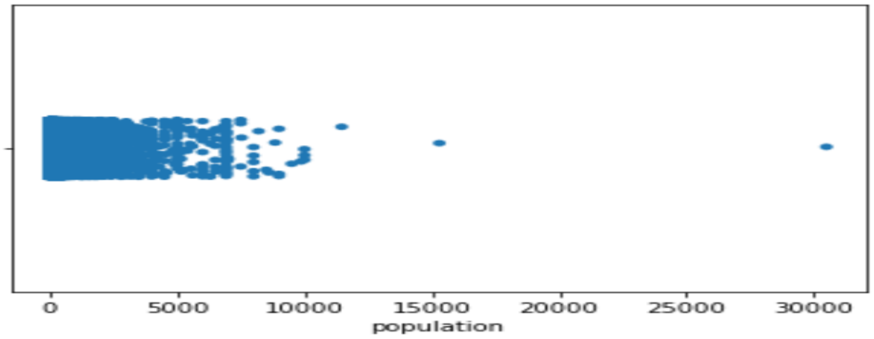
a. 1D scatter plots
b. Histogram
c. .PDF

d. CDF

e. Boxplot

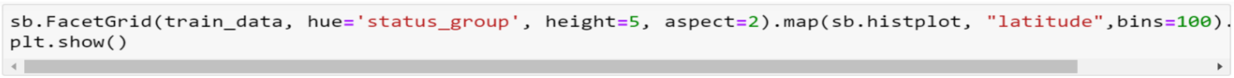
f. Violin plot

a. 1D scatter plots:

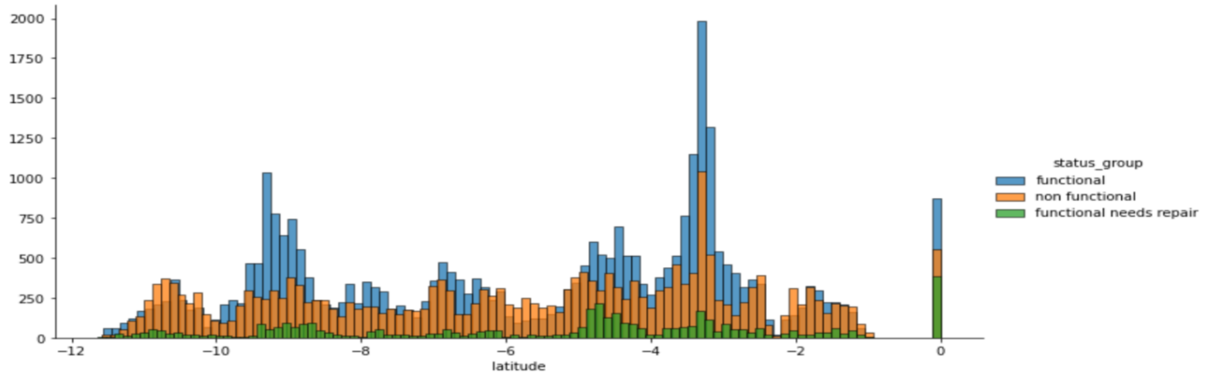


Scatter plots can able to represent the large quantity of data. In the above scatter plot, more data points are overlapped in the range of 0 to 7500 for population feature. We cannot draw any results using 1D scatter plot due to overlapping

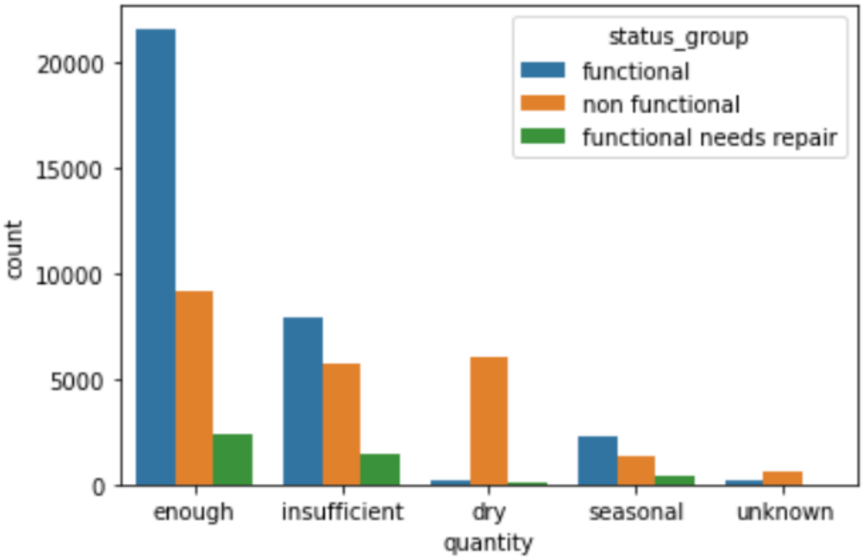
b.Histogram



```
sb.FacetGrid(train_data, hue='status_group', height=5, aspect=2).map(sb.histplot, "latitude",bins=100).
plt.show()
```

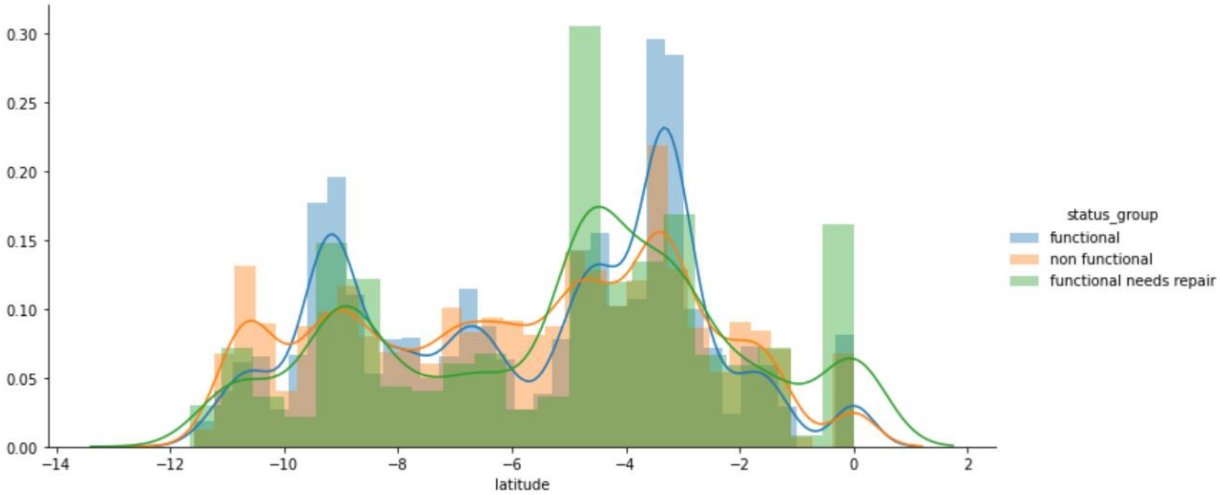


From the above histogram we can visualize that latitude values in between -4 and -2 has more points .Using the histogram, we can able to represent large amount of data in bins. Highest and lowest frequency intervals can be analyzed using histogram. Shape and spread of the continuous data will be visualized. But, we cannot able to visualize the density of the points in intervals and outliers' detection is not possible with histogram.



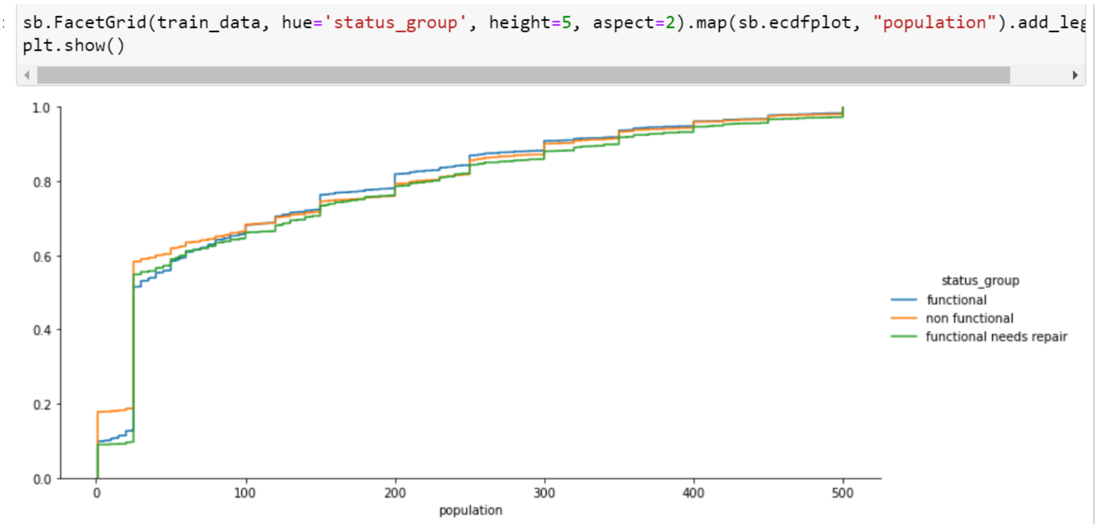
From the above plot we can observe that, water with enough quantity are functioning well and dry quantity water are non-functional.

c.PDF



PDF will give the probability of the points falling in intervals. It is better version of histogram. Using the PDF we cannot able to find out the outliers present in the data.From the above plot, more % of latitude points ,i.e 32% points are lies in the interval of -5 to -4. <1% points are lies in the interval of -24 to -12

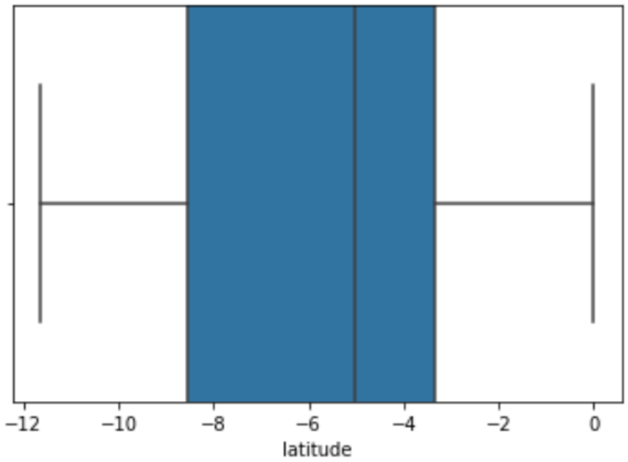
d.CDF



```
sb.FacetGrid(train_data, hue='status_group', height=5, aspect=2).map(sb.ecdfplot, "population").add_leg
plt.show()
```

Using the CDF plot, data points can be visualized along with their range of points falling in particular interval. From the above plot, we can able to see the most % of points interval. But, we cannot able to point out the outliers. From the CDF plot, 60% of the water points are with a population of 40.

e. box plot



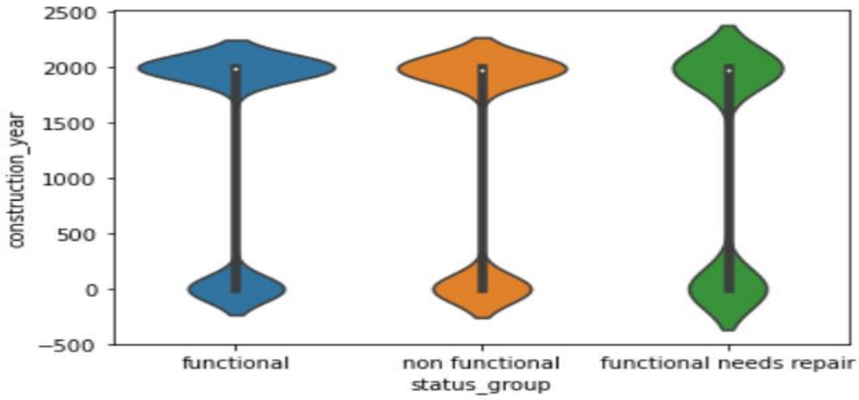
Irrespective of the distribution of the data, box plot will gives the visualization of data into 5 statistical points as minimum point, lower quartile, median, upper quintile and maximum point. Outliers can be detected using box plot by multiplying with 1.5 times of IQR and can be eliminated from the dataset. Box plot cannot able to represent the dataset distribution. From the box plot, min point =1960, max point=2013,IQR is 20 and Q1=1984,Q3=2004

Lower limit for outlier is =1984-1,5*20= 1954

Upper limit for outlier is =2004+1.5*20=2034

From the above boxplot, outliers are not present in the dataset.

f. violin plot

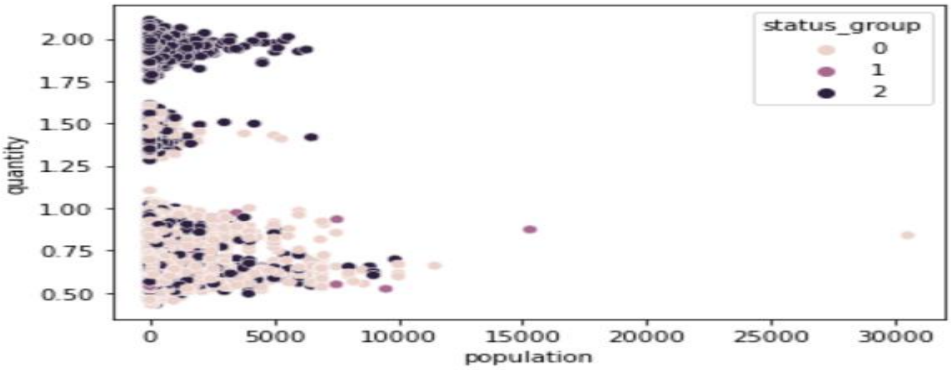


Along with the statistical points and outliers detection, violin plot also provides the distribution of data points in the dataset. From the violin plot, it is a non-Gaussian distribution.

Multivariate Feature analysis:

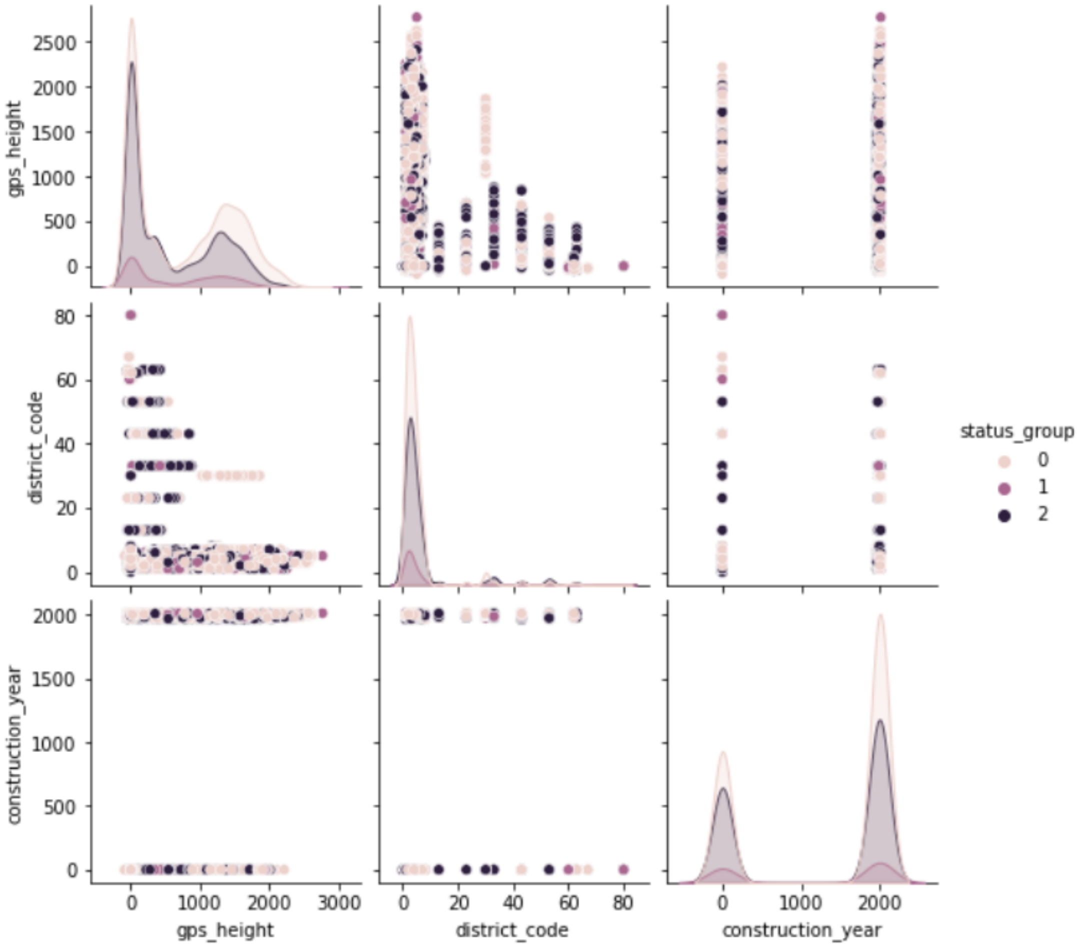
In multivariate feature analysis two or more features will be analyzed at a time.

1. Scatter plot



Data set is large, visualizing the features and feature selection with 2D scatter plot is difficult.

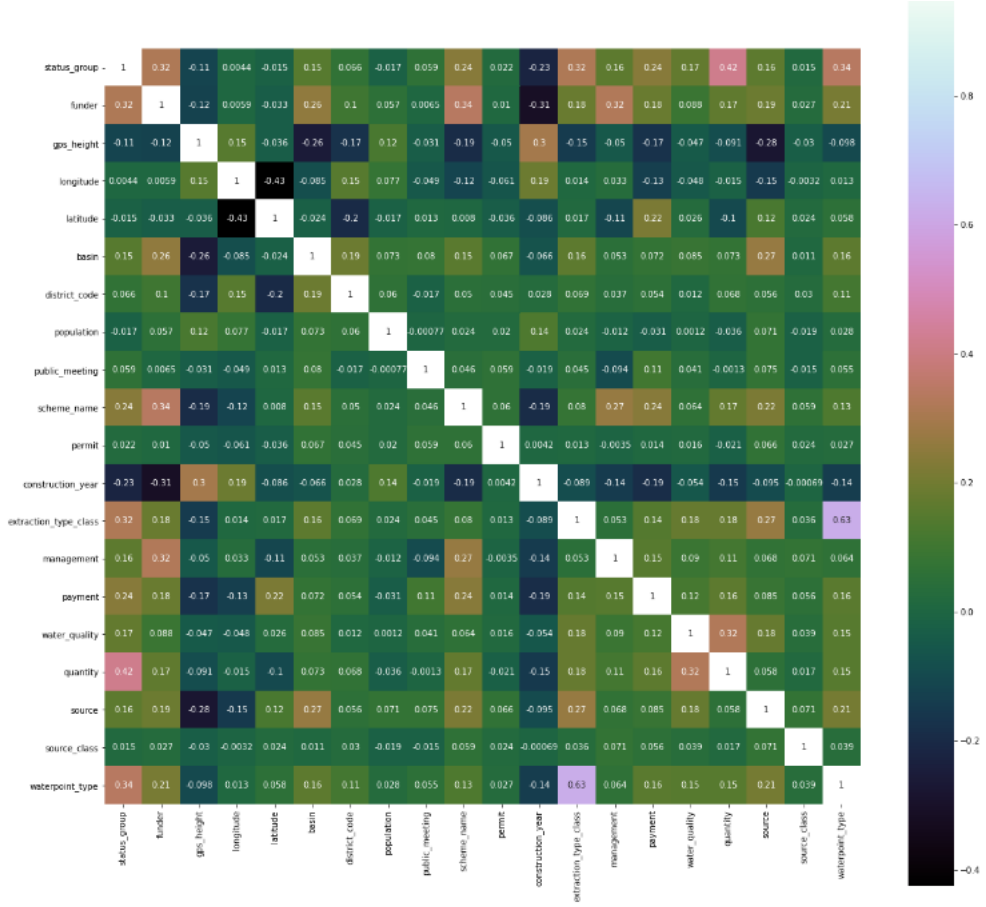
2. Pair plots



From the above pair plot with different numerical features, most of the pair of features is overlapping due to the more data points present in it. Increasing the scale of the plot and visualizing is difficult using pair plot. So, for feature selection pair plot is not useful if the data points are more.

3. Correlation matrix

Before performing the correlation matrix on the features, all the categorical features are converted into numerical features and correlation coefficient between all the pair of features is calculated and its range is from 0 to 1. Correlation coefficient near to 1 between the features has more correlation and suitable feature will be selected accordingly.



Form the correlation heat map there is a correlation between gps_height and construction_year. But, practically there is no relation between them. so, not considering that realtion

Feature encoding: Converting the categorical features to numerical features.

Label Encoding:

Assigning the values to label features based on their rank order. Firstly converting Target labels into numerical for modeling purpose.

1. Status_group

Target encoding:

Replacing a categorical value with the mean of the target variable. It avoids the higher dimensionality and sparsity in the features, which occurs due to one-hot encoding. But, due to target encoding data leakage and over fitting of model will occurs. To avoid this, Gaussian noise will be added to the encoded columns.

Features encoded with Target Encoding are:

1. Permit
2. Public_meeting

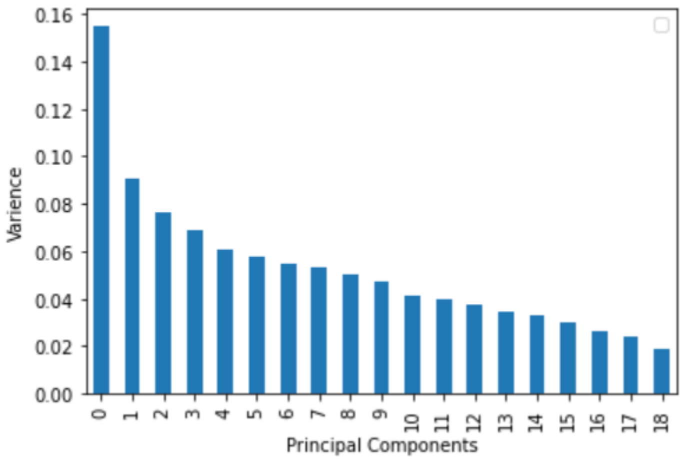
3. Water_quality
4. quantity
5. Waterpoint_type
6. Source
7. Source_class
8. Management
9. Management_group
10. Extraction_type_class
11. Basin
12. funder
13. scheme_name

After removing the features and feature encoding, data set shape changed from (59400, 40) to (59400,19)

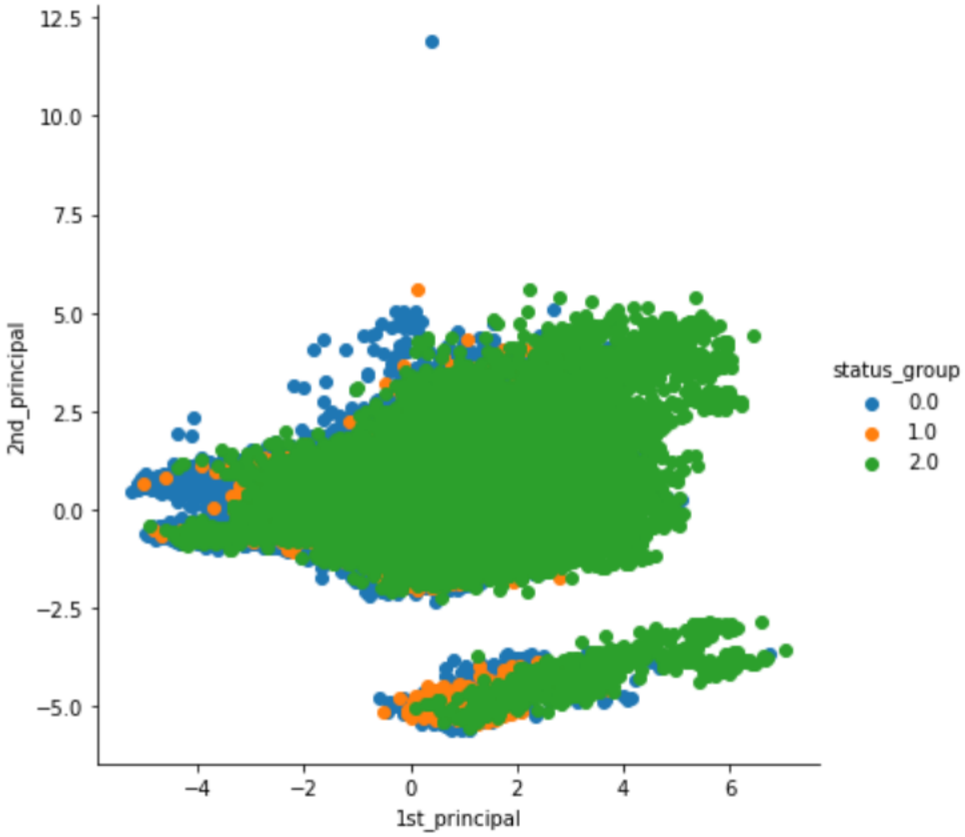
High dimensional data visualization:

1. Principal component analysis:

   Before performing the principal component analysis all the features in the dataset standardized to bring the all features with zero mean and unit standard deviation. Principal components and its variance to be plotted for selecting the no.of principal components holding the maximum variance.



From the above plot principal components vs variance, out of 19 principal components 16 are holding the 91% of variance of the data and some are having less variance of the data. Considering two principal components for visualizing the data



Advantages of PCA:

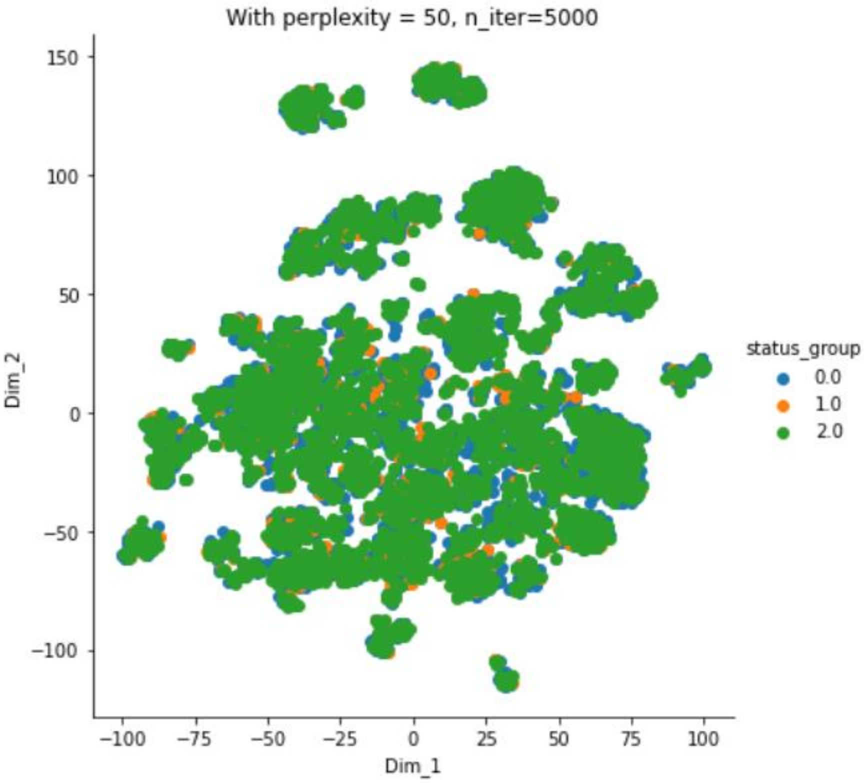
a) it removes the correlated features as all the principal components are independent to each other
b) Significantly improves model performance of high dimensional data by reducing the no.of features
c) It reduces over fitting by reducing the no.of features
d) It improves data visualization by converting from high dimensional to low dimensional using principal components.
e) It preserves global structure

Disadvantages of PCA:

a) principal components are less interpretable
b) information loss due to selecting the principal components with max variance
c) It doesn't preserve local structure

2. T-SNE

For performing high dimensional data visualization using T-SNE, standardization of data is required. It computes pairwise conditional probabilities for each data point.



Advantages of T-SNE:

a) It handles non-linear data efficiently
b) It preserves both local and global structures

Disadvantages of T-SNE:

a) Computationally complex and times complexity is more
b) It is Non-deterministic as different runs with same hyper parameters may produce different results
c) It requires hyper parameter tuning and produces noisy patterns