DIAMOND DATASET - PREPROCESSING

Short description about selected datasest (e.g. number of samples, the domain, main statistics)

```
index
              carat average_us_salary diamonds_mined \
count 52178.000000 51073.000000
                                53940.000000 53940.000000
mean 26979.725804
                   0.797823
                              39521.990100
                                              2.902669
std 15566.154681 0.473747
                             5486.892971
                                           1.325985
min
      1.000000 0.200000
                           30000.000000
                                          0.600000
                   0.400000
25% 13498.250000
                              34780.000000
                                             1.750000
50% 26989.500000
                   0.700000
                              39547.500000
                                             2.910000
75% 40458.750000
                   1.040000 44252.000000
                                             4.050000
max 53940.000000
                   5.010000
                              48999.000000
                                             5.200000
```

depth table price x y \

count 51866.000000 51030.000000 51844.000000 51027.000000 52208.000000

mean 61.750175 57.456332 3933.022047 5.731451 5.734517

std 1.433485 2.231611 3989.013631 1.121433 1.142543

 $\min \quad \ \, 43.000000 \quad \, 43.000000 \quad \, 326.000000 \quad \, 0.000000 \quad \, 0.000000$

25% 61.000000 56.000000 951.000000 4.710000 4.720000

50% 61.800000 57.000000 2401.000000 5.700000 5.710000

75% 62.500000 59.000000 5327.250000 6.540000 6.540000

max 79.000000 95.000000 18823.000000 10.740000 58.900000

Z

count 51532.000000

mean 3.538203

std 0.706057

min 0.000000

25% 2.910000

50% 3.530000

75% 4.030000

max 31.800000

The dataset is diamonds dataset, where I took the price as the target variable. Im trying to predict the price of the diamonds using other features.

It comes under mineralogy, we are encountering both numerical and categorical data

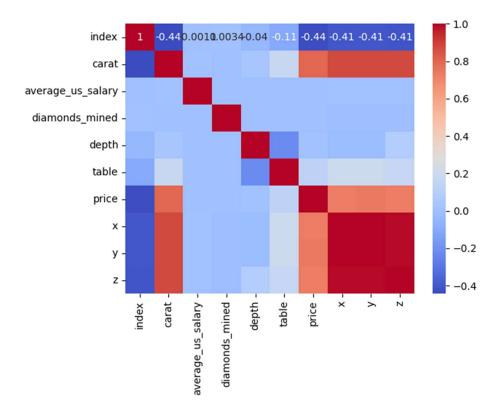
It contains 53940 entries and 13 features (variables)

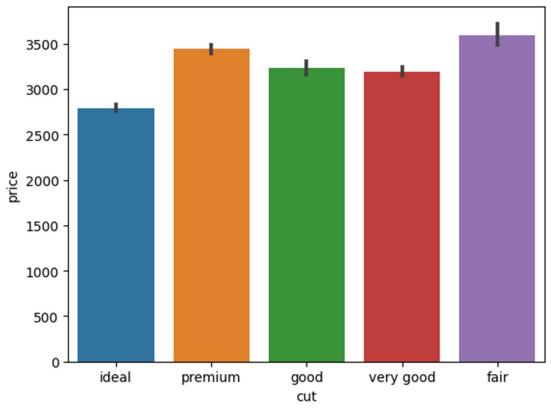
For preprocessing we used

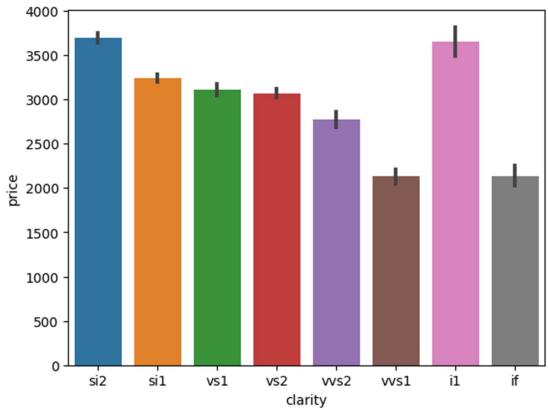
- mode imputaion for categorical data value_counts() and get the index0,
- mean imputation for numerical data
- find_quartile for finding the interquartile range
- corr() to find the correlation matrix
- Categorical().codes for converting the string values to categorical values
- Used fillna() to fill the computed values for mode where the column is empty
- Str.lower() to convert the string to lower case letter to handle the mismatched data
- Also used some basic functions like abs() to find the absolute value, df.mean() to find the mean and df.mode() to find the mode during preprocessing

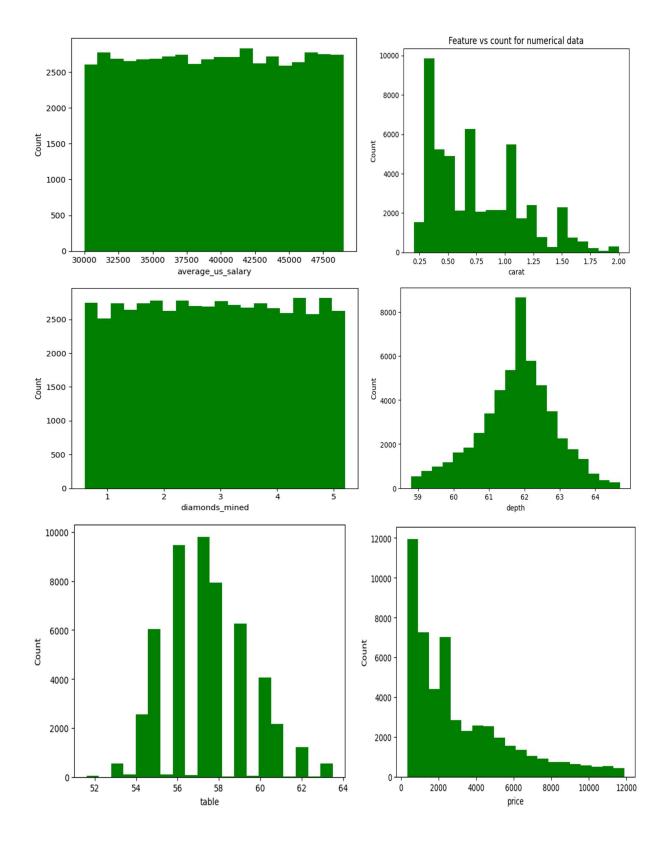
3. Provide graphs and your short description

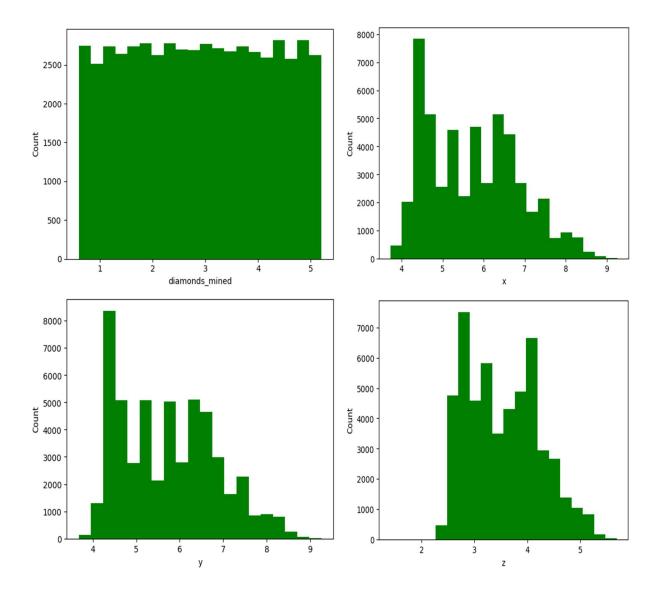
In visualization graphs we could have a prcise understanding on numeric data

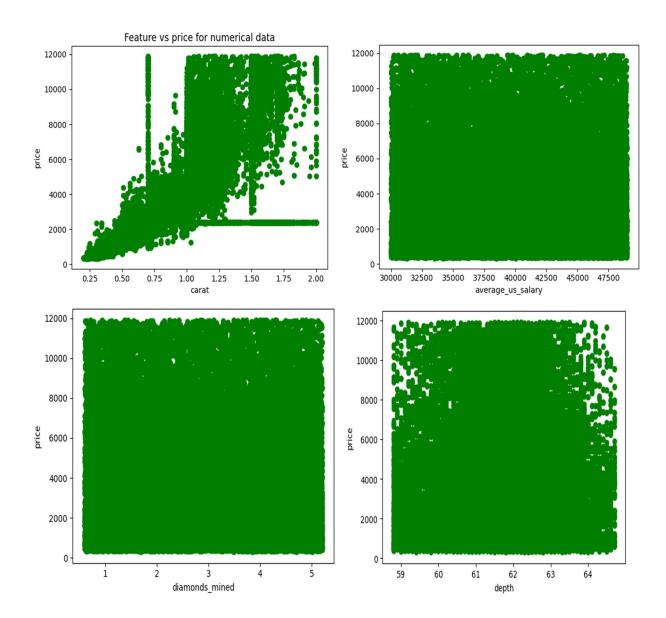


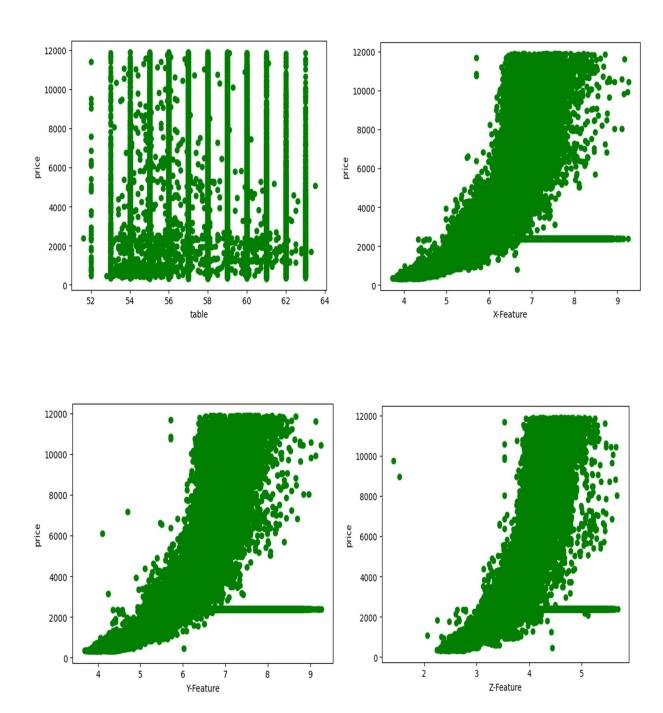




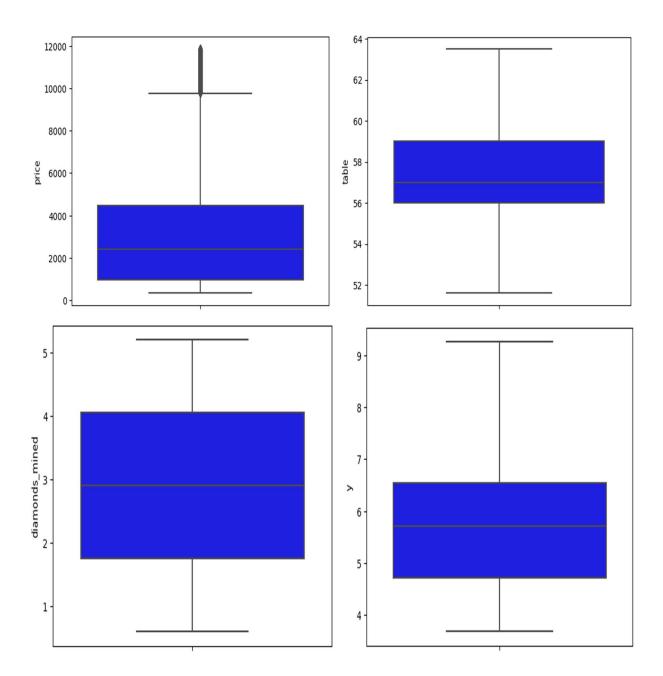


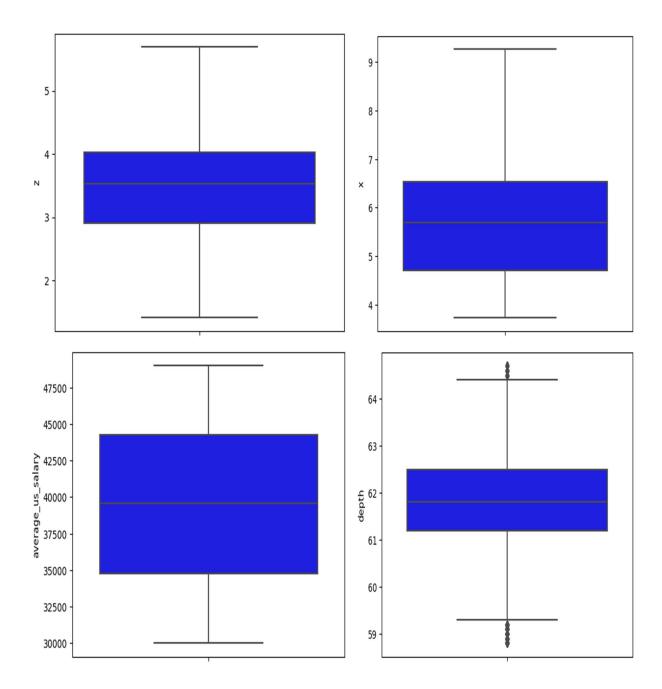






Here we could see that X, Y, Z feature is more related to the price value





INFERENCE FROM GRAPHS:

- Most of the diamonds have X between 4.5 to 6.5
- Most of the diamonds have Y between 4.5 to 6.5
- Most of the diamonds have Z between 3 to 4
- Most of the diamonds have price between 1000 to 4700
- Most of the diamonds have diamonds mined between 1.4 to 4
- Most of the diamonds have depth between 61 to 62.5
- Most of the diamonds have average_us_salary between 3500 to 4500

For categorical data

- Adelie species has the highest calorie requirement
- There are more male penguins than female
- Most of the penguins live in Torgersion island