First we import the data and read it through pandas.

To check what the data is about we use the head() function which gives us the top 5 rows of the data

Next we use the describe function to give us a summary of all the columns of the data. Mean , mode and percentile is all calculated and shown to us in this function.

Next we check for all the null NaN values in the dataset. The isnull() function returns a true/false -0/1 to show if the particular value is null or not. The sum function sums up all the positive values , i.e. all the null values in a particular column. This gives all the null values in the particular column

Next we print all the columns that actually have a null value by printing only those columns whose sum is greater than 0

By printing len(hd) we get to see the total rows are 1460 and thus we decide to drop the 3 columns poolqc , alley and miscfeatures because it is mostly filled with null values. Their individual null values is checked when we print the variable nd.

Since we cannot see all the columns we used the pandas function pd.set\_option("display.max\_columns", None) That helps us view all the columns

We see that miscval and poolarea columns are not useful anymore because we dropped their main columns and since they wont be effective for our calculations we decide to drop them

We once again see the columns with null values and try to do something about them

On reading the description of the data we understand that the null values in the columns related to the garage and fence are all null because they don’t have the respective amenity and thus could be useful for our models in the future so we decide to replace null with no garage and no fence respectively.

Checking to see the data type of column GarageYrBlt , since it comes out to eb float we will leave it for now.

The electrical column seems to only have one missing value. Since there is nothing particular assigned to nan for electrical we can assume there must be some sort of problem during data entry and thus we should drop this row because we cannot fill the value of this specific cell in any other way.

Now on checking the sum of null values in electrical we get 0 .

Next we will be dividing the columns i.e. our set of features into 2 parts , namely categorical features and numerical features. Categorical features will be the ones that are object datatype cause they string values and hence have categories. Numerical features will be the consist of the rest of the columns.

Numerical features can further be divided into discreet and continuous features. We make a condition that all the features that have less than 25 unique values will be discreet and rest will be continuous.

Next we will be going to the graph part of the project.

We have already found out the categorical features and we have those in a list ‘cafeatures’ , we run a loop through all the features in that list and individually make a bar graph with respect to our target feature ‘SalePrice’. We use the groupby feature for categorical features and then take the median to reduce the effect cause by outliers.

We graph the features in the discreet feature list similar to the categorical one because fundamentally they are treated the same , the only difference between them would be that categorical features have string label on the x-axis and discreet have numerical values on the x-axis.

We graph the continuous features using histogram. And make 25 bins as per our convenience, as observed we have a lot of outliers and our graphs are not normalized.

To fix the above problem we use a logarithmic transformation to normalize the data and while plotting the scatter plot ignore all the features that have a 0 that more or less represents null value. We can see corelation of features with respect to the sale price.

We use a box plot in a similar code as above to look out for outliers.