Slide-3

The following libraries were used to perform the data analysis. These include some libraries to plot graphs, some to transform data, and some libraries to run algorithms on the dataset.

Slide-4

- There are 4 files provided to use in the dataset. (Read the file descriptions from the slide).
- I used the “read_csv” function to read the data in the train.csv and test.csv files and stored them as dataframe in the notebook.
- Next, I used the .info and .head methods to basically get information about the datasets and the columns in those datasets. We displayed the data type of the columns and the values that they had etc.
- There are 81 columns in train data and 80 columns in test data.

Slide-5

- **Filter #1:** We created a filter logic that will cater the columns with null values. In this filter if there are null values column wise and if some columns have more number of null values than the initial data then we will choose a ratio and remove those columns that have null values to total dataset ratio more than a assumed ratio. For example, if we decide that our ratio will be 0.5 and we have 1000 records and out of those 600 records have null values. Therefore, we can drop that column.
- **Filter #2:** We created a filter logic that will cater the null values. In this filter we made a function to replace null values. We followed the following rules:
    - If the column has a data type of int or float then we will replace null values with the mean of that column.
    - If the column a data type of object then we will replace null values with the most frequently used object, mode, in that column.
- After investigating the training and testing dataset, it was found that some attributes in the training dataset have 3 classes while some attributes in the testing dataset had more or less than 3 classes. This inconsistency would later lead to inaccurate classifications when we will separate the classes into one hot encoders. We created a function that will be used to make one hot encoded format of all categorical variables in our dataset.
- Finally, the categorical values are in int form, therefore, we will be scaling the data using StandardScaler from sklearn.preprocessing. It will convert our data between 1 and -1.

Slide-6

- I used correlation matrix to see how much a attribute is correlated with the SalePrice (which is to be predicted).
- **Correlation coefficients** are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. There's is a limitation in using correlation coefficients as they are only for continuous values and not for classes so we actually created one hot encoder form which converted it into continuous form so that we can apply Correlation coefficients on it.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

  - **1** indicates a strong positive relationship.
  - **-1** indicates a strong negative relationship.
  - A result of **zero** indicates no relationship at all.

- Any attribute with a coefficient greater than **0.15** and smaller than **-0.15** is picked and used. I had to test different values before concluding onto this. Starting from a value of **0.6**, this value was narrowed down to **0.15** because it gave better results on this value. The columns that are most correlated with SalePrice are chosen and stored for later use.

Slide-7

- Choosing the ML Regression model was a tiresome process. I had to try four different regression models to basically find the best one that gave the least error score.
  - Artificial Neural Networks gave an error score of **0.18528.** After playing with parameters i.e 760 epochs, I concluded Ann with score of **0.18528** which was quite good.
  - I used Gradient Boosting Algorithm next but and got a score of **0.16394** but after changing parameters the score remained almost same.
  - Next I used Random Forest Regressor and i achieved a new score of **0.16282** which was slight improvement.
  - Next I tried XGBoost Regression and I achieved score of **0.14847** with 500 estimators and it was a great leap from Random Forest Regressor.

Slide-8

We will need to process our results by formulating the results in a csv file. We will have the house Id's and the predicted SalePrice side by side in a csv file.

Slide-9

Strengths

1. The final algorithms and analysis done on the dataset was based on four different approaches which gave a good understanding of the error scores and which algorithm predicted the sale prices most accurately.
2. Extensive data cleaning and preprocessing was done to make sure that the data is cleaned and does not have any discrepancies. Such as removing null values, one-hot encoding etc.

Weaknesses

1. Check for more features and attributes that might cause discrepancies.