Walpole philip

House price prediction-usa

TABLE OF CONTENTS

- INTRODUCTION
- DATA IMPORTATION
- DATA PREPROCESSING
- DATA VISUALIZATION

INTRODUCTION

The problem was to find predictions of house prices

With the dataset of USA gotten from kaggle.com I was able to solve the problem at hand

DATA IMPORTATION

Example

```
[ ]  import numpy as np
     import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     %matplotlib inline
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LinearRegression
     import datetime

[ ]  data=pd.read_csv('data.csv')
```

First we import the necessary tools needed for the model
Then we use the pd.read_csv()function to read the data set from kaggle.com after downloading

PRE-PROCESSING

Pre-processing routines **prepare the data for analysis**. Before we start the actual processing, the data has to be pre-processed to remove the detector effects. This is where you check for missing values,duplicated values and outliers.

EXAMPLE

```
[ ]  data.isnull().sum()
```

```
date             0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
sqft_above       0
sqft_basement    0
yr_built         0
yr_renovated     0
street           0
city             0
statezip         0
country          0
dtype: int64
```

```
[ ]  data.duplicated().sum()
```

```
0
```

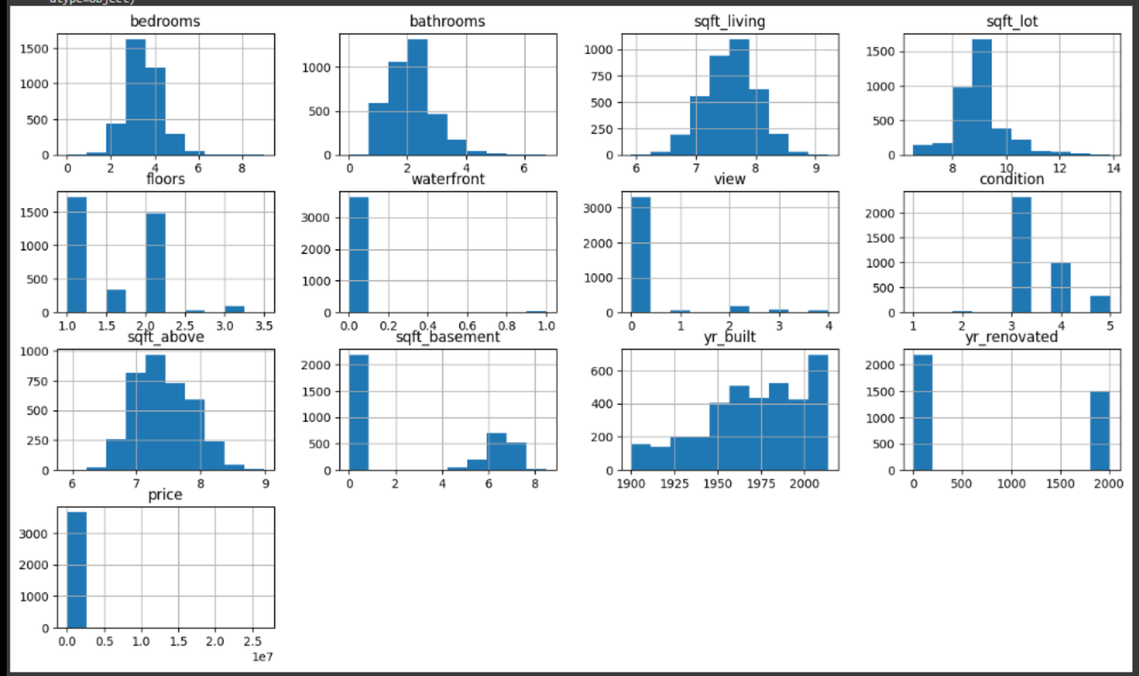
VISUALIZATION

Data visualization is **the representation of data through use of common graphics, such as charts, plots, infographics, and even animations.**
In the model there is the use of histograms ,heatmaps,scatterplots and distplot

HISTOGRAM

```
train_data.hist(figsize=(16, 9))
```

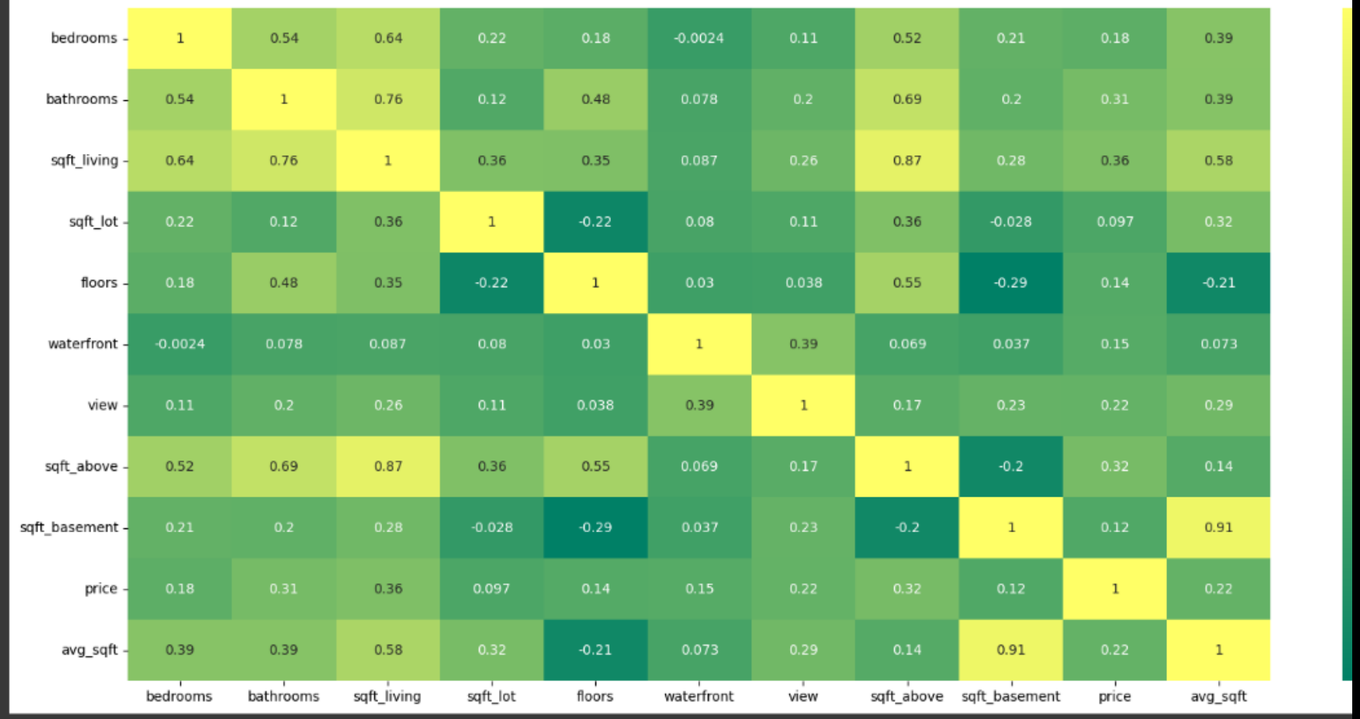
```
       <Axes: title={'center': 'sqft_living'}>,
       <Axes: title={'center': 'sqft_lot'}>],
      [<Axes: title={'center': 'floors'}>,
       <Axes: title={'center': 'waterfront'}>,
       <Axes: title={'center': 'view'}>,
       <Axes: title={'center': 'condition'}>],
      [<Axes: title={'center': 'sqft_above'}>,
       <Axes: title={'center': 'sqft_basement'}>,
       <Axes: title={'center': 'yr_built'}>,
       <Axes: title={'center': 'yr_renovated'}>],
      [<Axes: title={'center': 'price'}>, <Axes: >, <Axes: >, <Axes: >]],
     dtype=object)
```



HEATMAP

```
] plt.figure(figsize=(19,9))
  sns.heatmap(train_data.corr(),annot=True,cmap="summer")
```

```
<ipython-input-25-394d4c8fab76>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only vali
  sns.heatmap(train_data.corr(),annot=True,cmap="summer")
<Axes: >
```




```
] train_data
```

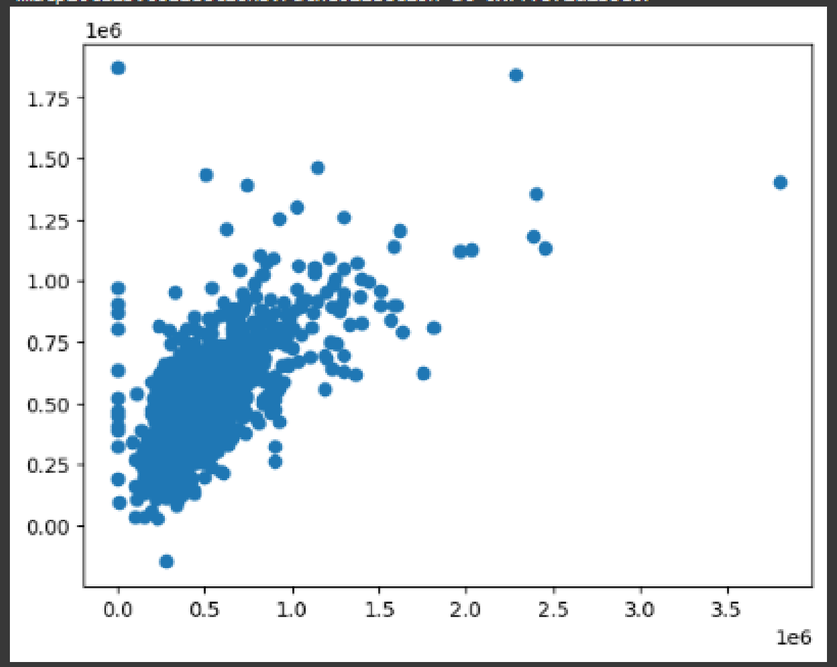
Purpose of Heatmaps

- to show correlation between variables

SCATTERPLOT

```
[ ]  plt.scatter(y_test,prediction)
```

```
<matplotlib.collections.PathCollection at 0x7f7572a2b5e0>
```



Scatterplots are used to show how different variables relate with each other

This scatterplot specifically was used to show the relation between all other variables to the price

CONCLUSION

- The model was a success though it's accuracy was quite low giving 40% in accuracy.
- It is not advisable to use it in prediction


THANK YOU