Bansilal Ramnath Agarwal Charitable Trust's

Vishwakarma Institute of Information Technology
Pune – 411048

Preliminary Project Report
On

Salary Classification and Prediction

Submitted By

Ajay Patil	323075 - U1510498
Abhishek Supsande	323076 - U1410418
Mayur Thakane	323078 - U1610217
Sanket More	323077- U1510580

Department of Computer Engineering
SavitriBai Phule Pune University, Pune
Year 2019-2020

CERTIFICATE

This is to certify that the following students of T.Y. Computer, Vishwakarma Institute of Information Technology, Pune

Ajay Patil	323075- U1510498
Abhishek Supsande	323076- U1410418
Mayur Thakane	323078 -U1610217
Sanket More	323077- U1510580

have successfully completed the project report on

House Price Prediction

in the fulfillment of the requirements for the project completion of T.Y. Computer in academic year 2019-2020.

Project Guide: Manish Mali

HOD:

Table of Contents

1. Title of the Mini Project
2. Objectives
3. Details of Data Set used
4. Steps performed while preprocessing of data.
5. Classifiers used for data modelling with its theory
6. Coding
7. Results
8. Observations/Inferences

1. **Title:** House Price Prediction Using Machine Learning

2. **Objectives:**

- Our main aim today is to make a model which can give us a good prediction on the price of the house based on other variables. We are going to work on a dataset which consists information about the location of the house, price and other aspects such as square feet etc. When we work on these sort of data, we need to see which column is important for us and which is not. We are going to use Linear Regression for this dataset and see if it gives us a good accuracy or not.

  - **Programming Languages**
  - DATA SCIENCE. THERE'S BATTLE OUT THERE HAPPENING IN THE
  - DATA SCIENCE TOOL.THE LATTER (MUCH DUE TO LIBRARIES
  - APPRECIATE IF HE WERE TO IMPLEMENT A MACHINE LEARNING

IMPORTANT FACTORS, FORTHAT PYTHON PREFERRED OVER
1) EASY TO LEARN
2) SCALABILITY
3) CHOICE OF DATA
SCIENCE LIBRARIES
4) PYTHON COMMUNIT

3. Details of Data Set: Dataset Specification

- The dataset has been taken from Kaggle

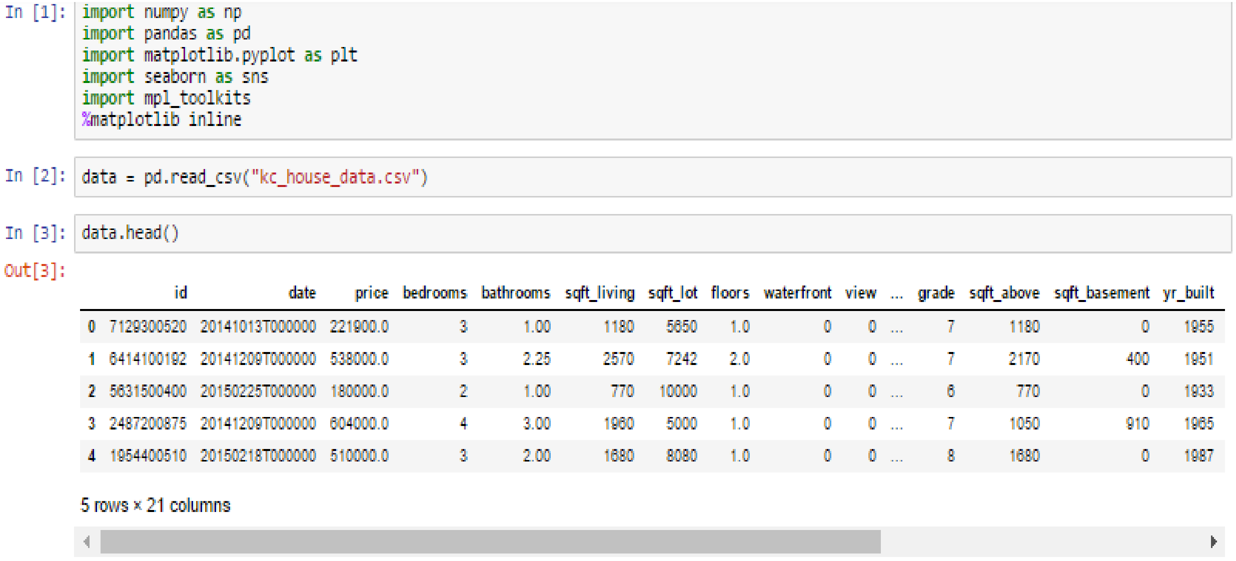
- This Dataset includes 21614 rows and 21 columns of information.
- The house data excel file contains our complete dataset.
- This data is collected through surveys and by estimation of analysts. We have total 21613 historical data prices for houses in USA.
- The dataset is a Structured dataset

.

- The dataset is uniform and consistent and the data have ratio scale of measurement.

4. Preprocessing of data: Data Pre-Processing Structured

- Load all the relevant python libraries. Load the excel file kc_house_data. There were some null values in that dataset that have been removed, this is an important step so that regression model can be run successfully. After that we observe that id and date column does not play a key role in price determination, so we need to drop these two columns. Now our data is clean and ready to be processed under Linear Regression Model in Python.
- We are going to see some visualization and also going to see how and what can we infer from visualization from the data set. First thing first, we import our libraries and dataset and then we see the head of the data to know how the data looks like and use describe function to see the percentile's and other key statistics.



```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import mpl_toolkits
        %matplotlib inline
```

```
In [2]: data = pd.read_csv("kc_house_data.csv")
```

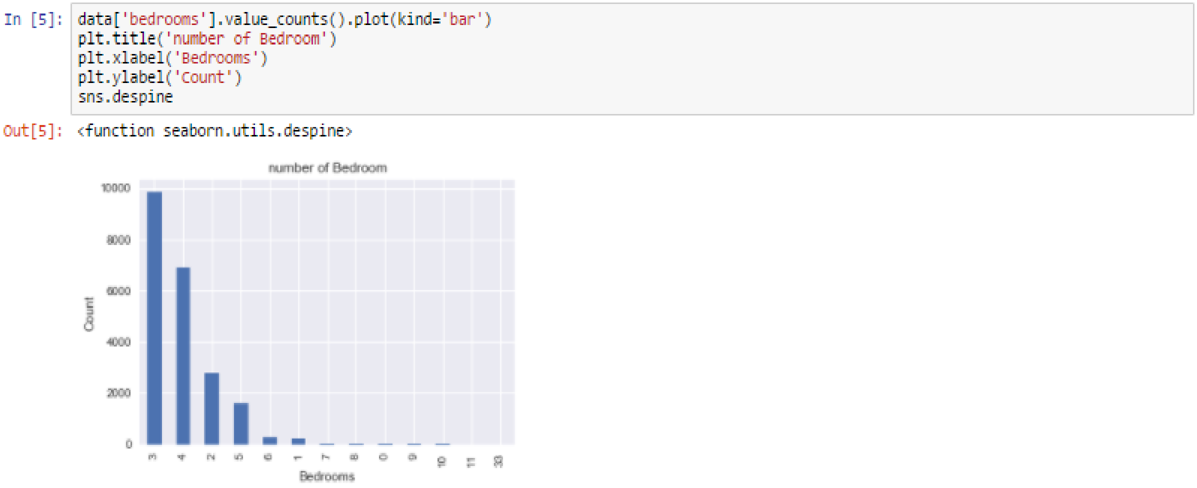
```
In [3]: data.head()
```

Out[3]:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987

5 rows × 21 columns

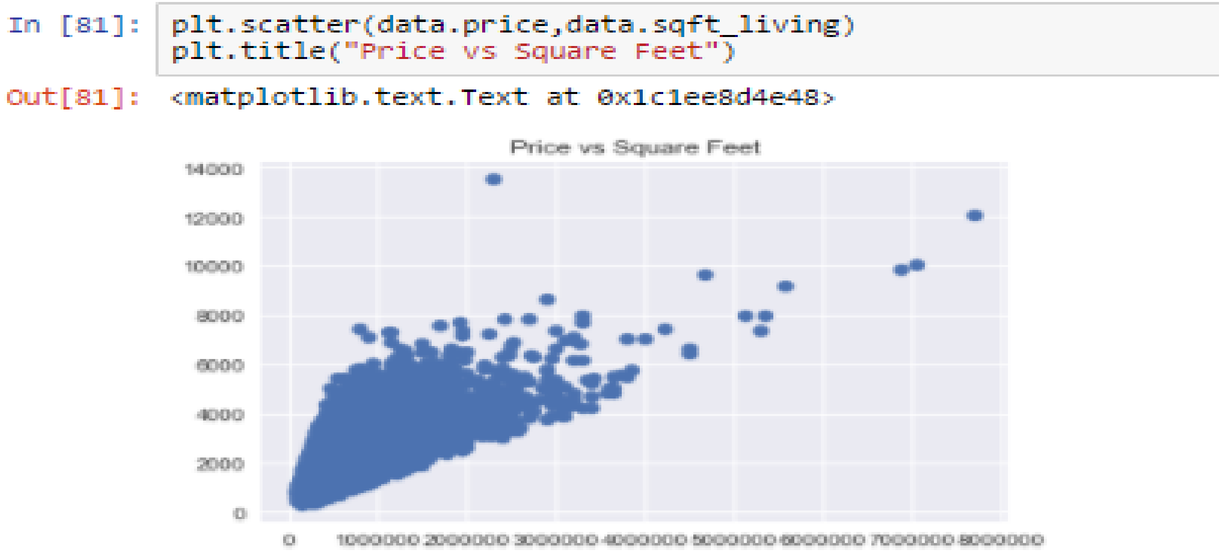
- Histogram are useful of statistics plotting type of bar plot that shoes the frequency or no of values compared to set of value in Python plotting package This Histogram is as per the Number of Bedroom wise. And count of it in each house. Using the libraries of **plt.title**. With x-axis and y-axis. It Plot a Histogram Graph with particular data provided.



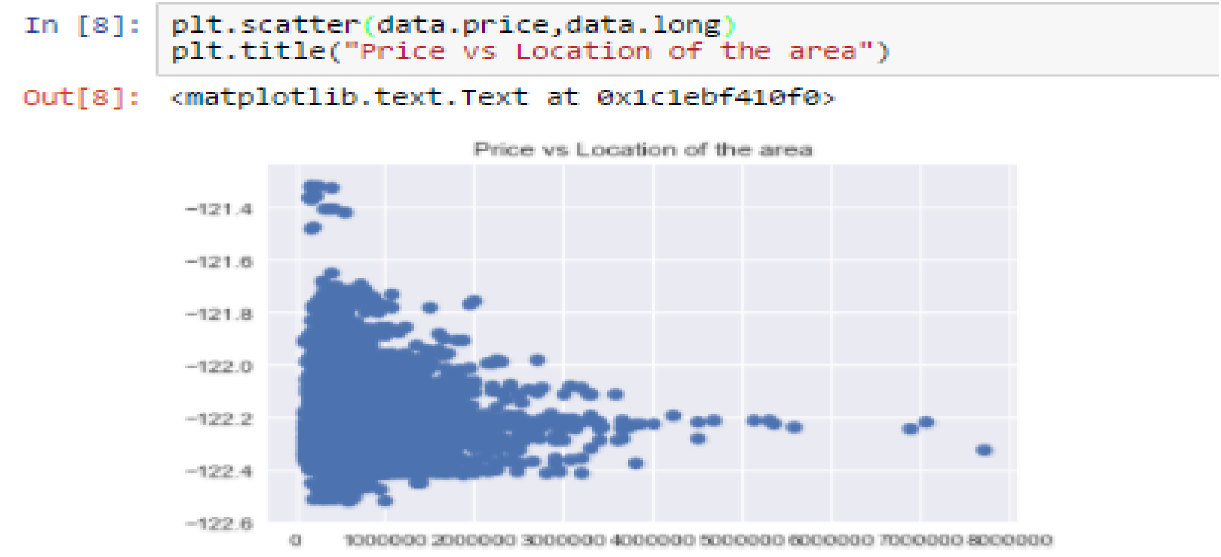
```
In [5]: data['bedrooms'].value_counts().plot(kind='bar')
        plt.title('number of Bedroom')
        plt.xlabel('Bedrooms')
        plt.ylabel('Count')
        sns.despine
```

```
Out[5]: <function seaborn.utils.despine>
```

- The plot that we used above is called scatter plot, scatter plot helps us to see how our data points are scattered and are usually used for two variables. From the first figure we can see that more the living area, more the price though data is concentrated towards a particular price zone, but from the figure we can see that the data points seem to be in linear direction we can also see some irregularities that the house with the highest square feet was sold for very less, maybe there is another factor or probably the data must be wrong. The second figure tells us about the location of the houses in terms of longitude and it gives us quite an interesting observation that -122.2 to -122.4 sells houses at much higher amount.



```
In [81]: plt.scatter(data.price,data.sqft_living)
         plt.title("Price vs Square Feet")

Out[81]: <matplotlib.text.Text at 0x1c1ee8d4e48>
```



```
In [8]: plt.scatter(data.price,data.long)
        plt.title("Price vs Location of the area")

Out[8]: <matplotlib.text.Text at 0x1c1ebf410f0>
```

5. **Data modelling:** **Classifiers used for data modelling: Phases of learning.**
    - Analysis Phase – This phase includes a detailed study of the dataset depending upon various parameters and makes various sparse of the dataset.
    - Algorithmic Phase – After analyzing the main work that comes in front is the processing of data under various conditions along with getting results with a very high accuracy, low precision, high recall. These terms will be elaborated further with the various algorithms.

- **Analysis Phase**
- We import our dependencies, for linear regression we use sklearn (built in python library) and import linear regression from it.
- We then initialize Linear Regression to a variable reg.
- Now we know that prices are to be predicted, hence we set labels (output) as price columns and we also convert dates to 1's and 0's so that it doesn't influence our data much. We use 0 for houses which are new that is built after 2014.
- We again import another dependency to split our data into train and test.
- After fitting our data to the model we can check the score of our data i.e., prediction. in this case the prediction is **73%.**
- Simple Regression. and if multiple predictor variable is present then multiple regression.

DATA SET = TRAIN SET + TEST SET

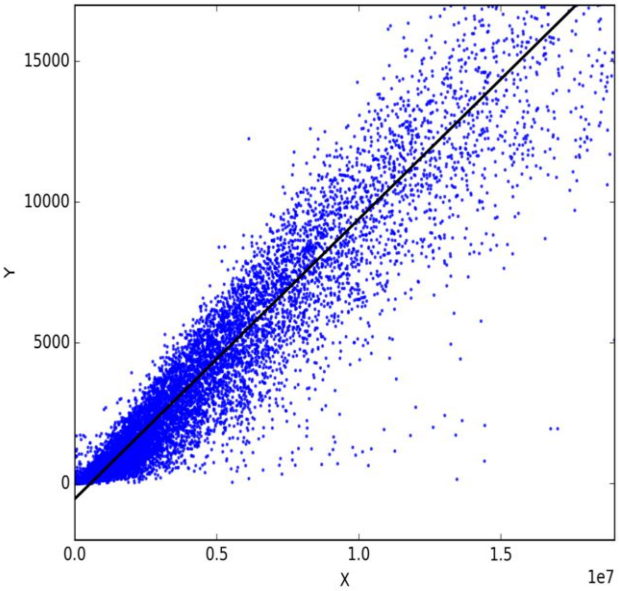
- Train Dataset – The dataset through which we are going to train our machine for finding out the fraud is called as train dataset.
- Test Dataset – The dataset on which we are going to check our algorithm for fraud detection is called as test dataset.
- We train data as 90% and 10% of the data to be my test data, and randomized the splitting of data by using random state. We have trained dat

Algorithmic Phase

- Linear Regression
- Decision Tree
- Random Forest

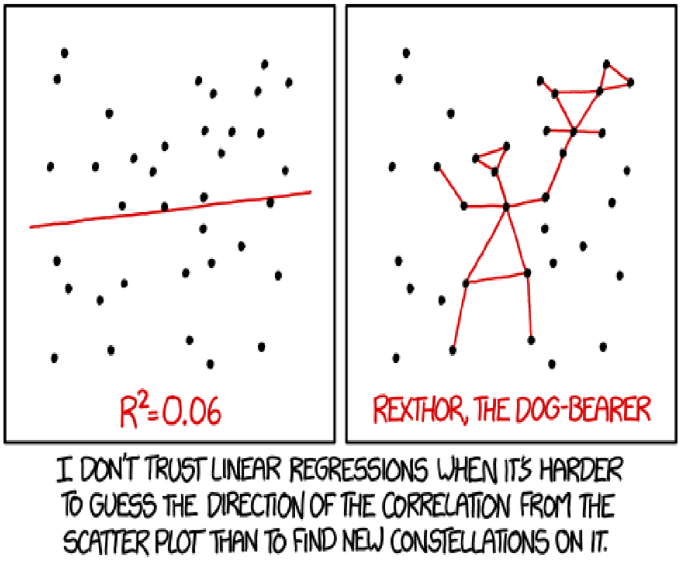
LINEAR REGRESSION

- In easy words a model in statistics which helps us predicts the future based upon past relationship of variables. So when you see your scatter plot being having data points placed linearly you know regression can help you
- Regression works on the line equation, y=mx+c, trend line is set through the data points to predict the outcome.
- The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called **Simple Regression. and if multiple predictor variable is present then multiple regression.**



Random Forest

- Random forest model, implemented in Random Forest Classifier.
- Based on decision tree
- If you input a training dataset with features and labels into a decision tree, it will formulate some set of rules, which will be used to make the predictions.
- Another important hyper parameter is **„max_features “**, which is the maximum number of features Random Forest is allowed to try in an individual tree. Sklearn provides several options, described in their documentation.

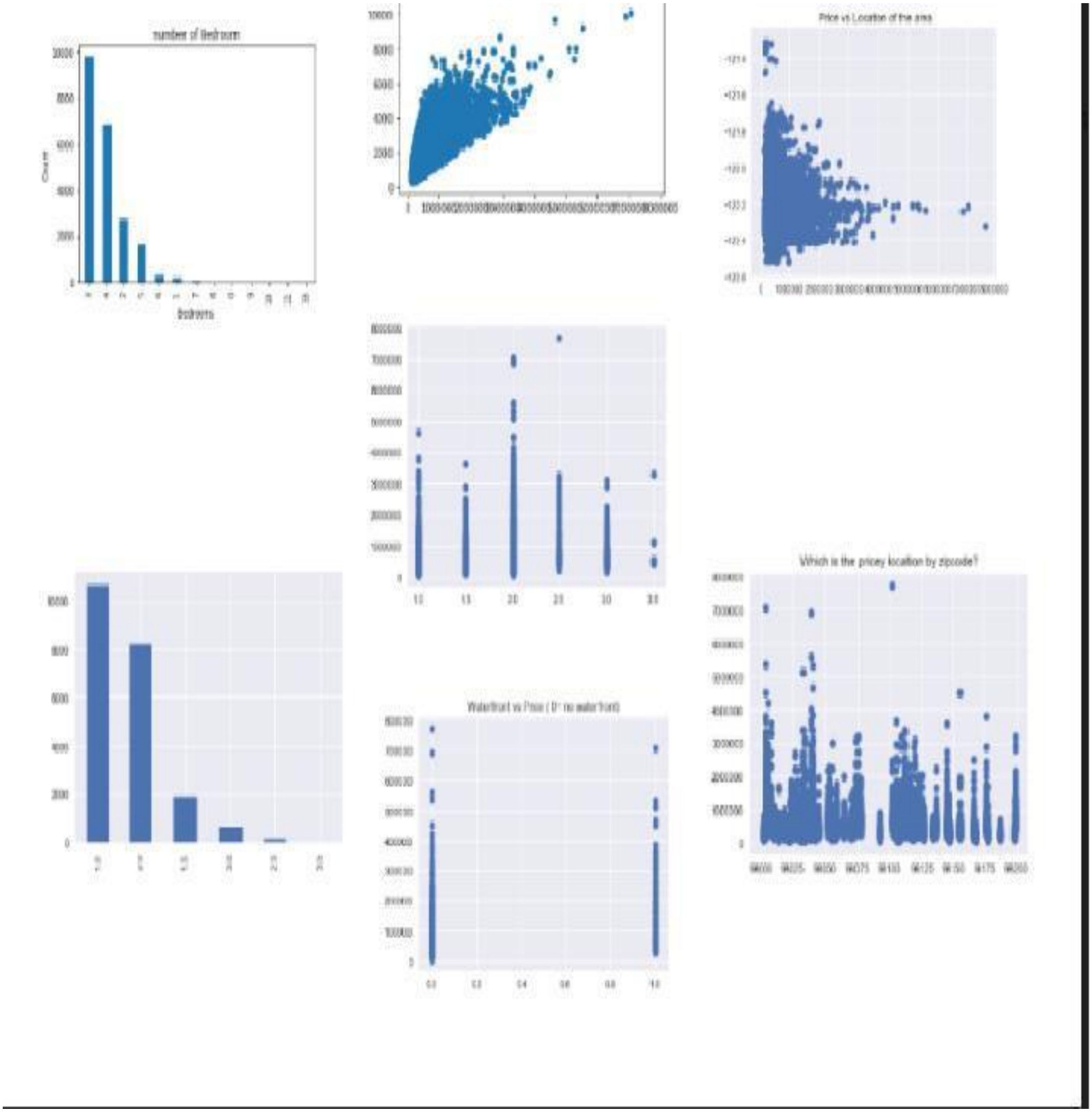

R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

6. **Code:**

```
    # -*- coding: utf-8 -*-
"""housesales.ipynb

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1lRkYMBgne_lcjmTd7FZa-
EsqD0-OOr-g
"""

# Commented out IPython magic to ensure Python compatibility.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import mpl_toolkits
# %matplotlib inline

data = pd.read_csv("kc_house_data.csv")

data.head()

data.describe()

data['bedrooms'].value_counts().plot(kind='bar')
plt.title('number of Bedroom')
plt.xlabel('Bedrooms')
plt.ylabel('Count')
sns.despine

plt.scatter(data.price,data.sqft_living)
plt.title("Price vs Square Feet")

plt.scatter(data.price,data.long)
plt.title("Price vs Location of the area")

plt.scatter(data.bedrooms,data.price)
plt.title("Bedroom and Price ")
plt.xlabel("Bedrooms")
plt.ylabel("Price")
plt.show()
```

```
sns.despine

plt.scatter(data.waterfront,data.price)
plt.title("Waterfront vs Price ( 0= no waterfront)")

train1 = data.drop(['id', 'price'],axis=1)

train1.head()

data.floors.value_counts().plot(kind='bar')

plt.scatter(data.floors,data.price)

plt.scatter(data.zipcode,data.price)
plt.title("Which is the pricey location by zipcode?")
```

7. Result: -Histogram/Scattered Plotting



8. Observations/Inferences:

- An accuracy score of 91.94%. Apply this technique on various other datasets and post your results. Try to put random seeds and check if it changes the accuracy of the data or not! Let me know if it does
- Random Forest provides high accuracy.
- Linear and Decision tree goes on similar path for accuracy
- Throughout this we made a machine learning regression project from end-to-end and we learned and obtained several insights about regression models and how they are developed.