Maharaja Surajmal Institute of Technology, New Delhi

DATA SCIENCE
(USING PYTHON)

"HOUSE SALES IN KING COUNTY, USA"

TRAINING PROJECT REPORT

SUBMITTED BY:

Prapti Singh

41396302818

CANDIDATE'S DECLARATION

We hereby declare that we have undertaken industrial training at "**WEBTEK LABS PVT. LTD.**" during a period from **11th May to 8th June 2020** in partial fulfilment of requirements for the award of degree of B.Tech (Electronics & Communication ENGINEERING) Maharaja Surajmal Institute of Technology, New Delhi. The work which is being presented in the training report submitted is an authentic record of training work.

PRAPTI SINGH _____

Student name Sig. of Student

ACKNOWLEDGEMENT

It gives us great pleasure to acknowledge the guidance, assistance and support of Ms. Mousita Dhar in making the Project and this Project report successful, which has been structured under her valued suggestion.

She has helped us to accomplish the challenging task in a very short period of time.

Finally, we express the constant support of our friends, family and professors for inspiring us throughout and encouraging us.

PRAPTI SINGH

SEMESTR: IV

(ECE)

CERTIFICATE OF APPROVAL

The project **"HOUSE SALES IN KING COUNTY, USA"** made by the efforts of **PRAPTI SINGH** is hereby approved as a creditable study for the **Bachelor of Technology** and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

Ms. Moushita
(Project In charge)

1.INTRODUCTION

1.1 About

Python:

- Python is a high-level, general-purpose, open source, strictly typed programming language. The language provides constructs intended to enable clear programs on both a small and large scale.
- Python was created By Guido van Rossum.
- The Python Software Foundation (PSF) is the organization behind Python.

Python versions:

- First released in 1991.
- Python 2.0 was released on 16 October 2000
- Python 3.0 was released on 3 December 2008

Current Versions:

- 3.6.3
- 2.7.14

Python features:

Some of the features of python include :-

- Easy to understand
- Dynamic
- Object oriented
- Multipurpose
- Strongly typed
- Open Sourced

Python is mainly used in many domains:

- Web Development
- Data Analysis
- Machine Learning
- Internet Of Things
- GUI Development
- Image processing
- Data visualization
- Game Development

IDLE:

IDLE is an integrated development environment for Python, which has been bundled with the default implementation of the language.

1.2 Anaconda

Anaconda is a open source Distribution for data science and machine learning using python. It includes hundreds of popular data science packages and the conda package and virtual environment manager for Windows, Linux, and MacOS.

1.3 Packages

1.3.1 NumPy

NumPy is the fundamental package for scientific computing with

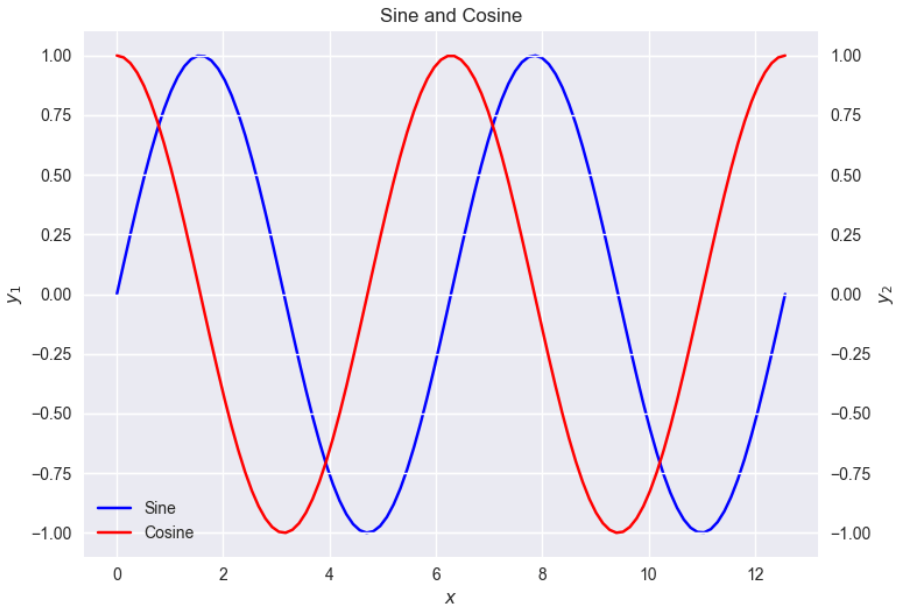
Python. It contains among other things:

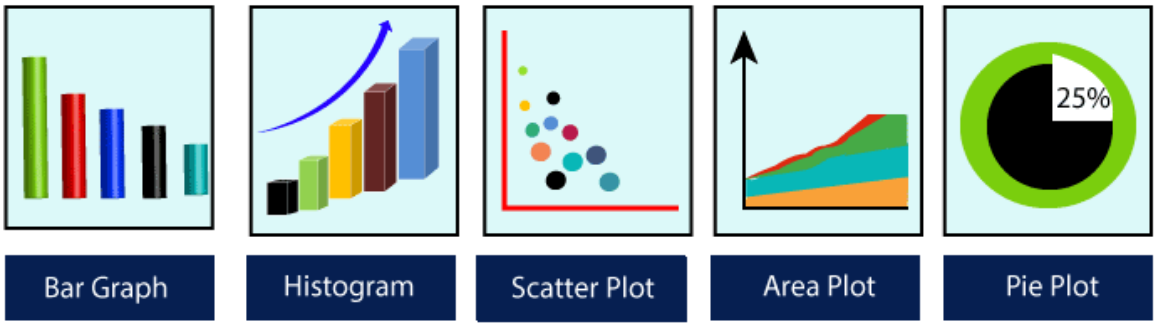
- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

1.3.2 Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.





Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

1.3.3 Scikit-learn

Scikit-learn provides machine learning libraries for python.Some of the features of Scikit-learn includes:

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

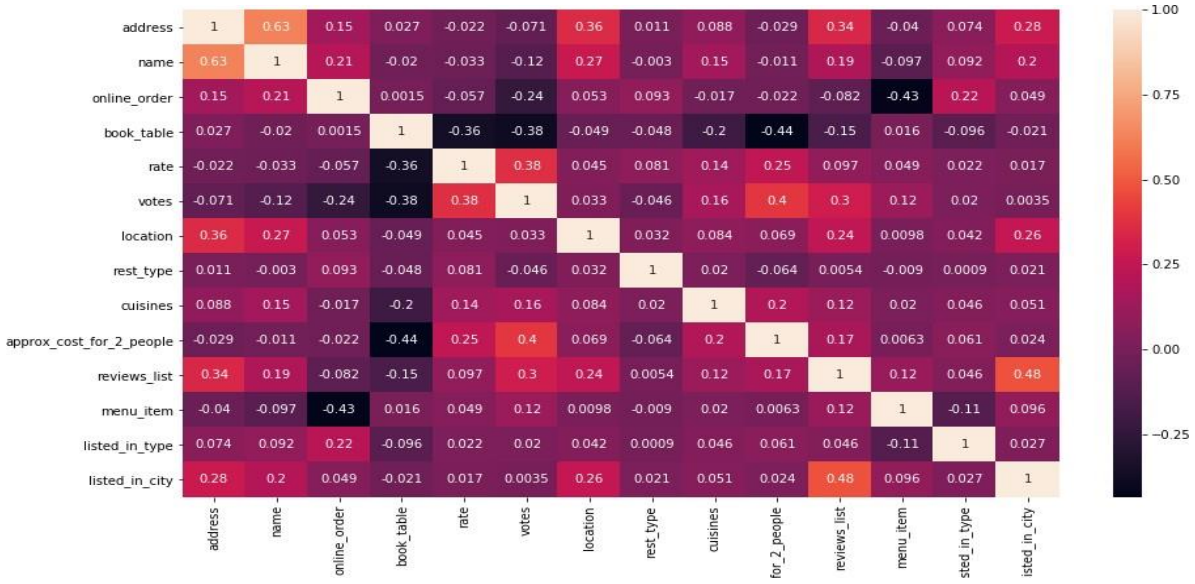
1.3.4 Pandas

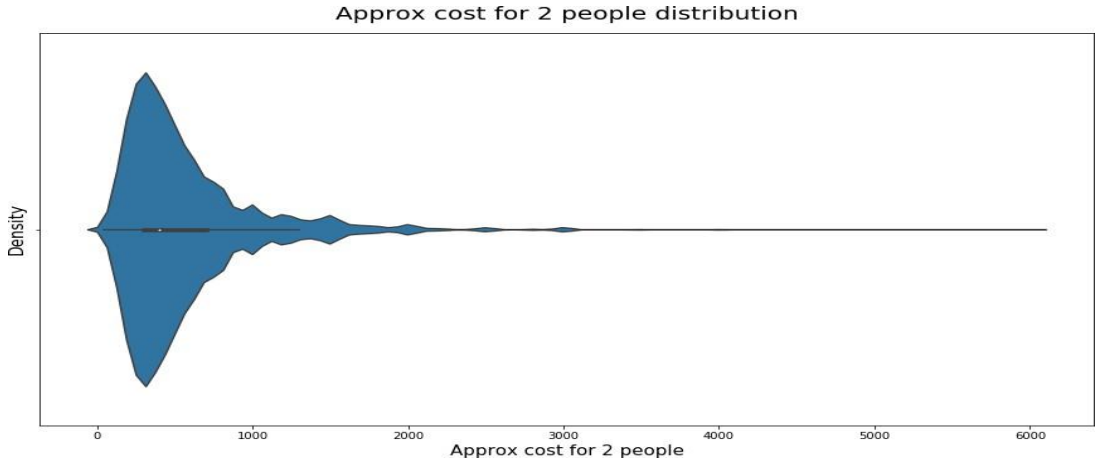
Pandas is an open source, BSD-licensed library providing high- performance, easy-to-use data structures and data analysis tools for the python programming language.

Pandas library is well suited for data manipulation and analysis using python. In particular, it offers data structures and operations for manipulating numerical tables
and time series.

1.3.5 Seaborn

Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics. E.g:-





2. TRAINING WORK UNDERTAKEN

2.1 COLLECTING DATA FROM KAGGLE

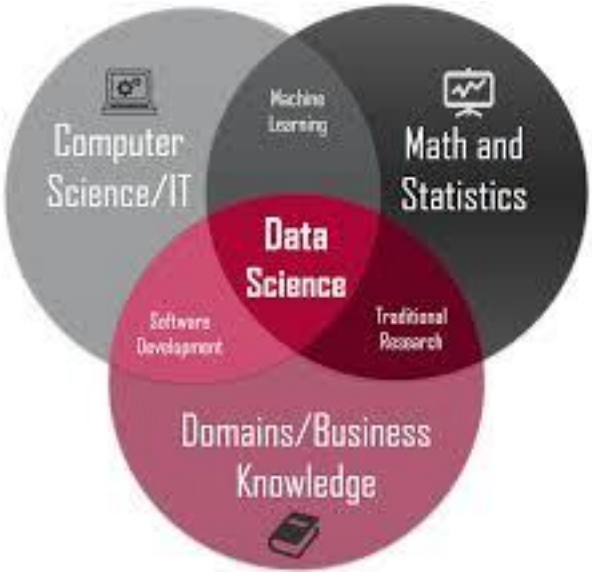
Kaggle is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowd sourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be most effective. On 8 March 2017, Google announced that they were acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyze cell images and identify nuclei.

2.2 DATA SCIENCE

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of

mathematics, statistics, information science, and computer science.

Turing award winner JiGray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, business intelligence, predictive modeling, or any arbitrary use of data, or used as a glamorized term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. Because of the current popularity of this term, there are many "advocacy efforts" surrounding the field..



DESCRIPTION OF MY DATA SET

- In this dataset the **sales price of houses in King County, Seattle** are present. It includes homes sold between May 2014 and May 2015.
- Before doing anything we should first know about the dataset what it contains what are its features and what is the structure of data.

2.3 SOURCE CODE & OUTPUT

1.IMPORT PACKAGES

```
2. import numpy as np
3. import pandas as pd
4. import matplotlib.pyplot as plt
5. import seaborn as sns
6. from sklearn.linear_model import LinearRegression
7. from sklearn.model_selection import train_test_split
8. from sklearn import linear_model
9. from sklearn.preprocessing import PolynomialFeatures
```

2. LOAD THE DATASET

```
df=pd.read_csv(r"c:/housing_price.csv")
df.head()
```

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

- Pandas head() method is used to return top n (5 by default) rows of a data frame or series

yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
1955	0	98178	47.5112	-122.257	1340	5650
1951	1991	98125	47.7210	-122.319	1690	7639
1933	0	98028	47.7379	-122.233	2720	8062
1965	0	98136	47.5208	-122.393	1360	5000
1987	0	98074	47.6168	-122.045	1800	7503

➢ The features got in the above tables are explained below:

- **id** :a notation for a house
- **date**: Date house was sold
- **price**: Price is prediction target
- **bedrooms**: Number of Bedrooms/House
- **bathrooms**: Number of bathrooms/bedrooms
- **sqft_living**: square footage of the home
- **sqft_lot**: square footage of the lot
- **floors** :Total floors (levels) in house
- **waterfront** :House which has a view to a waterfront
- **view**: Has been viewed
- **condition** :How good the condition is Overall
- **grade**: overall grade given to the housing unit, based on King County grading system
- **sqft_above** :square footage of house apart from basement
- **sqft_basement**: square footage of the basement
- **yr_built** :Built Year
- **yr_renovated** :Year when house was renovated
- **zipcode**:zip code
- **lat**: Latitude coordinate
- **long**: Longitude coordinate
- **sqft_living15** :Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
- **sqft_lot15** :lotSize area in 2015(implies-- some renovations)

3. DATA PRE-PROCESSING

1) TO FIND THE MISSING VALUES

```
df.isnull().sum()
```

```
id               0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
grade            0
sqft_above       0
sqft_basement    0
yr_built         0
yr_renovated     0
zipcode          0
lat              0
long             0
sqft_living15    0
sqft_lot15       0
dtype: int64
```

- Calling sum() of the DataFrame returned by isnull() will give a series containing data about count of NaN in each column.

2) CORRELATION BETWEEN THE FEATURES

```
df.corr()
```

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_b
id	1.000000	-0.016762	0.001286	0.005160	-0.012258	-0.132109	0.018525	-0.002721	0.011592	-0.023783	0.008130	-0.010842	-
price	-0.016762	1.000000	0.308350	0.525138	0.702035	0.089661	0.256794	0.266369	0.397293	0.036362	0.667434	0.605567	
bedrooms	0.001286	0.308350	1.000000	0.515884	0.576671	0.031703	0.175429	-0.006582	0.079532	0.028472	0.356967	0.477600	
bathrooms	0.005160	0.525138	0.515884	1.000000	0.754665	0.087740	0.500653	0.063744	0.187737	-0.124982	0.664983	0.685342	
sqft_living	-0.012258	0.702035	0.576671	0.754665	1.000000	0.172826	0.353949	0.103818	0.284611	-0.058753	0.762704	0.876597	
sqft_lot	-0.132109	0.089661	0.031703	0.087740	0.172826	1.000000	-0.005201	0.021604	0.074710	-0.008958	0.113621	0.183512	
floors	0.018525	0.256794	0.175429	0.500653	0.353949	-0.005201	1.000000	0.023698	0.029444	-0.263768	0.458183	0.523885	-
waterfront	-0.002721	0.266369	-0.006582	0.063744	0.103818	0.021604	0.023698	1.000000	0.401857	0.016653	0.082775	0.072075	
view	0.011592	0.397293	0.079532	0.187737	0.284611	0.074710	0.029444	0.401857	1.000000	0.045990	0.251321	0.167649	
condition	-0.023783	0.036362	0.028472	-0.124982	-0.058753	-0.008958	-0.263768	0.016653	0.045990	1.000000	-0.144674	-0.158214	
grade	0.008130	0.667434	0.356967	0.664983	0.762704	0.113621	0.458183	0.082775	0.251321	-0.144674	1.000000	0.755923	

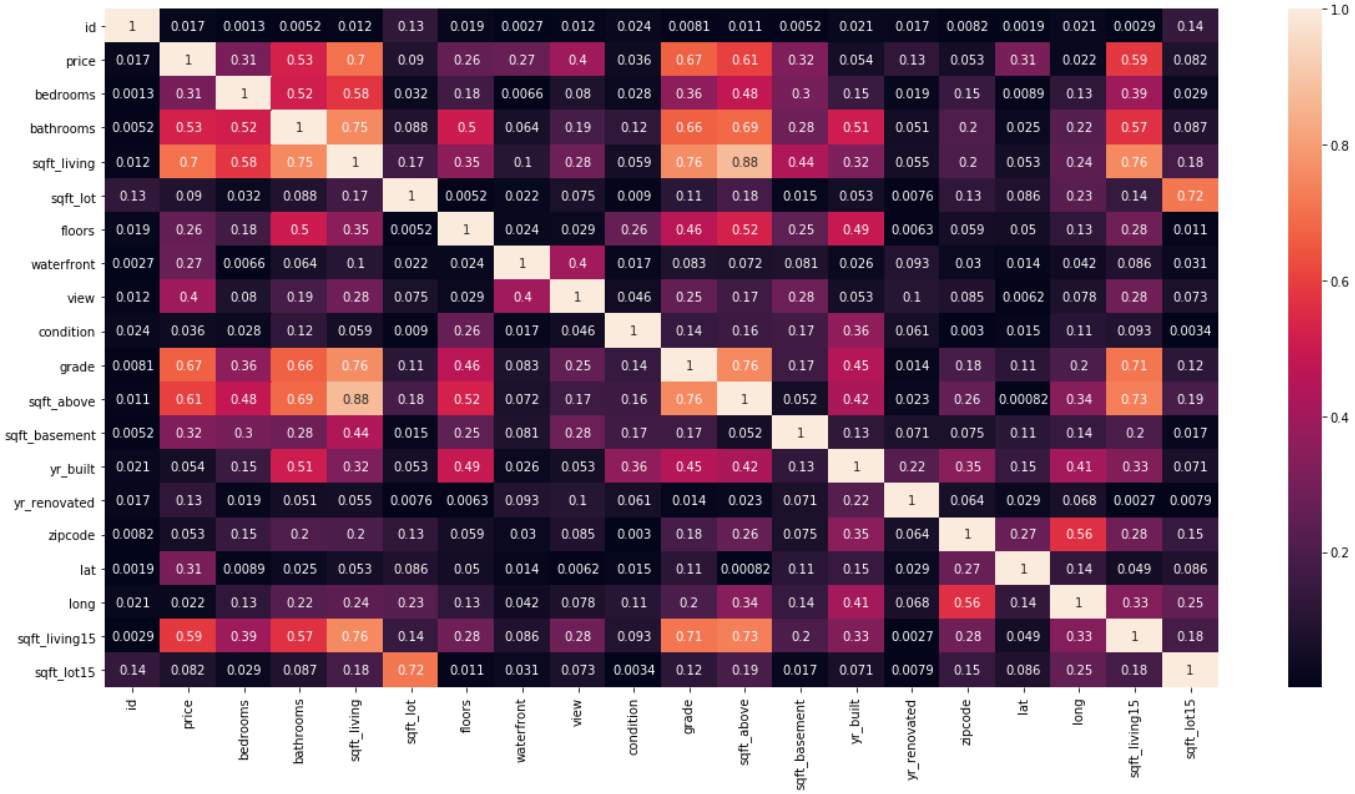
3) DATAFRAME INFO

Pandas ***dataframe.info()*** function is used to get a concise summary of the dataframe. It comes really handy when doing exploratory analysis of the data. To get a quick overview of the dataset we use the ***dataframe.info()***function.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21613 non-null  int64
 1   date           21613 non-null  object
 2   price          21613 non-null  float64
 3   bedrooms       21613 non-null  int64
 4   bathrooms      21613 non-null  float64
 5   sqft_living    21613 non-null  int64
 6   sqft_lot       21613 non-null  int64
 7   floors         21613 non-null  float64
 8   waterfront     21613 non-null  int64
 9   view           21613 non-null  int64
 10  condition      21613 non-null  int64
 11  grade          21613 non-null  int64
 12  sqft_above     21613 non-null  int64
 13  sqft_basement  21613 non-null  int64
 14  yr_built       21613 non-null  int64
 15  yr_renovated   21613 non-null  int64
 16  zipcode        21613 non-null  int64
 17  lat            21613 non-null  float64
 18  long           21613 non-null  float64
 19  sqft_living15  21613 non-null  int64
 20  sqft_lot15     21613 non-null  int64
dtypes: float64(5), int64(15), object(1)
```

4.STARTING DATA REGRESSION PART (PREDICTION)

- **Checking for correlation among all the x(inputs)**
- **The following fig. is a correlation matrix known as Heatmap.**



1.Prediction using Multiple Linear Regression

- Multiple linear regression (MLR), also known simply as multiple regression
- It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

```
Multiple Linear Regression
Mean Squared Error (MSE)  205244.56
R-squared (training)  0.655
R-squared (testing)  0.672
Intercept:  -32330182.91167577
Coefficient: [-2.62100082e+04 -3.47626774e+03  1.32069090e+02 -1.31506101e-01
 -3.22608866e+04  5.65731080e+05  6.81843494e+04  8.17559446e+04
  6.62246806e+01  6.58444098e+01  6.71765072e+05  4.60648979e+00]
```

R-SQUARED (training)= 65.5%
R-SQUARED (testing)=67.2%

2.Prediction using Polynomial Regression (degree=2)

- Polynomial regression is a special case of linear regression where we fit a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables.
- In a curvilinear relationship, the value of the target variable changes in a non-uniform manner with respect to the predictor (s).

```
POLYNOMIAL REGRESSION (degree=2)
Mean Squared Error (MSE)  175812.56
R-squared (training)  0.758
R-squared (testing)  0.759
```

R-SQUARED (training)= 75.8%
R-SQUARED (testing)=75.9%

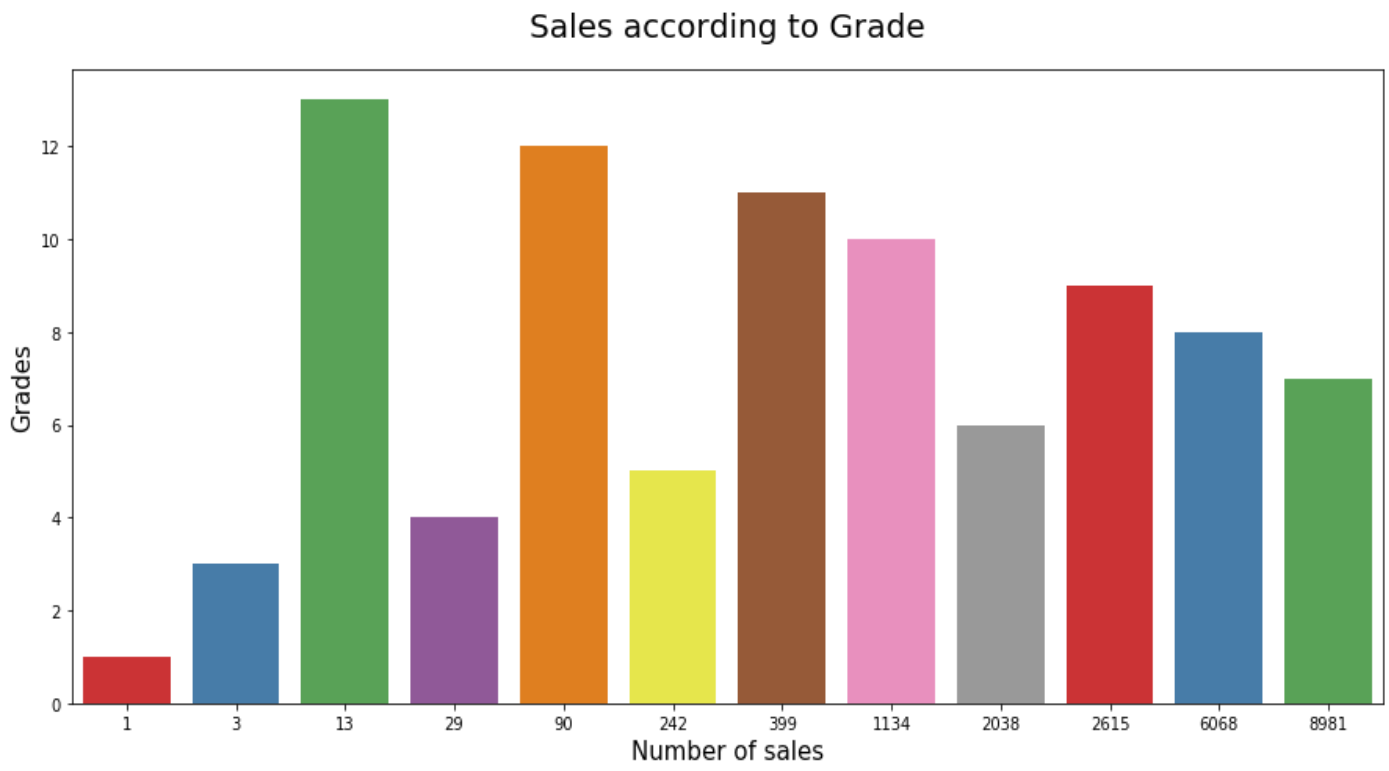
3.Predictions using Polynomial Regression (degree=3)

```
POLYNOMIAL REGRESSION(degree=3)
Mean Squared Error (MSE)  202646.78
R-squared (training)  0.776
R-squared (testing)  0.68
```

R-SQUARED (training)= 77.6%
R-SQUARED (testing)=68%

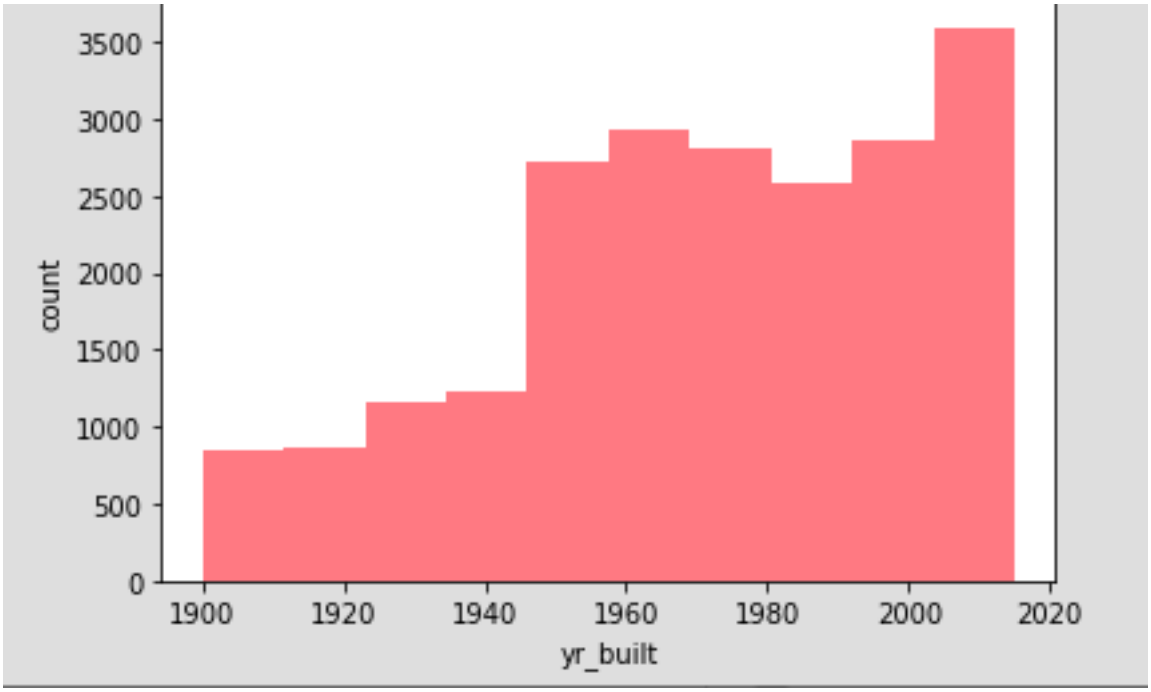
4.DATA VISUALIZATION

Grades assigned:



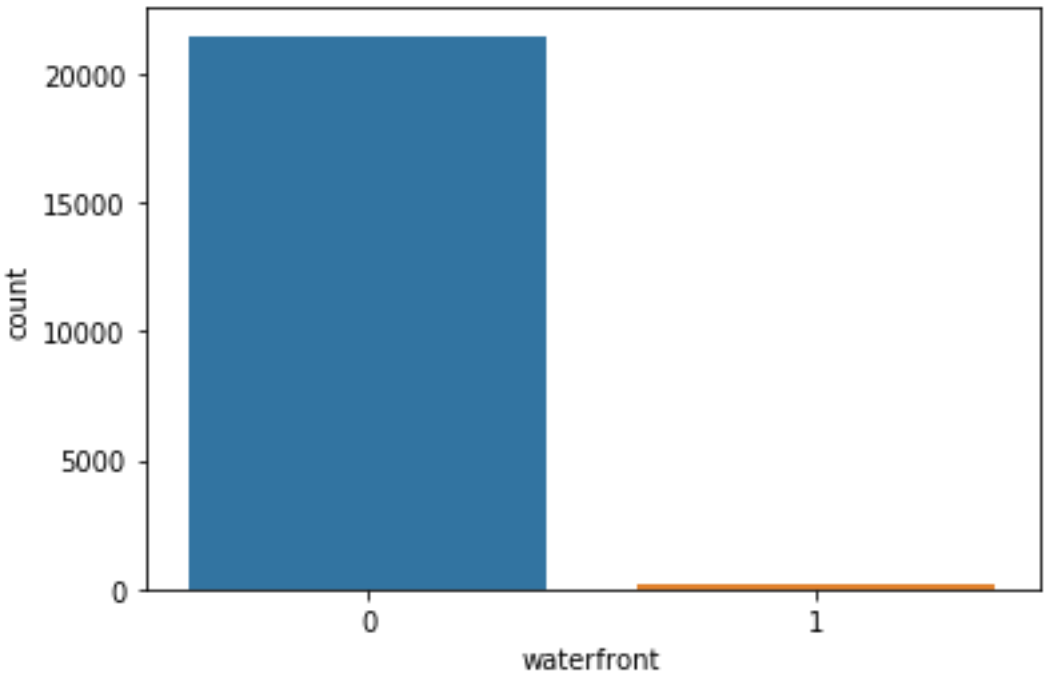
- **This depicts the overall grade given to the housing units in King County with the total counts of various grades.**

Number of houses built in particular year:



- **A lot of houses were built in the 2000s and 2010s.**

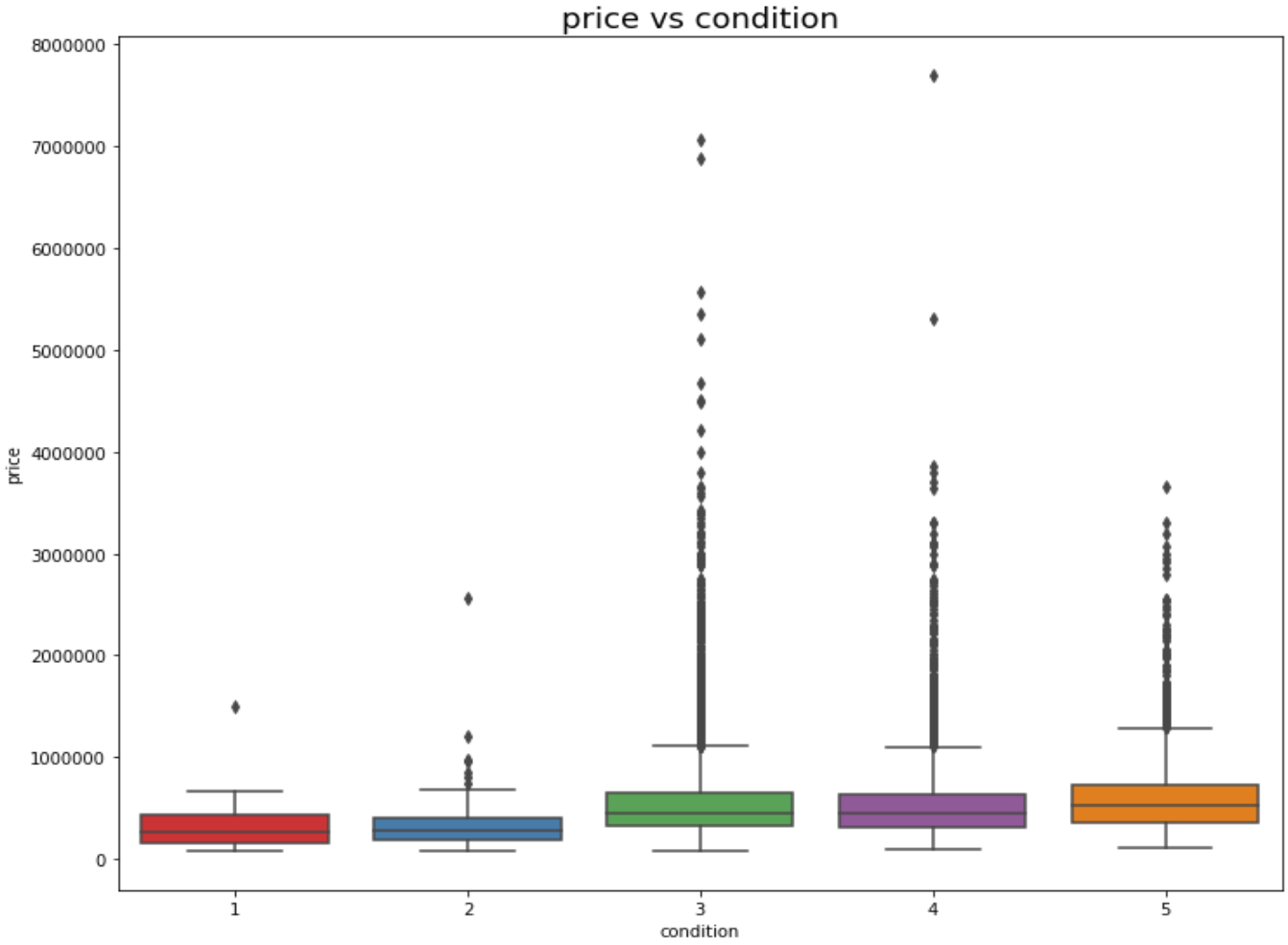
Waterfront- 0 or 1:



- House which has a view to a waterfront are given 1 and which do not have a view to water are given 0.

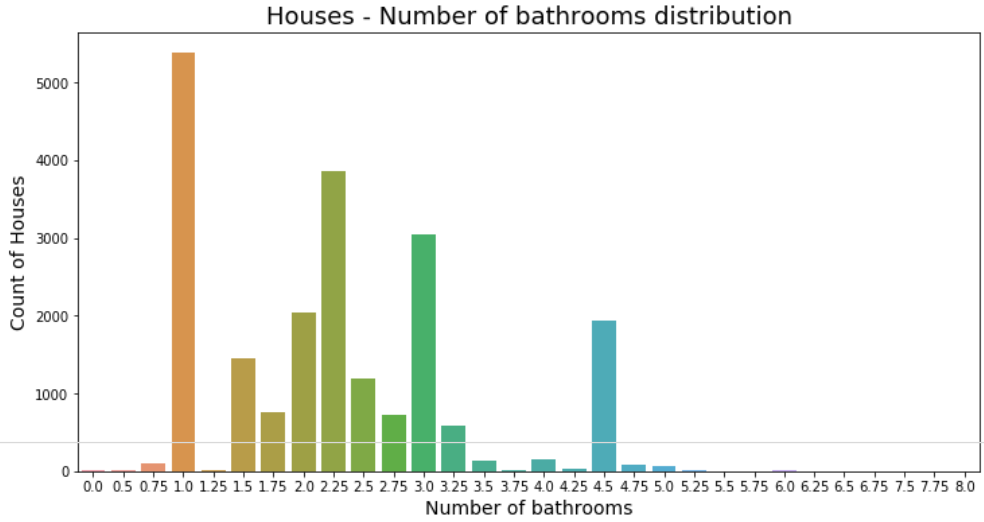
- Therefore, clearly all houses lack a water front view.

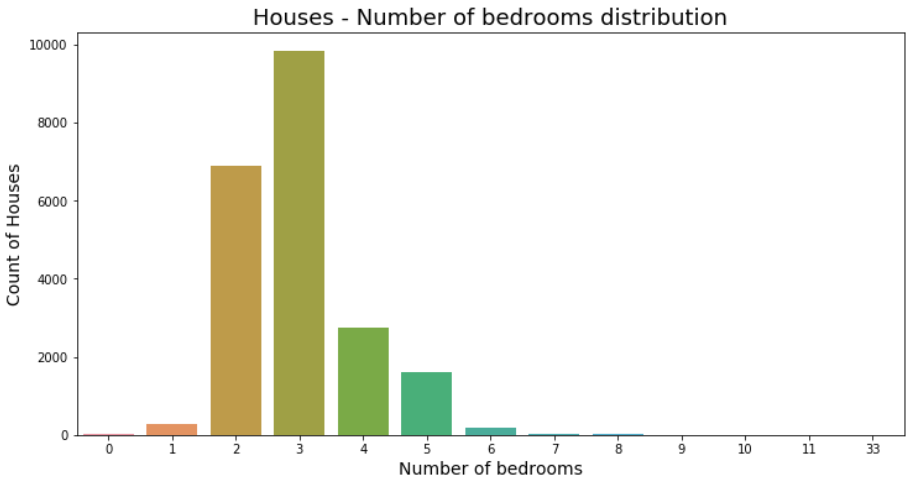
Prices vs condition boxplot:



- **We observe that a better condition doesn't necessarily imply a higher price.**
- **Again it would be to broad to generalize anything given the number of outliers.**

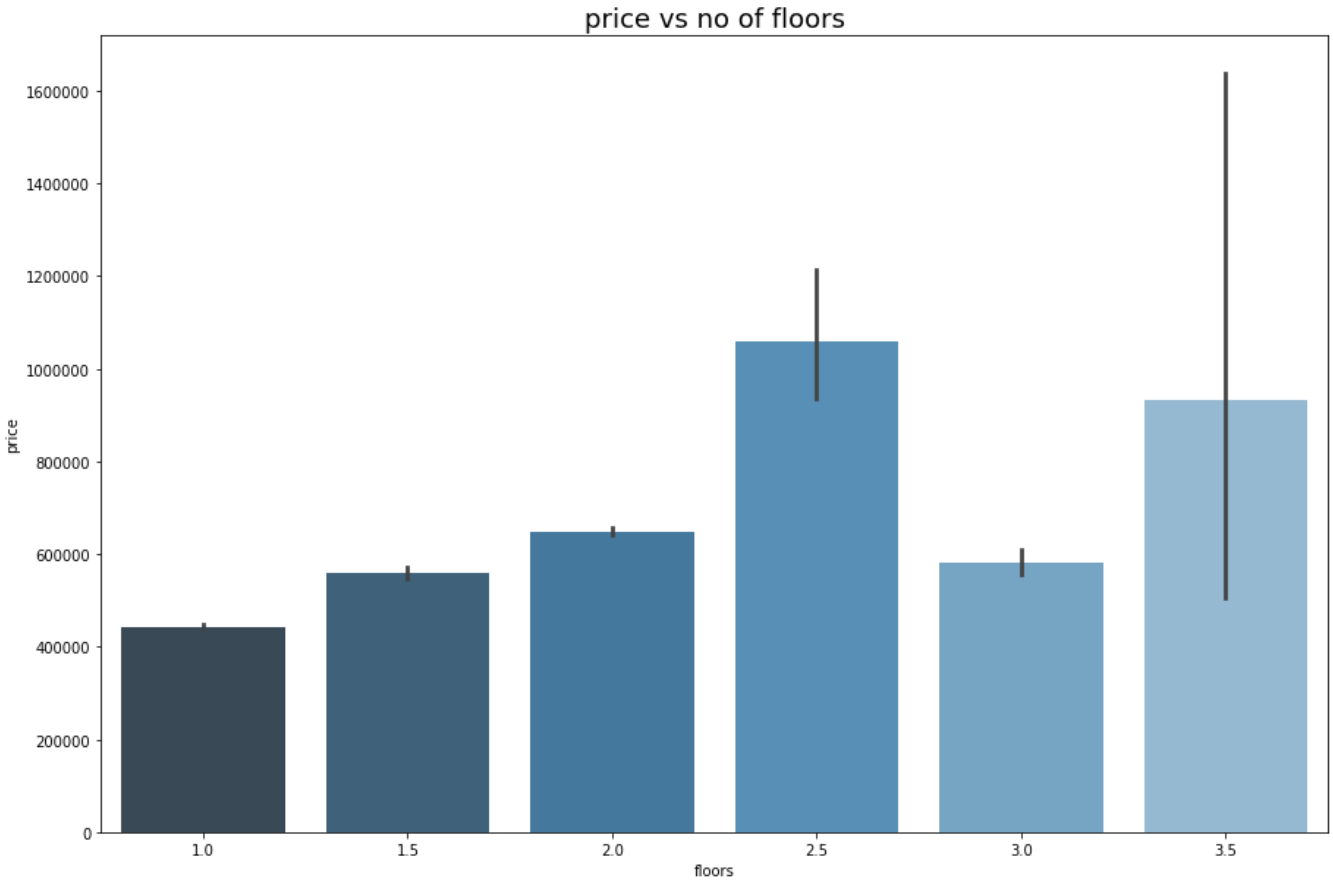
Frequency of number of bedrooms: and bathrooms





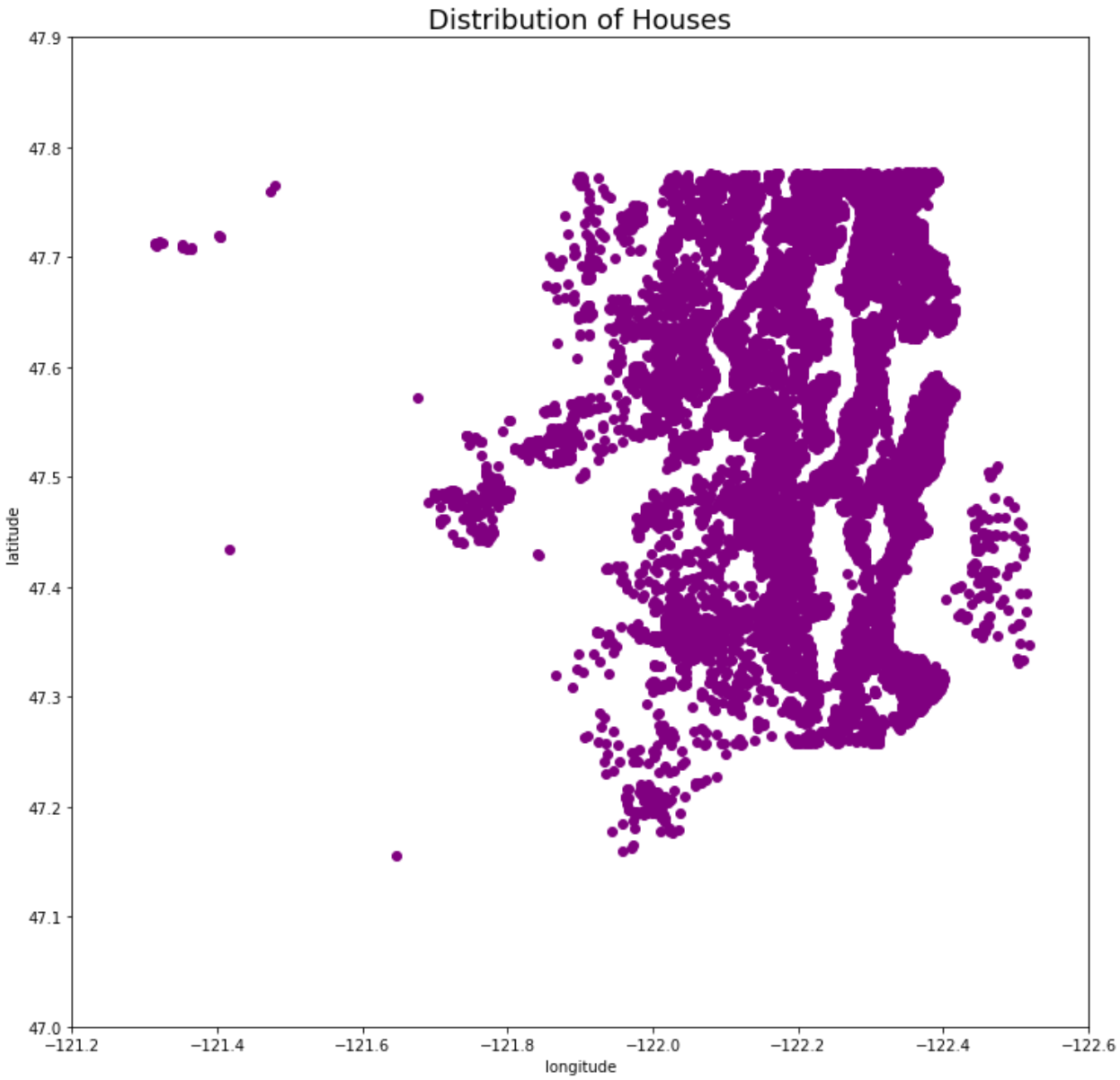
- **I plotted some bar chart to visualize the frequency the number of bedrooms and bathrooms occur in the dataset.**

Number of floors and prices:



- **As expected, although there is a postive correaltion between these two variables, there isn't an obvious trend.**
- **Penthouses and loft apartments in downtown Seattle might definitely be more expensive than a three-story suburban colonel.**

Showing the distribution of Houses:



- **This yield a very interesting result. It shows the closeness and placement of the houses.**

5. RESULTS AND DISCUSSION

- **Multiple Regressor – 65.5%**
- **Polynomial Regressor(deg 2) – 75.8%**
- **Polynomial Regressor(deg 3) – 77.6%**

But the diff. between training and testing data for Polynomial Regression (degree=3) is more, therefore we will not consider it.

BEST MODEL:
POLYNOMIAL REGRESSION (DEGREE=2)

CONCLUSION

- We first identified the independent variables and the dependent variable, price.
- After visualizing the dataset using graphics,we inferred various aspects
  - Most of the **houses are located in which area.**
  - Most of the houses were built in the 21st century.
  - We observed that the **housing in King County is primarily influenced by the living area and the grade of the construction materials** used to build the house followed by the number of bedrooms and the number of bathrooms.
  - **When it comes to prices and the number of bedrooms, bathrooms and floors, there are** a lot of exceptions because of the unique locations of the houses merged into one large dataset.
  - **Houses built with better construction materials cost more on an average though they do not guarantee a better condition of the house.**
  - For houses that have an excellent view of the waterfront, their prices are signifiantly higher than both those houses that totally lack a view or those with a compromised view.
- We built a **regression model using the Polynomial Method with around 76% of accuracy.**

REFRENCES:

www.kaggle.com
www.python.org
www.numpy.org
www.matplotlib.org

THANK YOU.