2020

PROJECT REPORT

Wande B. Adeyeye

adeyeyewande@gmail.com

9/25/2020

PROJECT REPORT REGRESSION ANALYSIS

KING COUNTY HOUSE PRICE

BY:

WANDE B. ADEYEYE

E-MAIL:

ADEYEYEWANDE@GMAIL.COM

King County House Prices Prediction Model. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The Goal of this analysis is to predict the price of housing in King County, based on the variables provided in the dataset.

Objectives:

I. Overview of Data

II. Data pre-processing

III. Data visualization

IV. Predictive Modeling

Overview of Data

ID	ID for each home sold
Date	Date for each home sold
Price	Price for each home sold
Bedrooms	Number of bedrooms in each apartment
Bathrooms	Number of bathrooms in each apartment
Sqft_living	Square footage living area in each home
Sqft_lot	Square footage of the land space
Floors	Number of floors
WaterFront	A dummy variable of either a house has a waterfront or not
Views	An index number from 0 to 4 on how good a property views was
Condition	An index from 1 to 5 on the condition of a property
Grade	An index from 1 to 13
Sqft_above	The square of the interior housing that is above ground level
Sqft_basement	The square of the interior housing that is below ground level
Yr_built	The year the house was built

Yr_renovated	The year the house was last renovated
Zipcode	The zipcode a house is in
Lat	Latitude
Long	Longitude
Sqft_living15	The square footage living area of interior housing living space of the nearest 15 neighbors
Sqft_lot15	The square footage of the land lot of the nearest 15 neighbors

The very first step in any Data Science and Machine Learning problem solving is asking questions and I mean asking the **right questions,** as a Data Scientist you must be capable of asking the right questions because if we get the question wrong i.e. if we asked the wrong questions then we will get our algorithm perfectly wrong because the questions we asked will determine how we will go about writing the algorithm to solve it.

FIRST STEP (Forming The Questions):

Imagine that you're living in the beautiful city King County in the United States and you're working in the real estate business and you have this friend who recently moved to the city. So you decide to meet up with them in one of the local coffee shops you two get talking and he asked how much a house cost here in the city, without giving much information like location, size of the house, features etc. what is the most truthful way of answering the question? Well it's simply the average house price in King County City and that's exactly what we're going to do in this project.

- Average home price in King County City in The United State
- Highlighting the important features in a home and how they influence price

SECOND STEP (Gathering the Useful Data):

After we've successfully formulated our questions, the next thing we're going to do is gather useful data that will help us solve the questions and Kaggle is one of the best place to do just that, the link to the

dataset is here. So, from here we move to Jupiter notebook to start forming our algorithm to solve the problem. Before we move on we need to ask ourself and clearify this six conditions:

1. Source of the Data
2. Description of the Data
3. Number of Data points
4. Number of features
5. Names of the features
6. Descriptions of the festures

Answering and understanding this questions gives us a base to know more about the data we're going to be working with and technics to approch it

Data pre-processing

Third Step (Data Cleanning, Explore & Visualise):

This two steps '**DATA CLEANNING**, **EXPLORE & VISUALISE**' will go together because as we clean our data, we need to visualise to see how and what we're doing. We've formulated our questions, we've gather our data so now its time to explore our dataset in depths. The first thing to do is using pandas Dataframe function to visualise the data, then explore futher analysing the whole dataset, each colums in relations to the price, checking for missing values, NAN values, use the value counts function to better understand a specific colum we do this in order to see the state of our dataset, the method of data cleanning depends solely on the state of the dataset.

Missing Value Detection: Missing data pattern was used to identify the missing data in the dataset. From the table below it can be observed that the data does not consist of any missing data for any of the variables which helps us creating better model.

Summary of other Data Inconsistencies: While exploring the data we found a few instances where the data between variables was inconsistent and didn't make logical sense. We chose to make the values consistent by exclude those instances from the data.

1. One observation with 33 bedrooms in 1620 Square feet with 1.75 bathrooms.

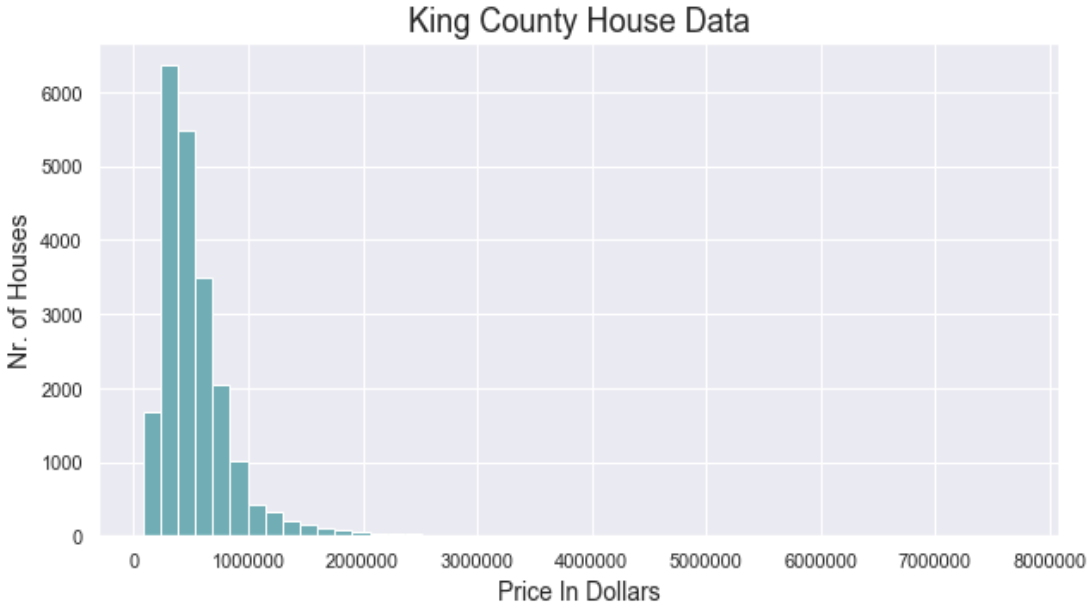
2. Ten observations with 0 bathrooms.

3. Thirteen observation with 0 bedrooms

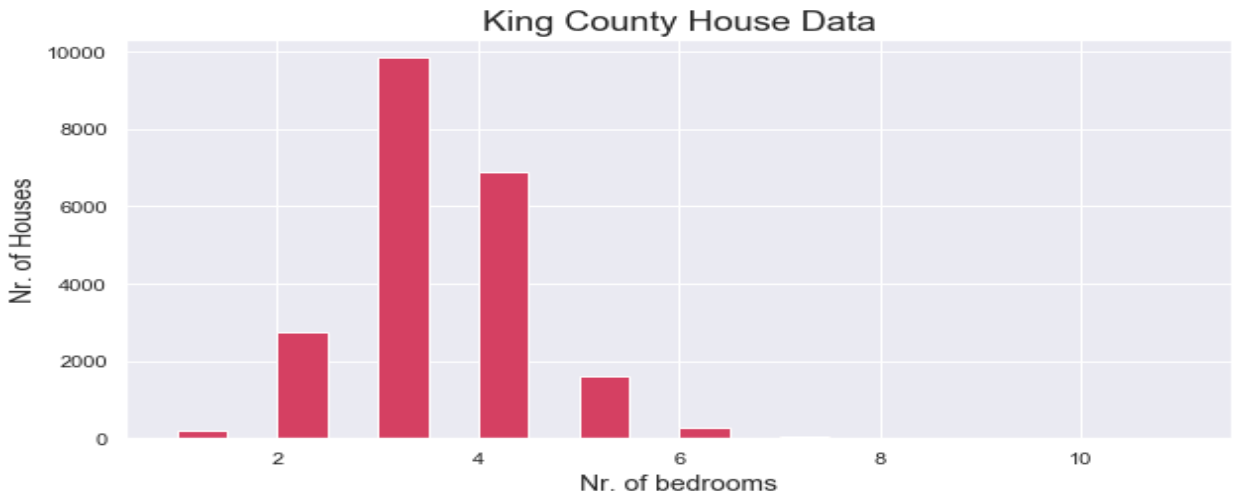
Since we don't have a direct access to where the data is originally from we deiced to drop those observations.

Data Visualising

One of the most important thing to do as part of explore is to visualize, it helps us make sense of our data and descover the pattern of our data. The two things we want to get sense from right now are the distributions and outliers and to do that we need to plot an histogram. First we plot the histogram for our price:



The histogram shows the frequencies, between the prices and the number of houses, the taller the individual bar the more occurrences they are in the dataset. The visilization for our price is quite good no outliers or whatsoever. The next plot below shows the numbers of bedrooms and we can see that there are around three bedrooms in each houses on average



Descriptive Statisics

We use descriptive satistics to quantitatively discribe and summarize features from our dataset to see how our features relate with one another. Using this methods we can spot outliers. Before we run our regression analysis, there is something we need to understand pior feeding our data into our machine learning algorithem as part of data exploration we need to understand to what extens our variables move together so we should look both to our correlation features and price and coorelations between different features

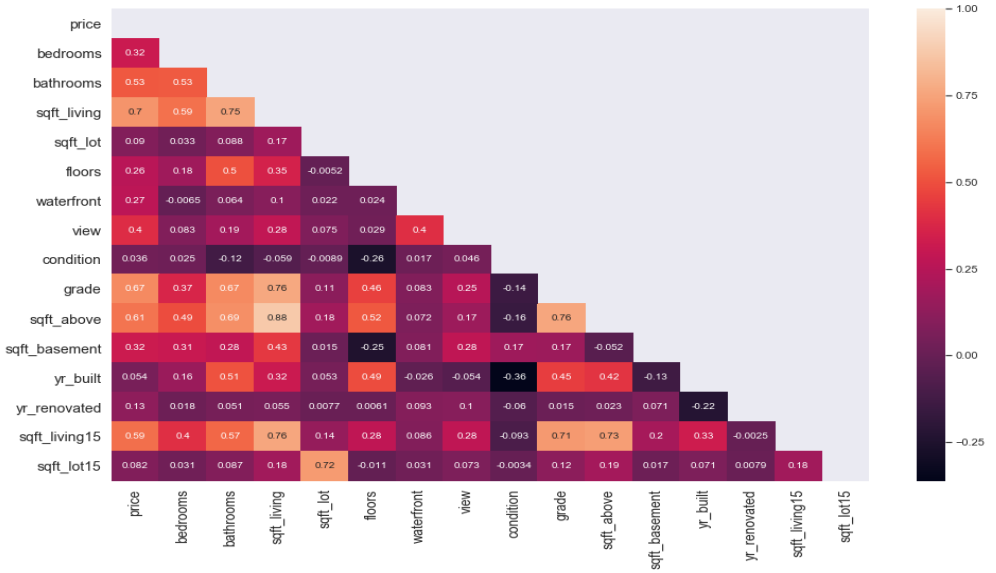
Correlation

Correlation is any statistical relationship, wether causal or not, between two random variables or bivariate data or any satistical association, through it commonly refers to the degree to which a pair of variable are linearly related. Correlation is between '-1 & 1', ' -1' is a perfect nagative corr. While '+1' is a perfect positive corr. And '0' means there is no correlation at all. We use correlation for basically two reasons, the strength of the correlation and the direction since our goal is to predict house prices, our models should include features that are correlated to house prices. It can be seen by calculating and visualisation

```
#correlation between property prices and the sqft_living
dataset['price'].corr(dataset['sqft_living'])
```

```
0.7019208734913488
```

We can see that 'price' and 'sqft_living' have a very high correlation which is a good thing. In other to see all the correlation of our dataset, we run correlation command on our dataset. One thing we notices is that there's a diagonal which is equals to '1', it's because a variable correlation with itself will be equals to one, we can simply ignore it because its not telling us much, we will as well ignore half of the arc splited by the diagonal because between the diagonal are telling us the same thing as show in the graph below:



Here, were looking for two things, the strength and the direction of the correlation and there's also a third thing we care about because we don't only have correlation between the features and the price only, we also have correlation between other features as well. If two features are perfectly correlated it can either be good or bad but it can be problematic for our regression model and it's called multicollinearity but also, correlation does not imply causation. Just because two things work together doesn't means one causes another. Another limitation for correlation is that it only check for linear relationships. Both sqft_above and grade have a very high correlation of 0.76, were using the defualt correlation which is pearson, one thing we need to know about this correlation is that, it's only valuable for contiuneous values, but grade is a categorical variable so our correlation is not valid for the grade feature.

Advanced Visualization

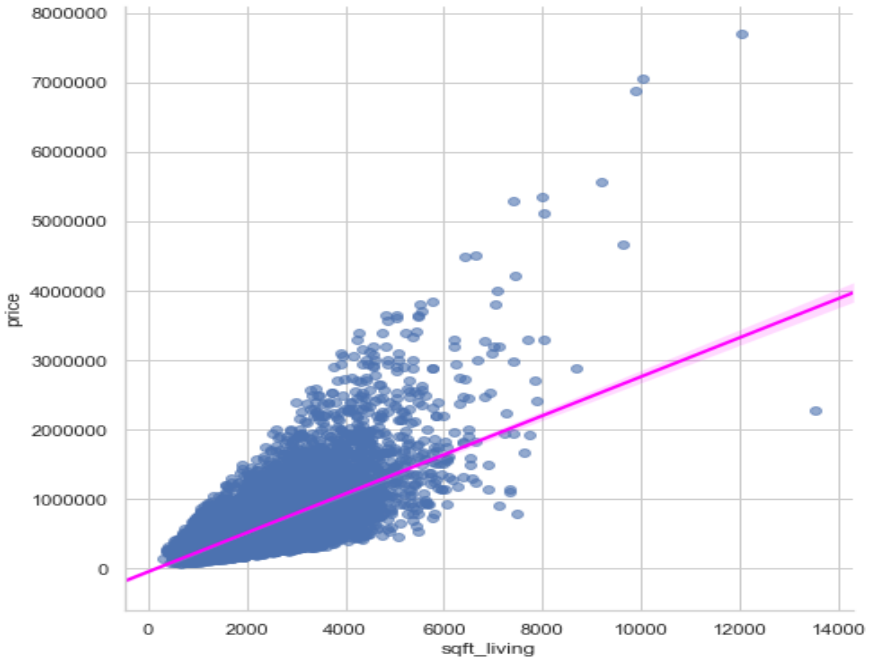
Now, we have three types of variables here:

1. Continuous: sqft_living, sqft_lot, Yr_built, Yr_renovation

2. Categorical: Bedrooms, Bathrooms, Waterfront, view, condition, grade, floors

3. Nominal: Zip code, lat, long

Using scikit-learn we will visualize multiple data in our dataset regression to see the relationship between our dataset. We're going to look at is the price and sqft_living, looking at the below graph we can see the positive correlation '0.7' ties nicely with the relationship on the scatter plot which means the living area have a clear relationship with our house price. From the graph below, we see that increase in living area also leads to increase in property price



To visualiz all our dataset, we will use seaborn pairplot function to visualize all of our data point and plot a regression line to each of the features to see their relationships in jupyter notebook.

Observation From Our Visualization:

1. The price of houses increases for houses with 0 - 2.5 (around 3) floors and then subsequently decreases
2. Houses with a greater number of floors have higher price.
3. Houses having a waterfront are valued higher
4. Prices increase as the number of view and grade increases.

Training Our Model

We will finally train our model, we're going to model our house price using three different techniques and see how our model passed. We're fitting more than one features in our model, the equation will look like this:

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

What this equation is telling us is that the estimate of our house price will be a linear combination of all of the features in our dataset so it's still a linear model.

Predictive modeling

Split Training and Test Data

One of the things we do in training our algorithms is to split up our data into two parts, we're going to shuffle our data and divide our dataset into training dataset and testing dataset, because we want our algorithms to learn those regression coefficients based only on the training dataset, and it means that we can use the other part of our dataset which hasn't been used for testing because with testing, we can see how our algorithms performs out of sample how it performs on a dataset that it hasn't seen yet.

After splitting our data into training and testing dataset, we will calculate the R-Squared for different regression model and pick the one that has a better performance.

Conclusion

We applied three models: Linear Regression, Decision Tree Regressor and Random Forest Regressor

As we can see Random Forest Regressor performed best (with accuracy ~ 0.87)

.....

This is my first ever project and I'm really glad to be able to put what I've learn together in this project, I learned a lot as well through the process of this task. My next one will be a clasification project and it will be much more improve/intuitive than this.

THANKS