Capstone Project 1: Project Proposal

I set out to use linear regression to predict housing prices in King County. The question I want to answer is Which features influence the price of a home and is it possible to predict a price given certain variables?

What is the problem?

- The problem is to build a model that will predict house prices with a high degree of accuracy given a dataset.

The Dataset

- The dataset I am using is the house sales in King County, USA located here https://www.kaggle.com/harlfoxem/housesalesprediction
- This dataset contains 19 house features, the price and the id with 21613 observations.
- The features of the dataset are as follows:
    - 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/house
    - 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
    - Yr_built: Built Year
    - Yr_renovated: year when house was renovated
    - Zipcode: zip
    - Lat: latitude coordinate
    - Long: longitude coordinate
    - Sqft_living15: living room area in in 2015 (implies some renovations) this might or might not have affected the lotsize area
    - Sqft_lot15: lot size area in 2015(impilies- some renovations)

Capstone Project 1: Project Proposal

The Approach

I will be doing my analysis on the data using the programming language Python and the Jupyter notebook. The Jupyter notebook is an powerful tool for interactive developing and presenting data science projects. A notebook integrates code and its output into a single document that combines visualizations, narrative text, and mathematical equations.

There are five steps in the Data Science Process and each plays its own unique part in solving the problem. The fives steps are:

1. Ask A Question: We ask an interesting question, a scientific goal. In my case, it is what I want to predict or estimate which is the housing price.
2. Get the Data: The first technical aspect of the project is get the data. This is by either creating a script to get the data from the source or download the data manually through application. My approach would be creating a script and reading the data onto my jupyter notebook. It is also through this step we get to clean the data and making our dataset presentable for further analysis.
3. Explore the Data: On this step get to know our data with the use of Python. The tools include matplotlib, numpy, pandas will allow for creating data visualizations to check for pattern, anomalies, and help us better understand the problem.
4. Model the Data: This step includes the machine learning aspect of Data Science. I will create a regression model using scikit learn, pandas, and other important tools if needed.
5. Communicate the Data: The last but certainly not least stage of the process is presentation. It is during this stage I will communicate my results with visuals, and writing.