# OVERVIEW

In this project I take the role of a data scientist tasked to model the formula of predicting a price of a house in the northwestern county, King County in Washington State in the United States of America.

The population was *2,269,675* in the [*2020 census*](https://en.wikipedia.org/wiki/2020_United_States_census) making it the most populous county in **Washington**, and the [13th-most populous](https://en.wikipedia.org/wiki/List_of_the_most_populous_counties_in_the_United_States) in the United States. The county seat is [Seattle](https://en.wikipedia.org/wiki/Seattle) also the state's most populous city.

# BUSINESS & DATA UNDERSTANDING.

The stake holder for this was a real estate firm found in the King county to assist them in price determination. Using the data “kc\_house\_data.csv” I am to make the predictive model.

The data has 21597 rows and 20 columns. I started of with the base model with “price” as the dependent variable and sqft\_living as the independent variable. Sqft\_living is the square footage of living space in the home. I started with it as it had the highest correlation with price and also showed a linear relationship with price.

## DATA UNDERSTANDING.

It has been loaded in with pandas as kc. Short for King County. It has 21597 rows and 20 columns.

No duplicate rows were observed. Only 3 columns were observed with missing values;

1. yr\_renovated 17.7% missing
2. waterfront 11.0% missing
3. view 0.29% missing

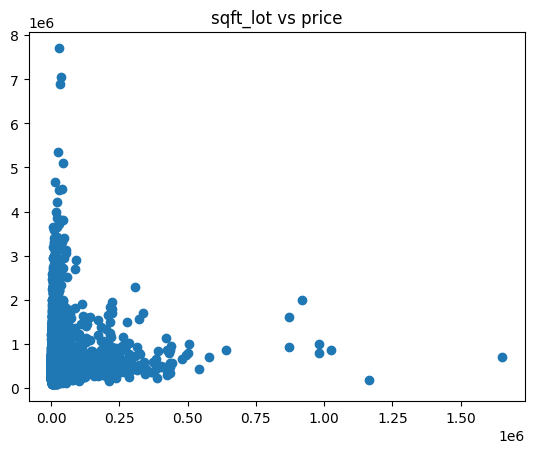
The kc column with the strongest correlation is the sqft\_living (Square footage of living space in the home) with a correlation of 0.701917. This shall be used for the baseline model.

The kc dataframe comprises of 3 dtypes;

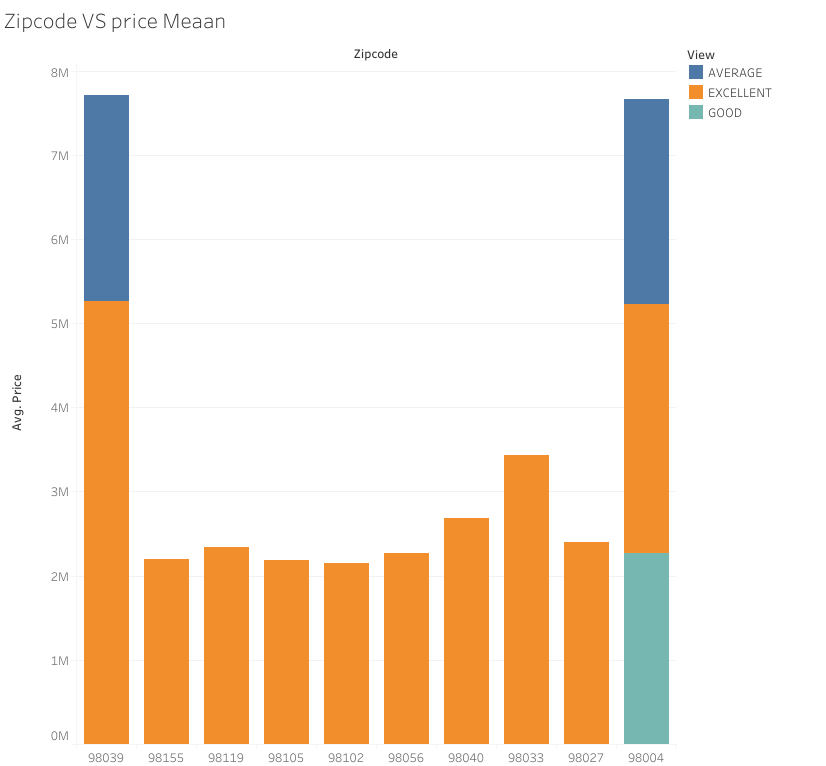
1. dtype('O')   
   t means: 'O' (Python) objects. Source.
2. dtype('float64')  
   Python float values are represented as 64-bit double-precision values.
3. dtype('int64')  
   The type int64 tells us that Python is storing each value within this column as a 64 bit integer.



Aside sqft\_living the other continuous cloumn that I used was sqft\_lot which is the Square footage of the lot.

It didn't have as strong of a linear relationship as sqft\_living but still maintained a linear relationship. But it was the only one that was not removed after removing the columns that could lead to multicellularity .

From the data, the top ten zip-code area with the largest mean of prices are depicted in the bar graph below. The top being Medina,Washington (98039) followed by Bellevue, Washington (98004).



My approach was that views surrounding a house from the data will greatly affect the price together with the condition of the house.

These were the only categorical data points I used.

# MODELING

The base model formula achieved up-to 95% confidence ;

y = ([282 - 291])Sqft\_living + ([-65,400 - -45,000]) USD + 175,000 USD

Multi-linear model of sqft\_living with sqft\_lot.

y = ([282 - 291])Sqft\_living + ([ -0.3 - -0.05])sqft\_lot + ([-65,400 - -45,000]) USD + 175,000 USD

Multi-linear model of sqft\_living with sqft\_lot with views.

y = ([259 - 268])Sqft\_living - ([-0.41 - -0.23])Sqft\_lot + ([143,200 - 9,880])view\_Average + ([174,900 - 16,200])view\_Fair + ([212,400 - 13,800])view\_Good + ([600,900 16,600])view\_Excellent - 25,800 USD + 167,100 USD

# **REGRESSION RESULTS**

From the models created , I have decided to favor *Multi-linear model of sqft\_living with sqft\_lot* *with views.*

Since it explained the largest percentage, about 60% of the variance observed in price variable.

# **CONCLUSION.**

Despite the formula I favor, the errors are still probable and for a more accurate prediction of the price I recommend the real estate firm to create models for each zip uniquely as they have different pricing per zip-code with the highest being **Medina,Washington (98039)** with a mean price of **2,215,069 $** and the lowest zip-code Auburn, Washington (98002) with a mean price of **233,924 $.**