Objective:

* Explore the characteristics of the houses using some business intelligence tool.
* Build a model that will predict the price of a house based on features provided in the dataset.
* Understanding which factors are responsible for higher property value - $650K and above. The questions have been provided later in the document for which you can use tableau.

Mysql:

* Dataset consists of 21,420 homes in Kings County, WA that were sold in between May 2014 – May 2015.
* It contains 20 variables of which will determine what affects the price of a home
* Check variables for any discrepancy in data:
  + Bedrooms range 1-33
  + Bathrooms range 0.5 – 8
  + Floors range 1- 3.5
  + Years built range 1900 – 2015
  + Sq\_ftliving range 370sqft – 13,540sq\_ft
  + Basement – Majority of homes do not have a basement
  + Waterview – Majority of homes do not have a Waterview. Rated 0-1
  + View – Majority of homes do not have a view. Rated 0-4
  + Zipcode – More homes sold in 98103
  + Condition– Most homes condition were of average conditions. Rated 1-5
  + Grade – Most homes Grade were of average grade. Rated 1-13
* Explore data for discovery:
  + Average price for a 3 bedroom home = $466,527.48
  + Average price for a 4 bedroom home = $635,794.49
  + Average sqft\_living for a 3 bedroom home = 1,807 sqft
  + Average sqft\_living for 4 bedroom home = 2,556 sqft
  + Average price for a waterfront home = $1,662,524.18 vs $532,137.39

Python

* Cleaning the data
  + Checking for nulls – None identified
  + Checking data types – Integers/Floats
  + Checking for duplicates – None identified
  + Drop columns – Id and Date were dropped as they are not relevant to the model
  + Change column type to object type – Condition, Waterfront, View, Zipcode
  + Deleted row with 33 bedrooms
  + Used log and square root transformation to manage outliers
  + Used sns.barplot for further data visualization and discovery
  + Heatmap: sq\_ftliving and bathrooms are highly correlated to price. Sqft\_above and sqft\_living15 also have high correlations to the target but can be because they are highly correlated to other similar variables.
  + VIF – Elevated VIF; can be caused by some data point being read as numerical but are really categorical numbers. Also, some data points are highly correlated with similar variables.
  + Dropped 'sqft\_above','sqft\_basement','yr\_renovated','sqft\_living15','sqft\_lot15' due to being to closely related to other variables.
  + Modeling data with scaled data and encoded data.
    - Linear regression: 0.8229647177490109
    - Decision tree: 0.6593127096904975
    - KNN: 0.7751567788406417

Tableau

* Most expensive zip code: 98039
* Waterfront properties are 3x more than properties without waterfront
* View = 4 has an avg cost of 1.4 million
* Avg price by condition of home is approximately 600K
* Most homes in the data set were built between 1900-2000 with an avg price of 520k
* Increase in grade of home the avg price goes up
* Majority of homes are average grading = 7
* There is a direct relationship between increased living area and increased price of home
* 2.5 floors has a higher avg price than that of 3.5 floors