### Capstone Project 1: Data Wrangling

### *by Mehreteab Kidane*

### Data Sources: This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Let’s get started by reading the dataset I’ll be working with and deciphering its variables. For this Capstone project, I’ll be using a [Kaggle dataset](https://www.kaggle.com/harlfoxem/housesalesprediction)house price patterns. Kaggle is a great community of data scientists analyzing data together – it’s a great place to find data to practice the skills covered in this project.

The dataset contains a detailed set of information about house pricing features and the main problem statement here is to determine the house price that should continue to sell, and how to predict it. The file contains the observations of house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The end solution here is to create a model that will predict house price – I’ll perform EDA on this data to understand the data better.

Let’s analyze the dataset and take a closer look at its content. The aim here is to find details like the number of columns and other metadata which will help us to gauge size and other properties such as the range of values in the columns of the dataset.

*Read the data into a data frame*

data = pd.read\_csv('. ./input/kc\_house\_data.csv')

All the exploratory data analysis like shape, info, head and describe has been performed to get an overview of the features and to look on the target variables.

**Missing values:** I also used the drop cleaning function to reduce the dataset by dropping columns that won't be used during the analysis.

*Example: data.drop (['id', 'date'], axis = 1, inplace = True)*

**To check if there are any null values in the dataset:** The answer to these questions is important for practical reasons because missing data can imply a reduction of the sample size. This can prevent us from proceeding with the analysis. Moreover, from a substantive perspective, we need to ensure that the missing data process is not biased and hiding an inconvenient truth. The good news is our data set contains no missing values.

*#missing data*

total = df\_house.isnull().sum().sort\_values(ascending=False)

percent = (df\_house.isnull().sum()/df\_house.isnull().count()).sort\_values(ascending=False)

missing\_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])

missing\_data.head(20)

# **Outliers Detection**: Outliers were detected and analyzed using the outlier box plots. From the outlier box plot I inferred that the data consist of many outliers for the target and price variables. However the outliers for the price variable corresponded to outliers for Numbers of bedrooms, number of bathrooms and square feet living. .

prices %>%

filter(price>=150000) %>%

ggplot(aes(Year,price,col=Year))+geom\_point(position="jitter")+

ggtitle("Outliers Are Lost in the Crowd")

# __results___19_1.png