**Data Challenge**

The dataset for kings county sales prices from May 2014-May 2015 (taken from Kaggle dataset/also used in Coursera). The datafile is named ‘kc\_house\_train\_data.csv’. We see that there’s many features in the data that can help us predict the price of a house in this area. Split this data to training and test data with an 80/20 split.

**First Part** – (single feature)

Create a linear regression with each feature.

Which single feature has the smallest MSE? Which one has the largest MSE?

**Second Part –** (multiple feature)

On the training data, perform a linear regression for different the different features:

**1 feature** - 'sqft\_living'

**3 features** - 'sqft\_living', 'bedrooms', 'bathrooms'

**19 features** – (all the features) - 'bedrooms', 'bathrooms', 'sqft\_living', 'sqft\_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft\_above', 'sqft\_basement', 'yr\_built', 'yr\_renovated', 'zipcode', 'lat', 'long', 'sqft\_living15', 'sqft\_lot15'

Which one of these fits have the smallest MSE?

**Third Part –** (exploring data)

Based on the training data – on average which zip code has the most expensive house prices? How many zip codes on average have at least a one million dollar value? Which zip code on average has the least expensive house prices?