**Data Challenge**

The dataset for kings county sales prices from May 2014-May 2015 (taken from Kaggle dataset/ also used in Coursera), which is divided into a test and training CSV.

We see that there’s many features in the data that can help us predict the price of a house in this area.

**First Part** –

On the training data do a regression for different features and different feature sizes

Such as

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

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

**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 the data the best? Why? Add visualizations

**Second Part-**

Implement a regression to prevent overfitting. Is this a better fit?

.