# Kings County Housing Price Prediction

## Project Objective:

You'll clean, explore, and model this dataset with a multivariate linear regression to predict the sale price of houses as accurately as possible.

## Data Review (need to do more research on meaning of variables)

Data Dictionary:

Column Names and descriptions for Kings County Data Set

\* \*\*id\*\* - unique identified for a house

\* \*\*dateDate\*\* - house was sold

\* \*\*pricePrice\*\* - is prediction target

\* \*\*bedroomsNumber\*\* - of Bedrooms/House

\* \*\*bathroomsNumber\*\* - of bathrooms/bedrooms

\* \*\*sqft\_livingsquare\*\* - footage of the home

\* \*\*sqft\_lotsquare\*\* - footage of the lot

\* \*\*floorsTotal\*\* - floors (levels) in house

\* \*\*waterfront\*\* - House which has a view to a waterfront

\* \*\*view\*\* - Has been viewed

\* \*\*condition\*\* - How good the condition is (Overall )

\* \*\*grade\*\* - overall grade given to the housing unit, based on King County grading system

\* \*\*sqft\_above\*\* - square footage of house apart from basement

\* \*\*sqft\_basement\*\* - square footage of the basement

\* \*\*yr\_built\*\* - Built Year

\* \*\*yr\_renovated\*\* - Year when house was renovated

\* \*\*zipcode\*\* - zip

\* \*\*lat\*\* - Latitude coordinate

\* \*\*long\*\* - Longitude coordinate

\* \*\*sqft\_living15\*\* - The square footage of interior housing living space for the nearest 15 neighbors

\* \*\*sqft\_lot15\*\* - The square footage of the land lots of the nearest 15 neighbors

Questions: how to model geolocation with house price?

For floors, condition, view, treat them as numerical value, is that okay?

I decide to drop view, yr\_renovated.

For grade, condition, is it better to use binning or keep as they are now?

Will right categorization improve model results?

Blog post: how feature transformation improved model results and why?

First didn’t bin categorical variables.

Then group them together.

Compare the adjusted R-square

Floors can do either way. New kind of house if floors more than 5.

Grade: if less than 10 unique values, probably categorical.

Zipcode: classification (svm); focus on lat and long. Lat and long. Leave out.

Explore before wiping out the outliers using Z score.

Outlier.

Normalized, categorized.

Video present findings to business people with slides - 5 min

Jupyter + non technical piece. Focus on one aspect – data visualization and data transformation.