# DS 6021 Project Proposal: Analysis on King County House Prices

Group 8: Dima Mikhaylov, Connie Cui, Peumali Surani Withanage, Mani Shanmugavel

Computing ID: agp7dp, qqv3uu, upp2dh, fdf7gn

## Introduction and Data

For our project, we have selected a housing dataset of properties located in King County, Washington in the United States and spans from May 2014 to May 2015. King County is the most populous of the three counties in the state of Washington and approximately two thirds of its population resides in the Seattle suburbs. This dataset was obtained from Kaggle (https://www.kaggle.com/harlfoxem/housesalesprediction?select=kc_house_data.csv) and contains approximately 21,600 observations with variables such as house price, number of bedrooms, number of bathrooms, number of floors, whether or not the property is waterfront, the year the property was built, square footage of the lot, etc. We hope to utilize this dataset to help analyze the relationships between these variables for future insight into predicting house prices.

| bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | grade | sqft_above | sqft_basement | yr_built | yr_renovated | sqft_living15 | sqft_lot15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1.00 | 1180 | 5650 | 1.0 | 0 | 0 | 3 | 7 | 1180 | 0 | 1955 | 0 | 1340 | 5650 |
| 3 | 2.25 | 2570 | 7242 | 2.0 | 0 | 0 | 3 | 7 | 2170 | 400 | 1951 | 1991 | 1690 | 7639 |
| 2 | 1.00 | 770 | 10000 | 1.0 | 0 | 0 | 3 | 6 | 770 | 0 | 1933 | 0 | 2720 | 8062 |
| 4 | 3.00 | 1960 | 5000 | 1.0 | 0 | 0 | 5 | 7 | 1050 | 910 | 1965 | 0 | 1360 | 5000 |
| 3 | 2.00 | 1680 | 8080 | 1.0 | 0 | 0 | 3 | 8 | 1680 | 0 | 1987 | 0 | 1800 | 7503 |

**Figure 1: Part of the 'x' dataframe containing all predictor variables**

Before we can actually begin our analysis and model fitting on our data, we first need to conduct some data cleaning. For our team, this process includes checking the dataset for missing values and outliers and deciding on how to proceed with them moving forward. We also plan to decide which variables we plan to drop in our dataset moving forward, and randomize and

shuffle our data. We will most likely impute any missing values with the median or mean of the variable in order to preserve all our current data.

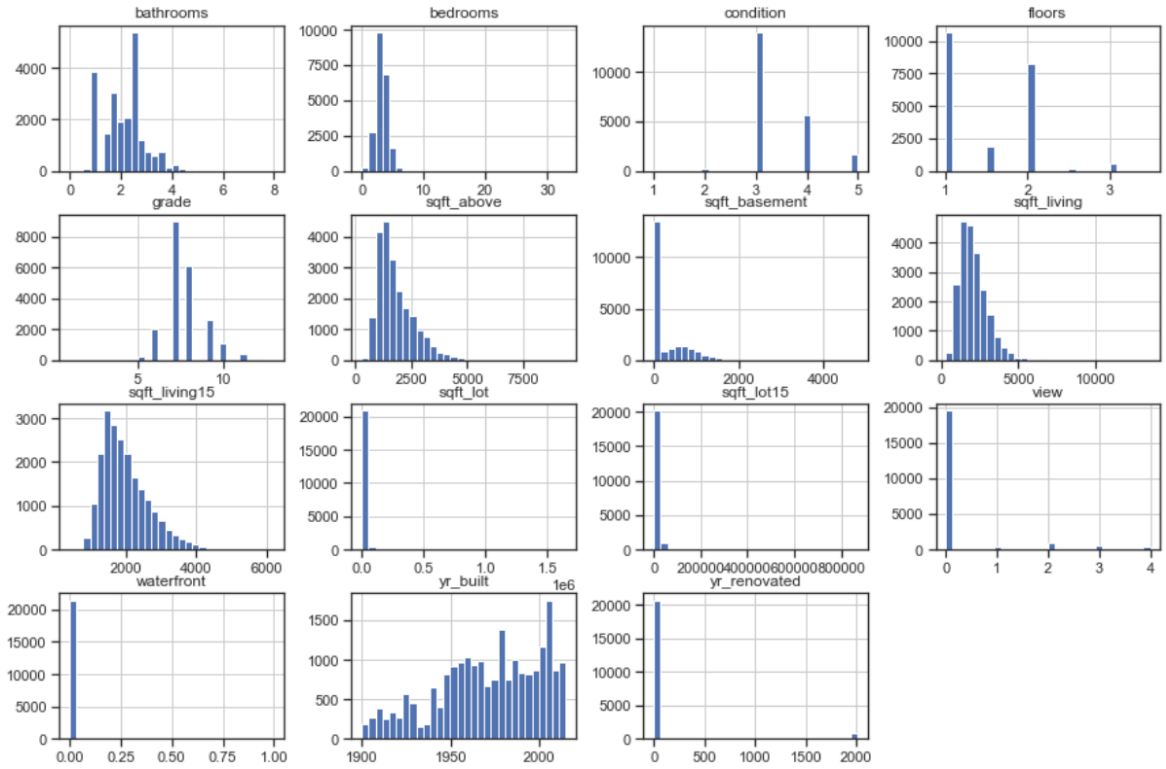


**Figure 2: Histograms of all potential predictor variables in the dataset**

## Objectives and Goals

The following are the objectives and goals of the project:

1. Data wrangling for efficient data analysis, i.e. transforming data and dealing with missing data (imputation);
2. Using data visualizations to explore how price is related to the other variables;
3. Build a multi-linear regression model with the house price of the unit area as the response variable;
4. Assess regression assumptions using graphical and inferential methods;

5. Assess multicollinearity and its impact on the variance of regression coefficients;
6. Build a logistic regression model to classify high-price and low-price homes;
7. Determine the predictor variables that would associate classifying high-price and low-price houses;
8. Identify the predictive ability of the logistic regression based on confusion matrix and ROC curve.

**Methodology**

As shown on the plot below, some of the selected predictor variables are strongly correlated with each other:

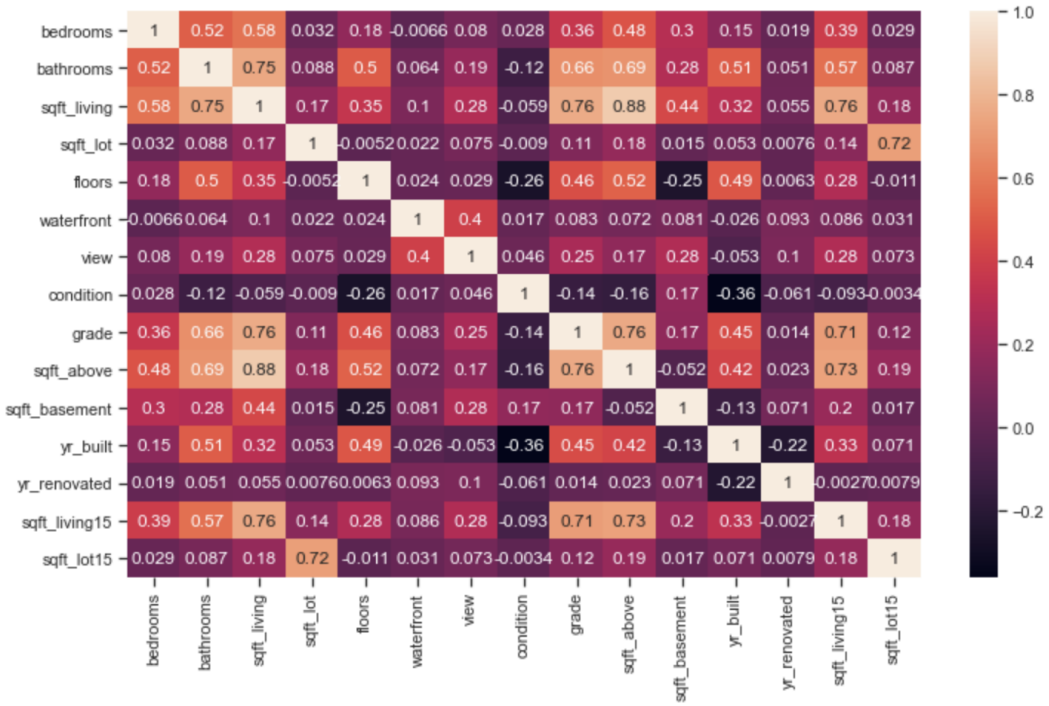


**Figure 3: Pairwise correlations of the selected predictor variables**

Multiple linear regression (MLR) will be used to examine relationships between several predictor variables and response variable price. First, MLR requires well behaving data, namely observations are assumed to be independent (shaffling) and missing values need to be either

imputed or eliminated. Second, residuals of a fitted MLR model need to have constant variance and overall mean of zero. Scatter plot of residuals, autocorrelation plot, and normal probability QQ-plot will be used to gain insights. Possible remedial measures shall include Box-Cox response variable transformations, as well as predictor variables transformation for linearity, presumably with some kind of convex functions. Third, VIFs and partial F statistics will be used to test for possible multicollinearity. Fourth, high leverage observations will be assessed using measures of influence. Finally, automatic search procedures, such as forward selection, will be used to choose the best subset model in terms of Akaike Information Criterion (AIC). Analysis of Variance (ANOVA) will be used to assess the levels of variability in the final MLR model and test for significance of various combinations of the predictor variables. The best estimated MLR coefficients will be reported in terms of corresponding p-values, as well as confidence intervals.

Given that the dataset for the project is ungrouped data, we will divide housing prices into two categories; high price ($1 million and up) and low price (below $1 million). We create an indicator variable as 1 being the high price and 0 being the low price. Hence, the price indicator variable will be our response variable, and checking the odds of being a high-price house is associated with other predictor variables. Moreover, we will examine each predictor variable for possible associations by two-way tables and visualization techniques.

In order to assess the predictive ability of the model, we will split the data set into two as training and testing datasets. After identifying the set of predictor variables, we use the training dataset to estimate the logistic regression, and for each coefficient , z-score and p-value will be examined to determine their significance. We may want to consider more than one combination

of predictor variables, $\Delta G^2$ = null deviance - residual deviance will be used to determine the usefulness of one model over the other.

To evaluate the model's predictive ability, we use the model in the testing dataset to identify the true positives, true negatives, false positives, and false negatives. The confusion matrix can be used to achieve this objective. Moreover, we will use the ROC curve and area under the ROC curve (AUC) to determine the model's predictive ability. The selected classification model will be used to assess feature importance to determine variables that contribute the most to predicting high price homes. Preliminary features are shown on the plot below.

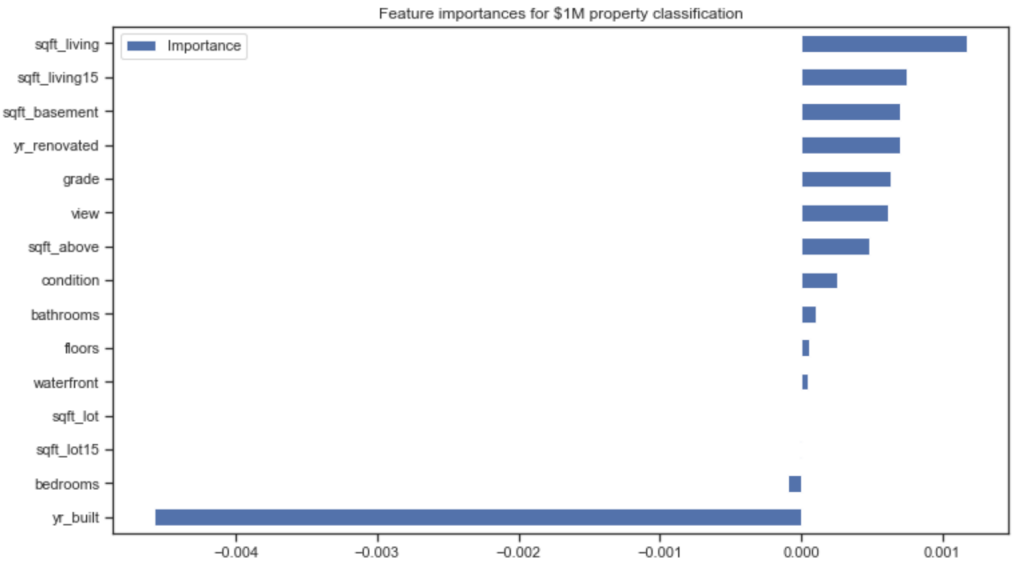


**Figure 4: Feature importances of selected predictor variables**