# Predicting house prices with the king county housing dataset

By Grace Anyango

# Overview

- I have been tasked to analyze data of King County houses based on certain features. So my goal is to analyze sale prices of houses so that they can use to make useful decisions. I have looked at the data and after careful examination I have decide on the features that I will use for this particular project for example building grade, square feet of living space, location.

# The Data

- This project uses the King County House Sales dataset, which can be found in kc_house_data.csv in the data folder in this assignment's GitHub repository. The description of the column names can be found in column_names.md in the same folder. As with most real world data sets, the column names are not perfectly described, so you'll have to do some research or use your best judgment if you have questions about what the data means. This file contains data for 21597 homes built in KC from 1900 to 2015. Each home in the set contains information regarding features such as zip code, square footage, number of bedrooms and bathrooms, number of floors, condition and more.

# Process

- Data understanding
- Data cleaning and preprocessing
- Building Linear models
- Interpreting the results
- Recommendations and conclusions

# Correlations

## Correlations with Price

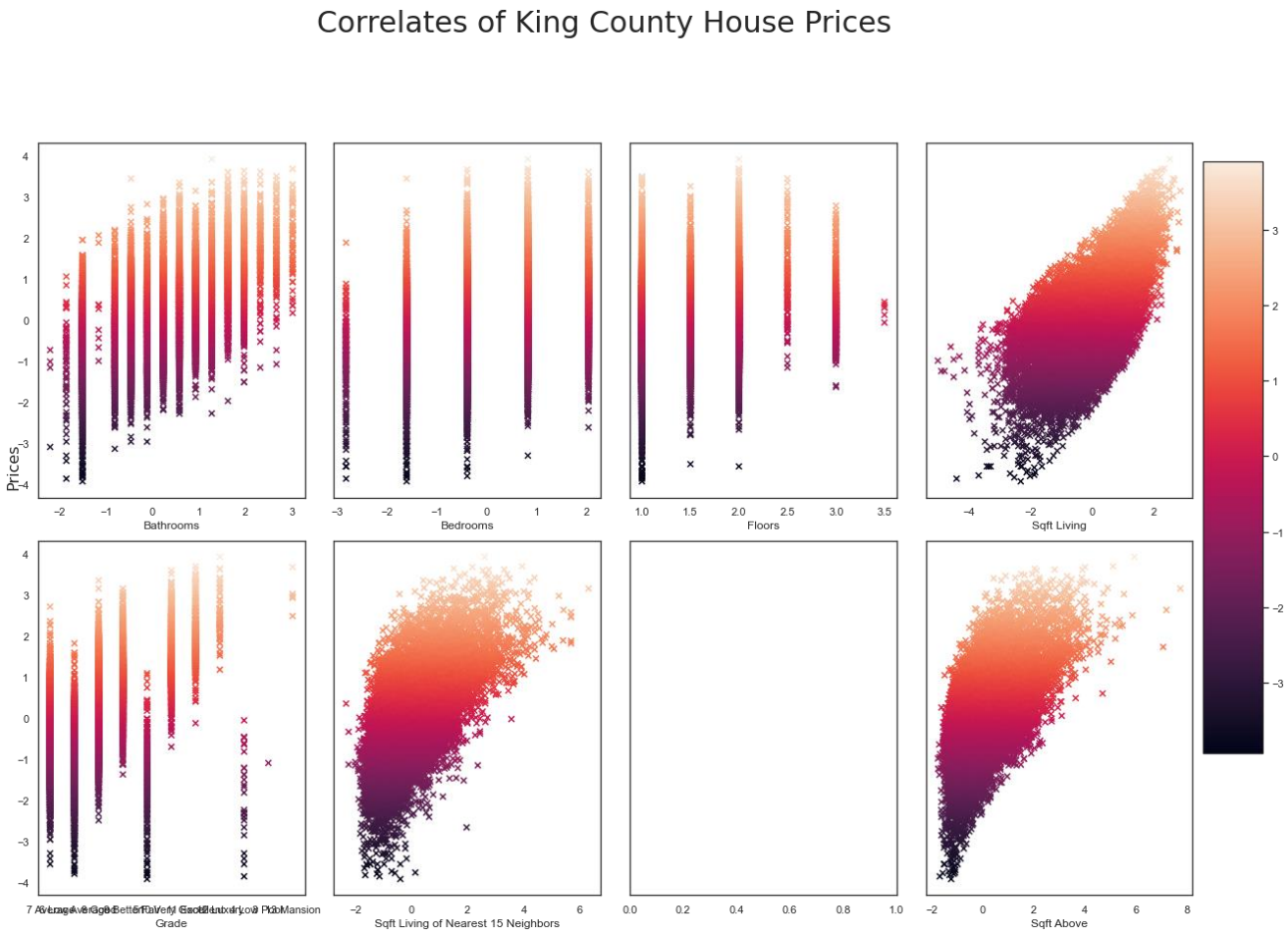
| | Correlations | Features |
|---|---|---|
| **2** | 0.677596 | sqft_living |
| **4** | 0.593674 | sqft_living15 |
| **3** | 0.578363 | sqft_above |
| **1** | 0.489138 | bathrooms |
| **0** | 0.302105 | bedrooms |

- Notes

Features that were highly correlated with price were considered for inclusion in the model.

# Notes

View, bedrooms and floors were excluded from the model



# 1.Which features are highly most correlated with price?

- Generally, any correlation above .7 is considered high. While there are no correlations with price above .7 in the dataset, there are several features with moderately strong correlations. Sqft_living, sqft_living15, sqft_above and bathrooms have the highest correlations with price.

# Question 2: Which features have the strongest correlations with other predictor variables?

| | Correlations | Features | |
|---|---|---|---|
| **0** | 0.868369 | [sqft_living, sqft_above] | |
| **1** | 0.868369 | [sqft_above, sqft_living] | |

# Question 3:What combination of features is the best fit, in terms of predictive power, for a multiple regression model to predict house prices?

- Grade, sqft_living and bathrooms are the best fit for a multiple regression model. These features are highly correlated with price, have relatively low multicollinearity, and can together account for more than half of the variability of price.
- All multiple regression assumptions are satisfied with these features included.

# Conclusions and limitations

**CONCLUSIONS**

- I managed to build a multivariate predictive model with an  R Squared of nearly 46%.
- Together, square footage, grade and bathrooms are the best predictors of a house's price in King County. Homeowners who are interested in selling their homes at a higher price should focus on expanding square footage and improving the quality of construction. When expanding square footage, homeowners should consider building additional bathrooms, as this analysis suggests that number of bathrooms is positively related to price.
- All the P values are below our alpha that is 0.05 so we reject null hypothesis and conclude that there is Significant relationship between target variable price and independent variables.

LIMITATIONS

- Given that some of the variables needed to be log-transformed to satisfy regression assumptions, any new data used with the model would have to undergo similar preprocessing. Additionally, given regional differences in housing prices, the model's applicability to data from other counties may be limited. Given that outliers were removed, the model may also not accurately predict extreme values.
- Future analysis should explore the best predictors of the prices of homes outside of King County, as well as homes with extreme price values.

Thank you