Analysis of Ames Home Sales

Prepared by Jean Brewster

Table of Contents

Background / Problem Statement

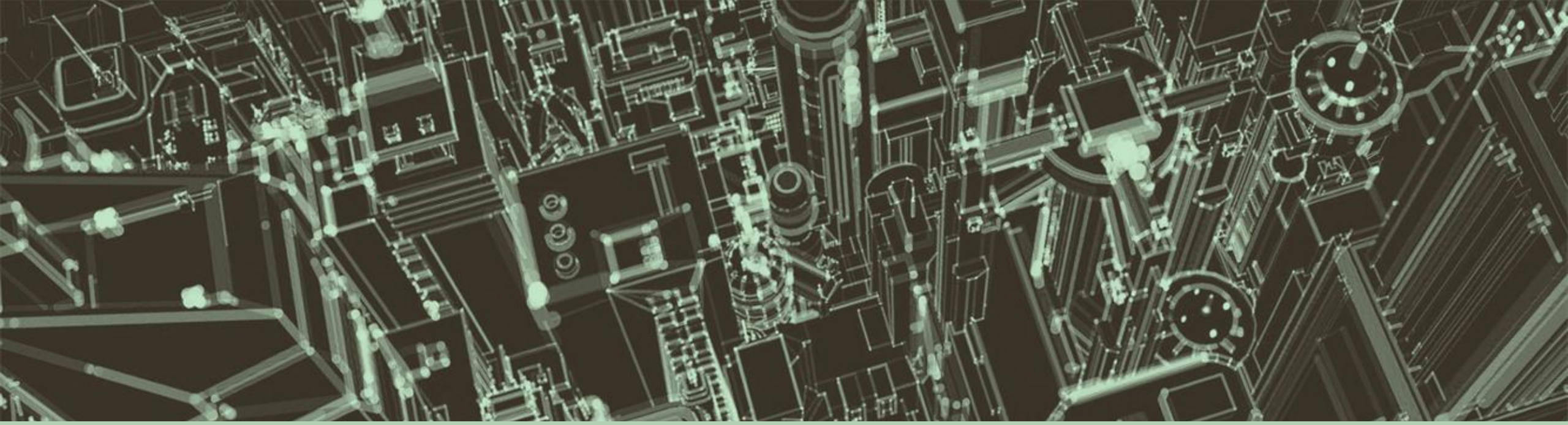
- I am consulting with a new real estate investment firm in Ames, IA. The firm wants to understand key predicters impacting the prices of home sales in the market between 2006-2010. Their objective is to acquire undervalued homes for the purpose of selling them for a profit.

Methodology

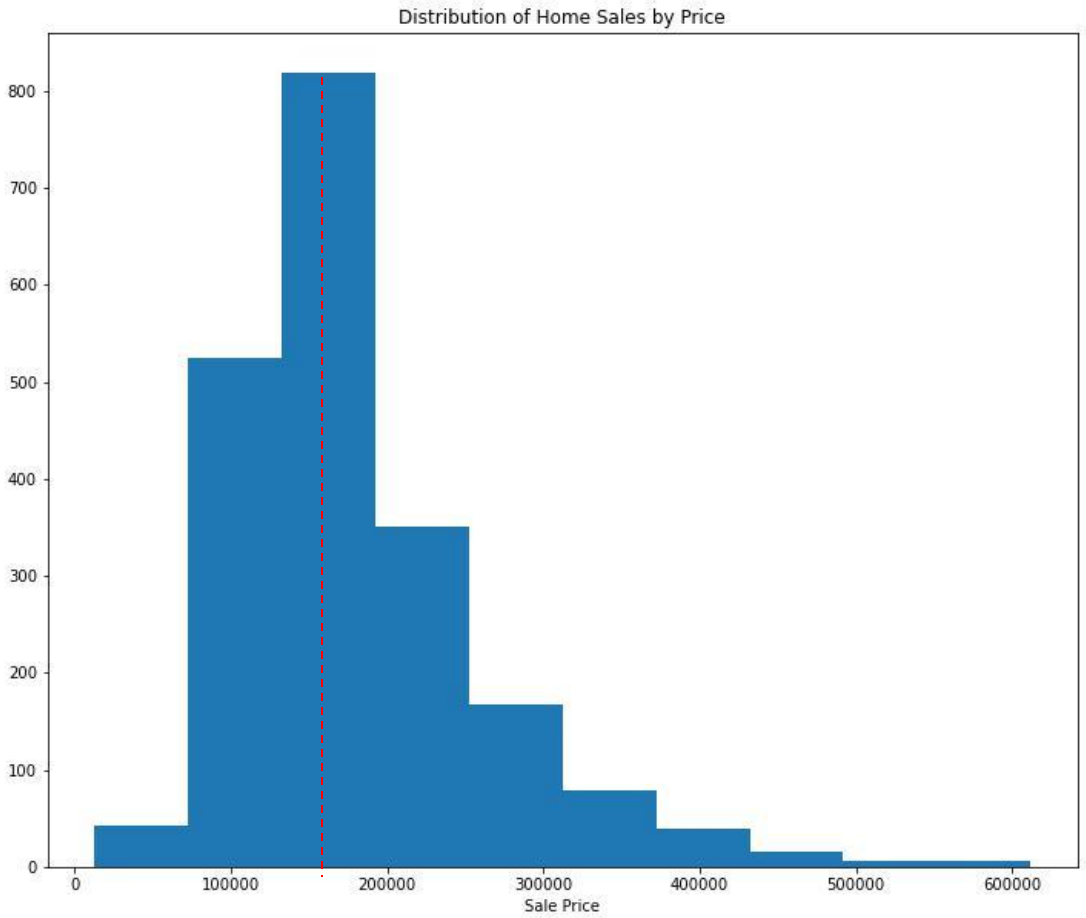
- Obtained data from the Ames Assessor's Office for individual residential properties sold in Ames, IA from 2006 to 2010
- Cleansed and performed exploratory data analysis (EDA) on numerical and categorical features:
  - Checked for missing values across dataset
  - Converted categorical features into numbers and included statistically significant variables in model
  - Included numerical variables with a strong correlation to sale price
- Developed and evaluated regression models to predict home sale prices

Model Development / Evaluation

- Identified numerical and categorical features for inclusion in model and imputed missing values
  - Included the mean averages for missing values for numerical features
  - Included the mode for missing values for categorical features
- Generated several regression models using the training data. Utilized following to evaluate the models
  - Train-test split
  - Cross-validation
  - R-squared



Data Observations

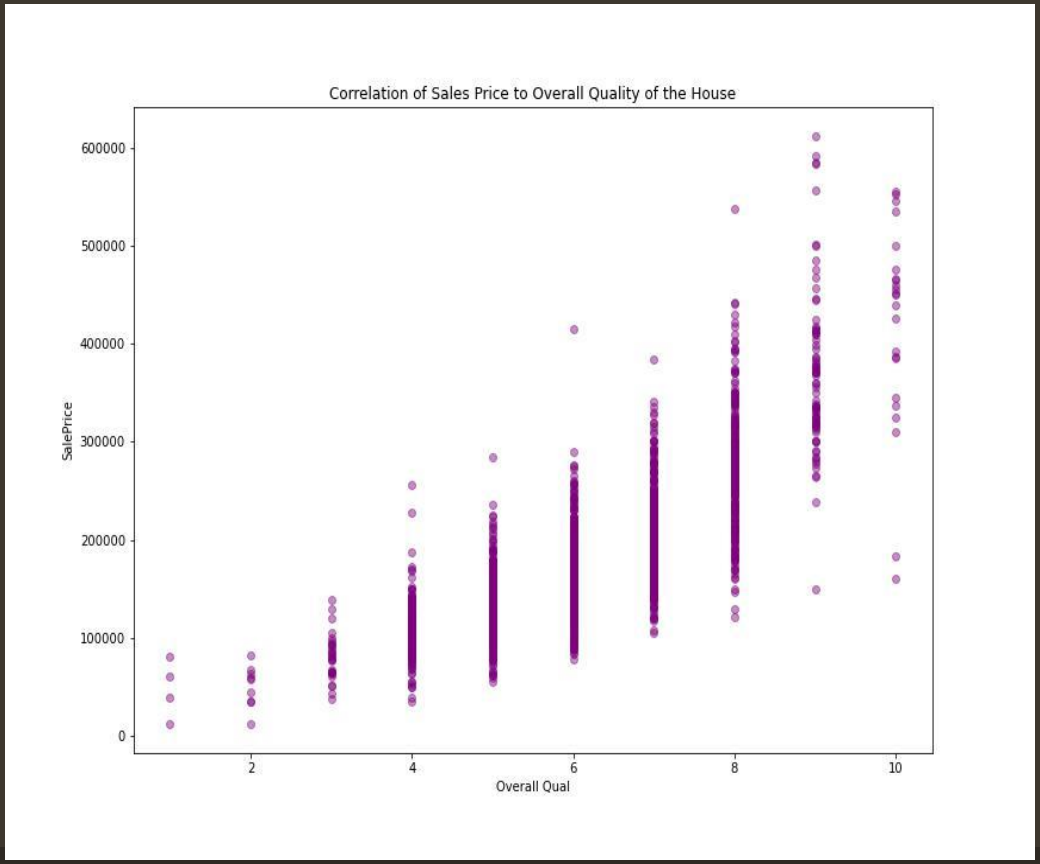


Price of home sales in Ames is skewed right

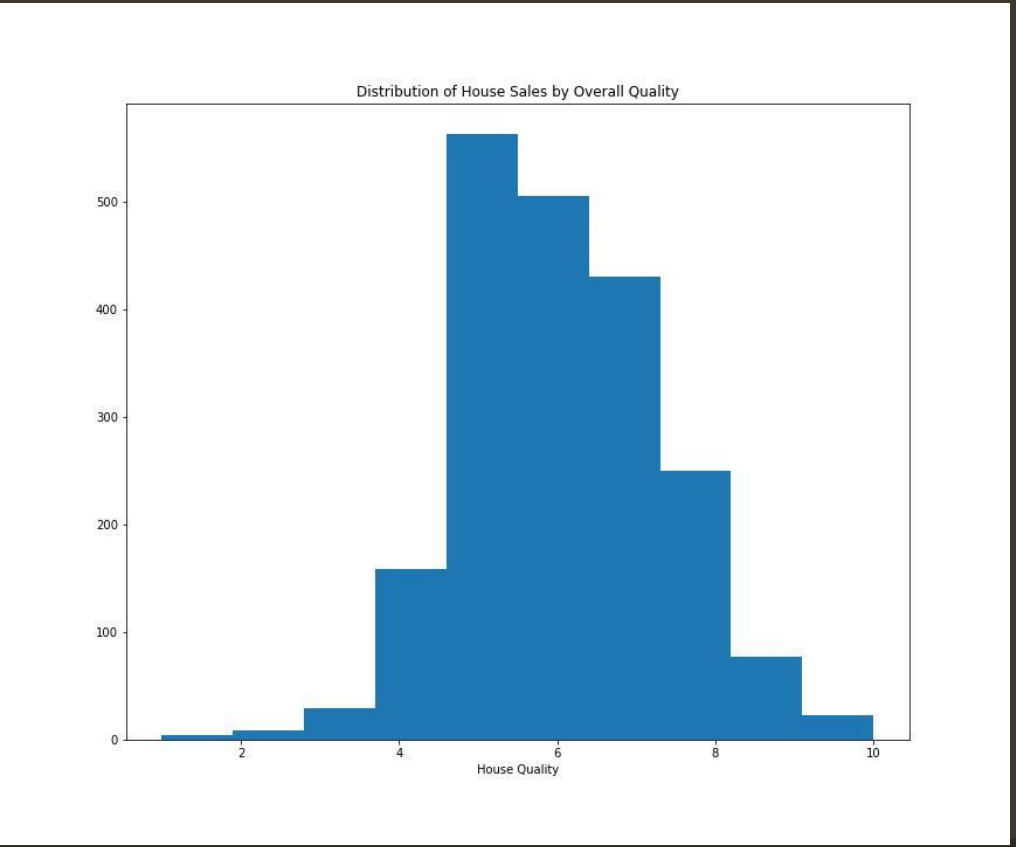
- Mean of sales prices is $182K
- The most expensive house sold for $611K

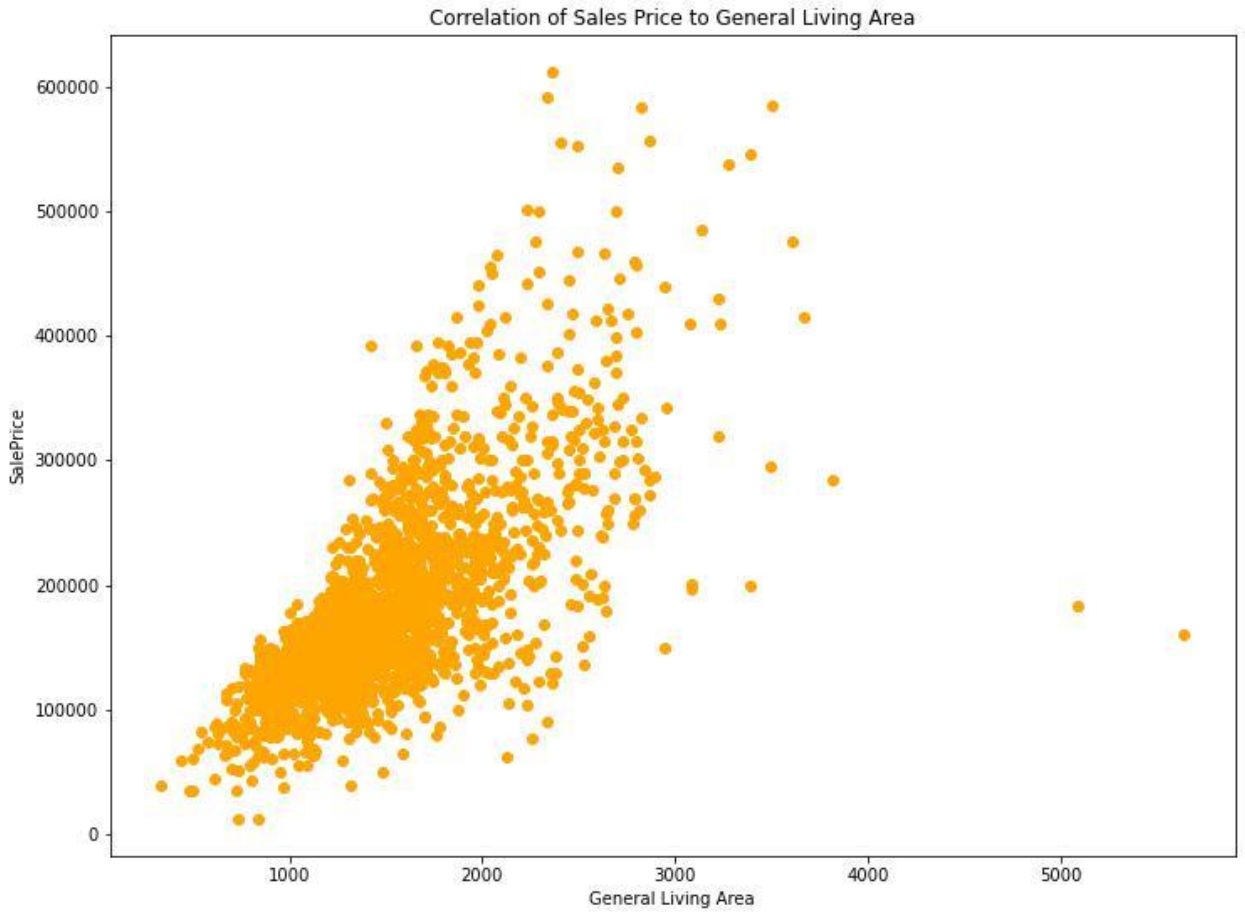
Overall quality of the house has the strongest correlation to sale price

As to be expected, the higher the overall quality, the higher the sale price



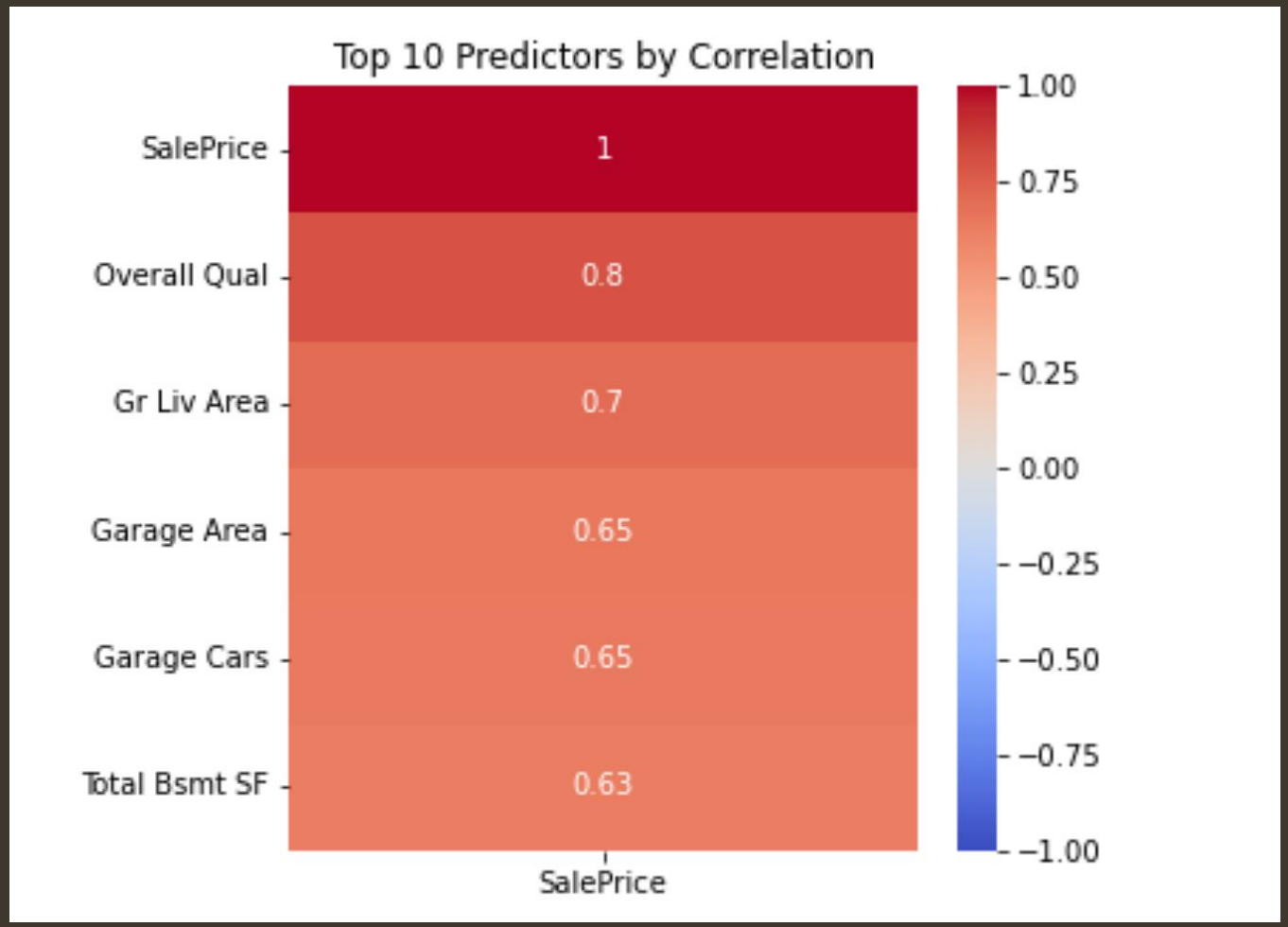
Most of the homes that sold had overall quality between 5 (Avg) – 8(Very good)





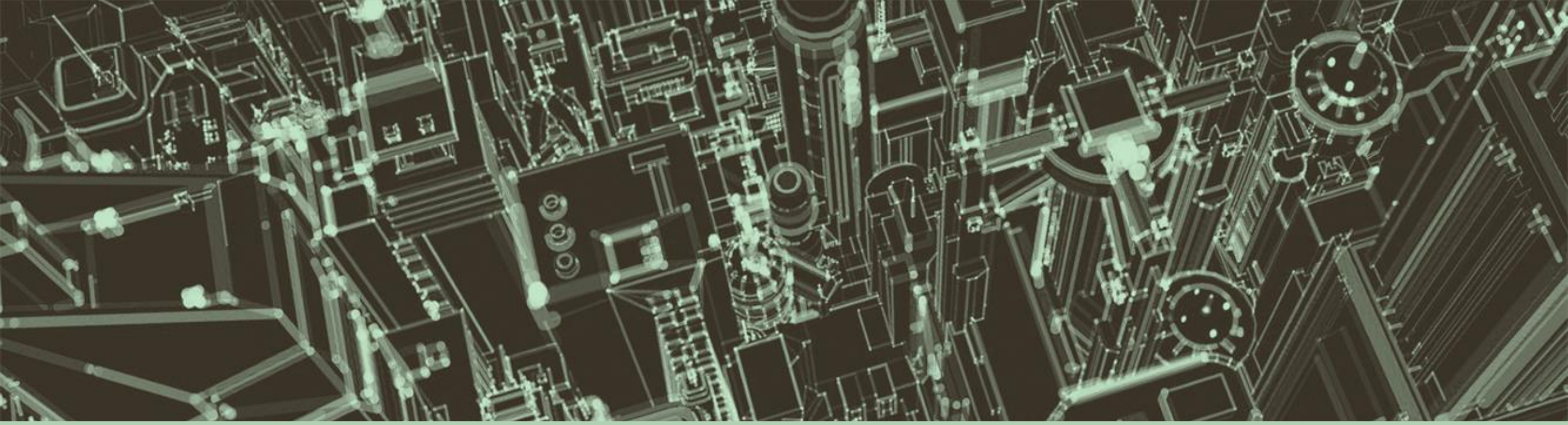
General living area has a mostly linear relationship with sale price

- Sale price increases as the general living area or square feet of a house increases



Top Ten Predictors based on correlation

- Overall quality of the house, living area, garage and total basement have strongest positive correlation to sale price
- The quality of the following features have strong negative correlations and impact the sale price:
  - Exterior of house
  - Kitchen
  - Basement
  - Finished / unfinished garage



Results / Recommendations

Model Scores / Evaluation

- Four models were developed and applied to cleaned train and test data. The test data was used only to produce results to Kaggle
- Lasso and Ridge models yielded the best results and have good predictive value, with more than 92% of variability in home sale prices explained by the independent variables

R-Squared	Model 1 (Baseline*)	Model 2 (All High Correlated**)	Model 3 (Lasso)	Model4 (Ridge)
Train	0.8178	0.8352	0.9228	0.9225
Test	0.8180	0.8211	0.8575	0.8612

* Strongly correlated numerical variables (X >= 50%
** Numerical & categorical predicters with strong correlation

Conclusions / Recommendations

- Use of the Lasso and Ridge regression models transformed the independent variables by shrinking unnecessary features and removing extreme outliers, yielding an R-squared value of 0.92.
- Identifying properties with overall good quality and high living area (square ft) will lead to the highest home prices.