**House Prices in Ames, Iowa**

**Group Members**

*August 10, 2020*

Aparna Nagarajan

Emmanual Ogwal

Sirisha Yellajosyula

Yezhou Jiang

Sean Xu

Steven Choi

**Introduction & Objectives**

Over the last 20+ years, real estate prices in both North America and globally have experienced unprecedented growth. Low interest rates and unemployment rates have led to a boom in residential house prices creating new financial barriers to entry. By analyzing real estate prices from Ames, Iowa (sourced from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>), this project aims to provide insight into the predictive value of housing related attributes to predict home prices while examining questions such as,

1. The average house price
2. Trends in housing data
3. The contribution of temporal (e.g., year built, month of sale, etc) and structural attributes to house prices

We’ve all heard that “Location, location, location” are the three most important factors in finding and unlocking the value of real estate. Here, we will examine and conclude on the validity of this statement and more.

In the following pages, we will share,

* Approach to prepare the Ames, Iowa housing data to support analysis and modelling which includes,
  + How we went about addressing data gaps for both quantitative and qualitative/categorical attributes.
  + Featured engineering to enhance and improve analysis and model accuracy.
* Multivariate analysis to identify independent variables correlated to house prices (our dependent variable). This includes correlation analysis of both quantitative and qualitative attributes.
* And finally, our conclusion.

Based on our initial observations and our own experiences with real estate, we hypothesize house prices are best predicted by,

* Neighbourhood
* Total square footage of the living space
* Age of the home

**Data Preparation**

Data collection and quality

The dataset used in the analysis of the Ames house prices was obtained from Kaggle online data science community platform. “*Kaggle* is the world's largest data science community with powerful tools and resources to help you achieve your data science goals” (Kaggle, 2020). A total of 1460 home prices records were retrieved from the training set in Kaggle to aid our analysis. To access this dataset, a short online registration with Kaggle company was required after which access was granted.

These records have 81 features recording data regarding various aspects of a house. The overall quality of these features are good. However, some of the dataset attributes required thorough cleaning and validation and sometimes exclusion from the assessment.

Handling missing values

There are 14 attributes with missing data. Preliminary analysis of the rate of missing data for these features revealed 5 of those with more than 30% missing data namely, PoolQC, MiscFeature, Alley, Fence and FireplaceQu. Imputing missing values for the above would not be an accurate representation of the dataset. Hence these were dropped from further analysis.

This left us with 3 numerical and 9 categorical attributes with missing values.

* The 9 categorical attributes were categorized into 3 distinct groups; electrical, basement and garage related attributes.
  + An observation was made that a missing value in bsmtQual, bsmtCond, bsmtExposure, bsmtFinType1, and bsmtFinType2, actually means there is no basement, which can be proved by an indicator of total basement square feet (TotalBsmtSF) being 0. All such missing entries were marked as ‘No Basement’.
  + Similarly, missing values in GarageYrBlt, GarageFinish, GarageQual, GarageCond and a 0 garage area actually means there is no garage. All such entries were marked with ‘No Garage’.
  + Rest of the truly missing electrical, basement and garage data was filled with their respective modes.
* The 3 numerical features with missing values were GarageYrBlt, LotFrontage and MasVnrArea.
  + All the missing GarageYrBlt were due to the house not having a garage which was marked as ‘No Garage’.
  + LotFrontage was analyzed to be related to the neighborhood and the number of cars. More cars the garage can fit the more LotFrontage it has. Also every neighborhood that has a fairly distinct spread of LotFrontage values. Hence a grouped median was used to impute missing LotFrontage entries.
  + MasVnrArea attribute is inferred to be 0 for those houses with no masonry veneer area. This was detected by using the attribute MasVnrType with values None. Rest of the missing entries were filled with the median values.

Engineering new attributes

Based on the observation from the raw dataset and real life house purchasing experience, it was assumed that the total square feet of the property, total bathrooms, swimming pool, and fireplace also has an impact on the house price. Hence, 'TotalSF', 'Total\_sqr\_footage', 'Total\_Bathrooms', 'Total\_porch\_sf', 'haspool', 'has2ndfloor', 'hasbsmt' and 'hasfireplace' attributes are introduced for further analysis.

After dropping and imputing missing values and engineering new attributes we arrived at a clean dataset with 1460 rows and 85 features which was used in the analysis and modelling phase.

Tools and codes used for data preparation

Various python and pandas functions were used in data preparation. To name some, drop, groupby, mode, median. Lambdas were extensively used to stitch it all together. We used the “drop” function to exclude ineligible variables from our analyses. The code used for data preparation can be found in Appendix 1.

Challenges faced during the project data validation phase

The dataset in itself has a lot of attributes which required a lot of reading and understanding of the domain to analyse patterns for imputing missing data. Our dataset is solely structural and temporal in nature which captures just the structural characteristics of the house such as lot size, living area and such. The data lacked other attributes (which when present) that could be useful in our analyses. Factor variables that were outside the structural attributes of the houses in Ames, Iowa such as the inflation rate in the USA at the time of our analysis was lacking. A more robust analysis could also be achieved if we had comparable house prices dataset of a similar economic region to Ames, Iowa. A key missing data would be the population of the city, and the rate of in-migration of settlers from other US cities and states. If these data were available, we could also observe how internal immigration would drive up home prices in Ames, Iowa as a factor that is not entirely structural. Lastly, our analysis focuses on a relatively short period of time and may not be representative of house prices in a longer time frame.

**Analysis**

Data analysis in this project was done in three chronological phases. We’ve looked into the trend of data, correlations between different attributes, used different methods to analyse and find patterns.

Univariate data analysis

Under this phase, exploratory and descriptive assessment of the Ames house prices data was conducted. Descriptive statistics such as the mean and median house prices were assessed and reported. A trend analysis based on the year built of the houses was assessed in order to produce the relationship of house age and house value in Ames city. The average house sale price in Ames city was $180,921.19(range $34,900 to $755,000, std. dev=$79,442).

Bivariate data analysis

The bivariate assessment of the Ames house prices was centered around finding out two-way associations between the factor attributes of the houses and their sale prices. Pairwise correlations and correlation matrix between the target variable (sale price) and the covariates (factor variables/attributes) were assessed and reported. Figure 1 below shows the heatmap of correlations between the variables which includes numerical and categorical variables.

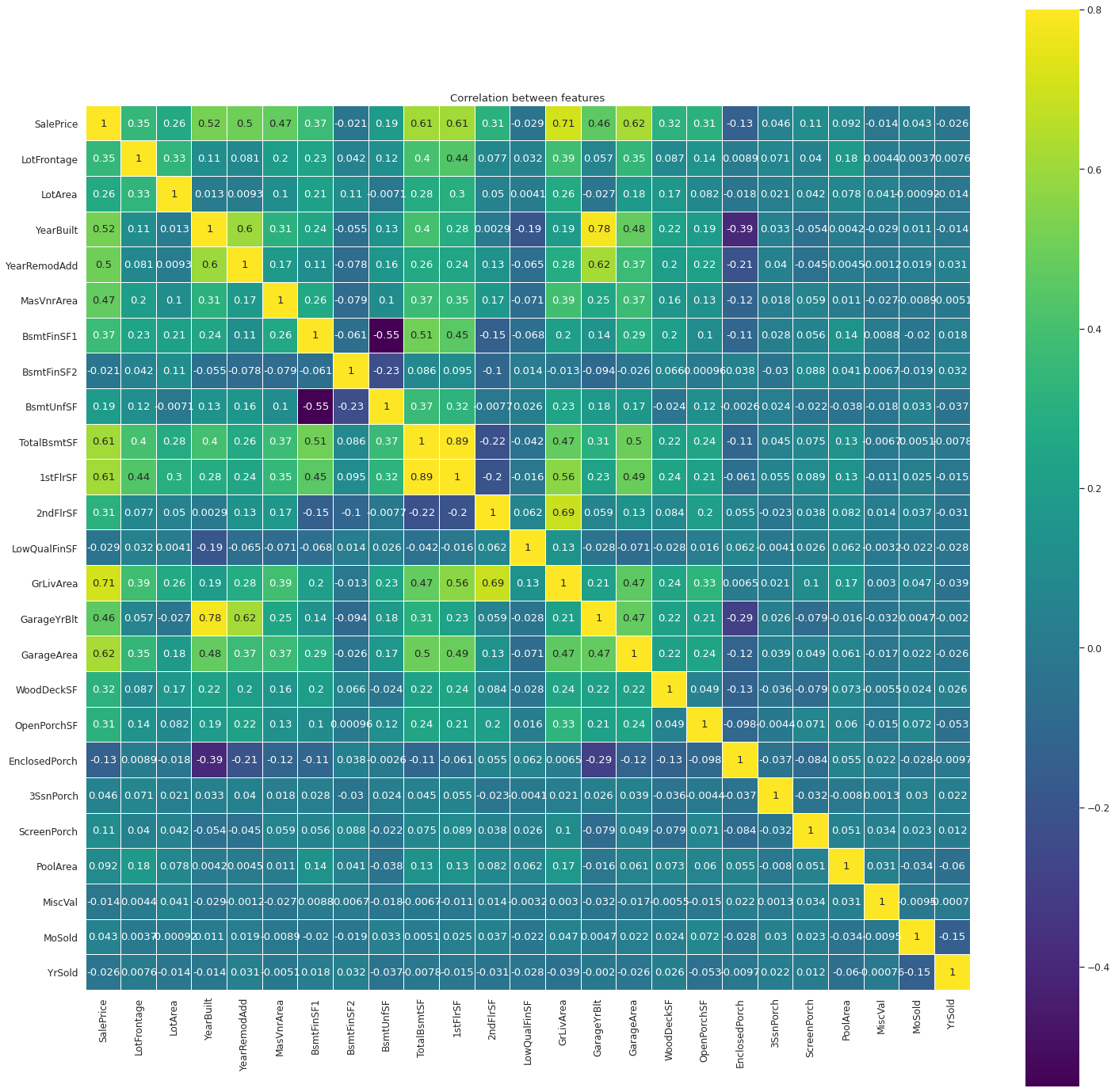


Figure1.

In total, there are 79 explanatory variables (excluding ID, and SalePrice). But it is too much work to understand each of them, and also not all of these variables will be selected as model inputs. Hence, based on the correlation level between target variables and explanatory variables, we only need to understand a few of explanatory variables.

From the correlation analysis of numerical variables, the correlation between sale price and year built, sale price and year renovated are telling that newly built or newly renovated homes seem to have higher prices than earlier ones. Except for YearBuilt and YearRemodAdd, TotalBsmtSF, 1stFlrSF, GrLivArea and GarageArea also showed positive correlation with sale price as well. The increase of these four areas increases the sale price.

For the categorical variables, OverallQual, TotRmsAbvGrd and Neighborhood are all having correlation with sale price.

Multivariate data analysis: regression analysis

Under this section, a regression analysis was performed in order to find out the predictors of house sale prices in Ames, Iowa. The analysis contains 4 dimensions that try to find out how different segmented variables, which were already proved having correlation with sale price, impact the sale price.

*Dimension 1: House overall condition: YearBuilt, OverallQual, YearRemodAdd*

The house overall condition determined by YearBuilt, OverallQual, YearRemodAdd therefore since they have correlation with SalePrice, the analysis between these variables are necessary. When these variables have been visualized, the impact was observable. Figure 2 and 3 below show how these three attributes impact SalePrice.

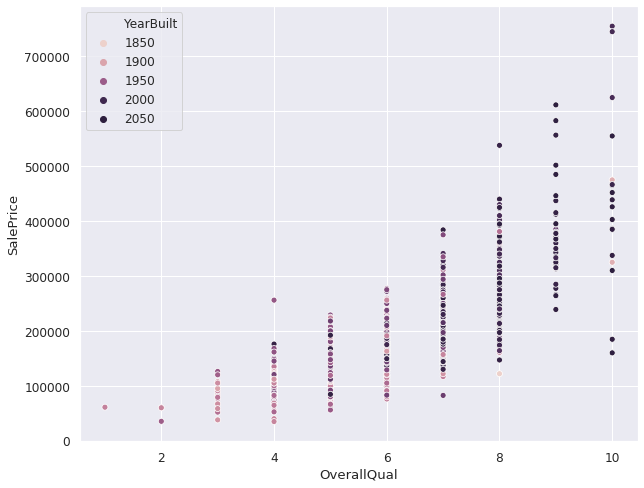


Figure 2.

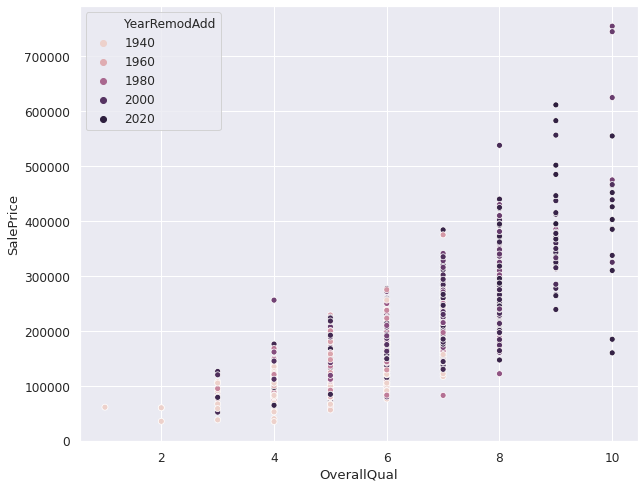


Figure 3.

Based on the observation and analysis, we see recently built or remodelled homes with better overall quality result in higher sale prices.

*Dimension 2: House inside condition: GrLivArea, TotalBsmtSF, 1stFlrSF, TotRmsAbvGrd, BsmtQual, KitchenQual*

If looking at the detailed attributes, the house inside condition, the outside condition and the community environment. These dimensions could impact the sale price of a house as well. This section provides the analysis result of the house inside condition. Again, please refer to Figure 4, based on the visualization and the impact is obvious, bigger living area(GrLivArea) and higher kitchen condition(KitchenQual) is positively impacting the SalePrice, which means the SalePrice will be higher. Meanwhile, bigger basement area, higher basement ceiling could boost the sale price as well.



Figure 4

*Dimension 3: House outside condition: GarageArea, Exterior1st, Exterior2st, ExterQual, GarageType, GarageCars*

These 6 attributes are all variables regarding the outside condition of the house, which are all correlated with the SalePrice either. According to the analysis, better exterior quality(ExterQual) has higher home prices. In particular the better quality vinyl and cement exterior have higher house price when looking into Exterior1st and Exterior2st. On the other hand different garage types(GarageType) reflect the different impacts on sale price. Attached and built in garages with a larger area seems to have a higher house price while detached garages seem to have lesser home prices though the area may be large. Figure 5 below shows how GarageType and GarageArea impact the sale price.

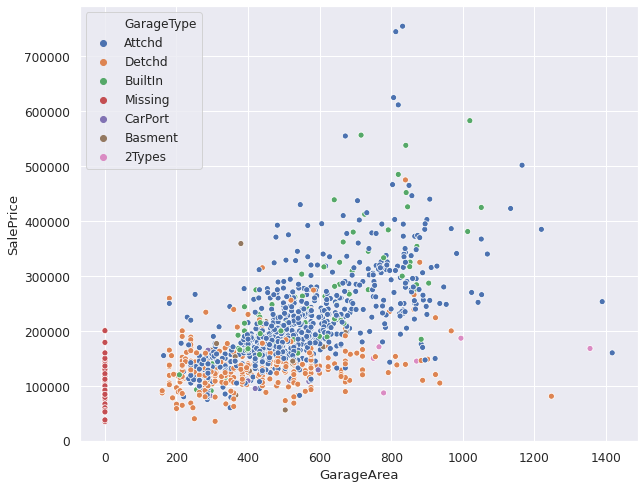


Figure 5

*Dimension 4: House community environment: Neighborhood, Condition1, Condition2*

The community environment definitely has impacts on the sale price of a house, which has also been proven that some of the attributes are correlated with sale price for example Neighborhood. Figure 6 below shows the scatter plot of how these attributes may or may not impact on sale price.

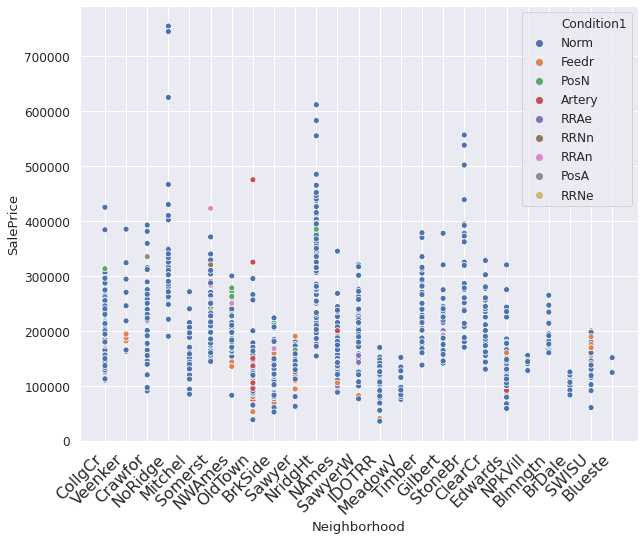


Figure 6

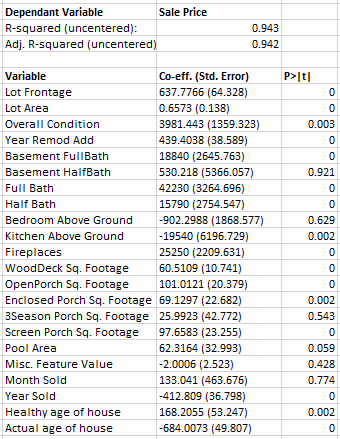
Based on the visualization, Neighborhood seems to have an impact on SalePrice with Condition1 and Condition2 which is referring to various proximities to main road or railroad, adjacent to arterial street or feeder street for example. But the impact is visually not obvious as the majority types of Condition1 and Condition2 appear at a very wide range of SalePrice.

**Model building**

Our intention is to run a regression model and deduce variable significance. Dummy variables were created for all categorical attributes. *Variance inflation factor (VIF)* for each variable was assessed to validate multi collinearity issues in the dataset. Using a threshold of 5 (VIF <5) we excluded all variables that were inflating the variance of the model. We ended up with 20 attributes that appear to reduce the variance.

OLS Model (Before transforming variables)

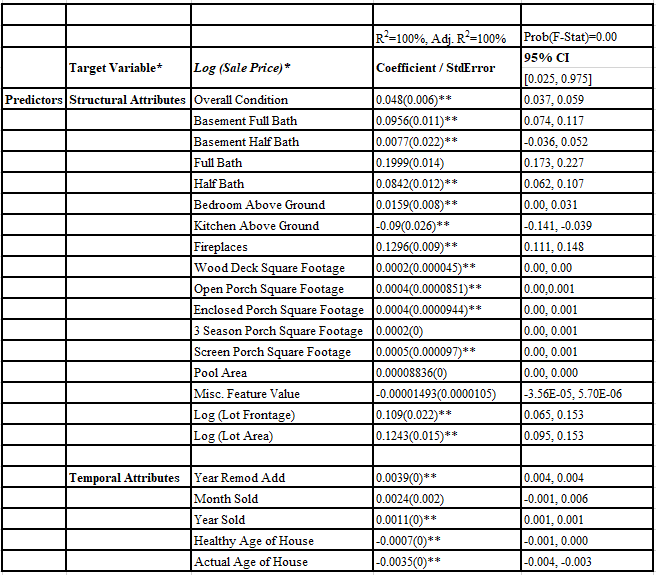
**Table1. OLS Regression Results for Predictors of House Sale Price in Ames, Iowa**

****

* Model Adjusted R Squared: 0.94
* Interpretation: Attributes like Lot Frontage Area, Lot Area, Overall Condition, Remodeled year, having a porch, Full bath (vs Half) in Basement and in the house, having the kitchen on the ground floor have contributed favorably to house price.

OLS Model (After transforming few continuous variables)

**Fig2.OLS Regression with Transformed Target and Predictors**

****

Notes: Prob(F-Stat)=General Model P-value; R2=R-Squared; \*Target Variable=Log(SalePrice); \*\*P-value<0.05: 95 C.I=95% Confidence Interval; StdError=Standard Error (in brackets); Adj=Adjusted

* Adjusted R Square value: 1.00
* Interpretation: Lot Frontage Area, Lot Area, Overall Quality, Overall Condition, Having a full bath (Basement and/or the house), additional features (Fireplace, Porch etc) all contributed positively to sales price. An interesting finding is that a remodeled house depreciates less significantly compared to a non-remodeled house i.e. all else being equal, a house which was renovated after it was built would command a higher price and depreciate slower compared to the non-remodeled house.

Regression Diagnostics

Running model diagnostics and heteroskedasticity tests, it was deduced that multicollinearity existed in the model. The Sale price, Lot Area and Lot Frontage variables were log transformed and subsequent model diagnostics resulted in acceptable limits of homoscedasticity.

**Conclusion**

The average home price in Ames, Iowa, which ranged from $34,900 and $755,000, was $180,921. Based on our analysis, home prices in Ames were influenced by,

* Neighbourhood
* Age of the home
* Quality of the home (e.g., overall quality, kitchen quality, exterior condition, etc); and
* Size of the home (e.g., total basement size, first floor size, ground living size and garage size)
* Additional features (e.g., full bath, fireplace, porch)

Neighbourhood

The number of neighbourhoods below and above the average price was evenly split with average prices in NoRidge, NridgHt and StoneBr exceeding $300,000. The average prices in BrDale, IDOTRR and MeadowV were closer to $100,000.

Age of the Home

Average home prices increased from $112,493 in 1900 to $269,220 in 2009.

Home Quality

Favourable quality attributes such as overall quality, kitchen quality, etc, showed correlations with price.

Size

Size attributes such as frontage lot size, ground floor size, etc, showed correlations with price.

Additional Features

Other home amenities such as having a full bath, fireplace, porch also exhibited correlation with price.

An interesting finding was the impact of renovations. A remodelled home (i.e., a house renovated after it was built) commanded a higher price and experienced slower depreciation compared with homes with no renovation history.

The above analysis was supported by the time and effort committed to data preparation. Although the overall quality of the original 1460 records and 81 attributes was good, the data did require cleaning, validation and enrichment. This required a good understanding and familiarity of both the data and real life comprehension of the subject (in this case real estate). This resulted in,

* Explaining away some missing values (e.g., when there was no garage, garage size was zero, other garage attributes were missing).
* Dropping of attributes (if more than 30% of the data was missing, we dropped the column)
* Inputting missing data using modes and medians
* Data enrichment by engineer attributes such as total square footage

From the 81 attributes, we dropped 5 attributes and added 9 for a total of 85 attributes.

At the beginning of our travels, we hypothesized that house prices were correlated with neighbourhood, total square footage of the living space and the age of the home and our final analysis concluded that quality and having additional features was also a factor.