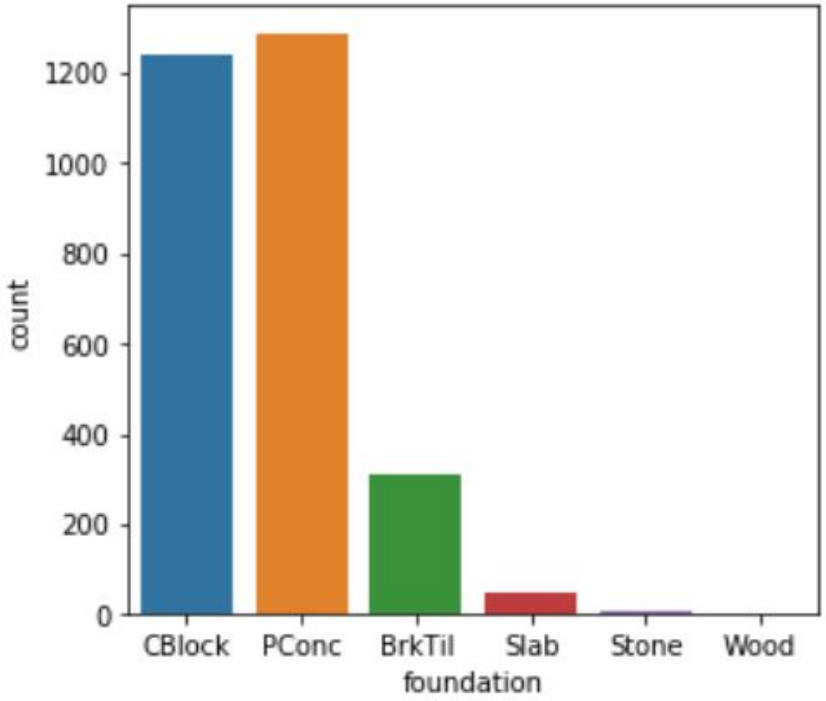
Prediction of Ames Housing Data

Project 2 Group 5:
Clarence - Jun Yu - Shu Jun

Overview

1. Problem Statement
2. Data Pre-Processing and EDA
3. Feature Selection
4. Prediction Model
5. Model Validation
6. Key Challenges
7. Interpretation of Results
8. Recommendations and Conclusions
9. Model Evaluation
10. References



Problem Statement

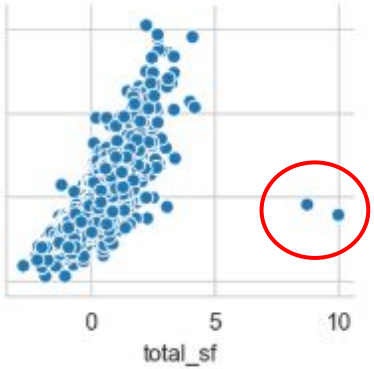
Having recently joined a property consultancy company in Ames, Iowa, we have been tasked with conducting a statistical analysis using the Ames Housing Dataset in order to determine which are the prominent factors that affect property prices in Ames.

Through this analysis, we hope to gain better insights so as to make recommendations and **give advice to homeowners on how to potentially increase a property's value**.

Data Pre-Processing and Exploratory Data Analysis

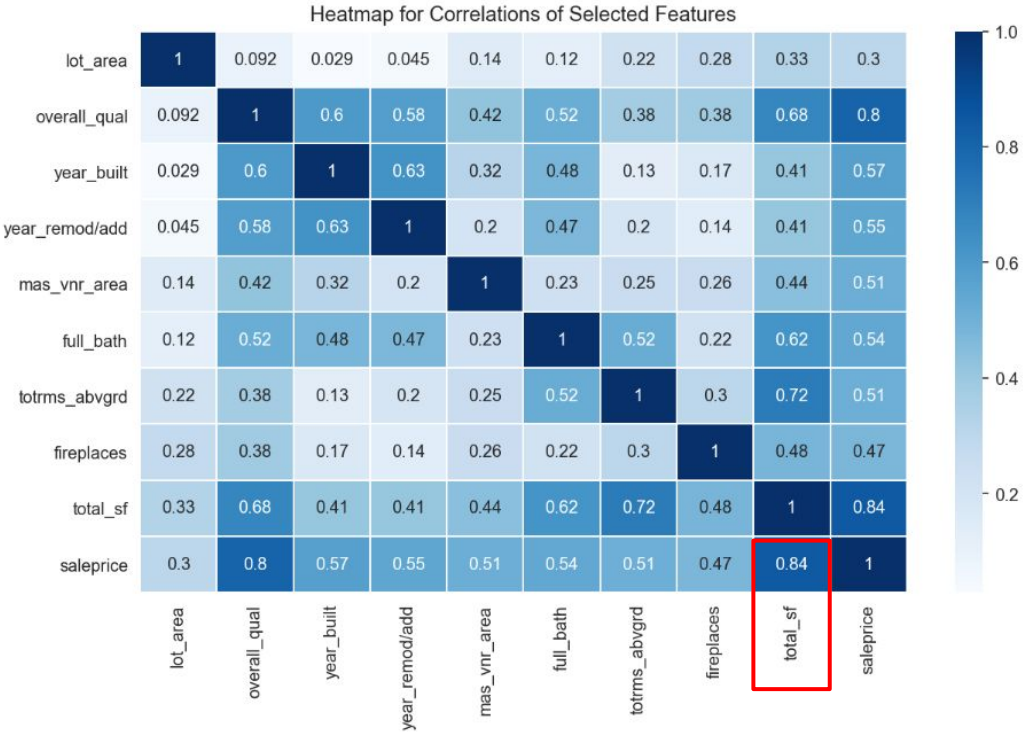
1. Replacing null values
   a. Understanding the Data Dictionary provided on Kaggle
   b. Most null values correspond to ‘NA’
   c. Cannot drop all rows with nulls as this would remove > 50% of dataset
2. Removing outliers
3. Scale numerical values

- Alley: Type of alley access to property
  - Grvl Gravel
  - Pave Paved
  - NA No alley access



Feature Selection - Numerical Features

1. Feature Engineering
   a. **Total area** from individual room areas
   b. **Age of property** from the time between the year it was built to the year it was sold
2. Features with higher correlations to saleprice were chosen

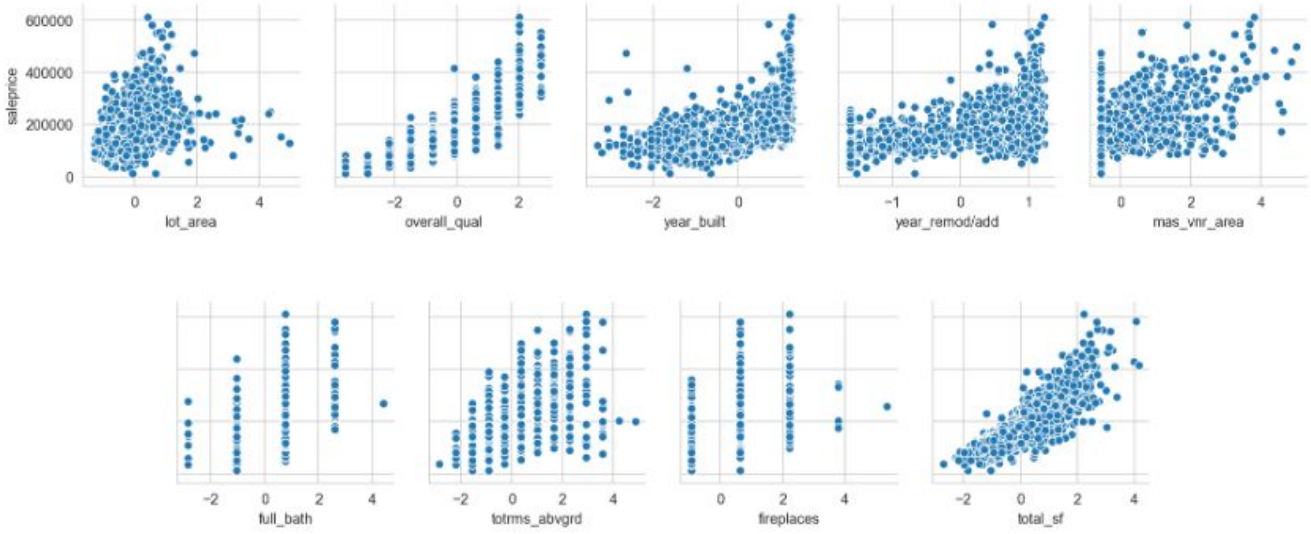


Heat Map for Correlations

Feature Selection - Numerical Features

3. Variance Inflation Factor to weed out features that were highly correlated with each other that may bias the model

	VIF Factor	features
0	1.3	lot_area
4	1.3	mas_vnr_area
7	1.4	fireplaces
3	2.0	year_remod/add
5	2.0	full_bath
2	2.1	year_built
6	2.3	totrms_abvgrd
1	2.9	overall_qual
8	4.6	total_sf



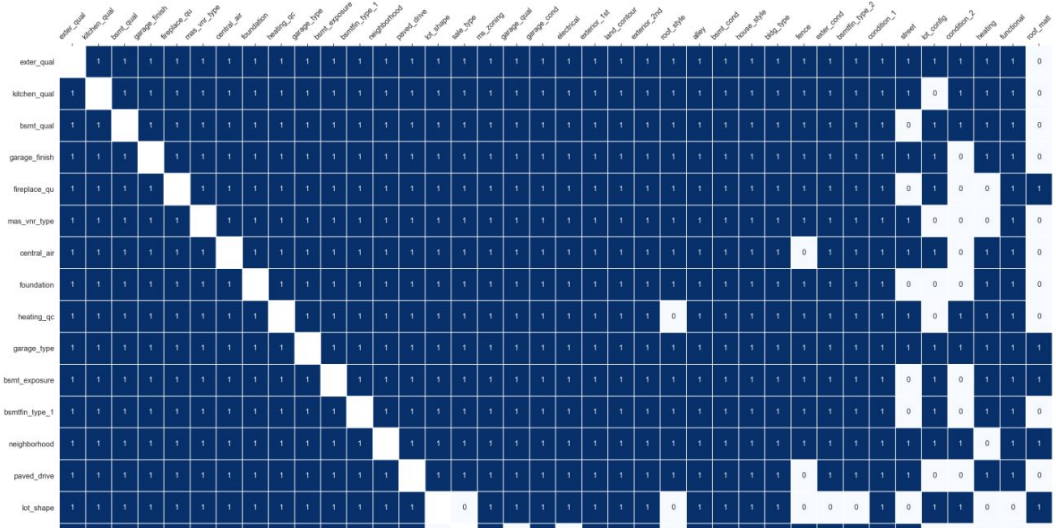
Scatter plots of selected features

Feature Selection - Categorical Features

1. Anova Test - Correlations of categorical features to saleprice
2. Chi-Sq Test - Correlations among categorical features

	F	p
utilities	1.401621	0.246436
land_slope	1.471841	0.229748
pool_qc	1.365022	0.243675
misc_feature	1.710732	0.144883

Anova Test:
Remove features that are not correlated to saleprice



Chi-Sq Test:
Remove correlated categorical features, with priority to keep features with higher correlations to saleprice

Prediction Model

1. OLS Linear Regression to remove statistically insignificant features
2. Ridge Regression on final selected features

Model Validation

1. Cross Validation to ensure that the model is robust and able to generalize well
2. Statistical metrics that mathematically quantify how well our model performs relative to the data

Key Challenges

1. Many ‘null’ values were actually ‘NA’ strings and not empty cells
    a. Modify function to load NA strings without them reading as empty cells
2. Many features were highly correlated to each other
3. Limitations of the Linear Regression Model - assumptions of linear relationships between features and target
    a. Consider exploring other more sophisticated machine learning models e.g. random forest

Interpretation of Results

	Variables	Coefficients
11	total_sf	39964.113566
10	overall_qual	19078.222516
3	lot_area	11896.253267
8	year_built	8044.195766
6	year_remod/add	6325.341230
4	mas_vnr_area	6092.438734
5	fireplaces	4535.606623
9	totrms_abvgrd	-3307.835839
7	full_bath	-4097.817934
1	Gd_exter_qual	-61418.508434
0	Fa_exter_qual	-67311.512496
2	TA_exter_qual	-71786.149201

1. **Total Sq Ft** has the highest positive effect on property sale price, followed by **Overall Quality** (material and finish) of house, and **Lot Area** (lot size in square feet)
2. Having a **good/fair/typical-average quality of the housing exterior** has a very big negative impact on property sale price, with TA quality having the biggest negative effect

Interpretation of Results

	Variables	Coefficients
11	total_sf	39964.113566
10	overall_qual	19078.222516
3	lot_area	11896.253267
8	year_built	8044.195766
6	year_remod/add	6325.341230
4	mas_vnr_area	6092.438734
5	fireplaces	4535.606623

Variables with Positive Impacts:

1. Area / Space
    a. Total Sq Ft, Lot Area, Masonry Veneer Area
2. Luxury / Quality
    a. Overall Quality, Fireplaces, Remodel Date
3. Age
    a. Year Built

Interpretation of Results

9	totrms_abvgrd	-3307.835839
7	full_bath	-4097.817934
1	Gd_exter_qual	-61418.508434
0	Fa_exter_qual	-67311.512496
2	TA_exter_qual	-71786.149201

Variables with Negative Impacts:

1. House features
    a. Full bathrooms, and total rooms, above grade
2. Exterior Quality
    a. Good / fair / typical-average

Recommendations and Conclusions

Recommend existing homeowners to focus on certain aspects for renovation:

1. Potential homebuyers in Ames value the concept of **"luxury"** and **"space"**
    a. BUT check on feasibility of such enhancements first!
    b. Also, costs of enhancements should be < potential gains of these enhancements (by looking at the coefficients of the variables).
2. **Remodelling** your home helps!
    a. Age of property can't be changed, but remodelling it and letting it appreciate over time does help
3. Having an **excellent housing exterior quality** fetches a premium on the property's sale price

Model Evaluation

Model isn't perfect and has a number of flaws:

1. Absence of any '**neighborhood/district**' variable
    a. Intuitively, properties at certain regions (near CBD or affluent shopping districts) would fetch higher prices than properties in the outskirts of the city
2. Another 2 variables that would intuitively affect sale price were dropped
    a. **Building Type** (type of dwelling, such as single-family detached or duplex)
    b. **House Style** (style of dwelling, such as one-storey or two-storey)
3. Need to further improve on categorical feature selection

References

https://www.kaggle.com/c/dsi-us-6-project-2-regression-challenge/data