House Prices in Ames, Iowa

Machine Learning Project

Outline

- Research Objectives / Target Audience
- Exploratory Data Analysis
- Customer Segmentation
- Conclusion, Recommendation, Future Work

Research Objectives

- Identify what variables of the housing dataset increase the sale price
- Segment potential customers utilizing income variable

Target Audience

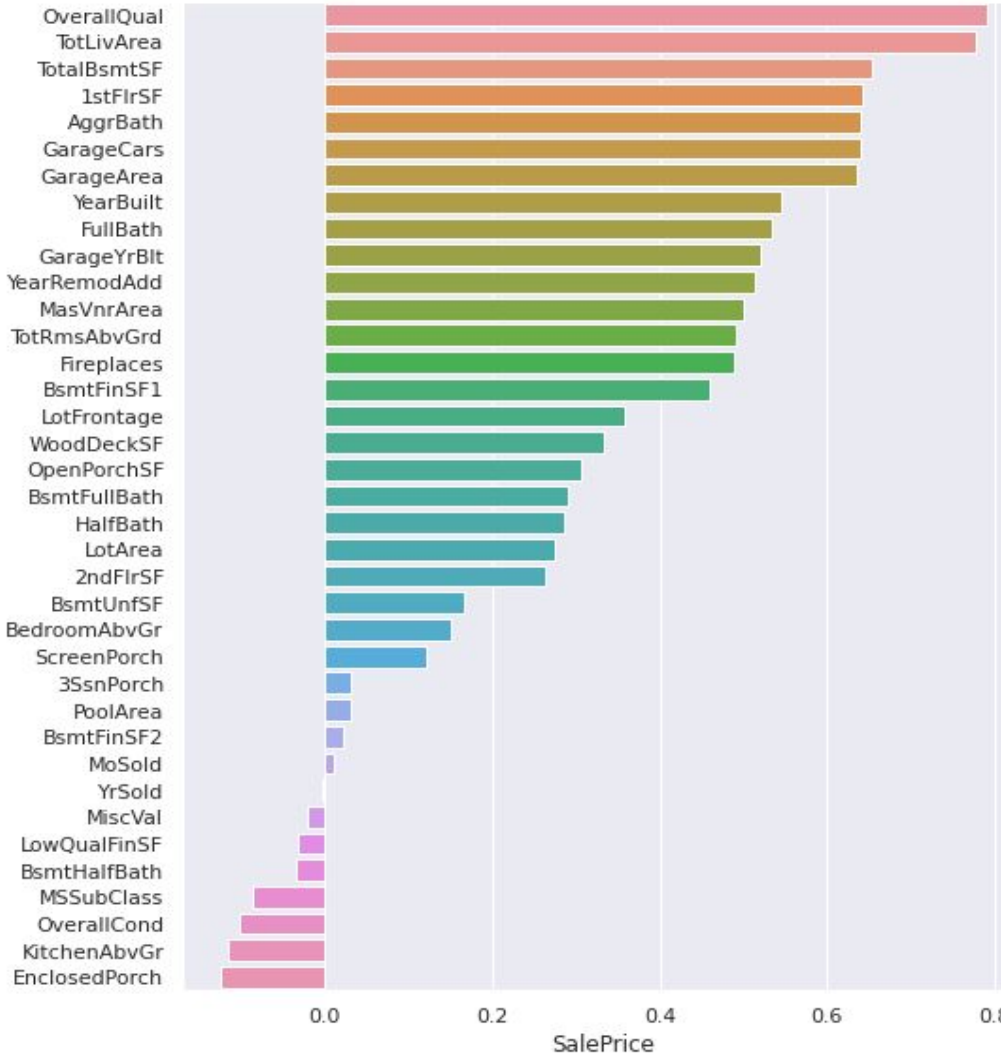
Individuals/Agencies interested in utilizing the results of this research for marketing purposes

Datasets used

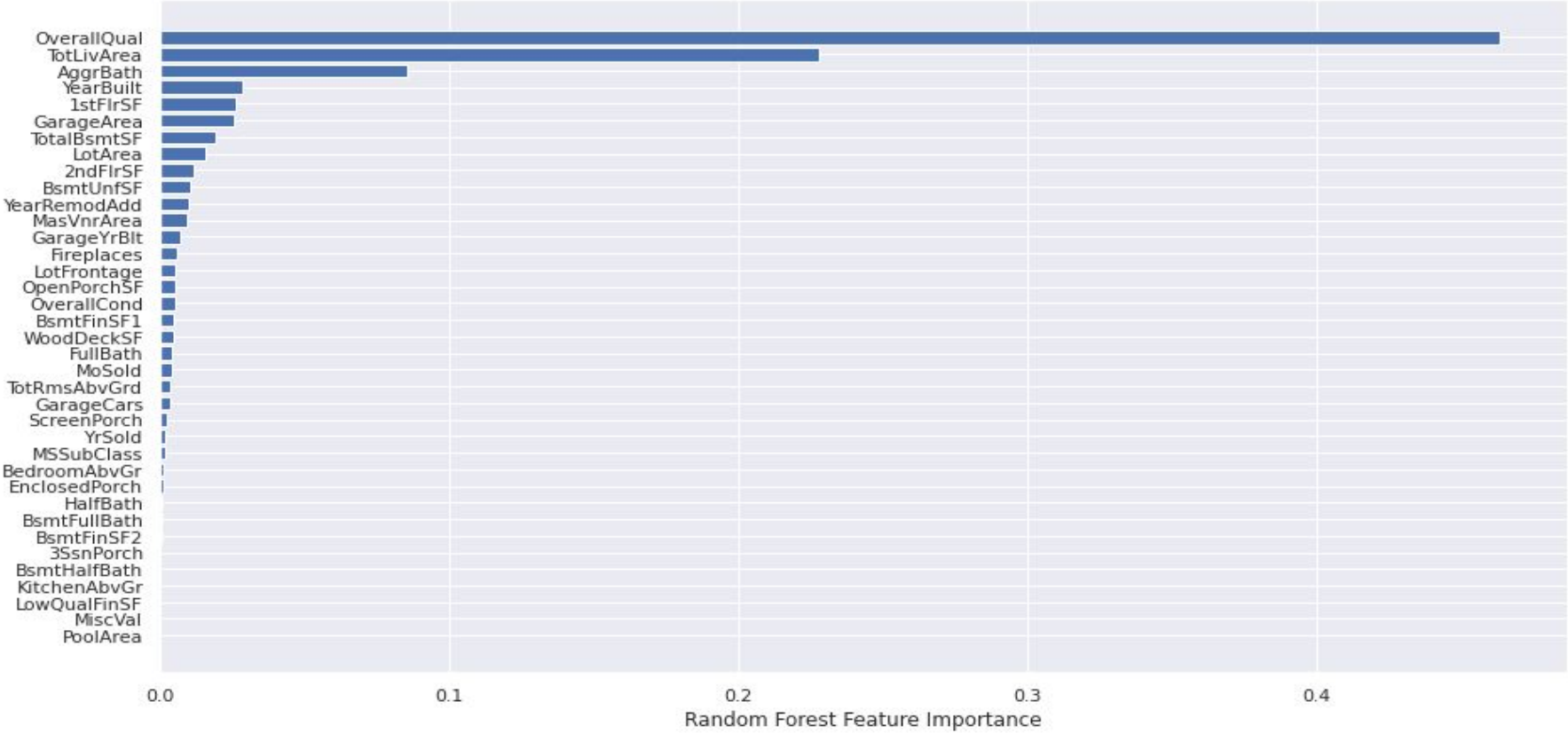
- Dataset `Ames_HousePrice` is collected from kaggle.com
- 2,580 observations, 81 variables
- From 2006 to 2010
- Dataset is collected `State_of_Iowa_Salary_Book` from data.iowa.gov
- 913,362 observations, 10 variables

Exploratory Data Analysis

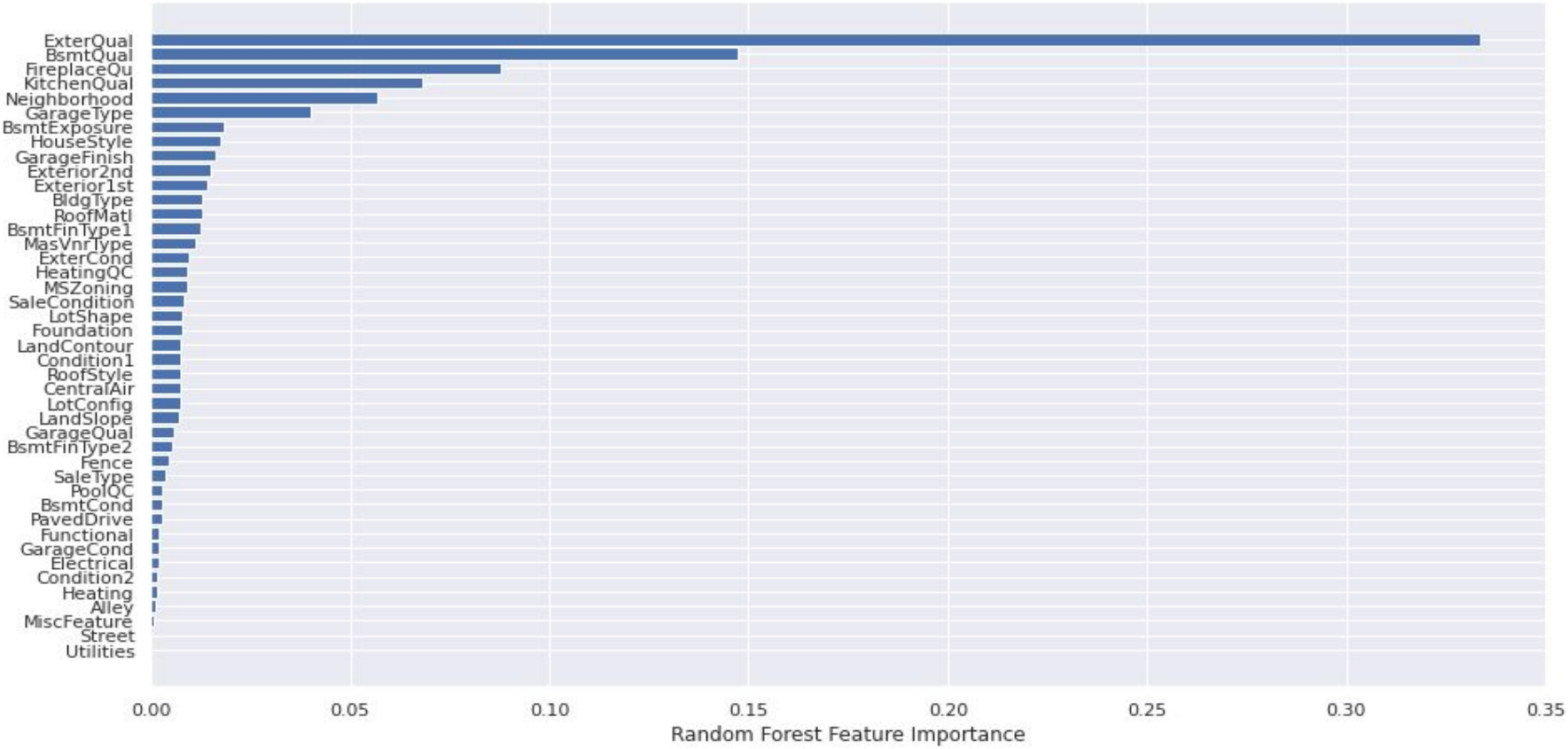
- GrLivArea was replaced with TotLivArea (1st Floor + 2ndFloor + Finished Basement)
- Simple Linear Regression to find the relationship between TotLivArea and Sales Price (**lm.score = 0.60,lm.coef_ = [81.44]** and **lm.intercept_ = 16869.54**)
- AggrBath was added (Full Bath + Half Bath + Bsmt Bath + Bsmt Half Bath)
- Simple Linear Regression to find the relationship between Aggr Bath and Sales Price (**lm.score = 0.41,lm.coef_ = [60249.49]** and **lm.intercept_ = 45160.46**)

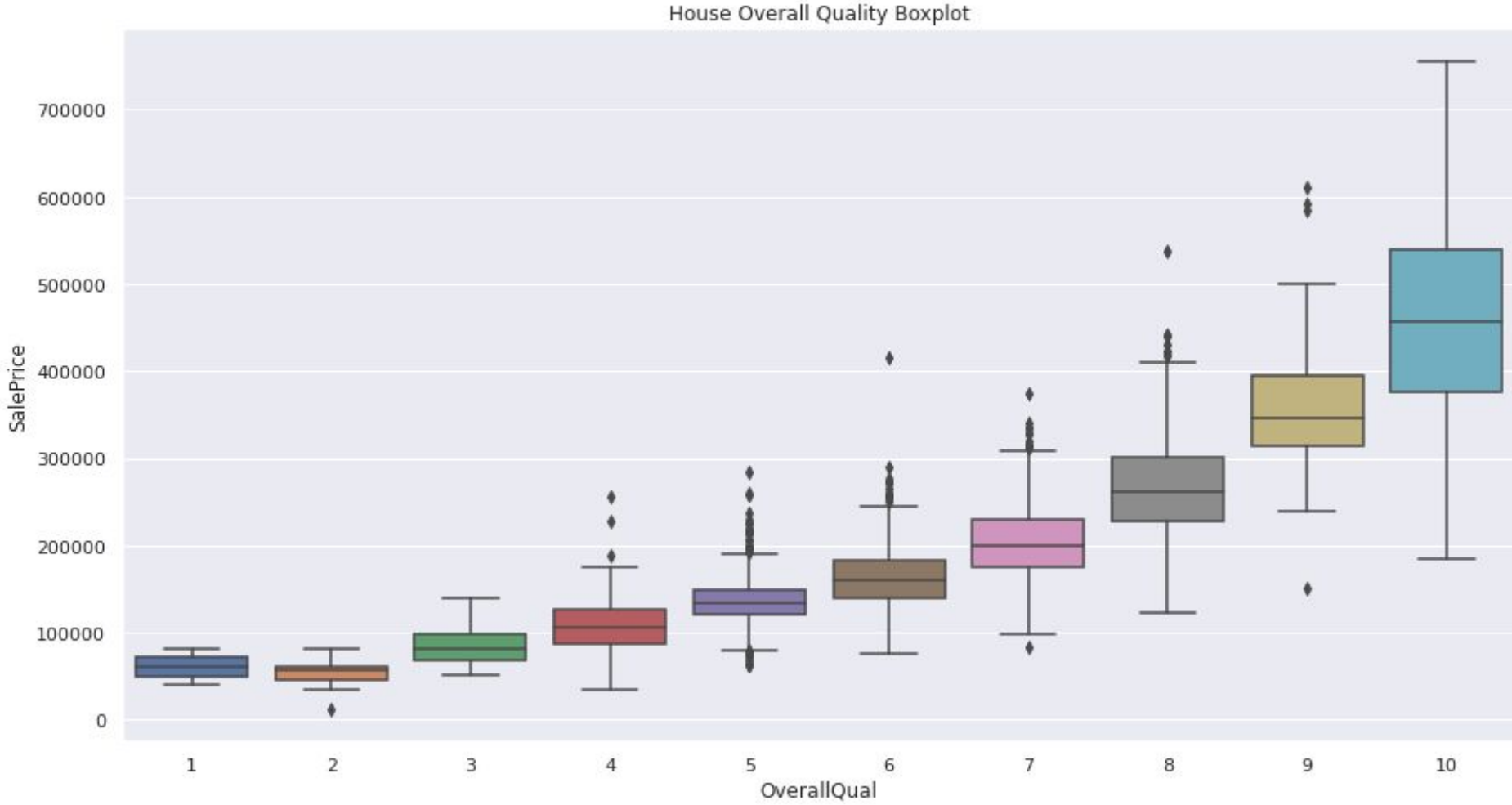


Random Forest Feature Importance (Numerical)

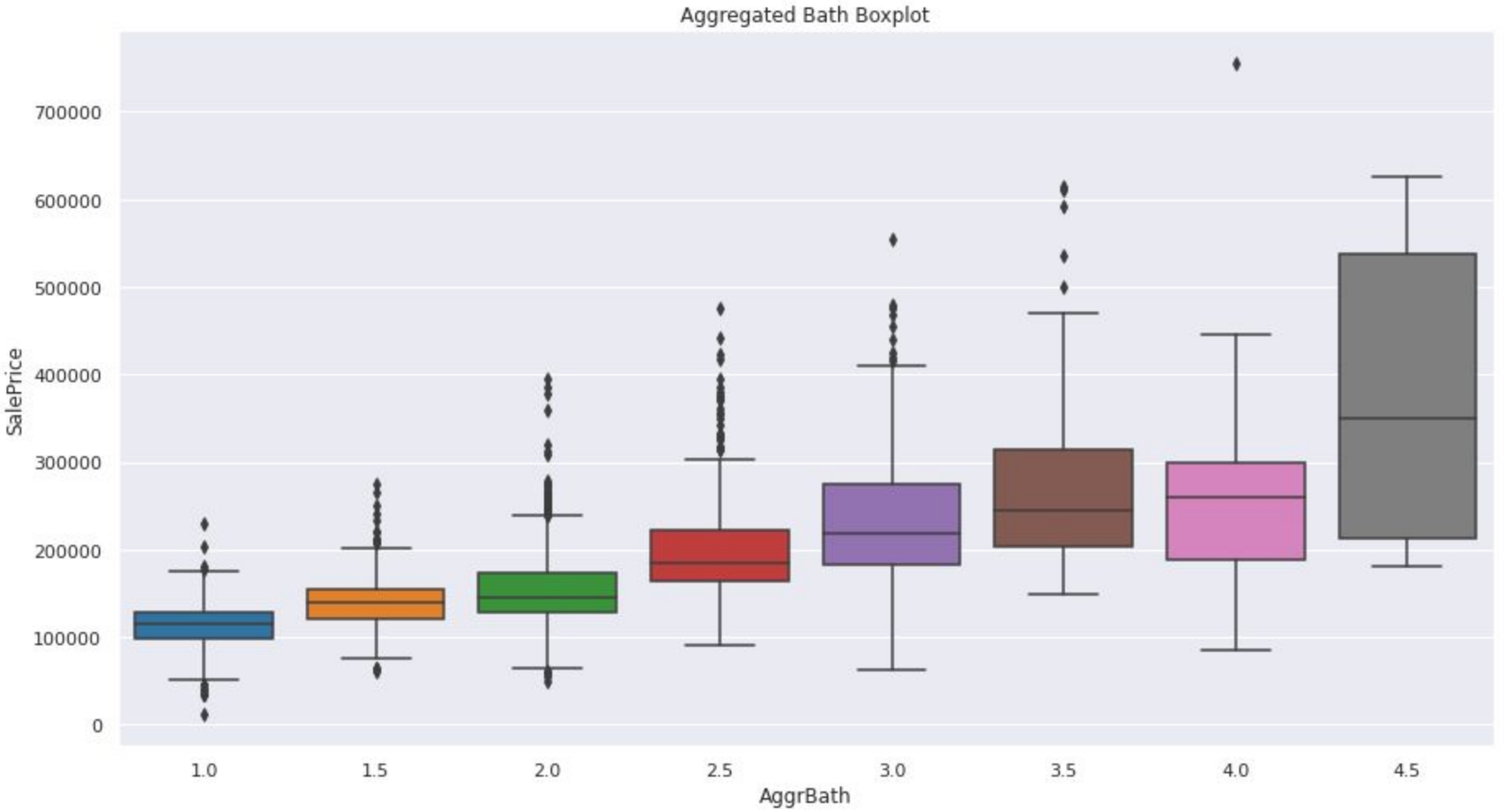


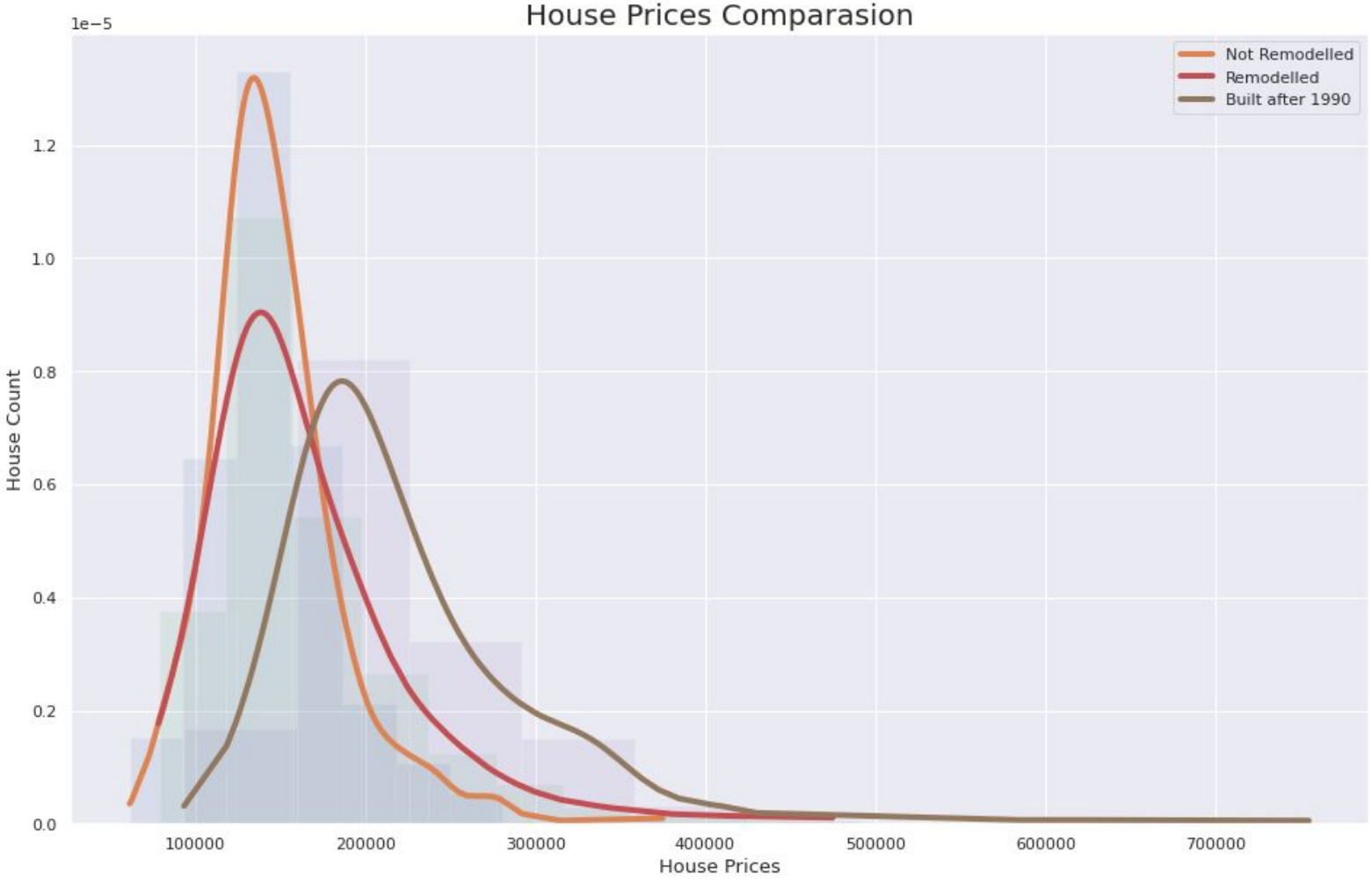
Random Forest Feature Importance (Categorical)

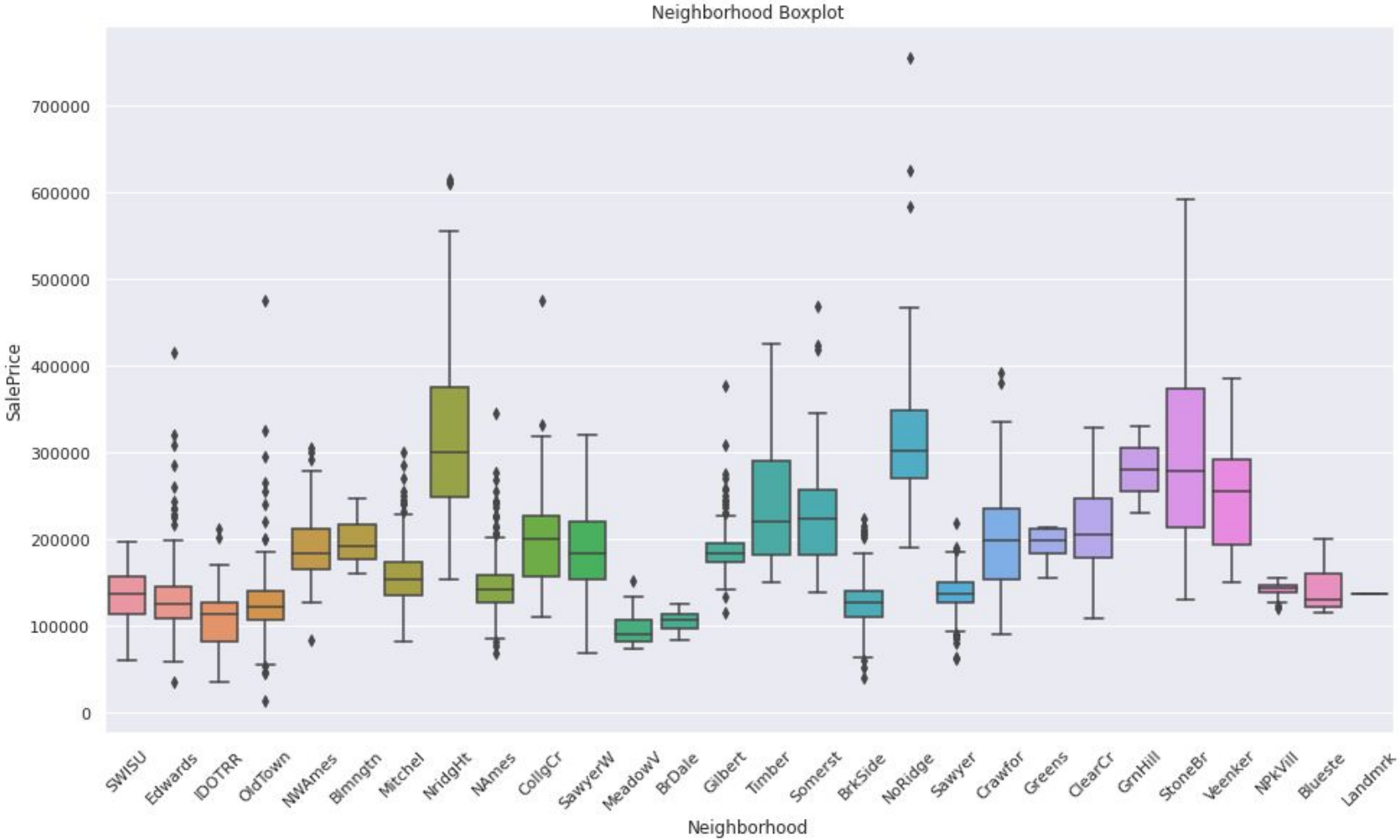



House Overall Quality Boxplot
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1
2
3
4
5
6
7
8
9
10
OverallQual


Sale Price per Total Living Area
Sale Price
700000
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
6000
7000
TotLivArea

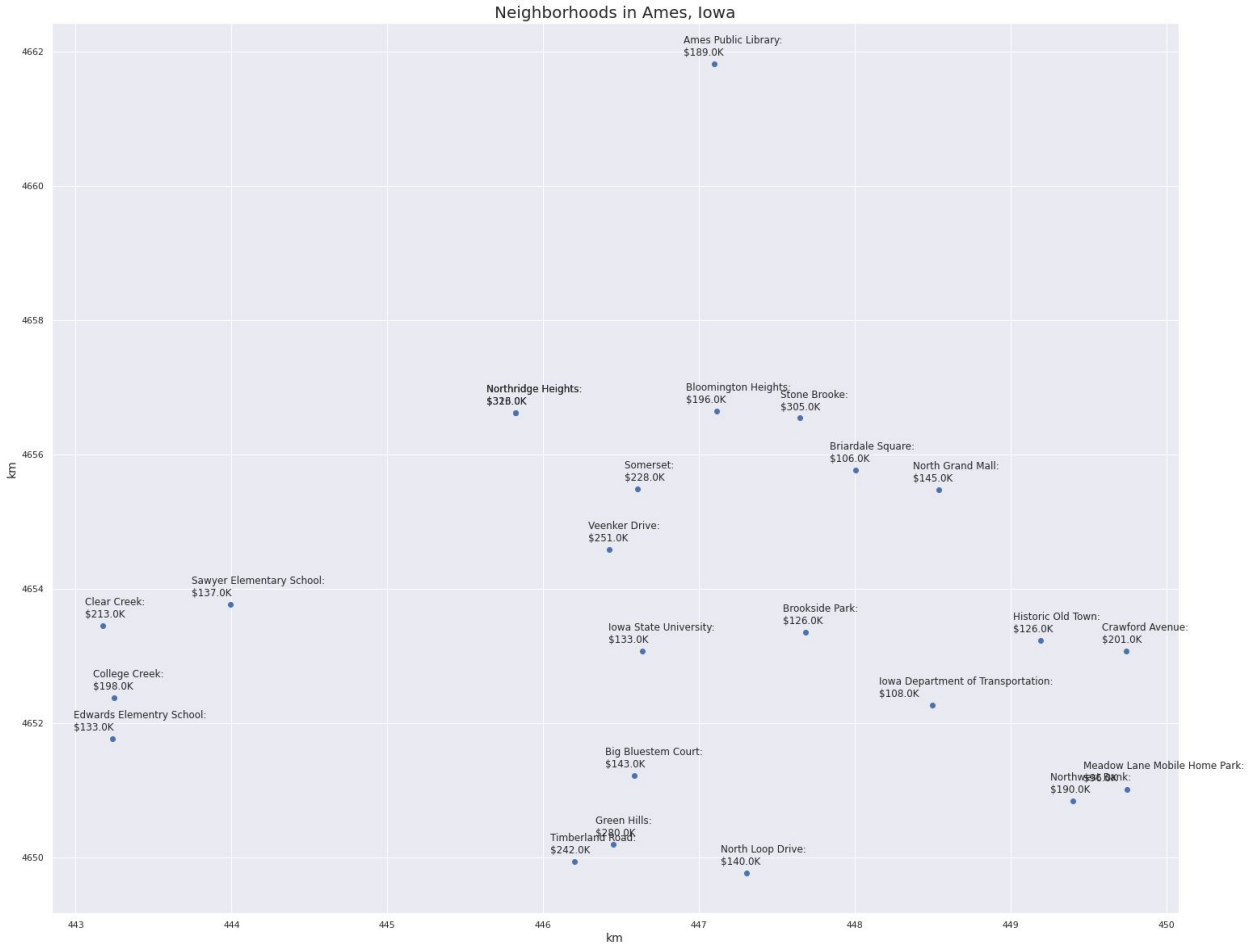

Aggregated Bath Boxplot
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
AggrBath


House Prices Comparasion
1e−5
Not Remodelled
Remodelled
Built after 1990
House Count
House Prices
0.0
0.2
0.4
0.6
0.8
1.0
1.2
100000
200000
300000
400000
500000
600000
700000


Neighborhood Boxplot
SalePrice
0
100000
200000
300000
400000
500000
600000
700000
SWISU
Edwards
IDOTRR
OldTown
MWAmes
Blmngtn
Mitchel
NridgHt
NAmes
CollgCr
SawyerW
MeadowV
BrDale
Gilbert
Timber
Somerst
BrkSide
NoRidge
Sawyer
Crawfor
Greens
ClearCr
GrnHill
StoneBr
Veenker
NPkVill
Blueste
Landmrk
Neighborhood

Average Price per Neighborhood

'Ames Public Library' - $189.0K
'Bloomington Heights' - $196.0K
'Big Bluestem Court' - $143.0K
'Briardale Square' - $106.0K
'Brookside Park' - $126.0K
'Clear Creek' - $213.0K
'College Creek' - $198.0k
'Crawford Avenue' - $201.0K
'Edwards Elementary School' - $133.0K
'Green Hills' - 280.0K
'Historic Old Town' - $126.0K
'Iowa Department of Transportation' - $108.0K
'Iowa State University' - $133.0K
'North Loop Drive' - $140.0K
'North Grand Mall' - $145.0K
'Northridge Heights' - $328.0K
'Northwest Bank' - $190.0K
'Iowa State University' - $133.0K
'Sawyer Elementary School' - $137.0K
'Somerset' - $228.0K
'Stone Brooke' - $305.0K
'Timberland Road' - $242.0K
'Veenker Drive' - $251.0K

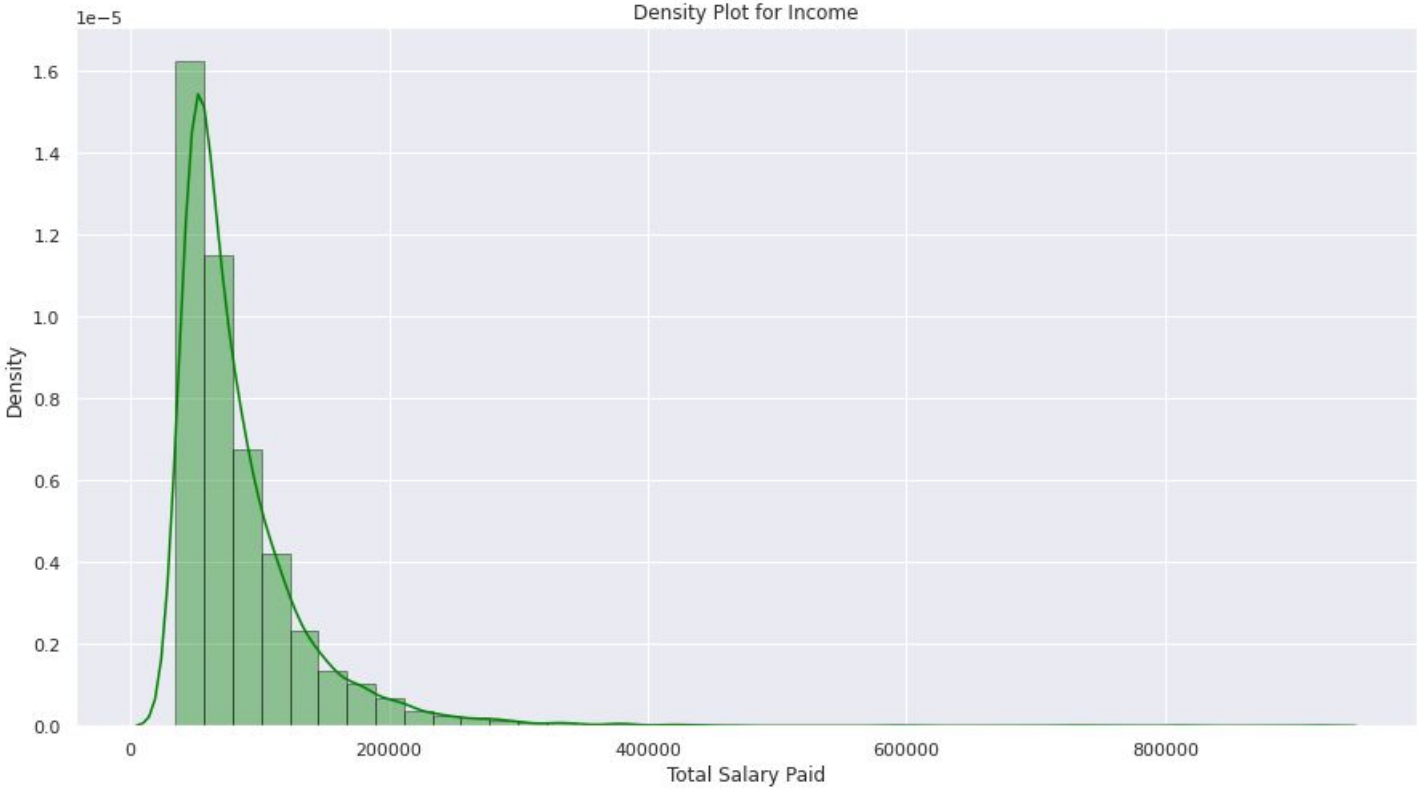


Iowa Salary Dataset

- 2 largest employers are Iowa State University and the Department of Transportation
- Total salary paid ranges from 1,000.49 (student) to 2,375,000.04 (head coach) for 2020
- For this research, I narrowed the total salary paid variable from 35,000 to 917,499.98

IOWA STATE UNIVERSITY

Customer Segmentation



Feature engineering for customer segmentation

- Monthly payments: the general rule state that an individual should spend around 28% of monthly income on the mortgage payment.

  Monthly payment = (Annual Income/12)*0.28

- Loan Amount: we make an assumption that the mortgage terms are 30 years and 6%.

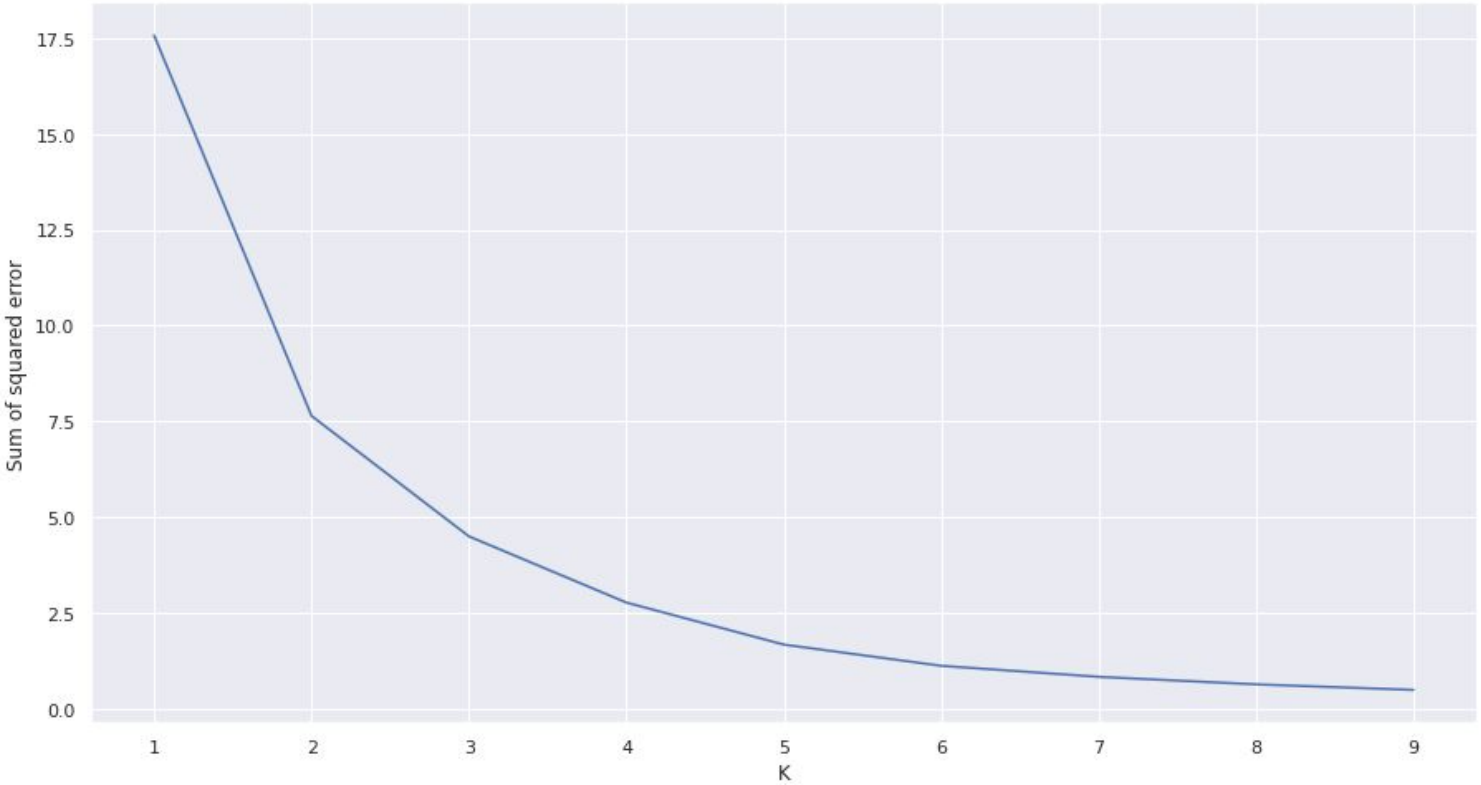
  Loan amount = (payment*(1 - (1+rate/12)**(-months)))/(rate/12)

- Home Price: we make an assumption that a 20% down payment is required for a mortgage loan.

  Home Price = (Loan Amount*0.1)/0.08


Income per Home Price
1e6
Home Price
0
1
2
3
4
0
200000
400000
600000
800000
Total Salary Paid

Customer Segmentation Using K Means Clustering



Customer Segmentation Using K Means Clustering

Cluster 1 - income range from $35,000.01- $76,014.87

Cluster 2 - income range from $76,108.69 - $131,095.15

Cluster 3 - income range from $131,576.79- $224,442.37

Cluster 4 - blue income range from $ 225,587.79- $462,453.87

Cluster 5 - black income range from $ 594,666.8- $917,499.98

Conclusion

The following clusters are the target demographic for marketing sales of homes:

Cluster 1 - home price range from $ 170,266.25 - $ 369,793.75 (house quality 5 to 8)

Cluster 2 - home price range from $370,250.00 - $637,746.25 (house quality 8 to 10)

Cluster 3 - home price range from $640,088.75 - $1,091,857.50 (house quality 10)

Cluster 4 - home price range from $1,097,430.00 - $2,249,725.0 (house quality 10)

Cluster 5 - home price range from $2,892,908.75 - $4,463,412.5 (house quality 10)

Recommendation/Future Work

Recommendation: other features can be incorporated into the analysis (such as Total Footage Square, Year Built, Neighborhood and etc.)

Future Work: given more time, I would build a predictive model that would incorporate individual/household incomes with the features from the housing price dataset

Questions?