Najaf Shaikh

PROJECT 2
AMES HOUSING

PROBLEM STATEMENT

This project aims to identify what the most important features are in the sale price of a house.

This information can be helpful for sellers who wish to increase the value of their home before selling and buyers who would want to better understand the price of a home before buying. In addition, it may also help investors and builders.

In this project, a Machine Learning model will be created to predict the house prices. This model will later be uploaded to Kaggle to receive a score on how well the RMSE compares to other competitors.

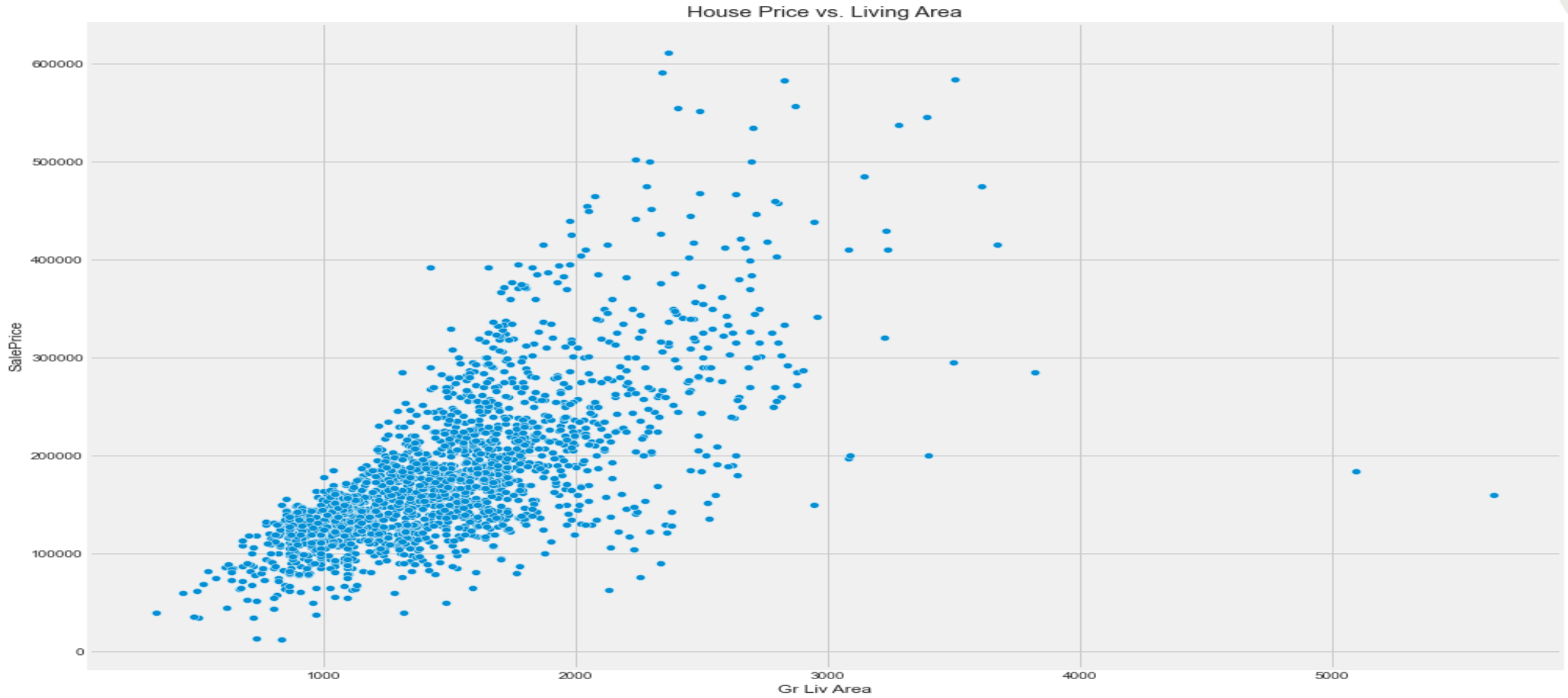
DATA SET

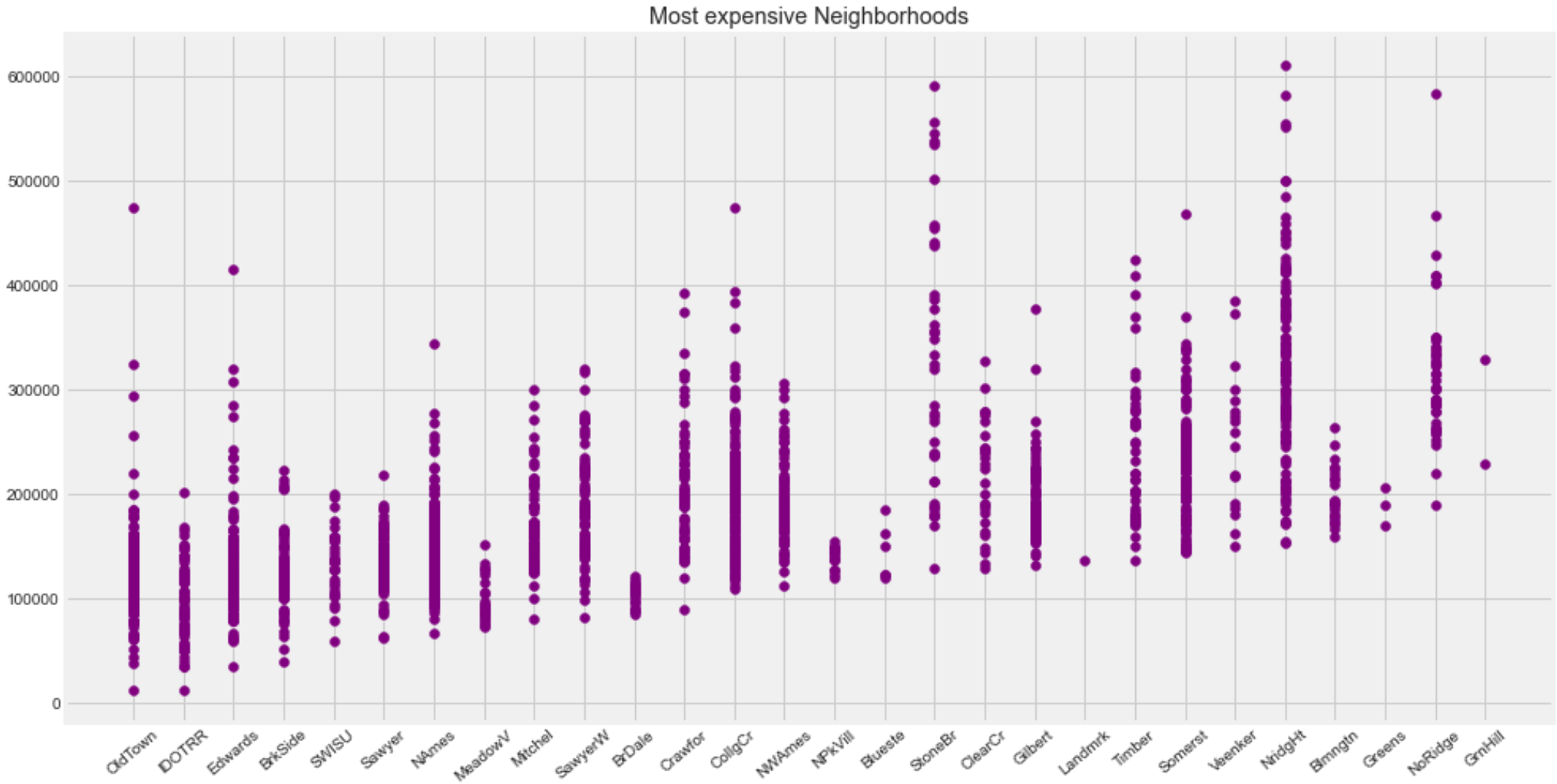
- Author: Dean De Cock

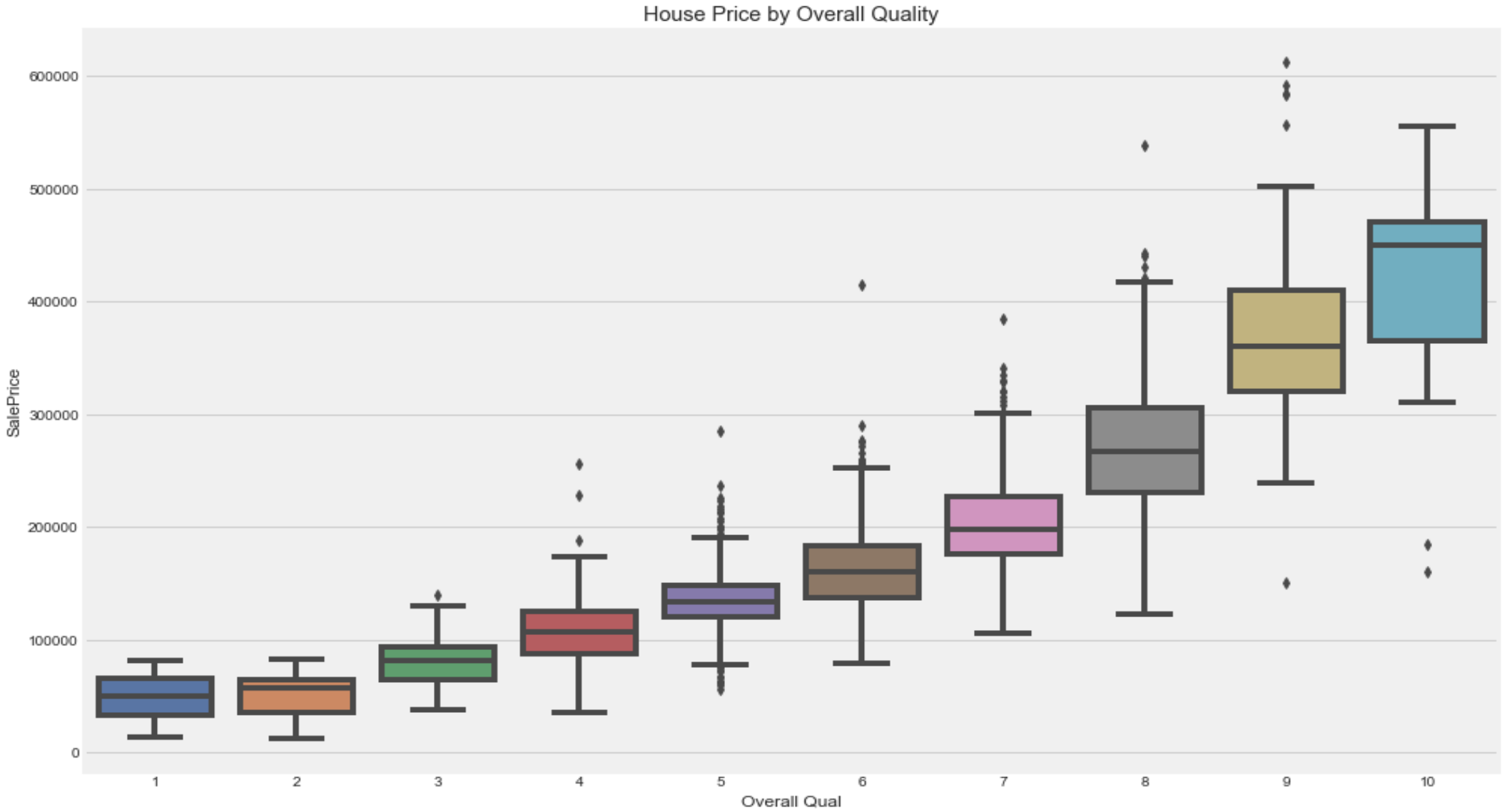
  "Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property"

- Observations
  - 2930
- Explanatory Variables
  - 23 Nominal
  - 23 Ordinal
  - 14 Discrete
  - 20 Continuous

LIVING AREA

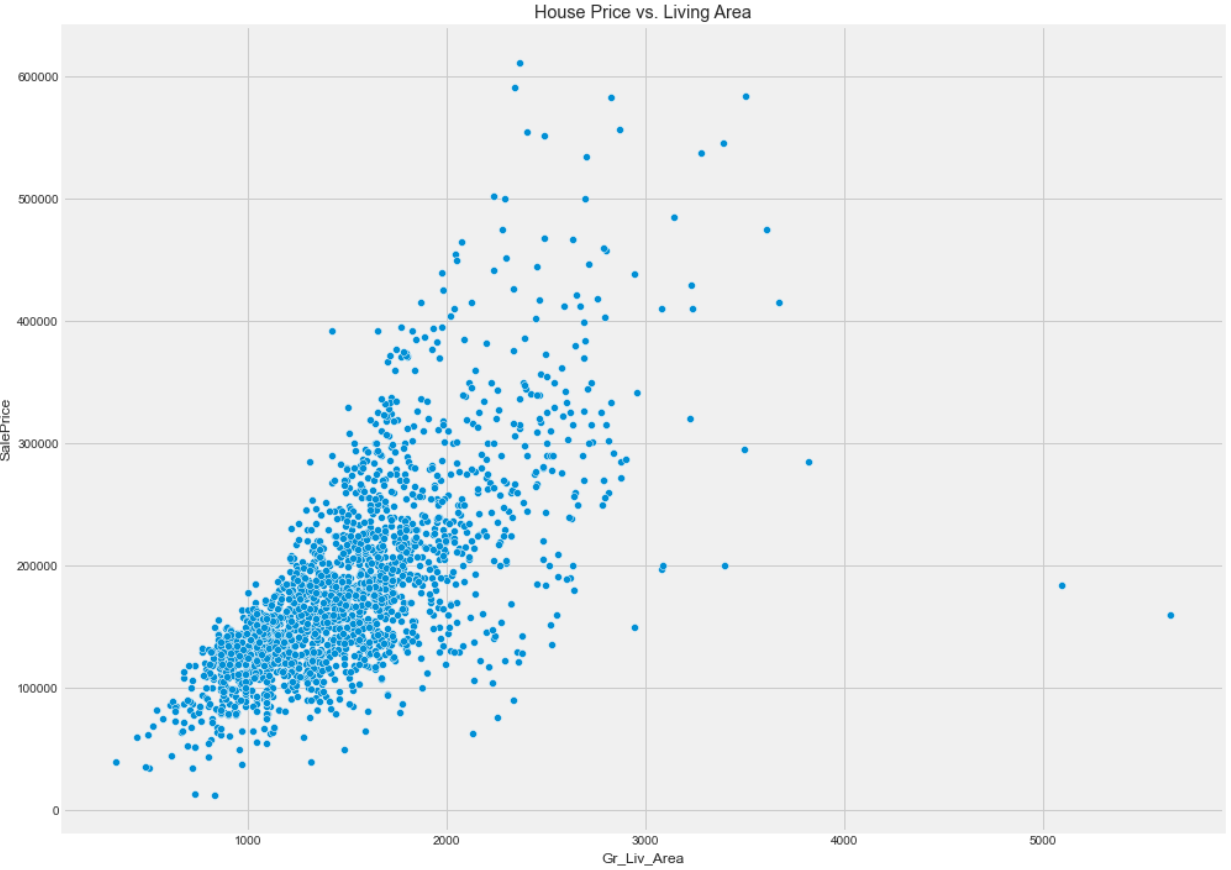

House Price vs. Living Area
SalePrice
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000
Gr Liv Area

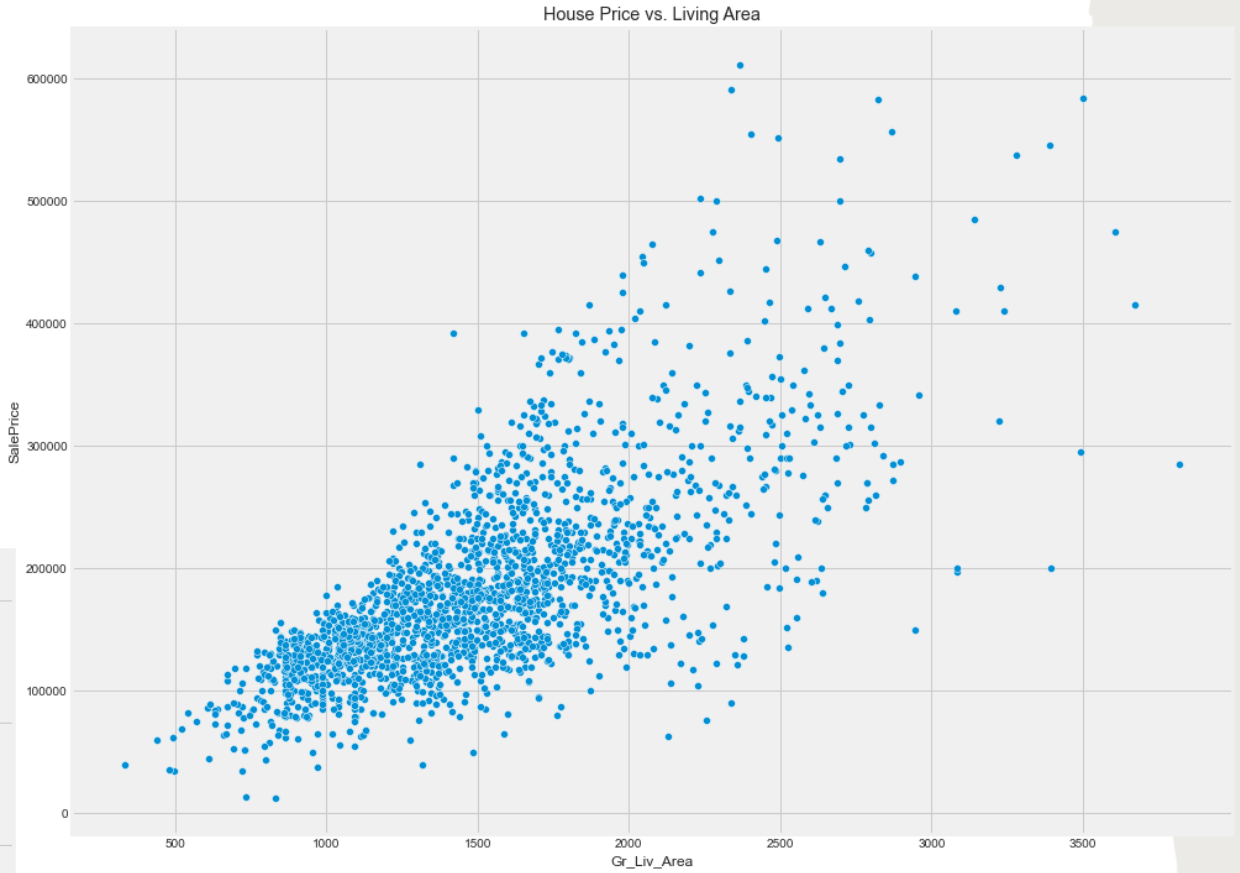

Most expensive Neighborhoods
600000
500000
400000
300000
200000
100000
0
OldTown
IDOTRR
Edwards
BrkSide
SWISU
Sawyer
NAmes
MeadowV
Mitchel
SawyerW
BrDale
Crawfor
CollgCr
NWAmes
NPkVill
Blueste
StoneBr
ClearCr
Gilbert
Landmrk
Timber
Somerst
Veenker
NridgHt
Blmngtn
Greens
NoRidge
GrnHill


House Price by Overall Quality
SalePrice
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
Overall Qual

OUTLIERS

WITH OUTLIERS





OUTLIERS REMOVED

"I would recommend removing any houses with more than 4000 square feet from the data set (which eliminates these five unusual observations)" – ***Dean De Cock***

MISSING VALUES

- Alley
  - Assuming no alleys
- Basement Features
  - Assuming houses do not have basements
- Garage Features
  - Assuming houses do not have basements
- Fence, Fireplace, Lot Frontage, Veneers, Miscellaneous Features, Pool, Electrical
- Assuming houses do not have these features

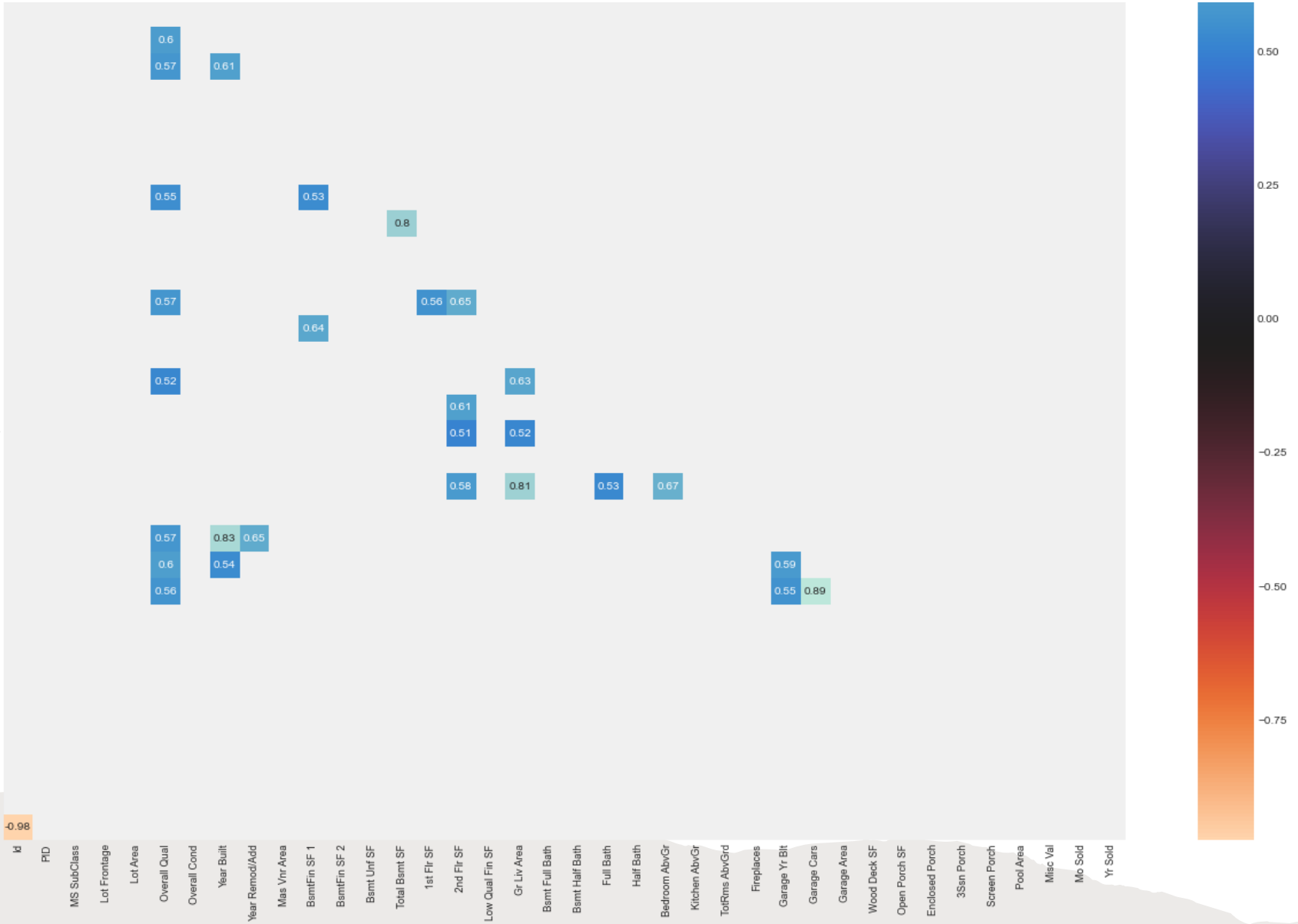
Pool QC	0.995562
Misc Feature	0.963810
Alley	0.932400
Fence	0.804712
Fireplace Qu	0.485490
SalePrice	0.299761
Lot Frontage	0.167293
Garage Finish	0.054285
Garage Qual	0.054285
Garage Cond	0.054285
Garage Yr Blt	0.054285
Garage Type	0.053602
Bsmt Exposure	0.028337
BsmtFin Type 2	0.027654
Bsmt Cond	0.027313
Bsmt Qual	0.027313
BsmtFin Type 1	0.027313
Mas Vnr Area	0.007853
Mas Vnr Type	0.007853
Bsmt Full Bath	0.000683
Bsmt Half Bath	0.000683
Garage Cars	0.000341
Garage Area	0.000341
Total Bsmt SF	0.000341
BsmtFin SF 2	0.000341
Electrical	0.000341
BsmtFin SF 1	0.000341
Bsmt Unf SF	0.000341
Kitchen Qual	0.000000

Garage Cars & Garage Area
= ***0.89***

Garage Year Built & Year built
= 0.83

1st Floor SF and Total Basement SF
= 0.81

Gr Liv Area & Total Rooms Above Ground
= 0.81



COLLINEARITY

ADDITIONAL STEPS

- Dropped Columns: Garage Year, Built, Garage Area, Total Rooms above ground, 1st floor SF
- Categorical features to ordered numbers
- Categorical features to dummies
- Feature Engineering: GR Living Area + Total Basement SF = AllSF
- Scaling

LASSO

- X_TRAIN SCORE = 0.91583…
- X- TEST SCORE = 0.905964…

Overall train score = 0.913348…

X-TRAIN RMSE =

Cross Val Score = 0.89257

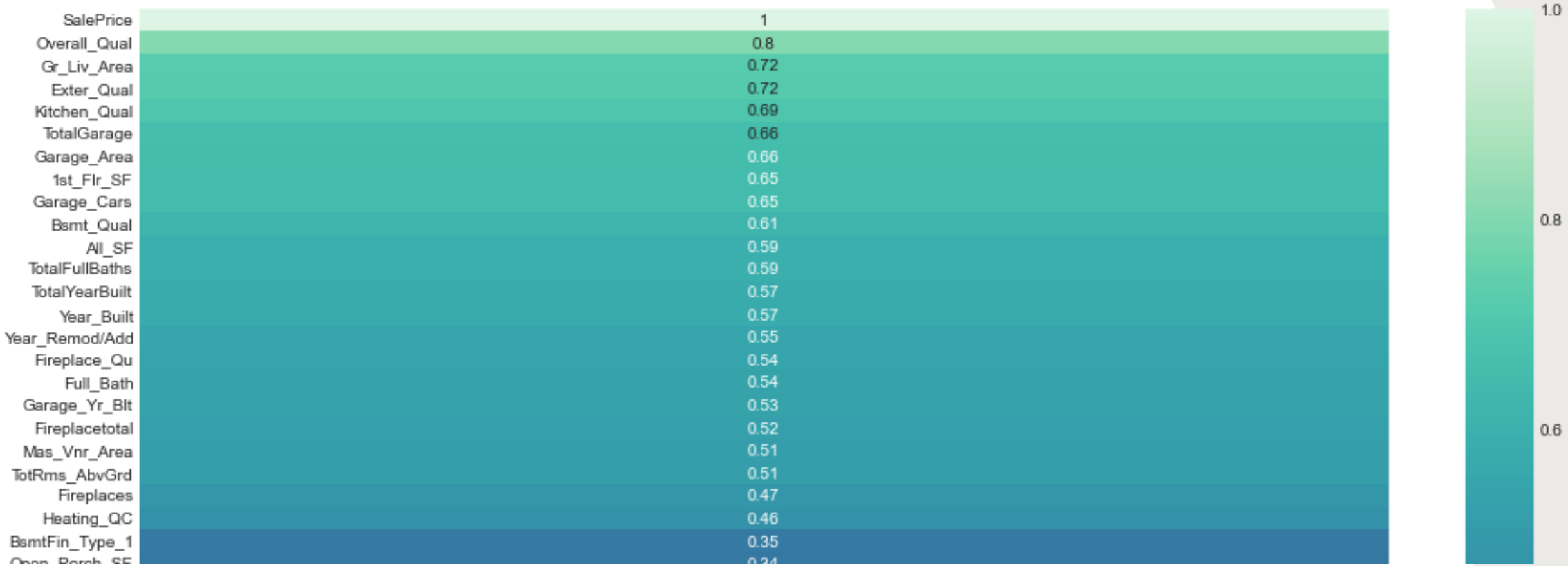
Overall RMSE = **23336.29**

Kaggle:

Score: 34049.95453

Public score: 38792.53943

HIGHEST CORRELATION



REFERENCES

De Cock, D. (n.d.). *Ames, Iowa: Alternative to the Boston Housing Data as an end of …* Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Retrieved July 18, 2022, from http://jse.amstat.org/v19n3/decock.pdf