Regression Challenge

General Assembly DSI 27

Ken

Background

Who am I? Who is my audience?

Employed by ERA

A talk with Real Estate Agents, Sellers and Buyers on Ames Iowa Housing

(For those who want to relocate from Singapore)

Problem Statement

Objectives

Create a model to produce housing price predictions based on past data

Top features correlated to prices

Data Set

From Kaggle:

- Train.csv (2051 rows, 81 columns)
- Test.csv (879 rows, 80 columns)

```
# reading in train.csv and test.csv
# set_option to remove ... for columns

train_df = pd.read_csv('../datasets/train.csv')
test_df = pd.read_csv('../datasets/test.csv')
pd.set_option('display.max_columns', 500)
```

```
# check shape

train_df.shape
```

```
(2051, 81)
```

```
# check shape

test_df.shape
```

```
(878, 80)
```

Workflow

EDA

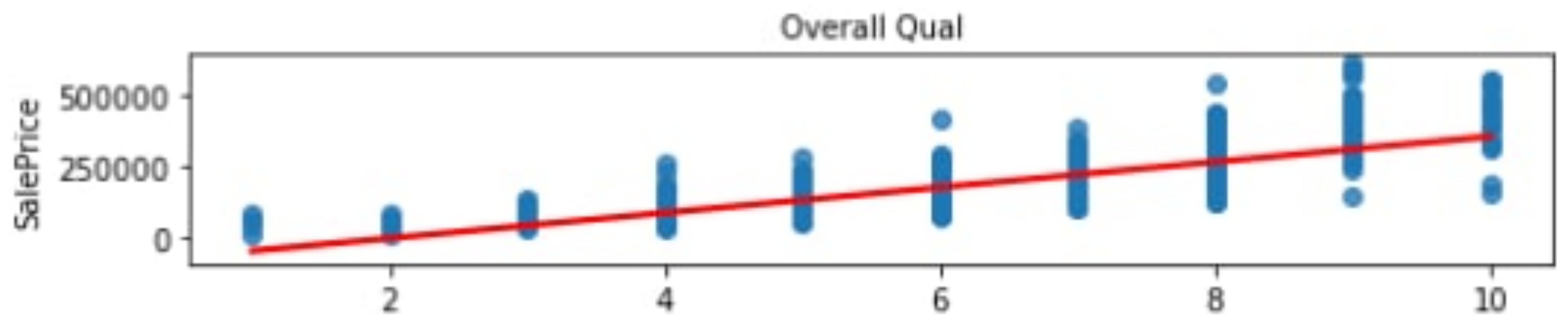
Data Cleaning
(Check/Fill Null Values)

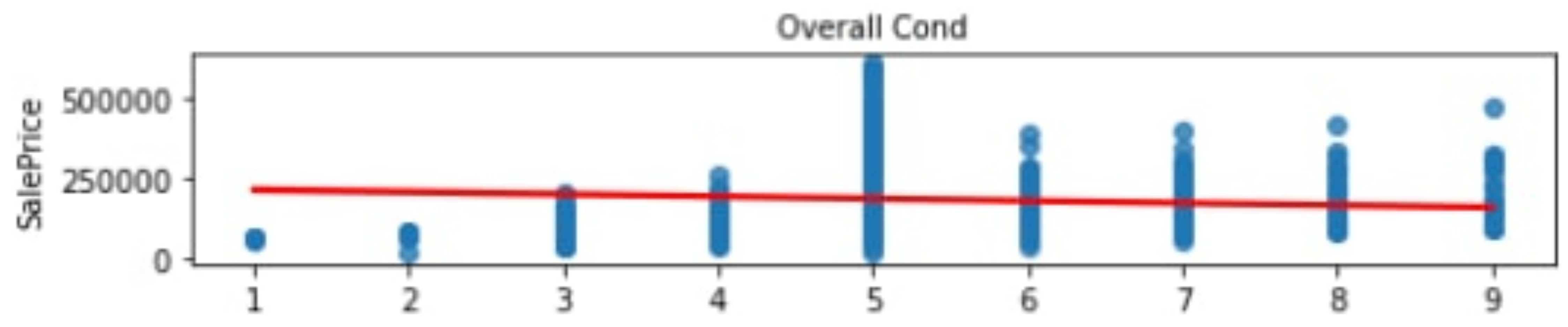
Feature Engineering

Modelling

Conclusion

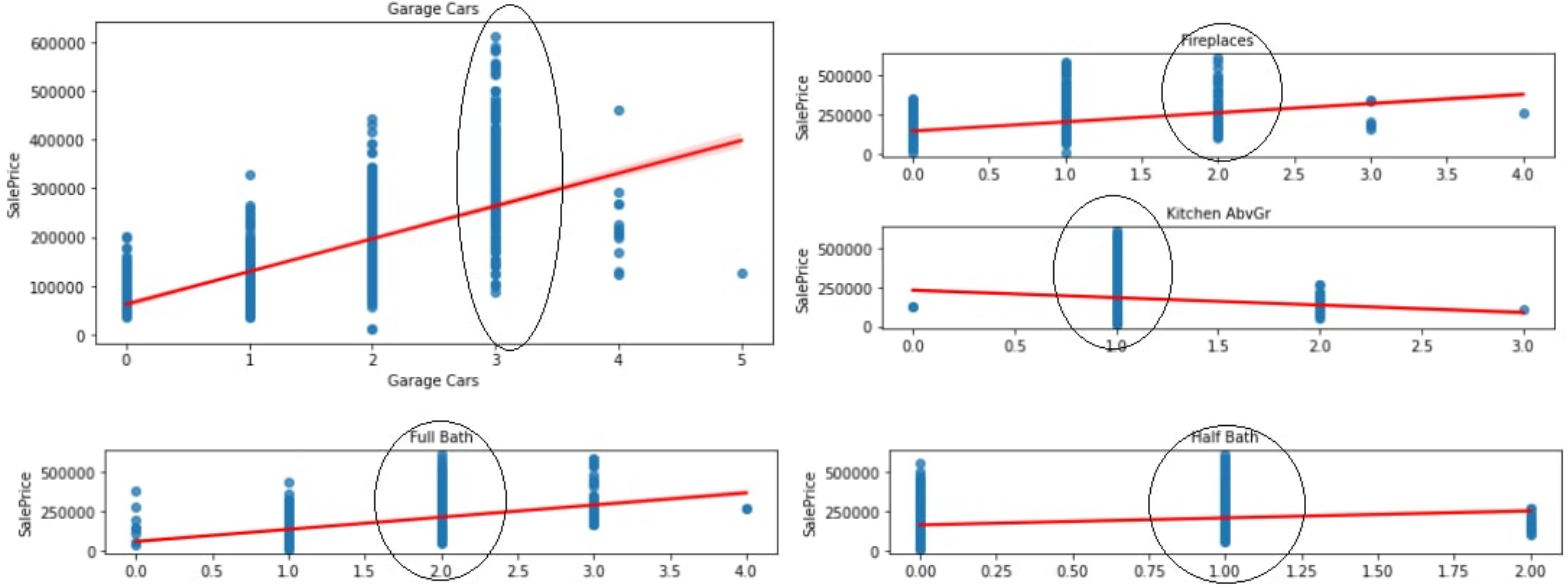
Overall Quad => Higher Price But not Overall Cond?





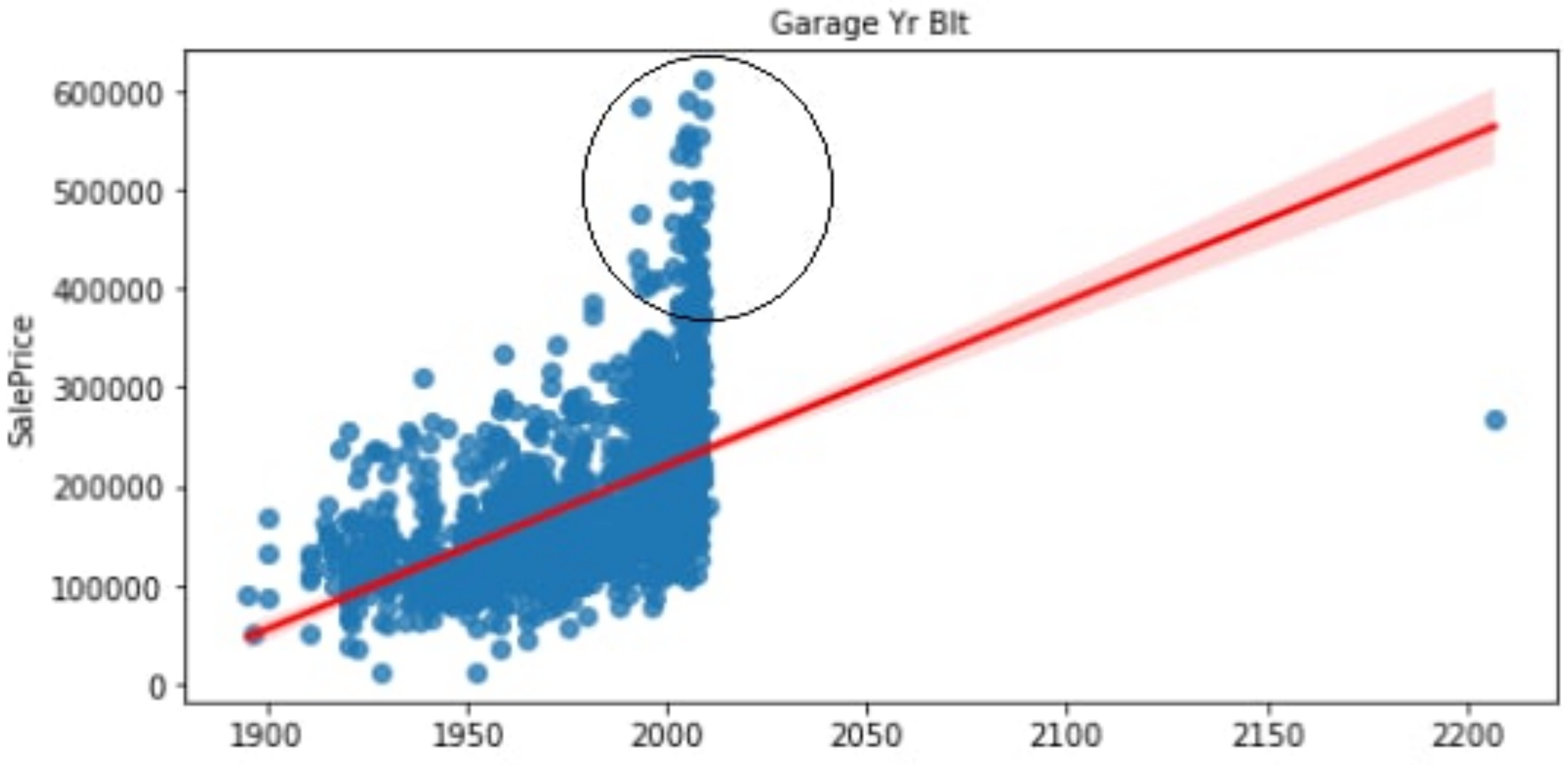
Doesn't Overall Qual affect Overall Cond?

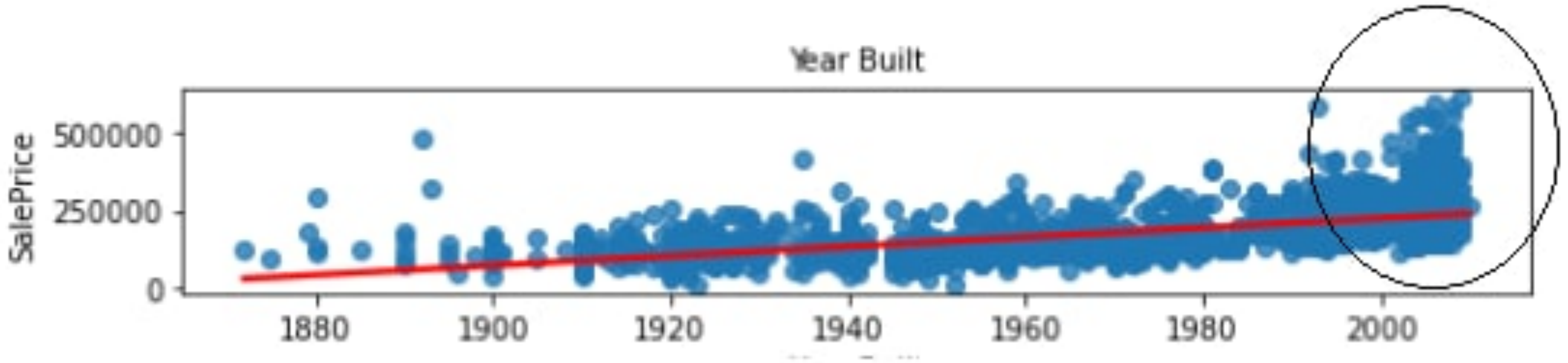
Facilities correlate with prices on specific quantity

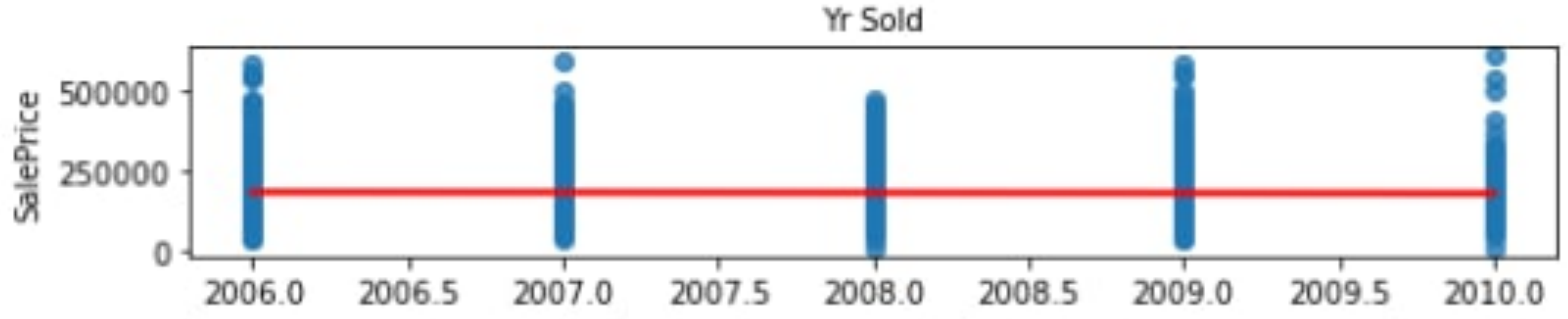


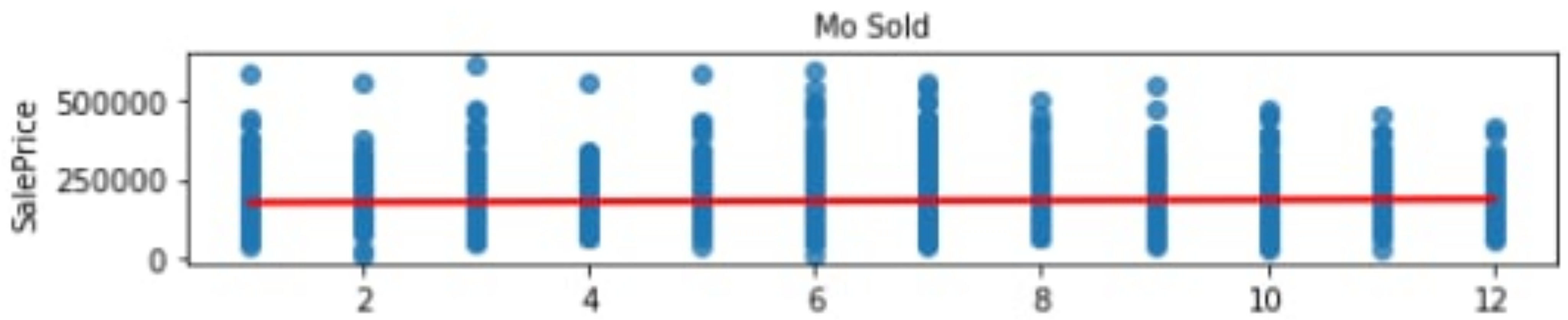
Too much of anything is bad

Prices shoot up from 2000, but no obvious trends for Mo Sold & Yr Sold



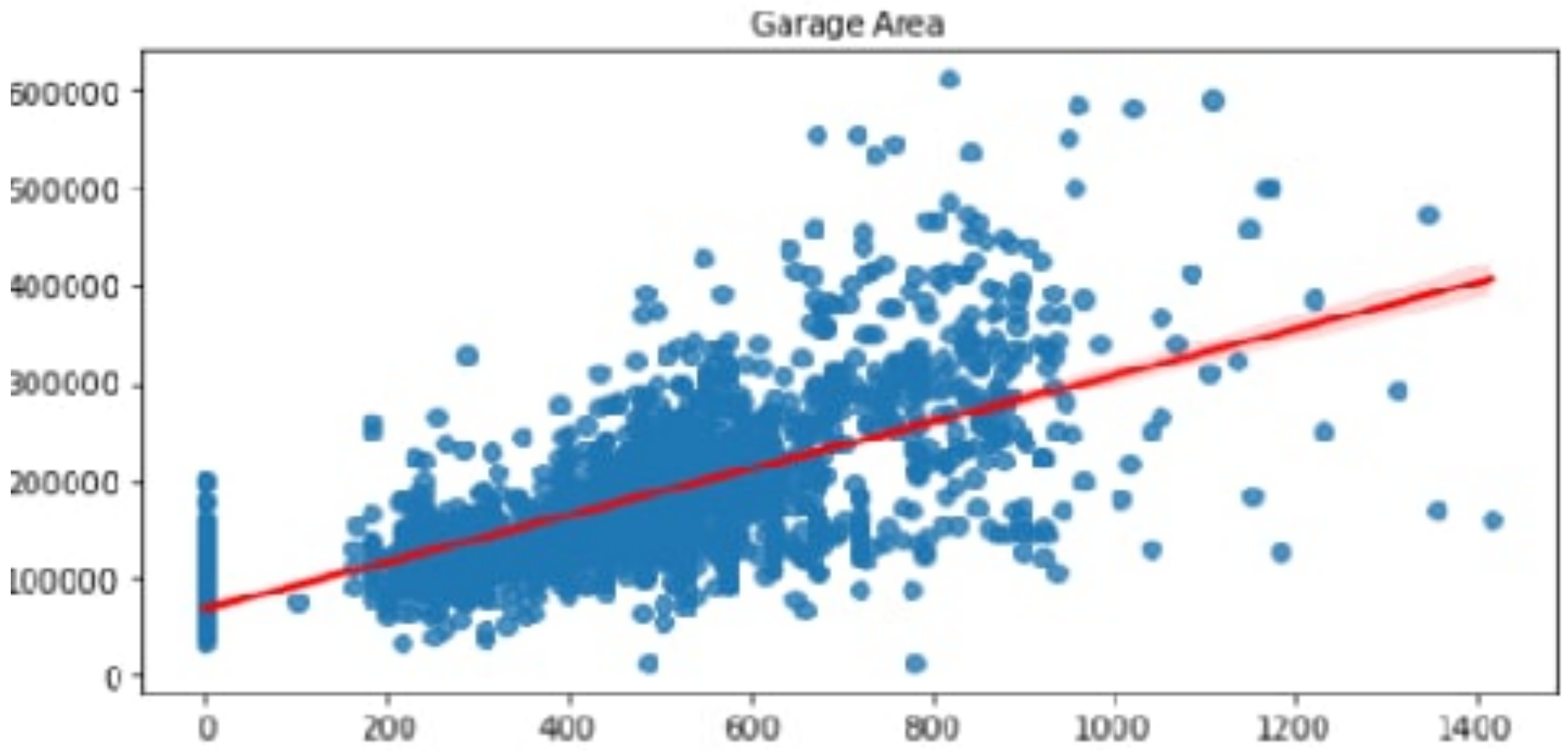


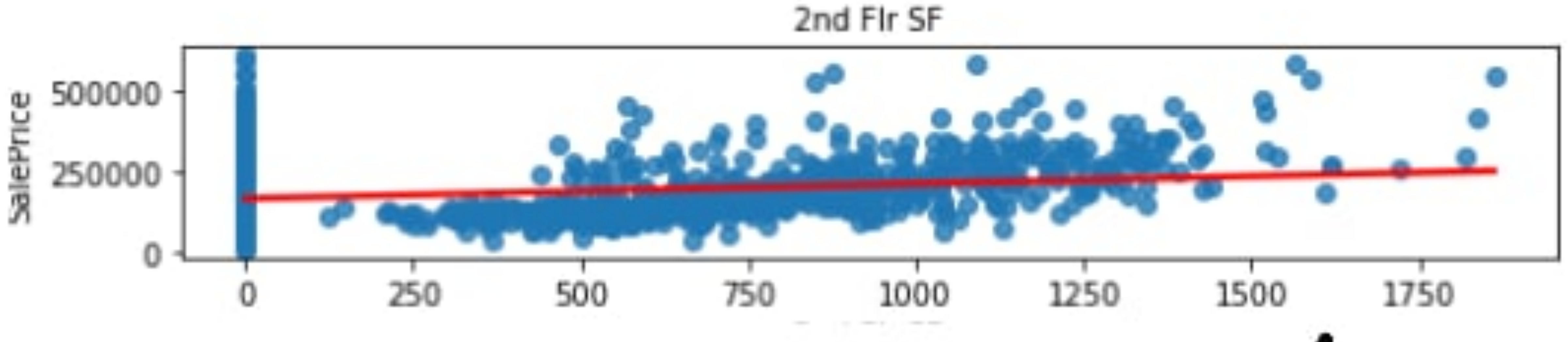


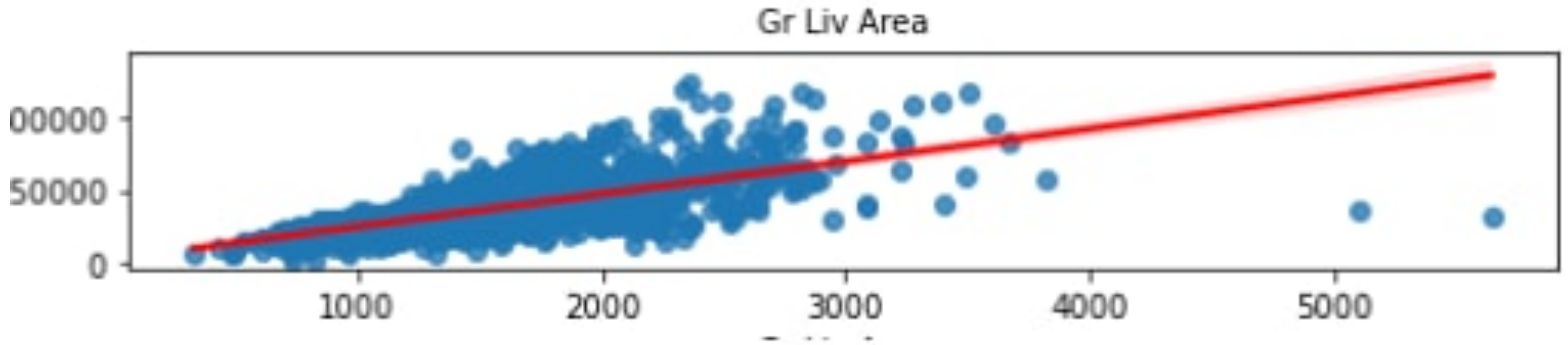


We don’t have to worry about GSS!

Larger sizes = More expensive?
Not really..

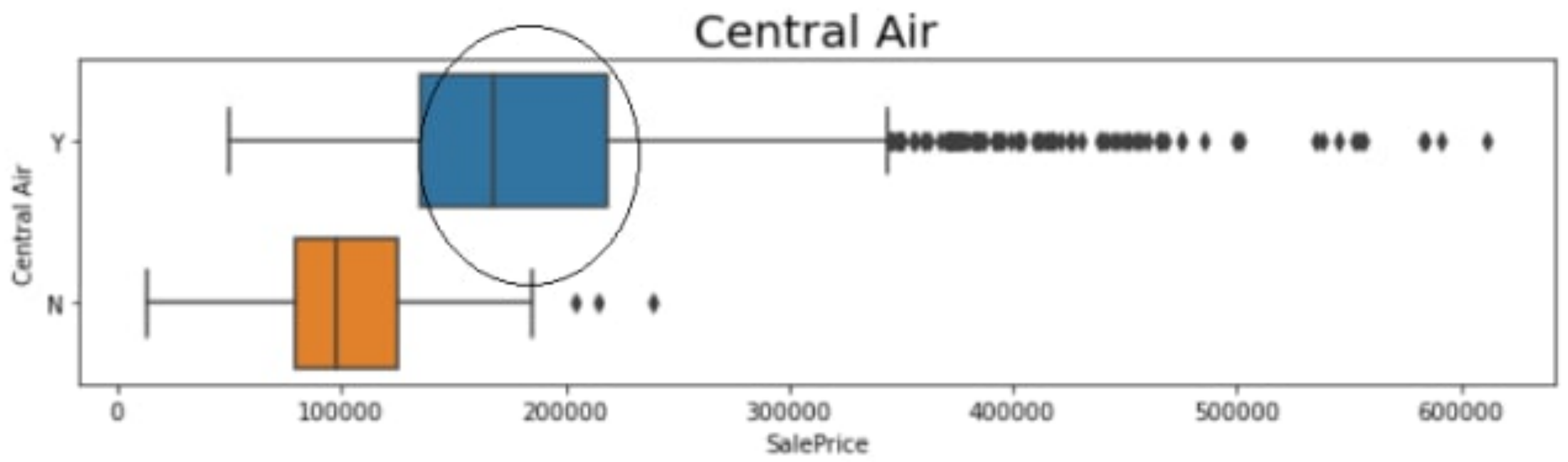


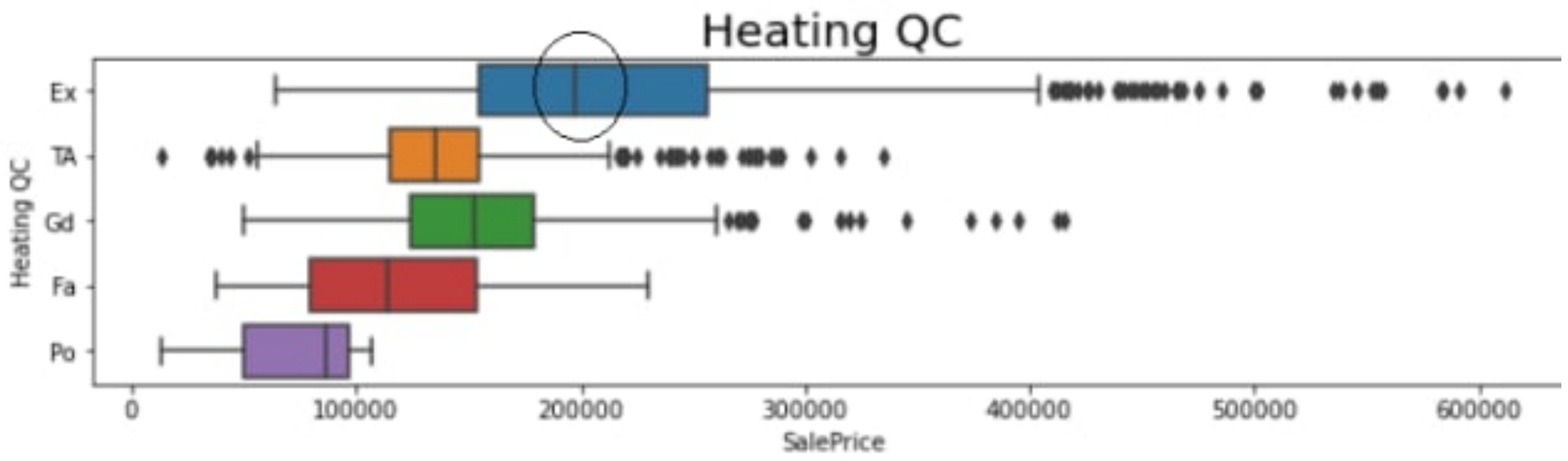


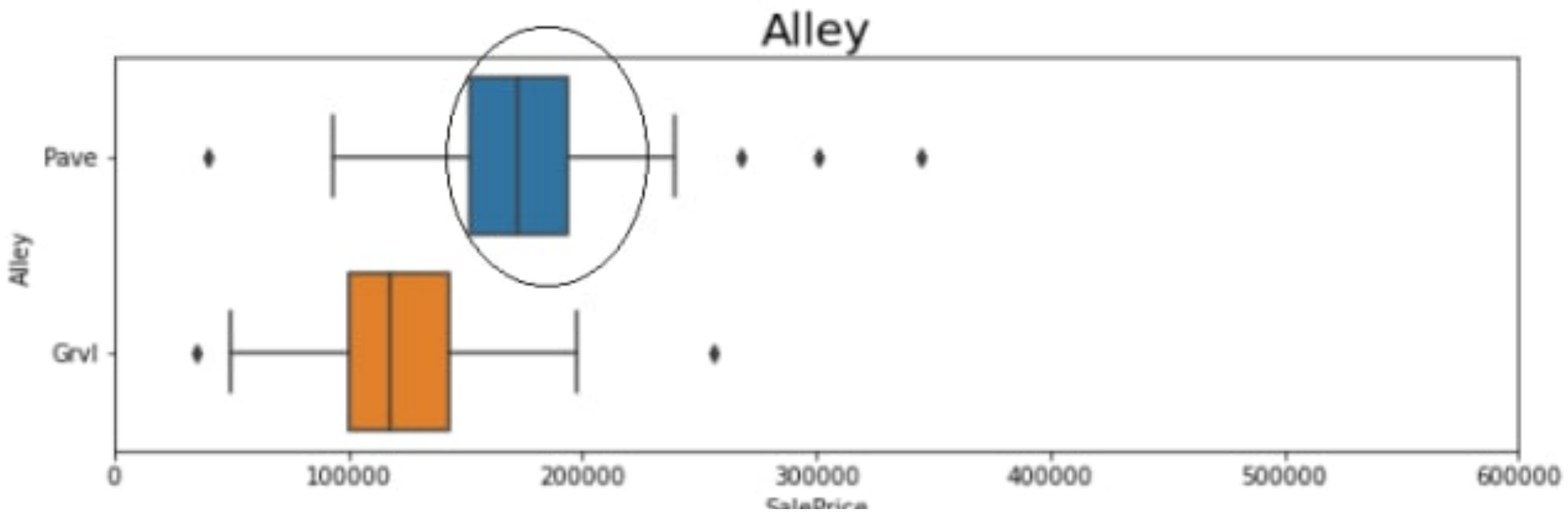


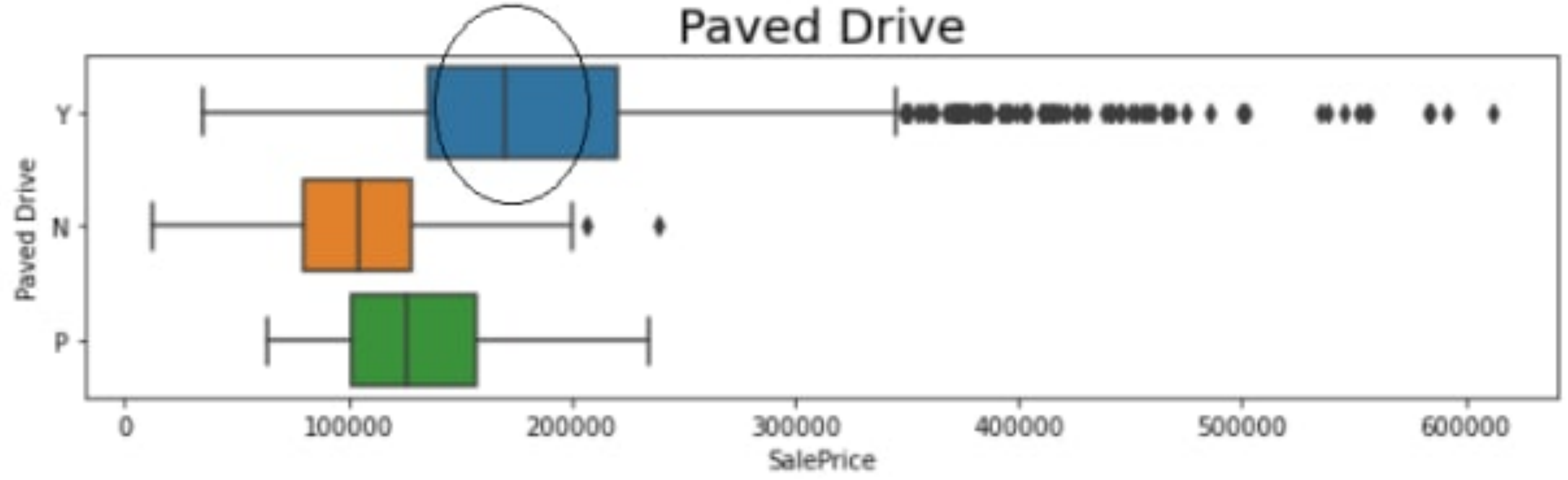
2-storey house but doesn't correlate?

People from Ames enjoy higher quality of life



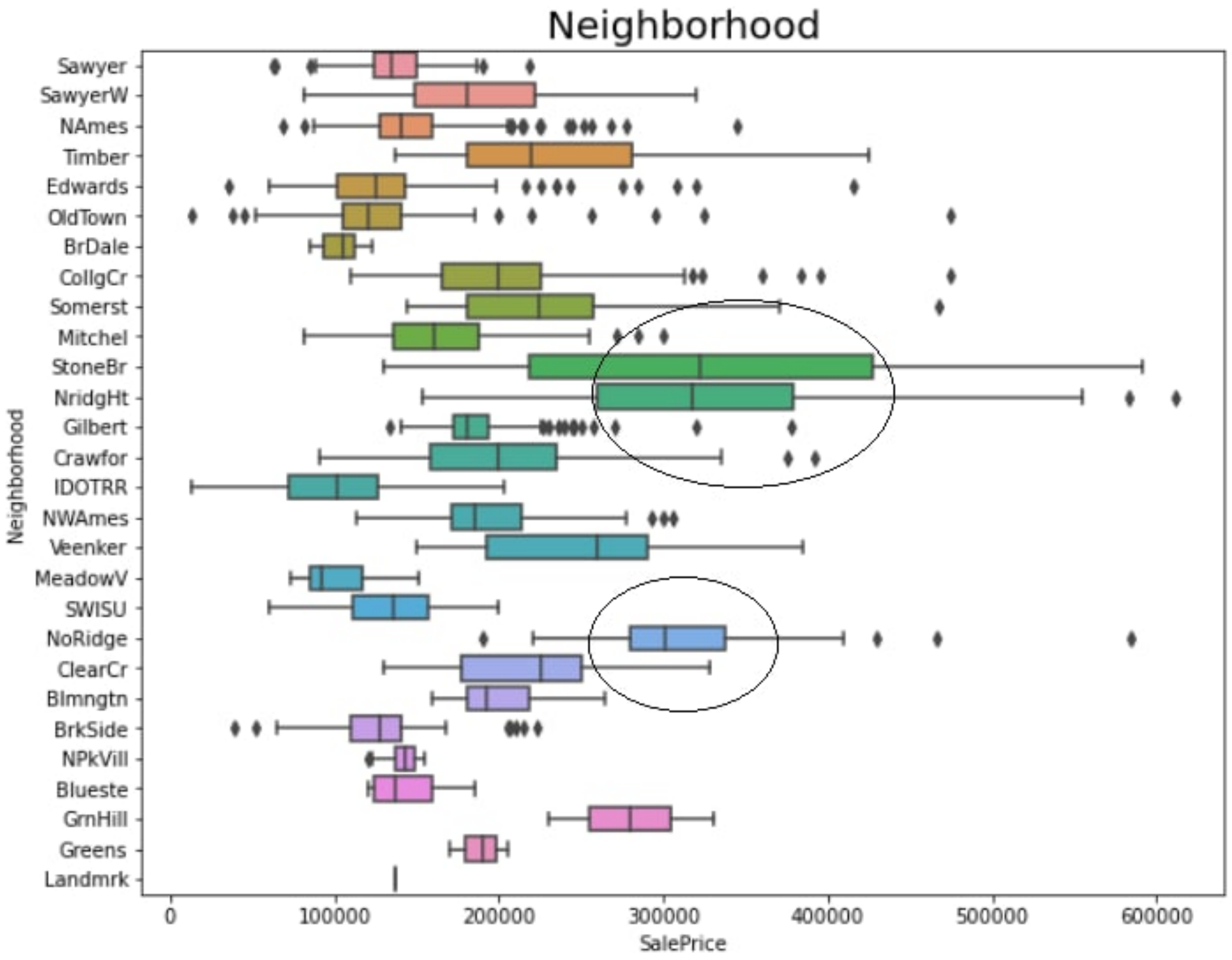






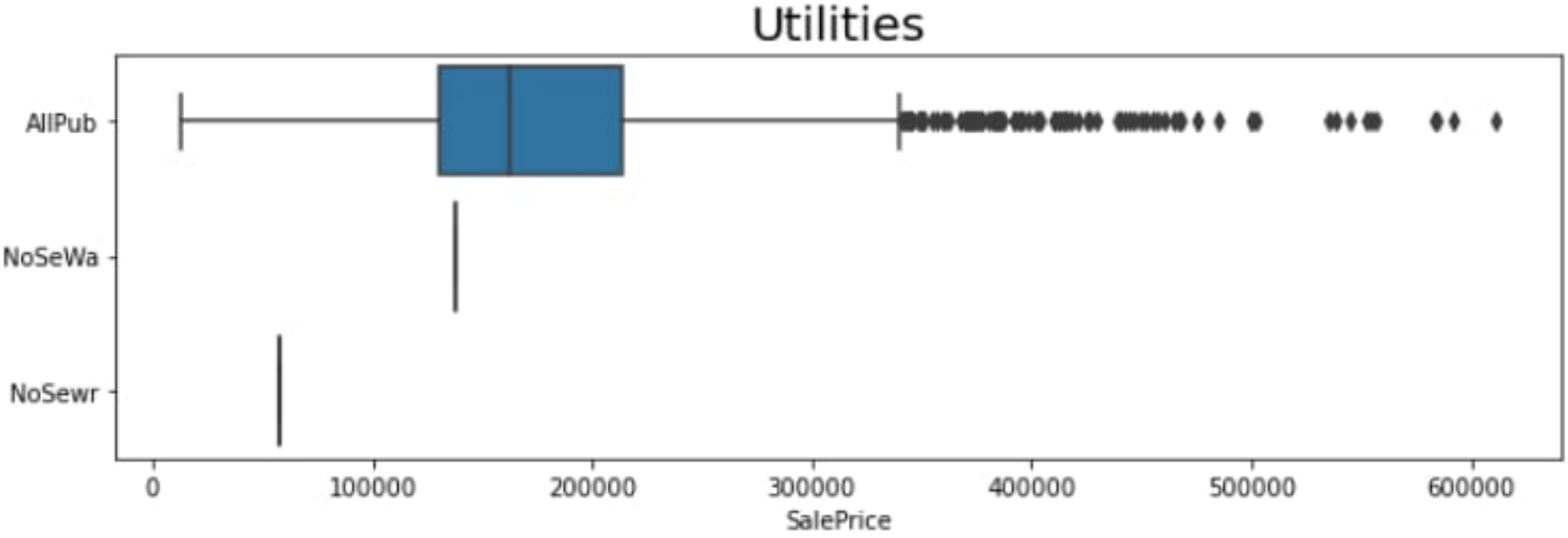
Despite outliers, we can tell the medians are high

Upper Bukit Timah of Ames



StoneBr, NridgHt and NoRidge

Some data may not be useful?



Data Cleaning and Feature Engineering

- Filling Discrete/Continuous Columns with 0/0.0
- Filling Categorical Columns with ‘None’
- Ordinate Categorical Columns
- Combine train_df and test_df
- Create dummies

(2929 rows, 207 columns)

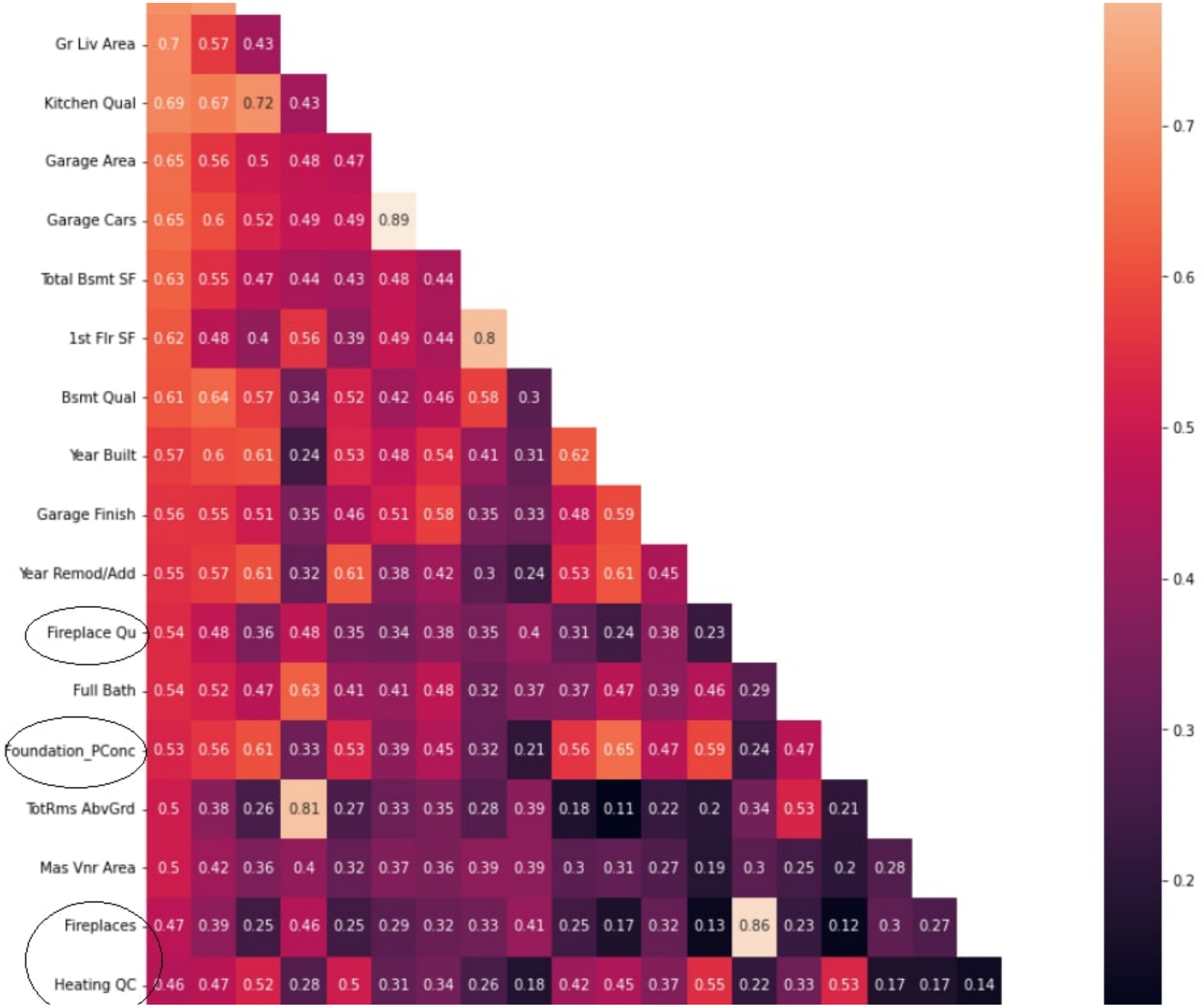
```
test_df['Pool QC'] = test_df['Pool QC'].fillna('None')
test_df['Pool QC'] = test_df['Pool QC'].map(ordinal_dict_pool_qc)


test_df['Fence'] = test_df['Fence'].fillna('None')
test_df['Fence'] = test_df['Fence'].map(ordinal_dict_fence)

test_df['Alley'] = test_df['Alley'].fillna('None')
```

`Garage Type`, `Garage Finish`, `Garage Qual`, `Garage Cond`

- Garage Type - NA represents No Garage
- Garage Finish - NA represents No Garage
- Garage Qual - NA represents No Garage
- Garage Cond - NA represents No Garage

```
train_df['Garage Type'] = train_df['Garage Type'].fillna('None')

train_df['Garage Finish'] = train_df['Garage Finish'].fillna('None')
```

Data Cleaning and Feature Engineering



Identify new correlations

Data Cleaning and Feature Engineering

Total SF = Total Bsmt SF + 1st Flr SF + 2nd Flr SF

Total Bath = Full Bath + Bsmt Full Bath +
(Half Bath + Bsmt Half Bath)/2

House Age = Year Remod/Add - Year Built

Drop Garage Area (Related to Garage Cars)

Drop Garage Yr Built (Similar to Year Built)

Drop Mo Sold/Yr Sold

Data Cleaning and Feature Engineering

```
In [118]: lasso_cv.coef_

Out[118]: array([-0.00000000e+00, -0.00000000e+00, -6.51144845e+03,  0.00000000e+00,
                  2.79166873e+03,  0.00000000e+00,  0.00000000e+00, -1.23418438e+02,
                  1.43231254e+04,  2.18515293e+03,  2.94191610e+03,  7.42130370e+03,
                  0.00000000e+00,  1.78393058e+02, -0.00000000e+00,  5.77929263e+03,
                  2.33591452e+03,  3.77364677e+03,  0.00000000e+00,  0.00000000e+00,
                 -0.00000000e+00,  2.16338328e+03,  0.00000000e+00, -0.00000000e+00,
                  1.35077010e+04, -0.00000000e+00, -0.00000000e+00,  5.93483077e+03,
                  2.28785932e+03,  8.27128290e+02,  1.88460673e+03,  2.27458031e+03,
                  0.00000000e+00,  5.21172134e+03,  0.00000000e+00, -0.00000000e+00,
                  0.00000000e+00,  1.20897111e+03,  0.00000000e+00, -0.00000000e+00,
                  0.00000000e+00,  3.50844984e+03,  0.00000000e+00, -3.82545757e+03,
                  0.00000000e+00, -6.48113181e+03, -0.00000000e+00,  0.00000000e+00,
                  0.00000000e+00,  0.00000000e+00,  0.00000000e+00, -3.44814384e+02,
                  0.00000000e+00,  0.00000000e+00, -0.00000000e+00,  2.47661179e+03,
                  5.19682805e+01,  0.00000000e+00,  1.25400170e+03, -0.00000000e+00,
                 -0.00000000e+00, -0.00000000e+00, -0.00000000e+00, -0.00000000e+00,
                  0.00000000e+00,  0.00000000e+00, -0.00000000e+00,  1.70701846e+03,
                 -1.67557048e+03, -0.00000000e+00, -0.00000000e+00,  3.03704063e+03,
                 -0.00000000e+00,  0.00000000e+00,  0.00000000e+00,  1.85176099e+02,
                 -0.00000000e+00,  0.00000000e+00, -0.00000000e+00,  4.74643697e+03,
                  9.25721929e+03, -7.95533506e+02, -0.00000000e+00, -0.00000000e+00,
                 -0.00000000e+00,  1.53358227e+03,  6.73275239e+03,  0.00000000e+00,
                  0.00000000e+00, -0.00000000e+00,  1.44270525e+03,  0.00000000e+00,
                  1.93259257e+03, -0.00000000e+00,  0.00000000e+00, -0.00000000e+00,
                 -0.00000000e+00,  0.00000000e+00, -0.00000000e+00,  8.65939601e+02,
                  4.17524622e+01,  0.00000000e+00, -0.00000000e+00,  0.00000000e+00,
                  0.00000000e+00, -0.00000000e+00, -0.00000000e+00, -1.03585146e+03,
                  0.00000000e+00,  0.00000000e+00,  6.34181269e+02, -0.00000000e+00,
```

Coefficients from cleaned train_df

	col	coef
1	Gr Liv Area	14824.986020
0	Overall Qual	13355.099107
2	Neighborhood_NridgHt	9854.146994
7	Misc Val	8758.278205
4	Neighborhood_StoneBr	7298.914406
6	MS SubClass	7071.280438
3	Exter Qual	7053.265099
5	Total Bath	6519.045652
9	Bsmt Exposure	6210.722005
8	Kitchen Qual	5611.552867
12	Neighborhood_NoRidge	5285.202739
11	Garage Cars	4993.785525
14	Pool QC	4472.986561
10	Total SF	4387.803985
18	Mas Vnr Area	4378.028353
15	BsmtFin SF 1	4203.257964
13	Sale Type_New	4168.229206
16	Screen Porch	4056.243491

Elimination through Lasso

Modelling

```
print("Linear X_train_sc RSME: ", np.sqrt(mean_
print("Linear X_test_sc RSME: ", np.sqrt(mean_
```

```
Linear X_train_sc RSME:  25065.339861612236
Linear X_test_sc RSME:  27369.77748892229
```

```
# with cross_val_score

lr_with_cv_score = cross_val_score(lr, X_train_
print('Linear Cross Val RSME:', np.sqrt(lr_wit
```

```
Linear Cross Val RSME: 31589.85329178855
```

```
In [154]: # calculate RSME

          print("Ridge X_train_sc RMSE:", np.sqrt(mean_s
          print("Ridge X_test_sc RMSE:", np.sqrt(mean_sc
```

```
Ridge X_train_sc RMSE: 25745.143371023085
Ridge X_test_sc RMSE: 27385.78143640325
```

```
In [155]: # with cross_val_score

          ridge_with_cv_score = cross_val_score(ridge, X
          print('Ridge Cross Val RSME:', np.sqrt(ridge_w
```

```
Ridge Cross Val RSME: 30234.070611169598
```

```
print("ElasticNet X_test_sc RMSE: ", np.sqrt(mean_s
```

```
ElasticNet X_train_sc RMSE: 26494.736587545307
ElasticNet X_test_sc RMSE: 27493.94750969941
```

```
# with cross_val_score

en_with_cv_score = cross_val_score(elasticnet, X_tr
print('ElasticNet Cross Val RSME:', np.sqrt(en_with
```

```
ElasticNet Cross Val RSME: 30201.149775451955
```

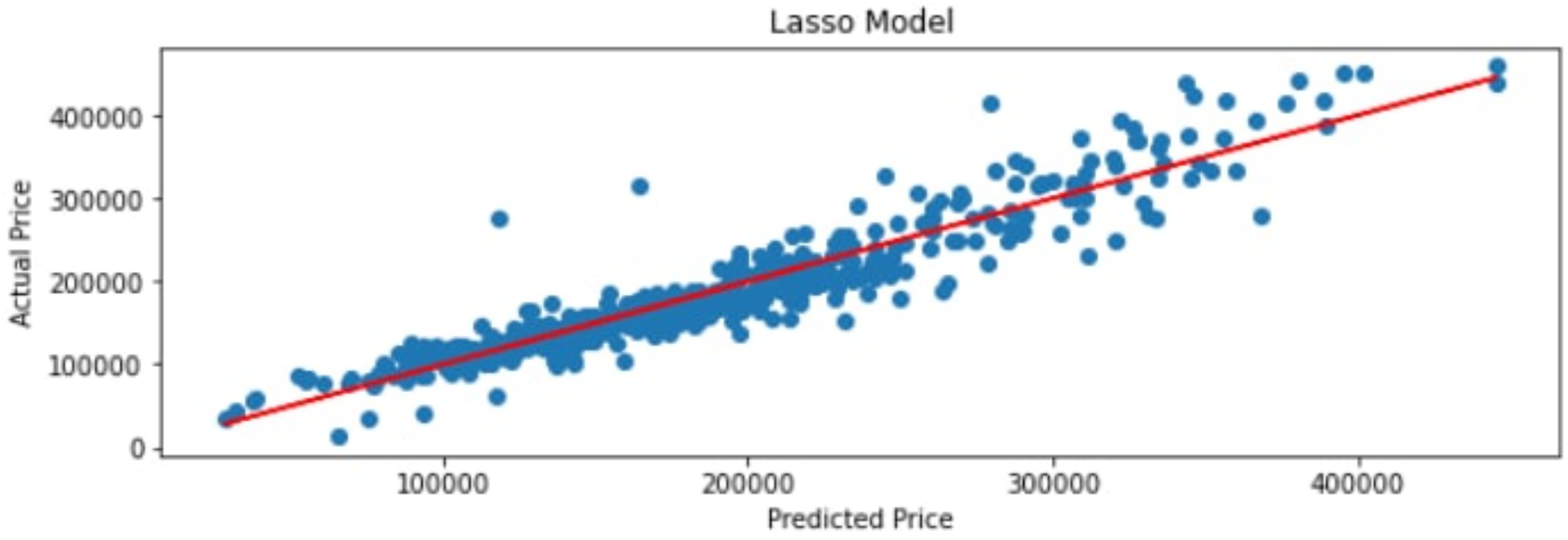
```
elasticnet.coef_[lasso.coef_ != 0]
```

Linear Regression

Ridge Regression

Elastic Net

Modelling

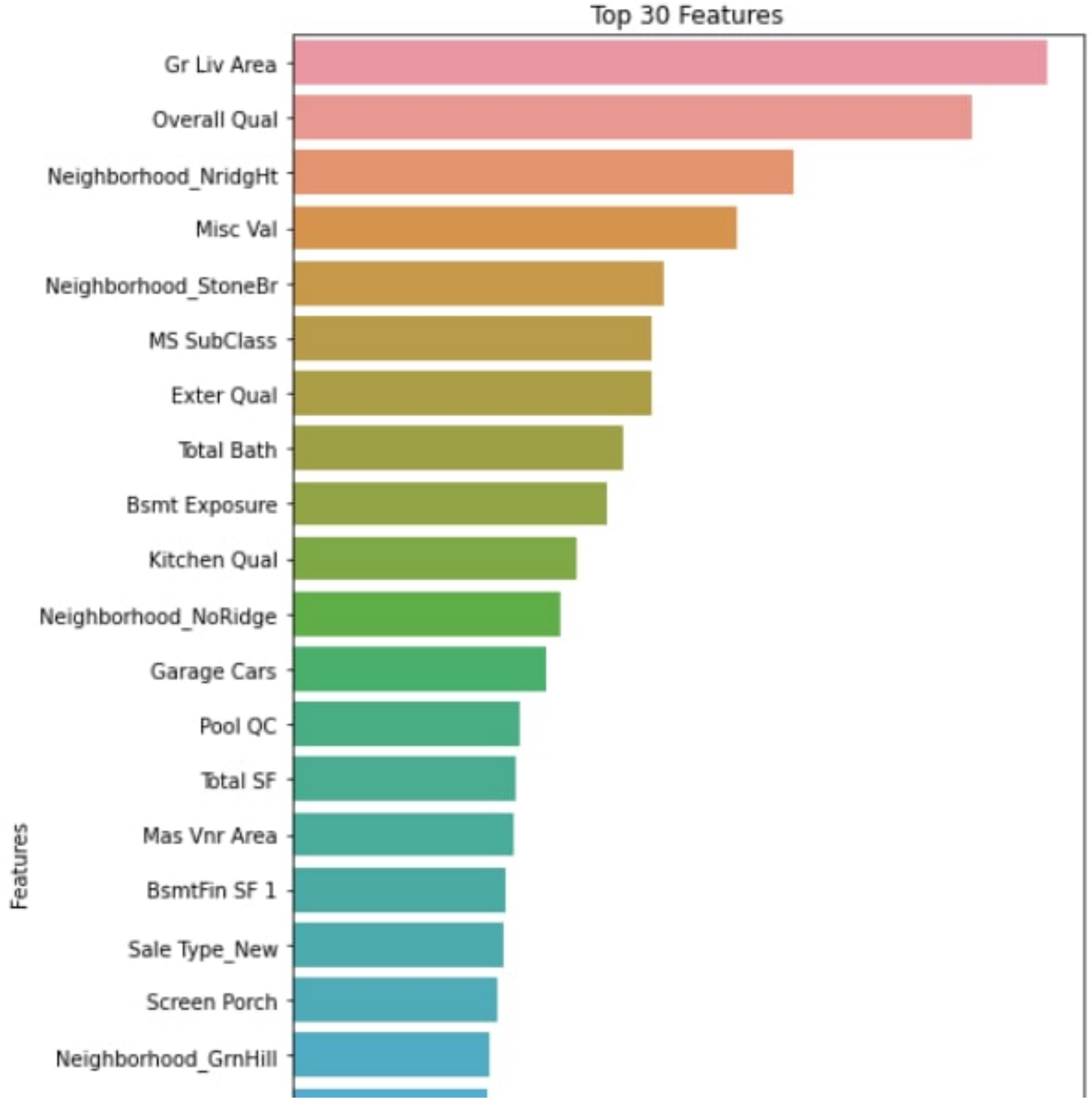



```
]: # calculate RSME

   print("Lasso X_train_sc RSME:", np.sqrt(mean
   print("Lasso X_test_sc RSME:", np.sqrt(mean_

   Lasso X_train_sc RSME: 25380.829651096446
   Lasso X_test_sc RSME: 26612.041524452397

]: # with cross_val_score

   lasso_with_cv_score = cross_val_score(lasso,
   print('Lasso Cross Val RSME:', np.sqrt(lasso

   Lasso Cross Val RSME: 31325.55036922845
```

Lasso Regression

Top Features

	col	coef
1	Gr Liv Area	14824.986020
0	Overall Qual	13355.099107
2	Neighborhood_NridgHt	9854.146994
7	Misc Val	8758.278205
4	Neighborhood_StoneBr	7298.914406
6	MS SubClass	7071.280438
3	Exter Qual	7053.265099
5	Total Bath	6519.045652
9	Bsmt Exposure	6210.722005
8	Kitchen Qual	5611.552867
12	Neighborhood_NoRidge	5285.202739
11	Garage Cars	4993.785525
14	Pool QC	4472.986561
10	Total SF	4387.803985
18	Mas Vnr Area	4378.028353
15	BsmtFin SF 1	4203.257964
13	Sale Type_New	4168.229206
16	Screen Porch	4056.243491
17	Neighborhood_GrnHill	3865.877436
28	Misc Feature_Gar2	3837.629567
31	Misc Feature_Othr	3246.811889
22	Roof Matl_WdShngl	3144.509574
26	Overall Cond	3024.170805
20	Roof Style_Hip	2992.603029
19	Lot Area	2832.145313
24	TotRms AbvGrd	2620.270655



NridgHt and StoneBr!

What to look out for?

Size of the interior of the house
Neighborhood (surroundings)
High ratings for conditions of the house
Misc Values of upgrades, additional facilities

Areas of Improvement

Learn and apply various techniques
ANOVA correlation coefficient?
Kendall's rank coefficient?
Variance Analysis?
Recursive Feature Elimination?

To reduce noise from too many features and amplifying signals to better our predictions

Thank you