Ames Housing Challenge

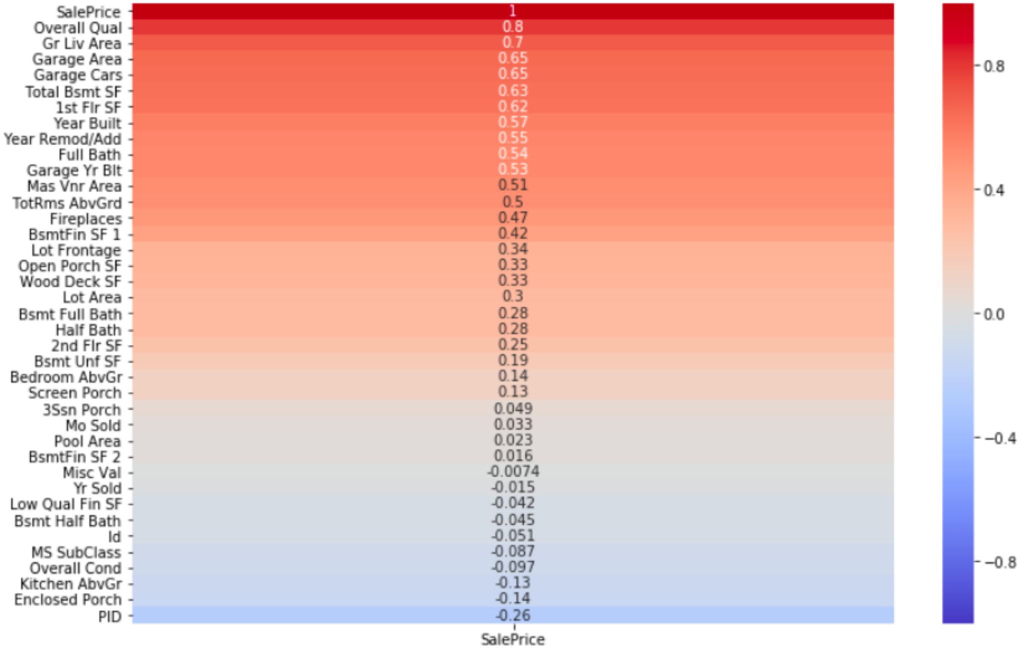
- Antony Paulson Chazhoor

Problem Statement

- Create a regression model based on the Ames Housing Dataset which will predict the price of a house at sale
  - Models are primarily Linear Regression Models, Lasso & Ridge
  - Evaluation Parameter : RMSE

- Why is this analysis important?
  - An analyst can help prospective buyers to understand in cost estimation
  - Overpriced or Underpriced(Mega Deals!!) Houses can be spotted

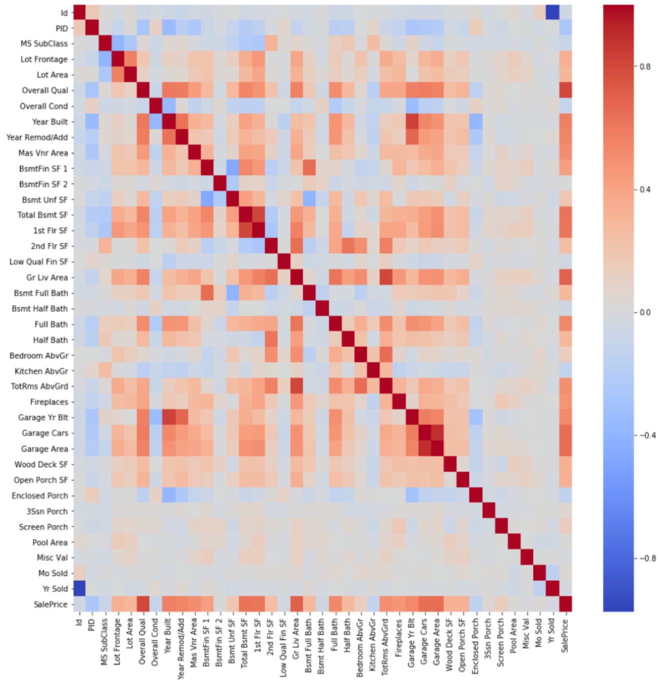
Exploratory Analysis



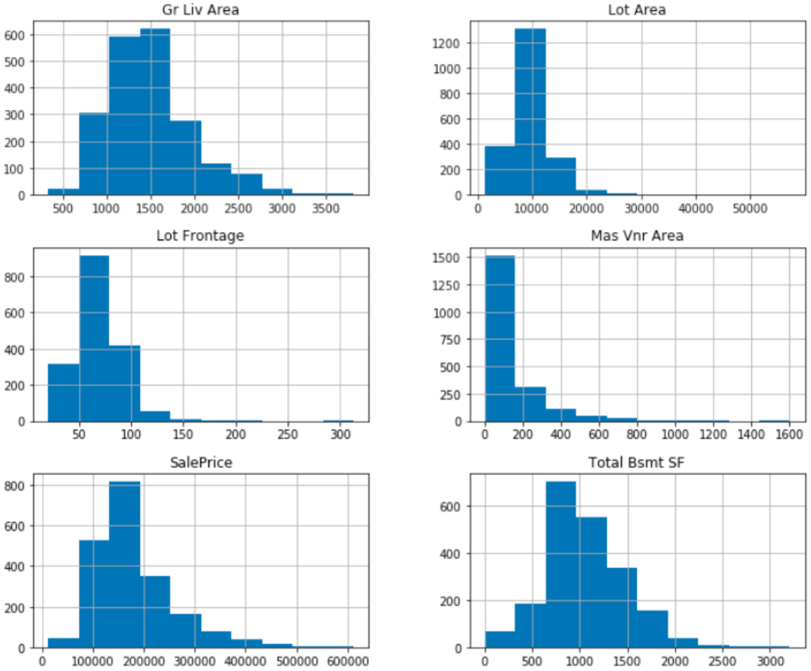
Predictors that are highly correlated with Target(Sale Price)

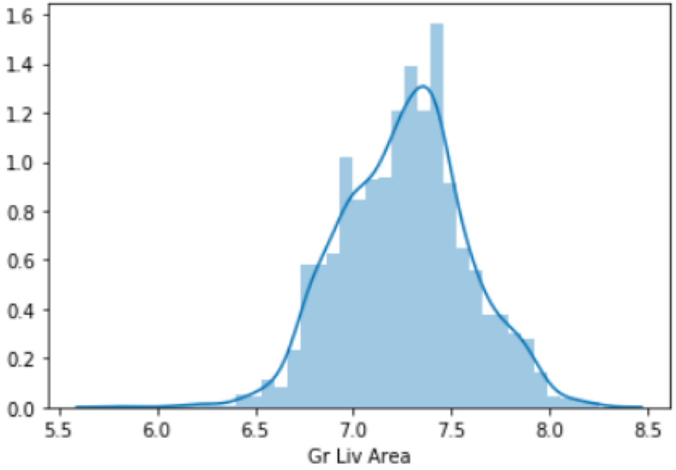
Exploratory Analysis(Contd)

Extremely correlated variables : Garage Area, Garage Cars



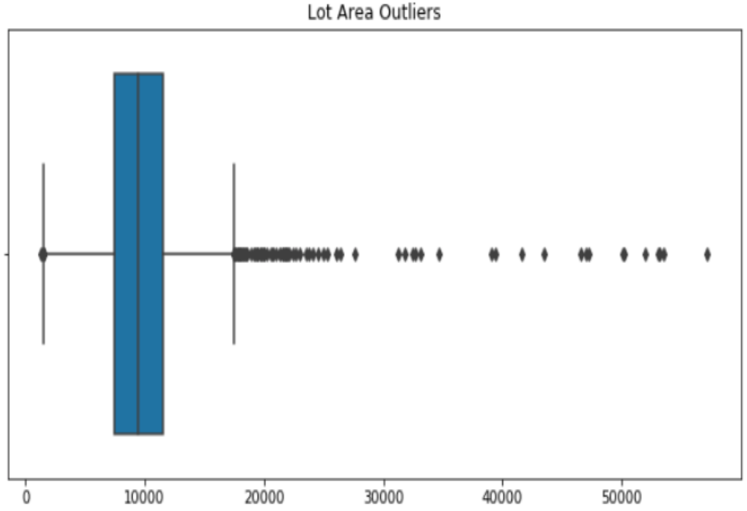
Right Skewed Data



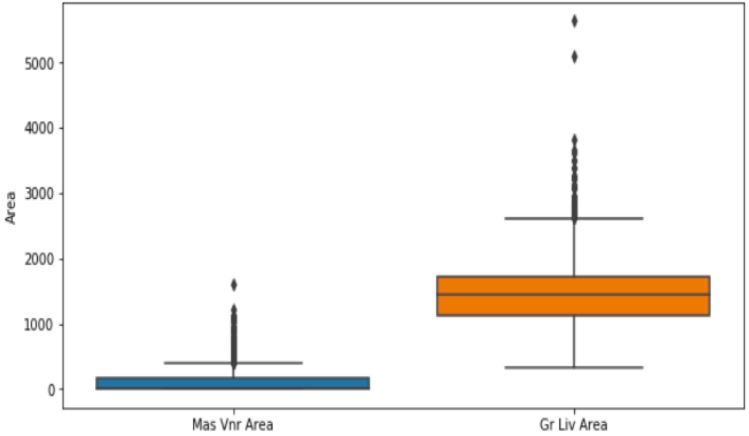


These can be brought to the normal distribution form through log transformations

Extreme Outliers in the Data



Lot Area Outliers



Removing these greatly improved RMSE Values

Handling Missing Values

- The Most tricky part in the EDA process

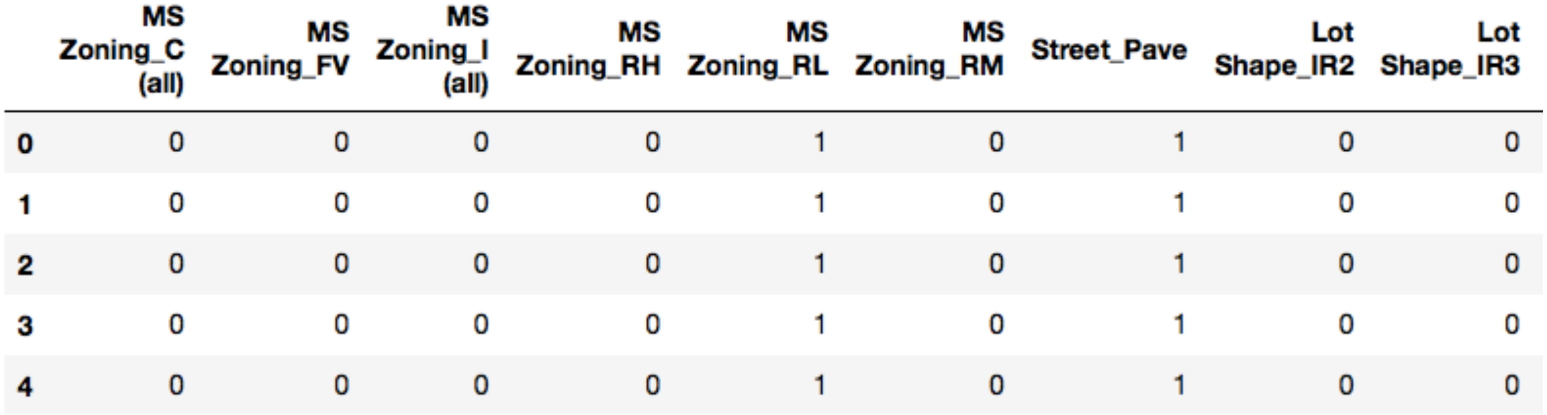
- I used three values to fill my null values:
  - ‘0’ : For numeric data that was missing
  - ‘Appropriate String’ for categorical predictors
  - ‘Forward Fill’ : Single missing value.

- Dropped columns with more than 1000 missing values

Preprocessing & Modelling Steps

- One hot encoding of categorical variables was absolutely crucial
- Log transformation of the predicted column SalePrice
- The next step would entail splitting and testing the model
- Scaling the data is necessary for application into Ridge, Lasso models
- First assessment based of the linear regression RMSE
- Cross Validation scores compared across Ridge, Lasso and Linear regression models
- Application of Lasso Model to the Scaled Test data

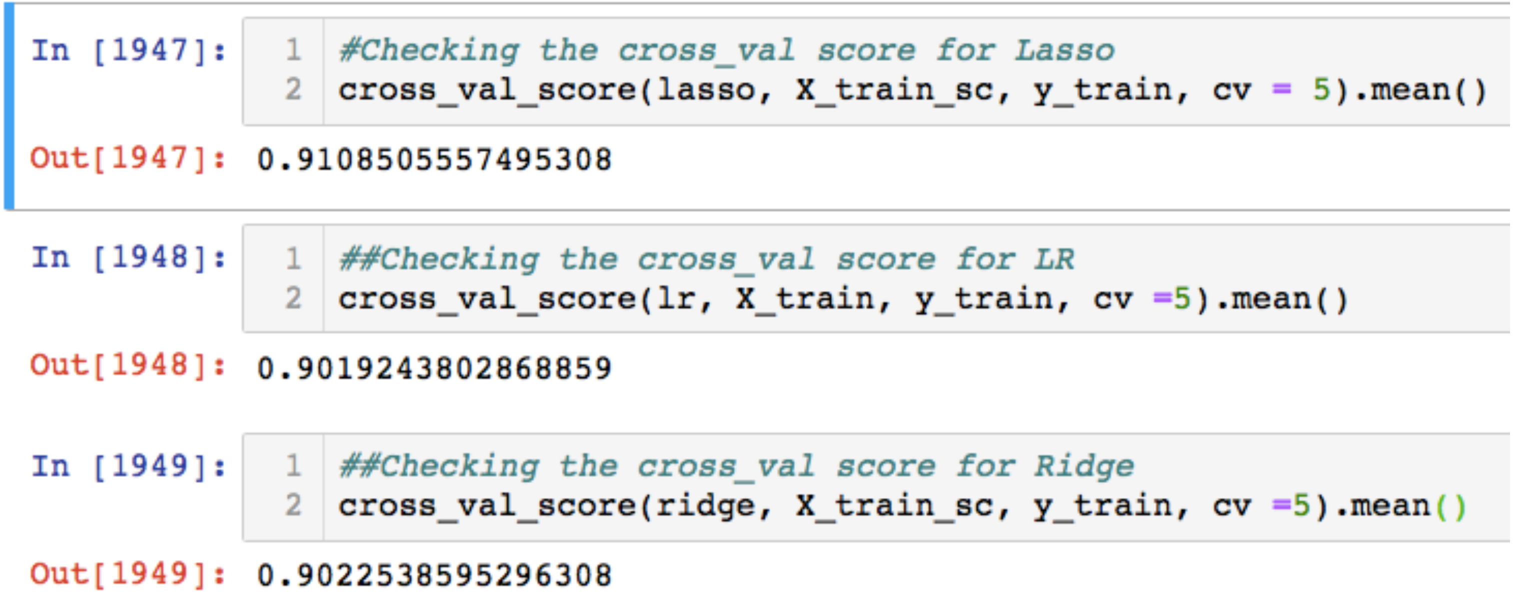
Creation of Dummy Variables



	MS Zoning_C (all)	MS Zoning_FV	MS Zoning_I (all)	MS Zoning_RH	MS Zoning_RL	MS Zoning_RM	Street_Pave	Lot Shape_IR2	Lot Shape_IR3
0	0	0	0	0	1	0	1	0	0
1	0	0	0	0	1	0	1	0	0
2	0	0	0	0	1	0	1	0	0
3	0	0	0	0	1	0	1	0	0
4	0	0	0	0	1	0	1	0	0

- The number of categorical variables: 37
- Number of dummy columns created : 214

Comparison of Cross validation Scores



```
In [1947]: #Checking the cross_val score for Lasso
           cross_val_score(lasso, X_train_sc, y_train, cv = 5).mean()

Out[1947]: 0.9108505557495308
```

```
In [1948]: ##Checking the cross_val score for LR
           cross_val_score(lr, X_train, y_train, cv =5).mean()

Out[1948]: 0.9019243802868859
```

```
In [1949]: ##Checking the cross_val score for Ridge
           cross_val_score(ridge, X_train_sc, y_train, cv =5).mean()

Out[1949]: 0.9022538595296308
```

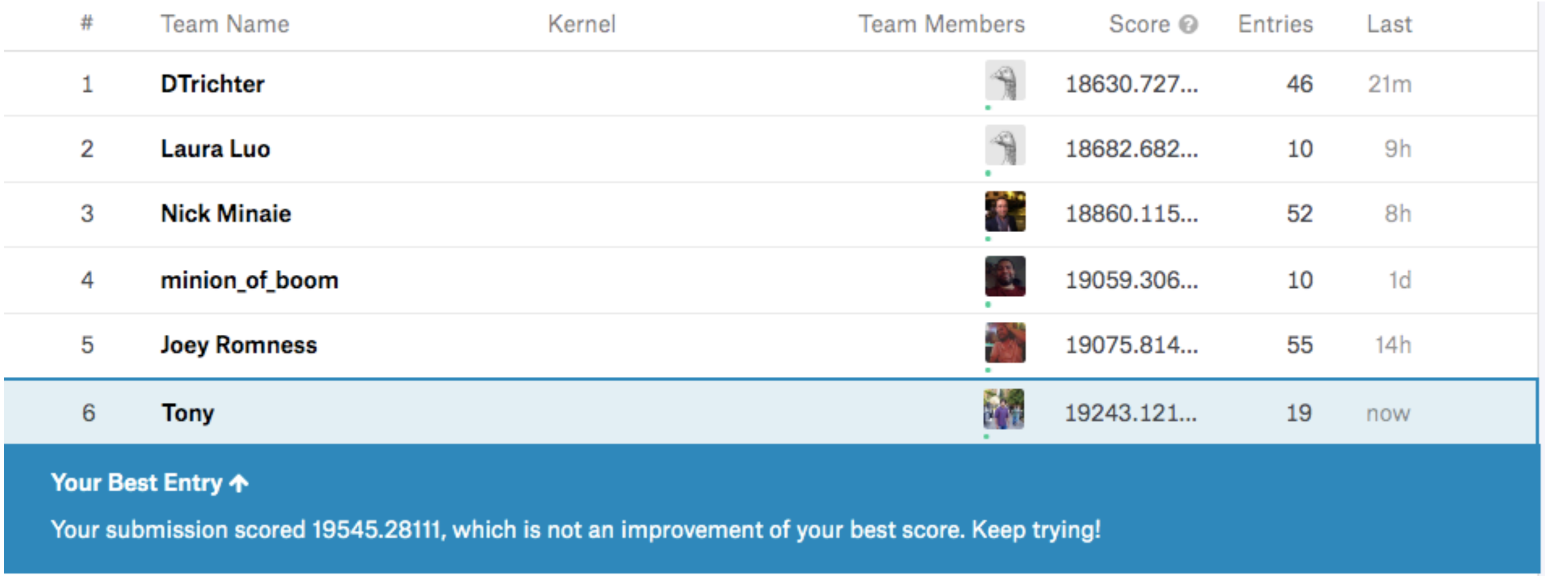
Results

- The Lasso model from sklearn provided the best accuracy among models
- About 91% of change in the Sales Price could be accounted for by the variables in my model.
- The most correlated columns to sales price were quality & Living area
- The lowest RMSE obtained by the model was approx. 19243.

Future Improvements

- The model can be improved further by implementing inferences from the EDA:
  - Log transformations
  - Dropping highly inter correlated columns
  - Excluding more outliers
  - Eliminating predictors which have very little effect on sale Price

Results of Predictions



#	Team Name	Kernel	Team Members	Score	Entries	Last
1	**DTrichter**			18630.727...	46	21m
2	**Laura Luo**			18682.682...	10	9h
3	**Nick Minaie**			18860.115...	52	8h
4	**minion_of_boom**			19059.306...	10	1d
5	**Joey Romness**			19075.814...	55	14h
6	**Tony**			19243.121...	19	now

Your Best Entry ↑

Your submission scored 19545.28111, which is not an improvement of your best score. Keep trying!