Project 3: Ames Housing Data

Noelia Lopez

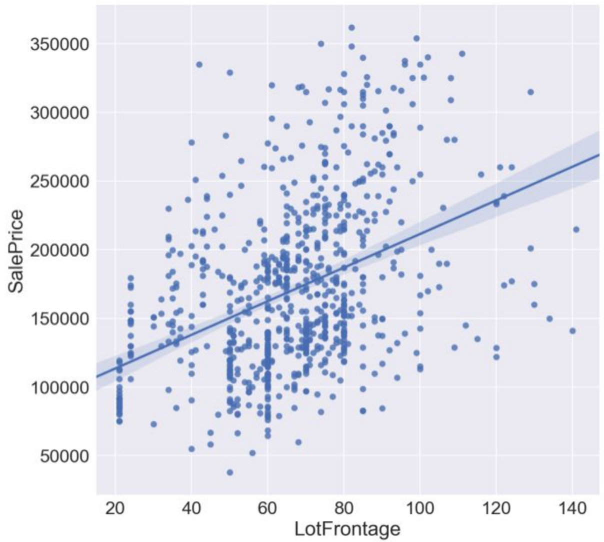
1 . Estimating the value of homes from fixed characteristics

Understanding the Data and data Cleaning:

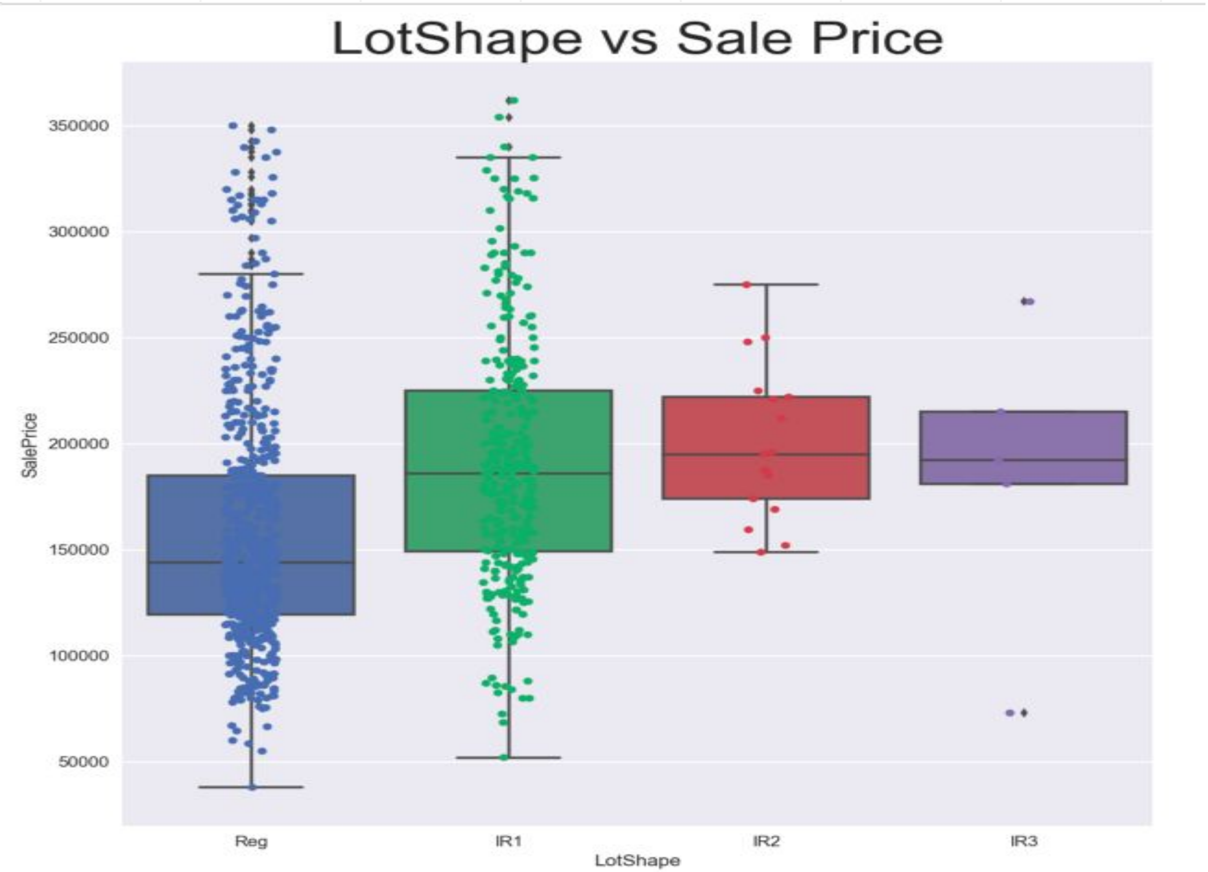
- Deal with NaN values,
- Visualize heatmap and scatter plot to see the correlation with Sales Price and analyze how important the Features are

```
]: #We check out how many null values there are in the data
   nulls = house.isnull().sum()
   nulls.sort_values(ascending=False).head(19)
```

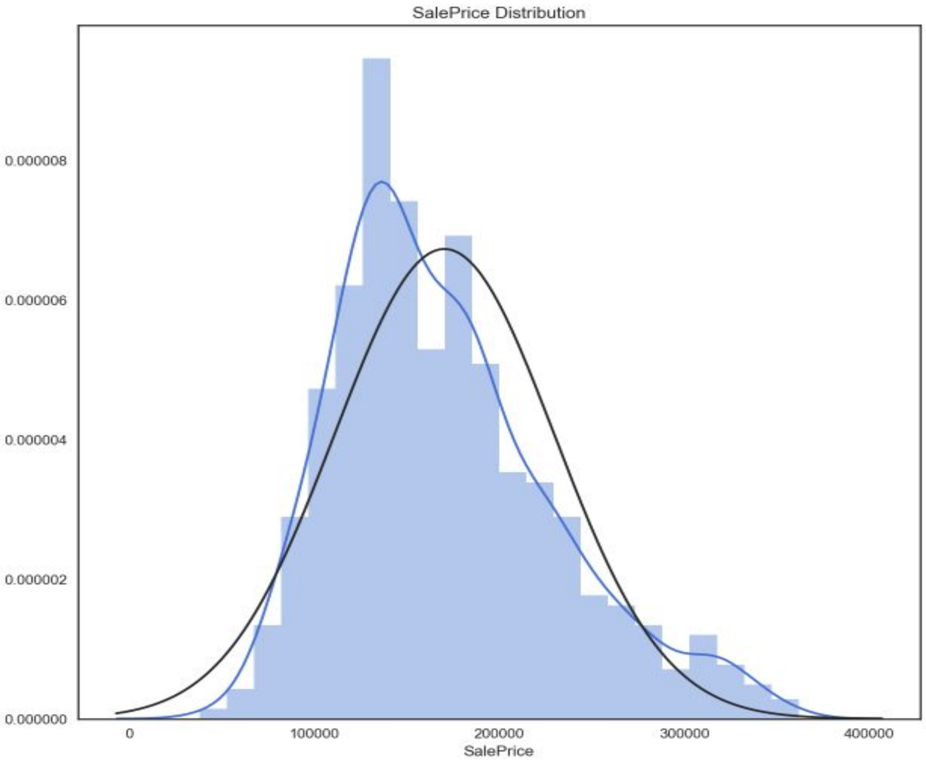
```
]: PoolQC          1453
   MiscFeature     1406
   Alley           1369
   Fence           1179
   FireplaceQu      690
   LotFrontage      259
   GarageCond        81
   GarageType        81
```



OUTLIERS


LotShape vs Sale Price
SalePrice
350000
300000
250000
200000
150000
100000
50000
Reg
IR1
IR2
IR3
LotShape

Sale Price Distribution




```
count        962.000000
mean      170308.391892
std        59317.974280
min        37900.000000
25%       128000.000000
50%       159217.000000
75%       202975.000000
max       361919.000000
Name: SalePrice, dtype: float64
```

3. Identifying fixed features that can predict price

In [889]:

```
# Load the data with just the fixed variables
fixed_columns= ['Neighborhood','YearBuilt','1stFlrSF','2ndFlrSF', 'PoolArea', 'LotConfig',
                'LotArea','MasVnrArea','MasVnrType', 'LandSlope', 'GarageCars','LowQualFinSF', 'BedroomAbvGr',
                'LotShape','LandContour','Foundation', 'TotRmsAbvGrd','BsmtExposure', 'BsmtFinSF1','TotalBsmtSF',
                'Exterior1st','Exterior2nd', 'GrLivArea', 'KitchenAbvGr', 'LotFrontage', 'OverallQual']
print len(fixed_columns)
fixed_df=house[fixed_columns]
fixed_dummies_df=get_dummies(fixed_df)
fixed_dummies_df.head()
```

```
26
```

Out[889]:

	YearBuilt	1stFlrSF	2ndFlrSF	PoolArea	LotArea	MasVnrArea	GarageCars	LowQualFinSF	BedroomAbvGr	TotRmsAbvGrd	BsmtFinSF1	TotalBsmtSF	GrLi
0	2003	856	854	0	8450	196.0	2	0	3	8	706	856	
2	2001	920	866	0	11250	162.0	2	0	3	6	486	920	
4	2000	1145	1053	0	14260	350.0	3	0	4	9	655	1145	
6	2004	1694	0	0	10084	186.0	2	0	3	7	1369	1686	
10	1965	1040	0	0	11200	0.0	1	0	3	5	906	1040	

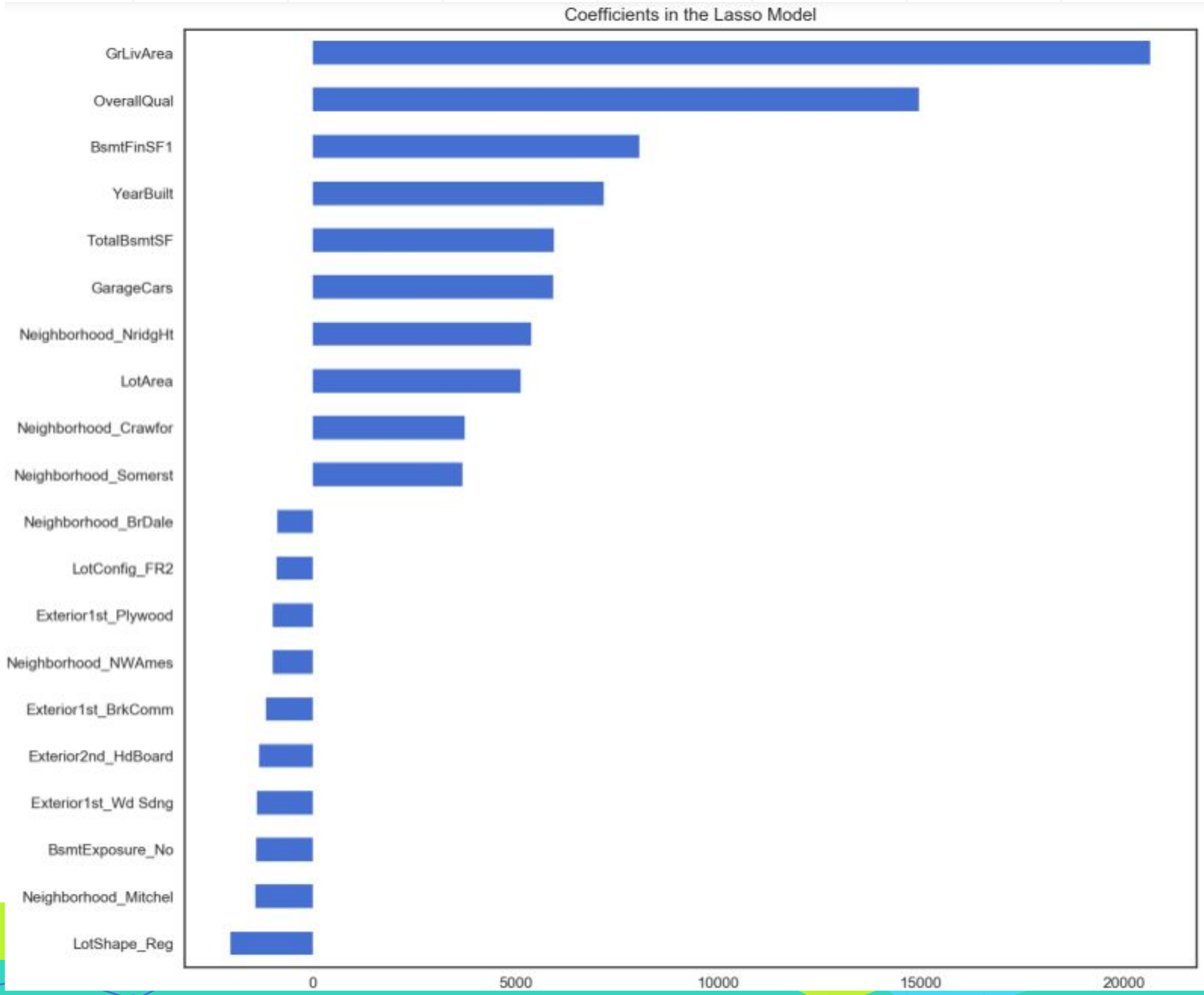
Future Selection: KBest, RFE, and Feature elimination using the lasso penalty

	mean score	std score
kbest	0.878287	0.017672
rfecv	0.885174	0.020112
Randomized Lasso	0.885174	0.020112

Prediction Model: Ridge, Lasso & Elasticnet

```
--------------------------------------------------------------------------
Ridge Regularization:
('R2 Train:', 0.89211624835303405)
('R2 Test:', 0.78392574164839646)
--------------------------------------------------------------------------
Lasso Regularization:
('R2 Train', 0.89341239658315419)
('R2 Test', 0.82526122209168329)
--------------------------------------------------------------------------
ElasticNet Regularization:
('R2 Train', 0.89362609147442795)
('R2 Test', 0.76025955169426951)
--------------------------------------------------------------------------
```

Lasso regularization has better model


Coefficients in the Lasso Model
GrLivArea
OverallQual
BsmtFinSF1
YearBuilt
TotalBsmtSF
GarageCars
Neighborhood_NridgHt
LotArea
Neighborhood_Crawfor
Neighborhood_Somerst
Neighborhood_BrDale
LotConfig_FR2
Exterior1st_Plywood
Neighborhood_NWAmes
Exterior1st_BrkComm
Exterior2nd_HdBoard
Exterior1st_Wd Sdng
BsmtExposure_No
Neighborhood_Mitchel
LotShape_Reg
0
5000
10000
15000
20000

2. Determine any value of *changeable* property characteristics unexplained by the *fixed* ones

Solution Approach

Run the model to find the predicted price

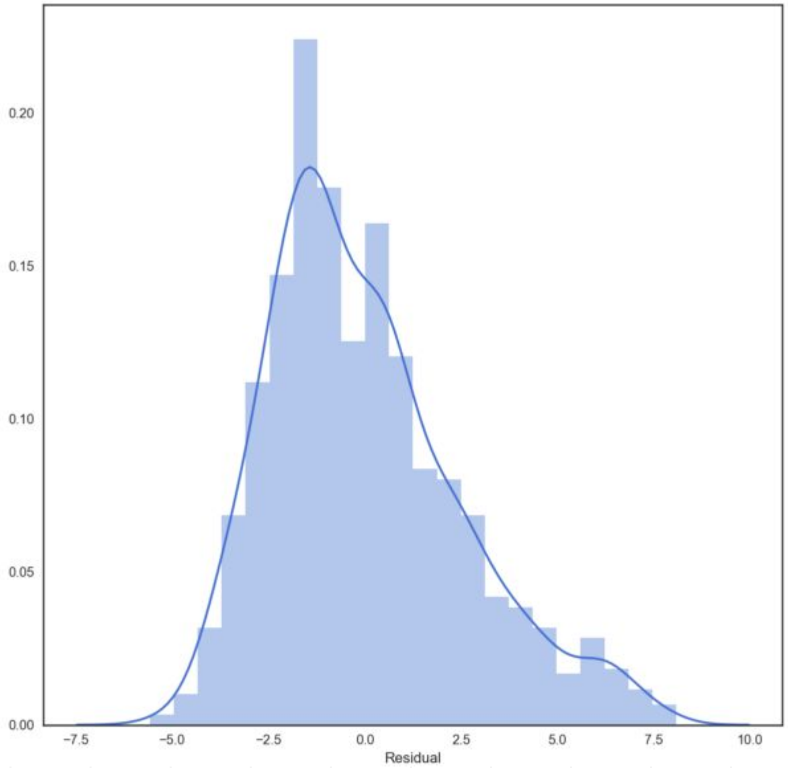
Find the differences between the predicted price and the real price= RESIDUAL

This RESIDUAL value will be the target variable for renovate-able features, and we will make another model that predict the residual value using an Ordinary Least Squares model

This way we can quantify the prediction made using renovatable features.

```
renovateable_columns = ['OverallCond','RoofMatl','Exterior1st','Exterior2nd','MasVnrType','ExterQual',
'ExterCond','Heating','HeatingQC','CentralAir','Electrical','KitchenQual','Functional',
'Fireplaces','GarageQual','GarageCond', 'SalePrice']
```

Residual Distribution



How much of the variance in price remaining is explained by these features

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.402e+04	2.54e+04	-0.551	0.582	-6.4e+04	3.59e+04
OverallCond	5213.5012	659.300	7.908	0.000	3919.551	6507.451
Fireplaces	1005.7153	992.127	1.014	0.311	-941.444	2952.875
RoofMatl_Tar&Grv	5.968e+04	2.85e+04	2.096	0.036	3793.806	1.16e+05
RoofMatl_WdShngl	-1e+04	1.76e+04	-0.569	0.569	-4.45e+04	2.45e+04
Exterior1st_BrkComm	1.229e+04	2.49e+04	0.493	0.622	-3.67e+04	6.12e+04
Exterior1st_BrkFace	1.317e+04	1.2e+04	1.100	0.272	-1.03e+04	3.67e+04
Exterior1st_CemntBd	-7026.4335	2.08e+04	-0.338	0.736	-4.79e+04	3.38e+04
Exterior1st_HdBoard	2186.1567	1.2e+04	0.183	0.855	-2.13e+04	2.57e+04
Exterior1st_ImStucc	1506.7466	2.18e+04	0.069	0.945	-4.12e+04	4.43e+04
Exterior1st_MetalSd	3583.0856	1.35e+04	0.266	0.790	-2.29e+04	3e+04

The Coefficient illustrates the dollar value change in Residual of one unit change in the variable. For OverallCond: Overall condition rating, every 1 unit increase equals to a $5213.50 increase in sale price

Dep. Variable:	Residual	**R-squared:**	0.224
Model:	OLS	**Adj. R-squared:**	0.168
Method:	Least Squares	**F-statistic:**	4.040
Date:	Thu, 12 Apr 2018	**Prob (F-statistic):**	2.40e-21
Time:	20:44:09	**Log-Likelihood:**	-10679.
No. Observations:	962	**AIC:**	2.149e+04
Df Residuals:	897	**BIC:**	2.181e+04
Df Model:	64		
Covariance Type:	nonrobust		

Using all renovatable features (67 features) ,This model has a R-squared valueof 0.224 , meaning that this model explains 22.4% of the variance in our dependent variable.

THANKS

Any Questions?