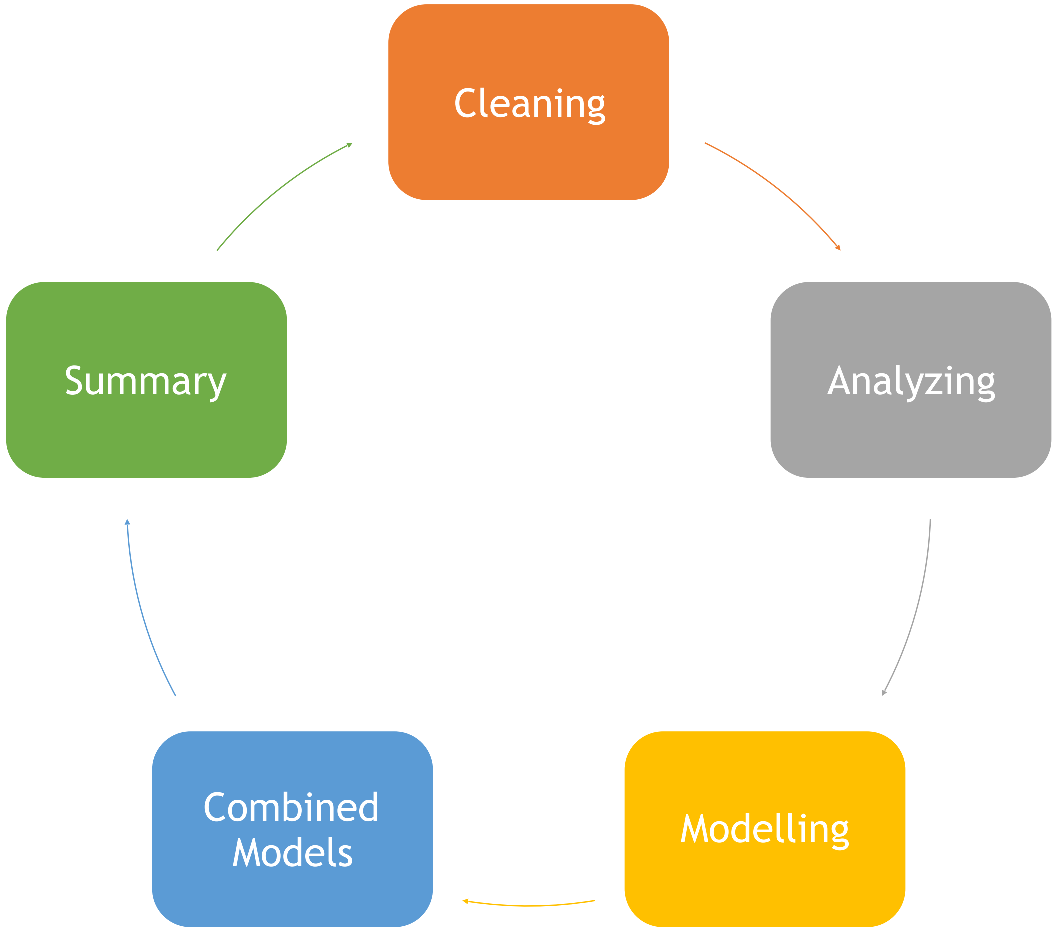
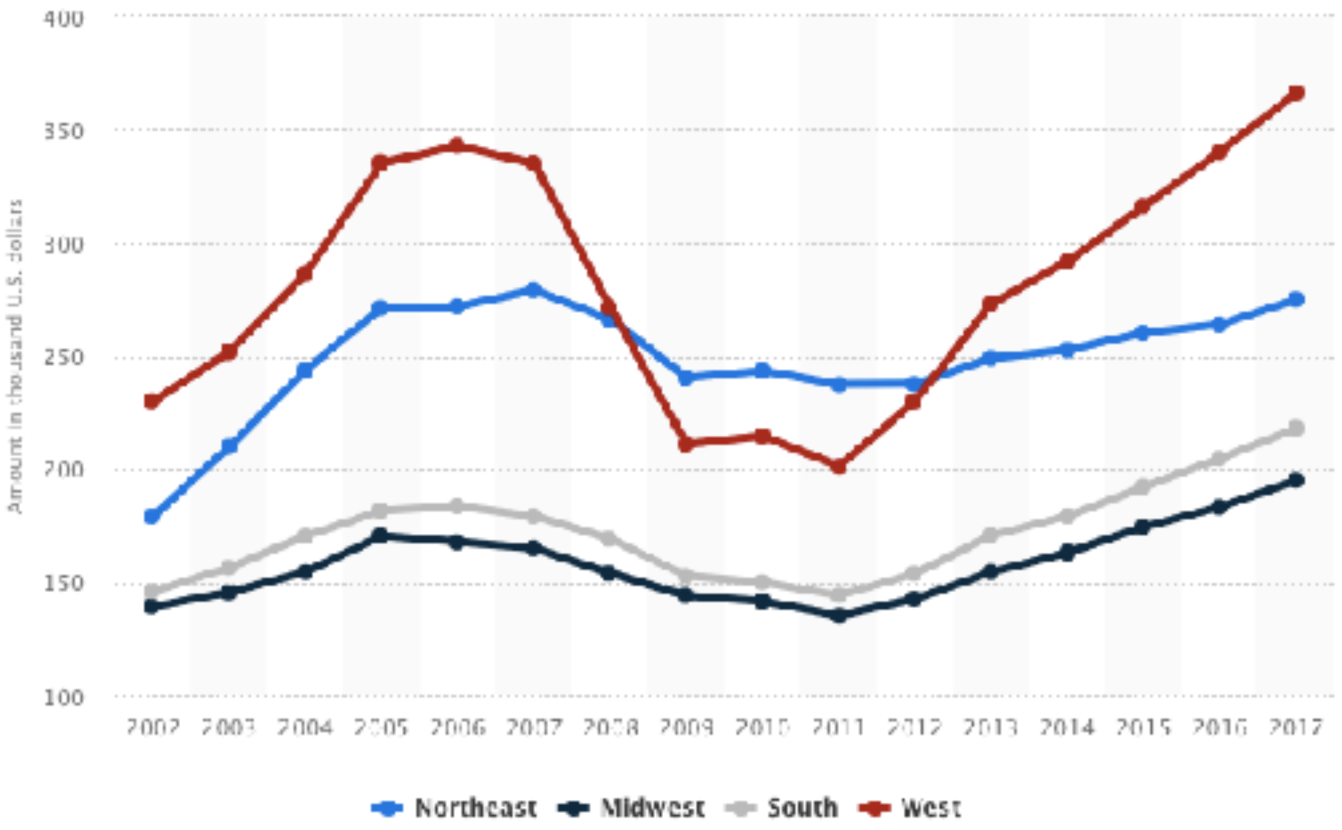

MACHINE
LEARNING

STEPS TO PREDICTIVE MODELLING



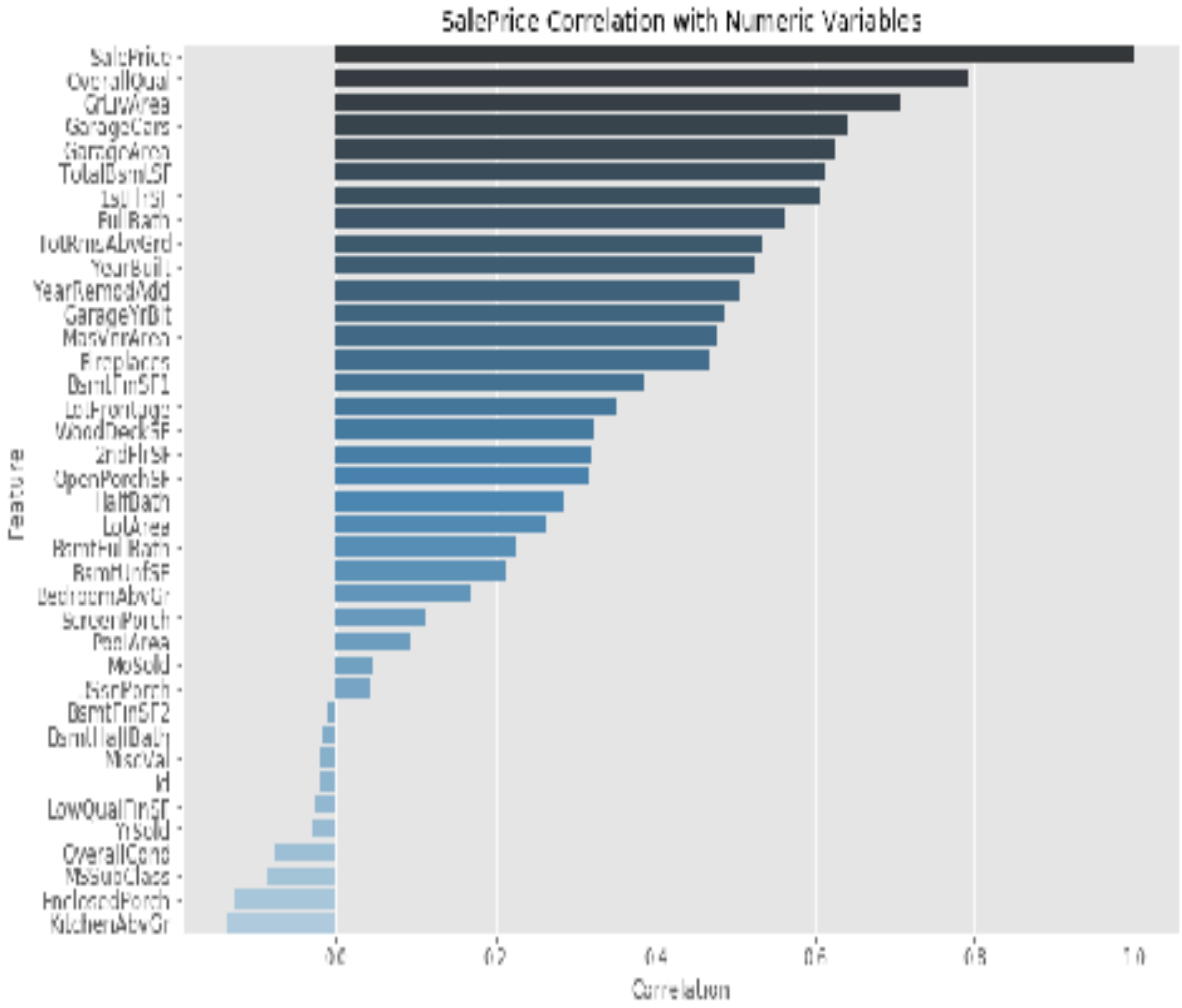
EXPLORATORY DATA ANALYSIS

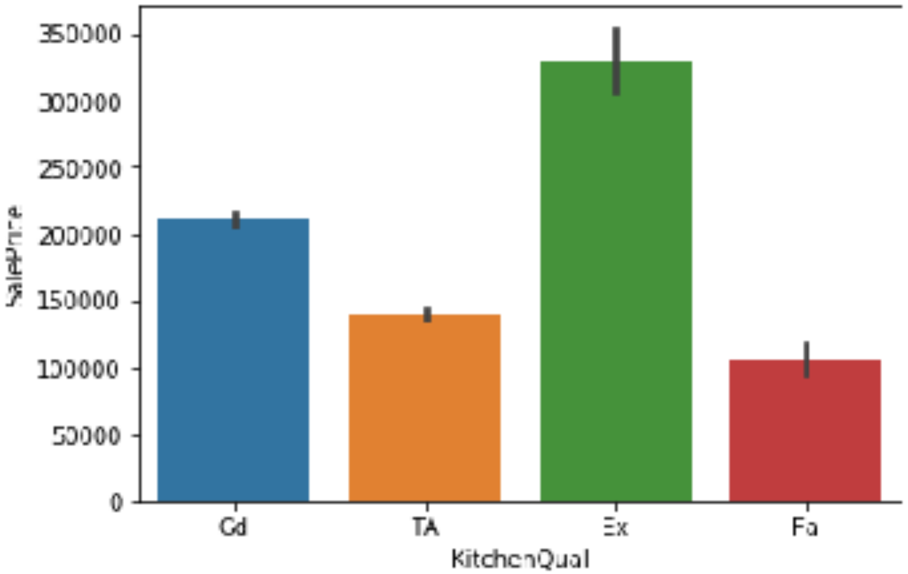
- 2930 observations taken from 2006-2010
- 80 variables related to property sales
- Kaggle Dataset: 37 as numeric & 43 as object type

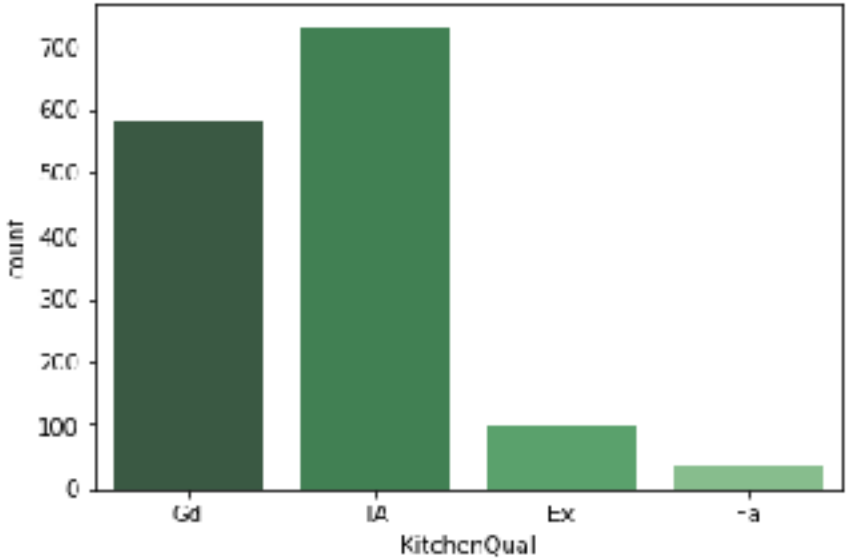


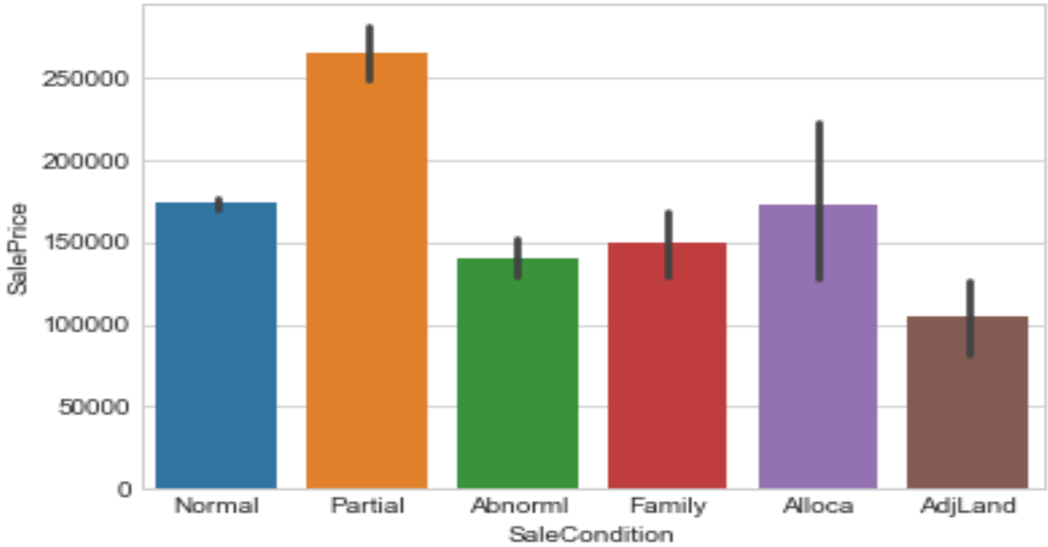
- Ames Housing Data:
  - 20 continuous variables → dimensions (sqft)
  - 14 discrete variables → quantify items occurring in house
  - 23 nominal variables → identify various types of conditions
  - 23 ordinal variable → rate various items in property

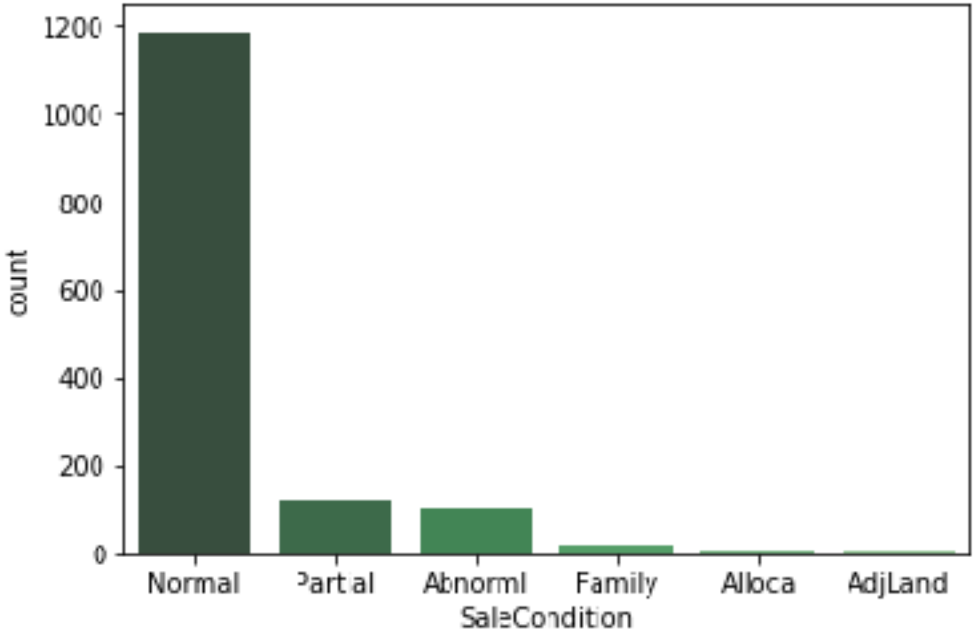
index	count	mean	std	min	25%	50%	75%	max
Id	1460.0	730.500000	421.610009	1.0	365.75	730.5	1095.25	1460.0
MSSubClass	1460.0	56.897260	42.300571	20.0	20.00	50.0	70.00	190.0
LotFrontage	1201.0	70.049958	24.284752	21.0	59.00	69.0	80.00	313.0
LotArea	1460.0	10516.828082	9981.264932	1300.0	7553.50	9478.5	11601.50	215245.0
OverallQual	1460.0	6.099315	1.382997	1.0	5.00	6.0		
OverallCond	1460.0	5.575342	1.112799	1.0	5.00	5.0		
YearBuilt	1460.0	1971.267808	30.202904	1872.0	1954.00	1973.0		
YearRemodAdd	1460.0	1984.865753	20.645407	1950.0	1967.00	1994.0		
MasVnrArea	1452.0	103.685262	181.066207	0.0	0.00	0.0		
BsmtFinSF1	1460.0	443.639726	456.098091	0.0	0.00	383.5		
BsmtFinSF2	1460.0	46.549315	161.319273	0.0	0.00	0.0		
BsmtUnfSF	1460.0	567.240411	441.866955	0.0	223.00	477.5		
TotalBsmtSF	1460.0	1057.429452	438.705324	0.0	795.75	991.5		
1stFlrSF	1460.0	1162.626712	386.587738	334.0	882.00	1087.0		
2ndFlrSF	1460.0	346.992466	436.528436	0.0	0.00	0.0		
LowQualFinSF	1460.0	5.844521	48.623081	0.0	0.00	0.0		
GrLivArea	1460.0	1515.463699	525.480383	334.0	1129.50	1464.0		




SalePrice
KitchenQual
Gd
TA
Ex
Fa
0
50000
100000
150000
200000
250000
300000
350000

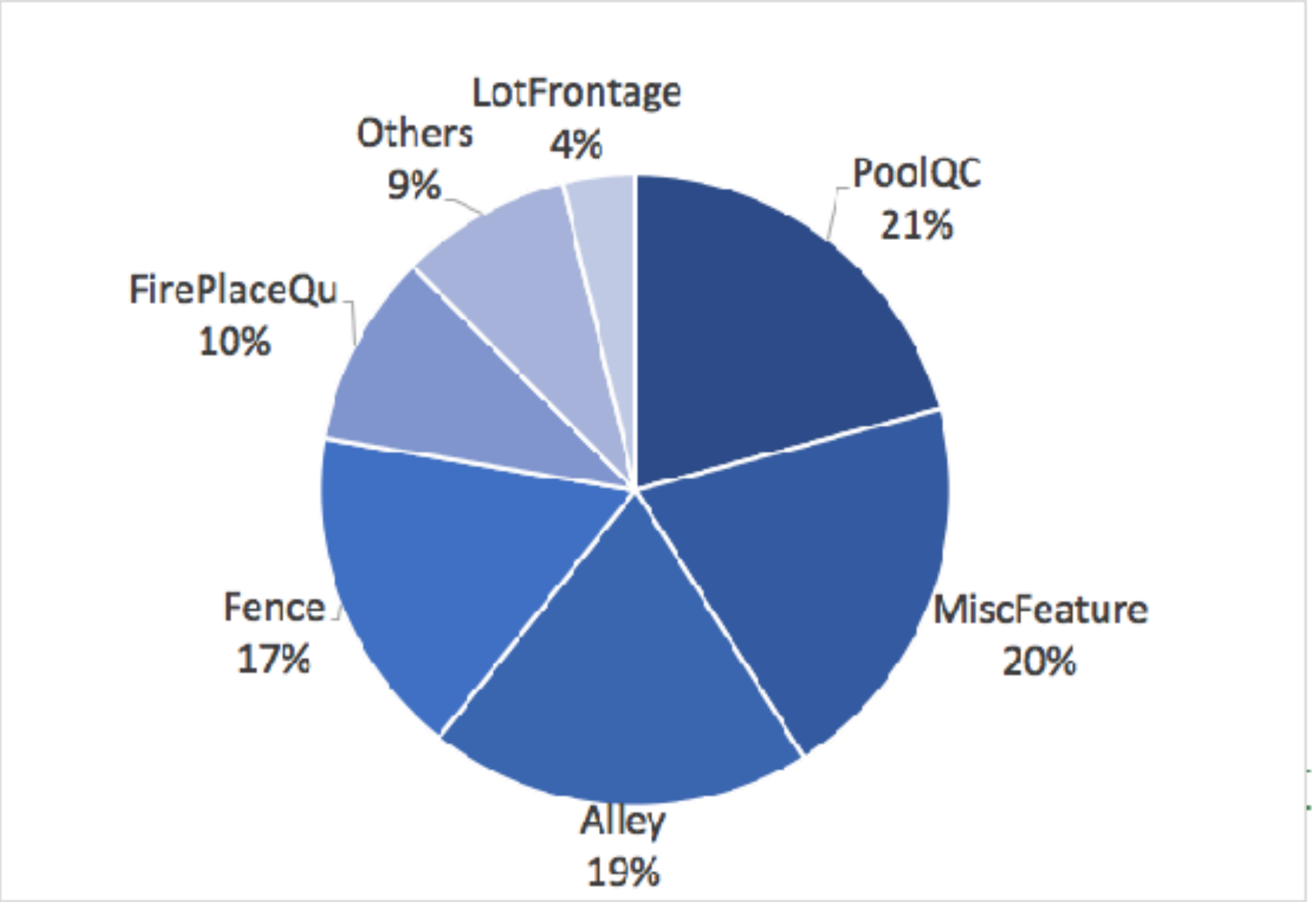

count
KitchenQual
Gd
Ex
0
100
300
400
600


SalePrice
SaleCondition
Normal
Partial
Abnorml
Family
Alloca
AdjLand
0
50000
100000
150000
200000
250000

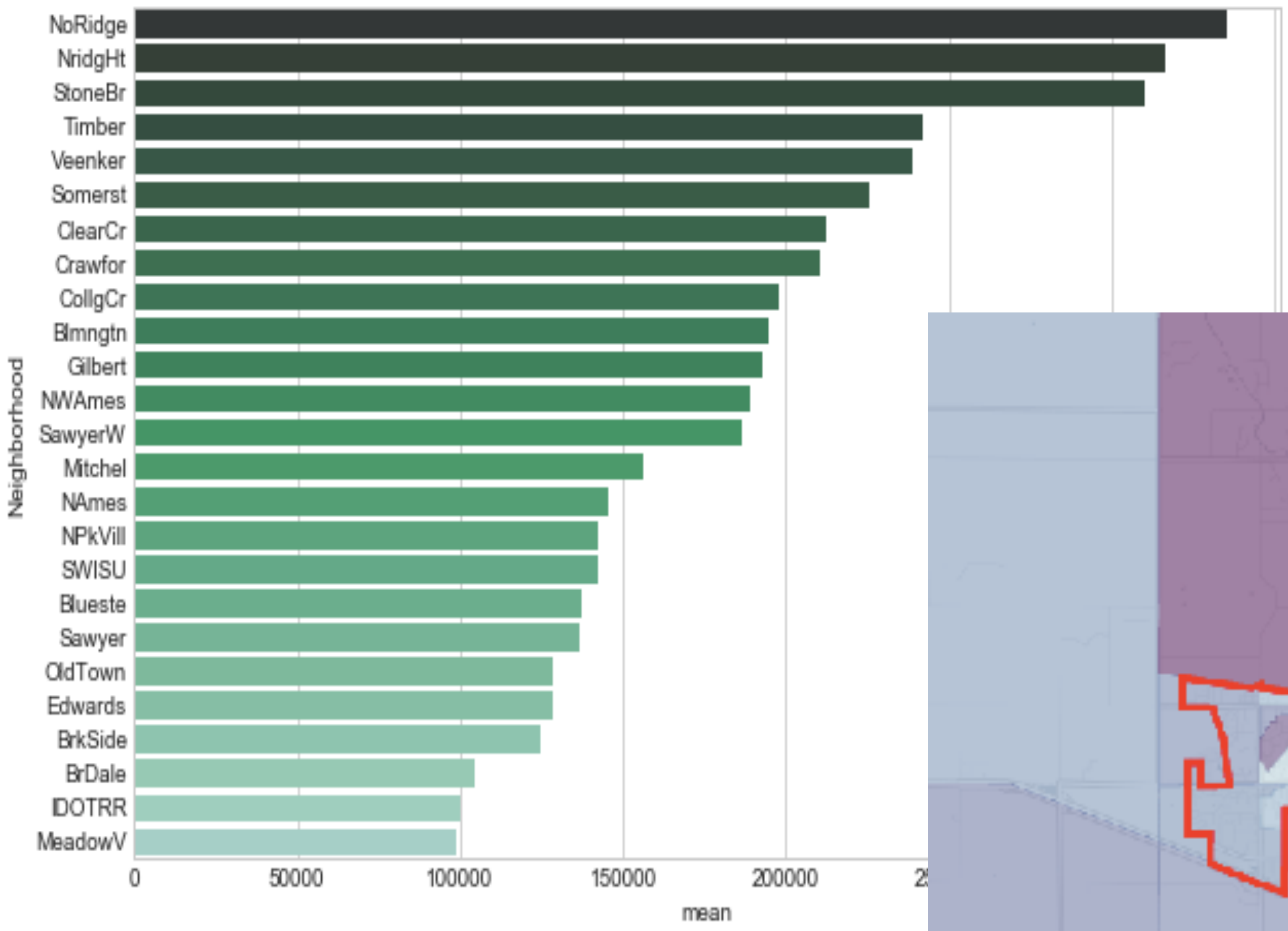

count
SaleCondition
Normal
Abnorml
Family
Alloca
AdjLand
0
200
400
600
800
1000
1200

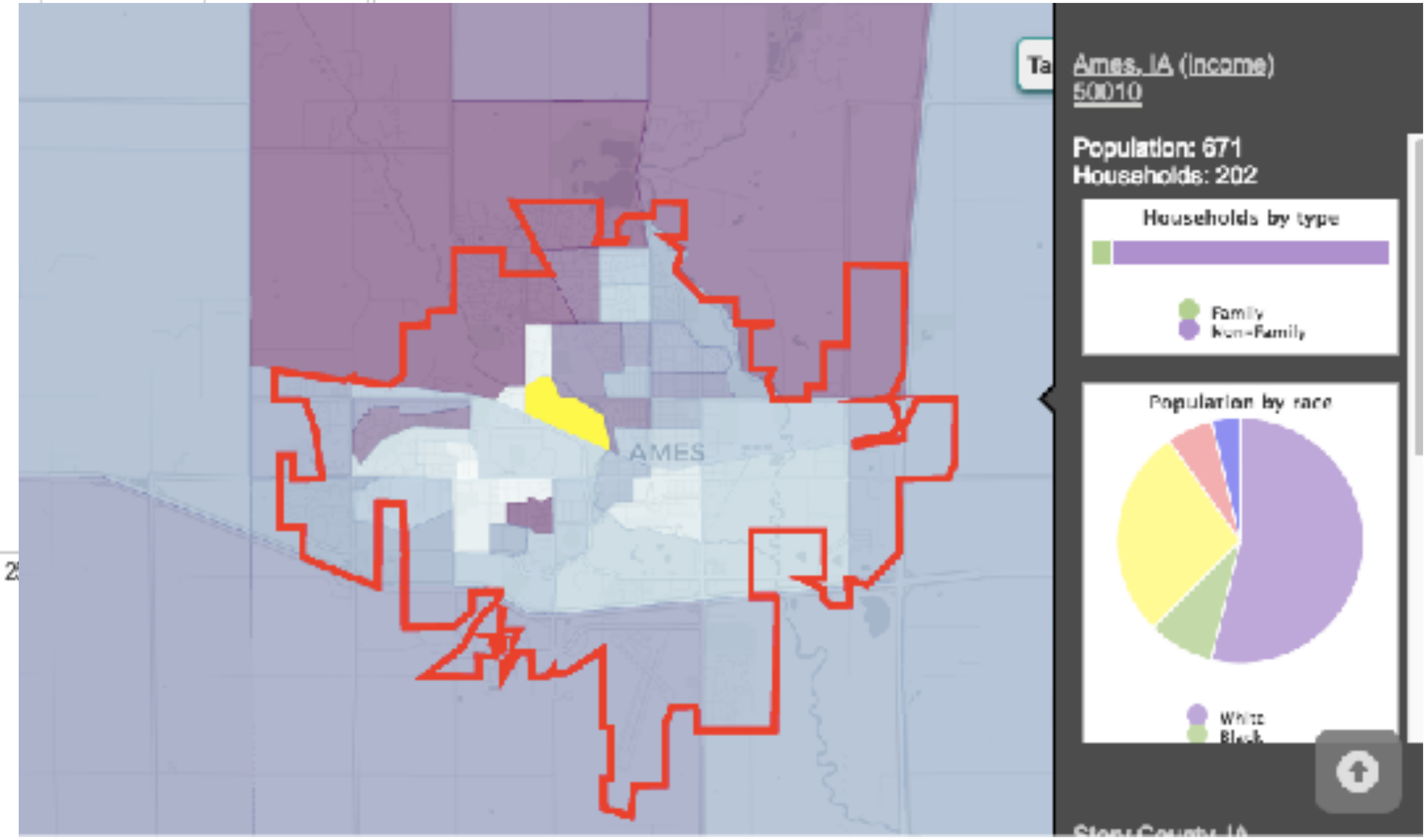
Missingness

Feature	Missingnes
PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FirePlaceQu	690
Others	609
LotFrontage	259
TOTAL	6965



Feature Engineering



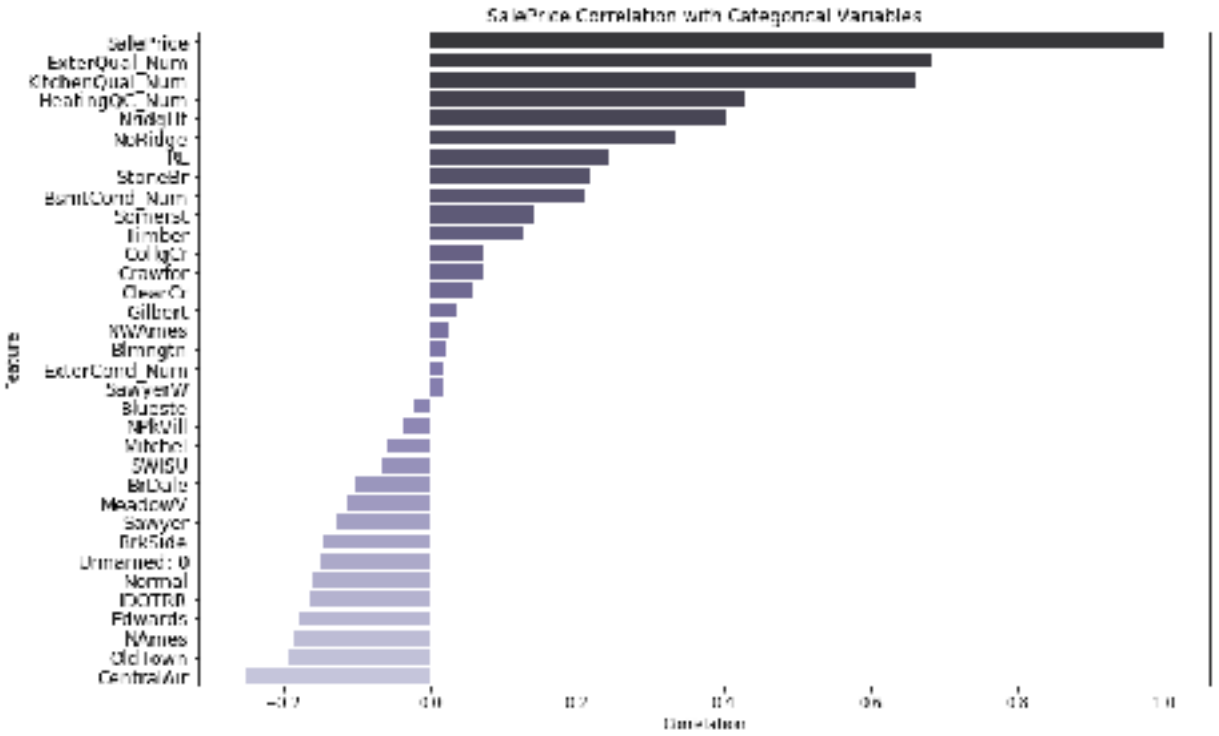


	FEATURE	VALUES	TRANSFORMATION
Ordinal	ExternalQual	Ex Gd TA Fa Po	5, 4, 3, 2, 1
	External Cond	Ex Gd TA Fa Po	5, 4, 3, 2, 1
	HeatingQC	Ex Gd TA Fa Po	5, 4, 3, 2, 1
	BsmtCond	Ex Gd TA Fa Po	5, 4, 3, 2, 1
	KitchenQual	Ex Gd TA Fa Po	5, 4, 3, 2, 1
Binary	SaleCondition	Normal/Abnormal/AdjLand/Alloca/Family/Partial	Dummified Normal(1) / Others(0)
	MSZoning	A/C/FV/I/RH/RL/RP/RM	Dummified RL(1) / Others(0)
	Central Air	No / Yes	Dummified Y(1) / N(0)
	Neighborhood	25 Categorical	Dummified 24 Variables

- Total of 9 variables were transformed
- Ordinal Variables were later on combined
- Neighborhood ”dummified” substantially increased number of features

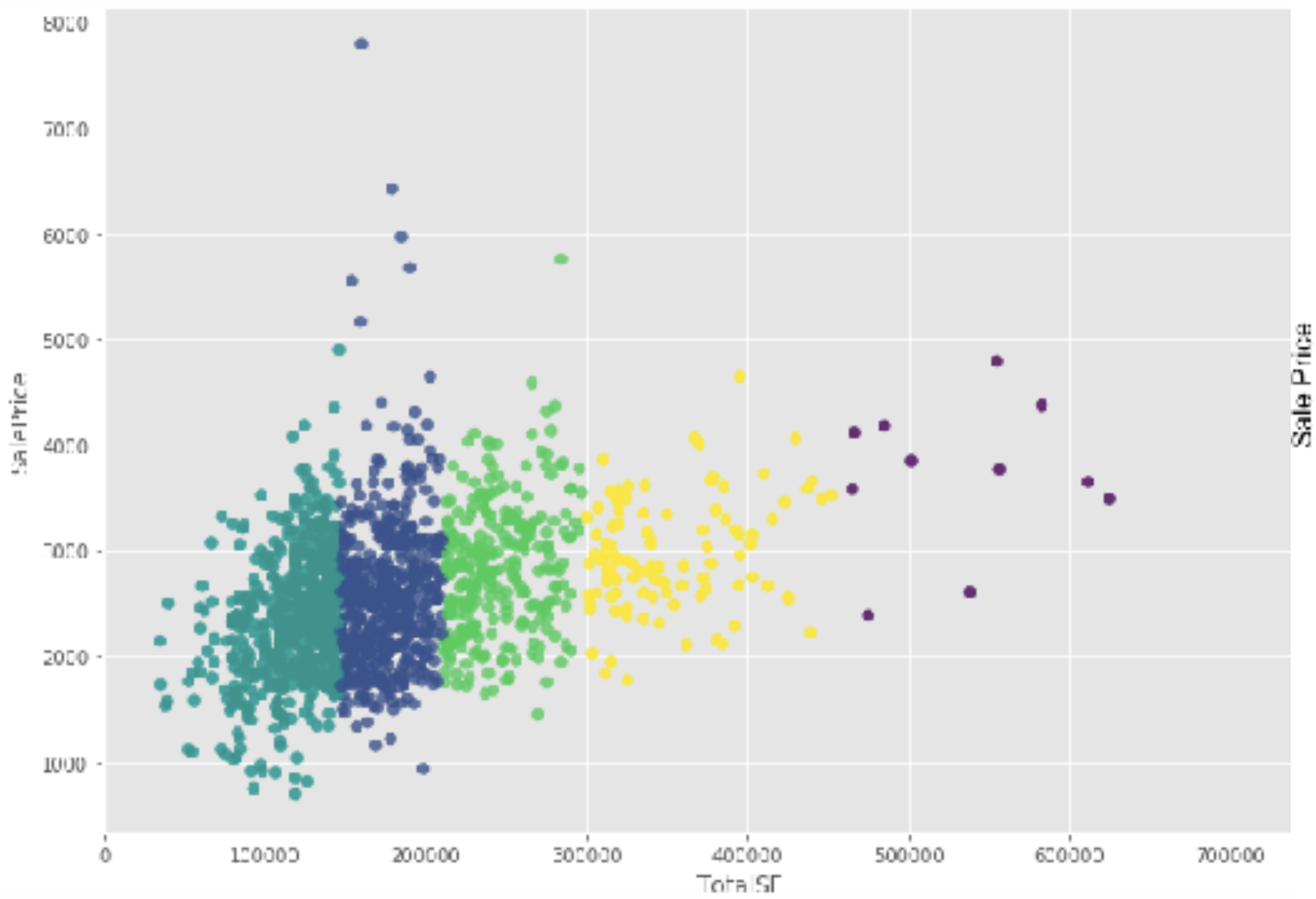
Random Forest and Correlations

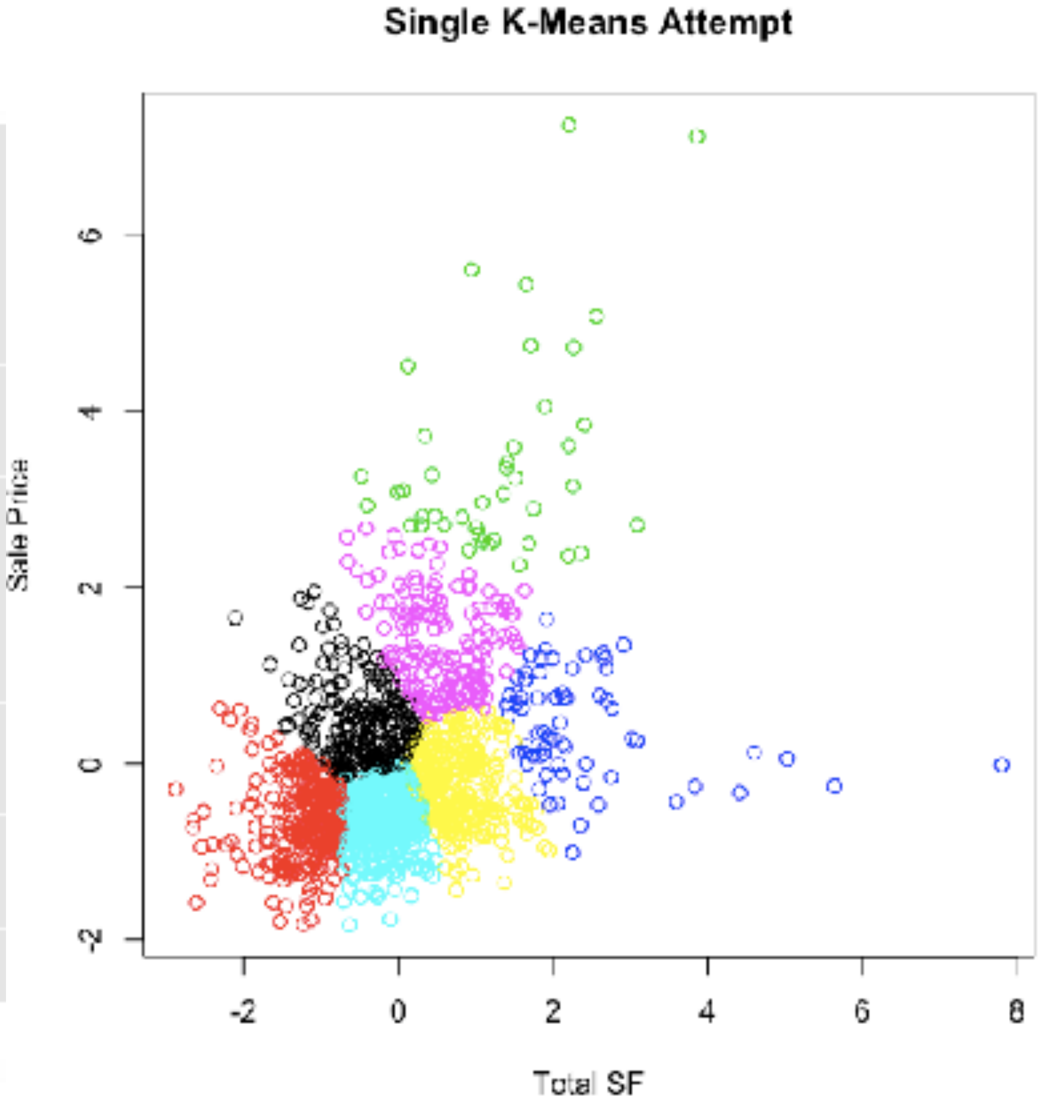
	FEATURE	ExtraTreesClassifier		RandomForestClassifier	
Ordinal	ExternalQual	0.0998	0.65	0.0619	0.54
	External Cond	0.1187		0.0900	
	HeatingQC	0.1628		0.1665	
	BsmtCond	0.1162		0.1022	
	KitchenQual	0.1490		0.1148	
Binary	SaleCondition	0.0908	0.15	0.0886	0.17
	MSZoning	0.0348		0.0408	
	Central Air	0.0275		0.0359	
	Neighborhood (25)	0.2004		0.2993	



Total SF Cluster Analysis

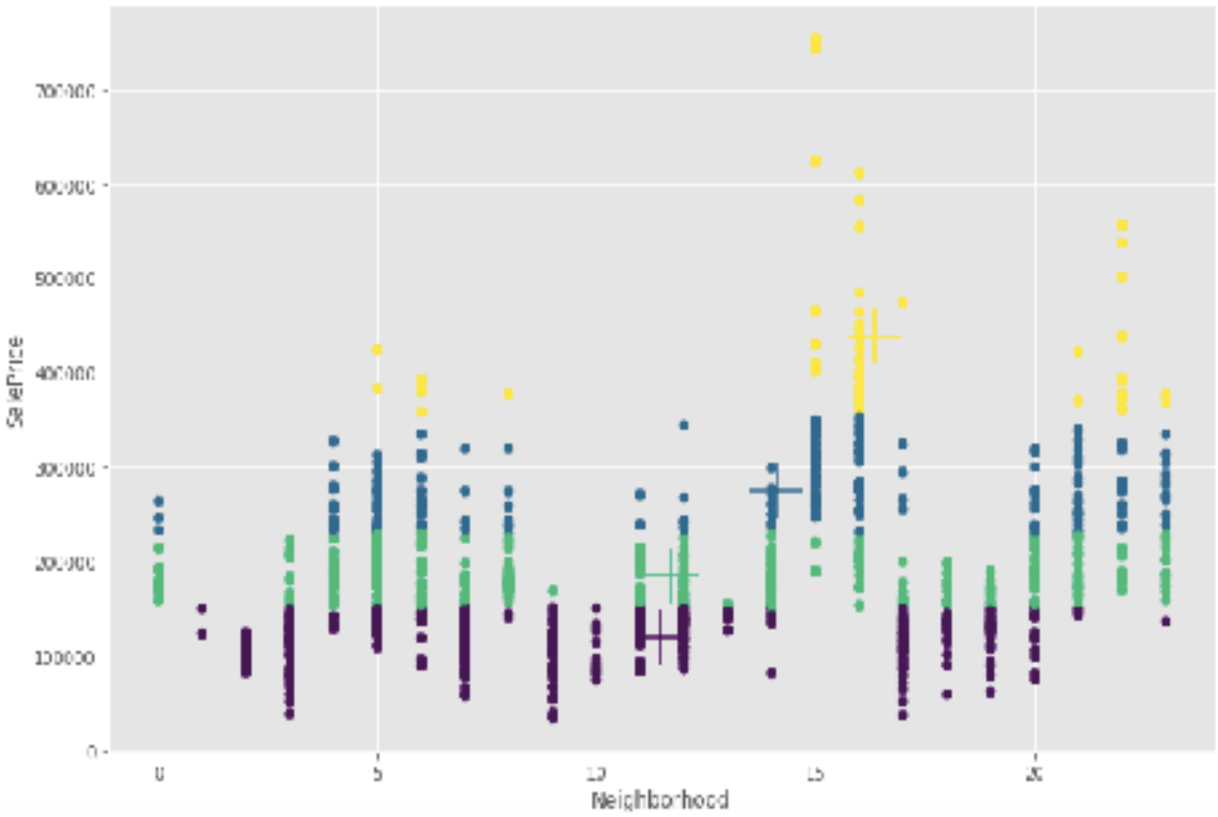
K-means Cluster



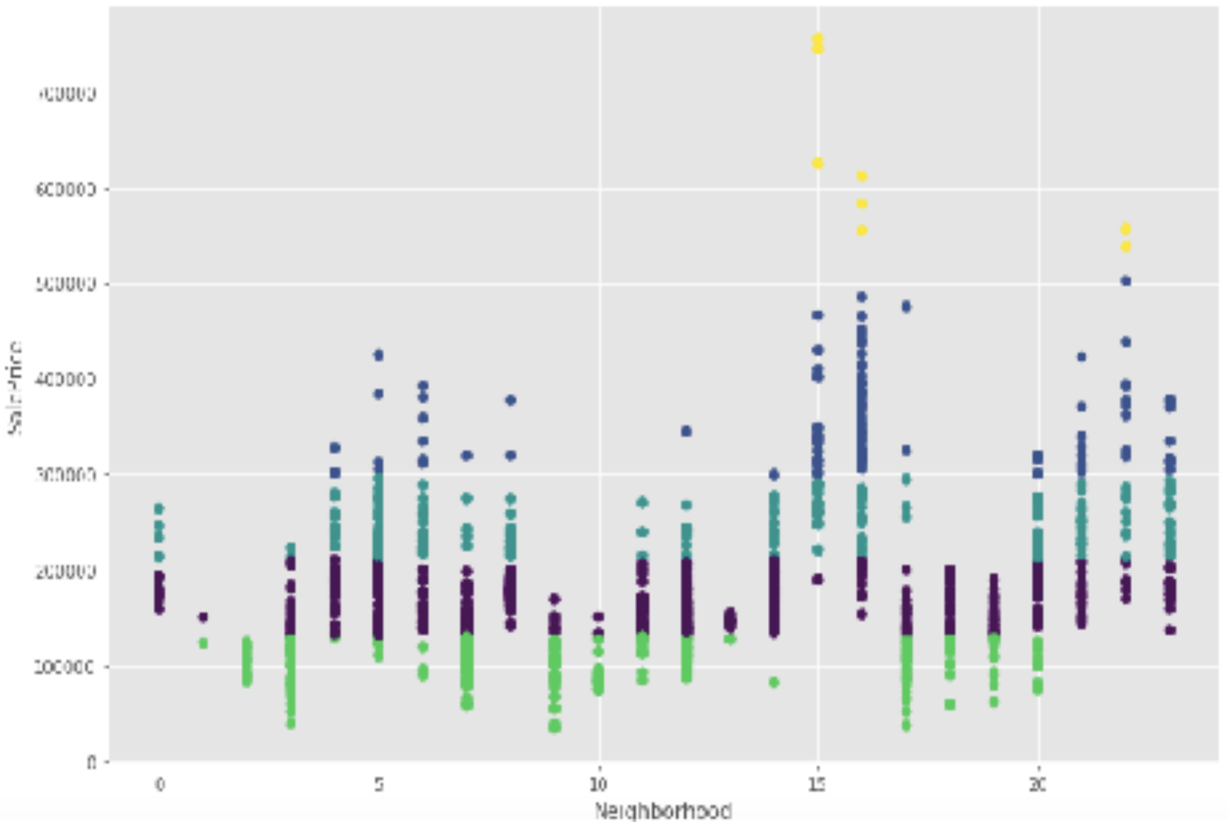


Neighborhoods Cluster Analysis

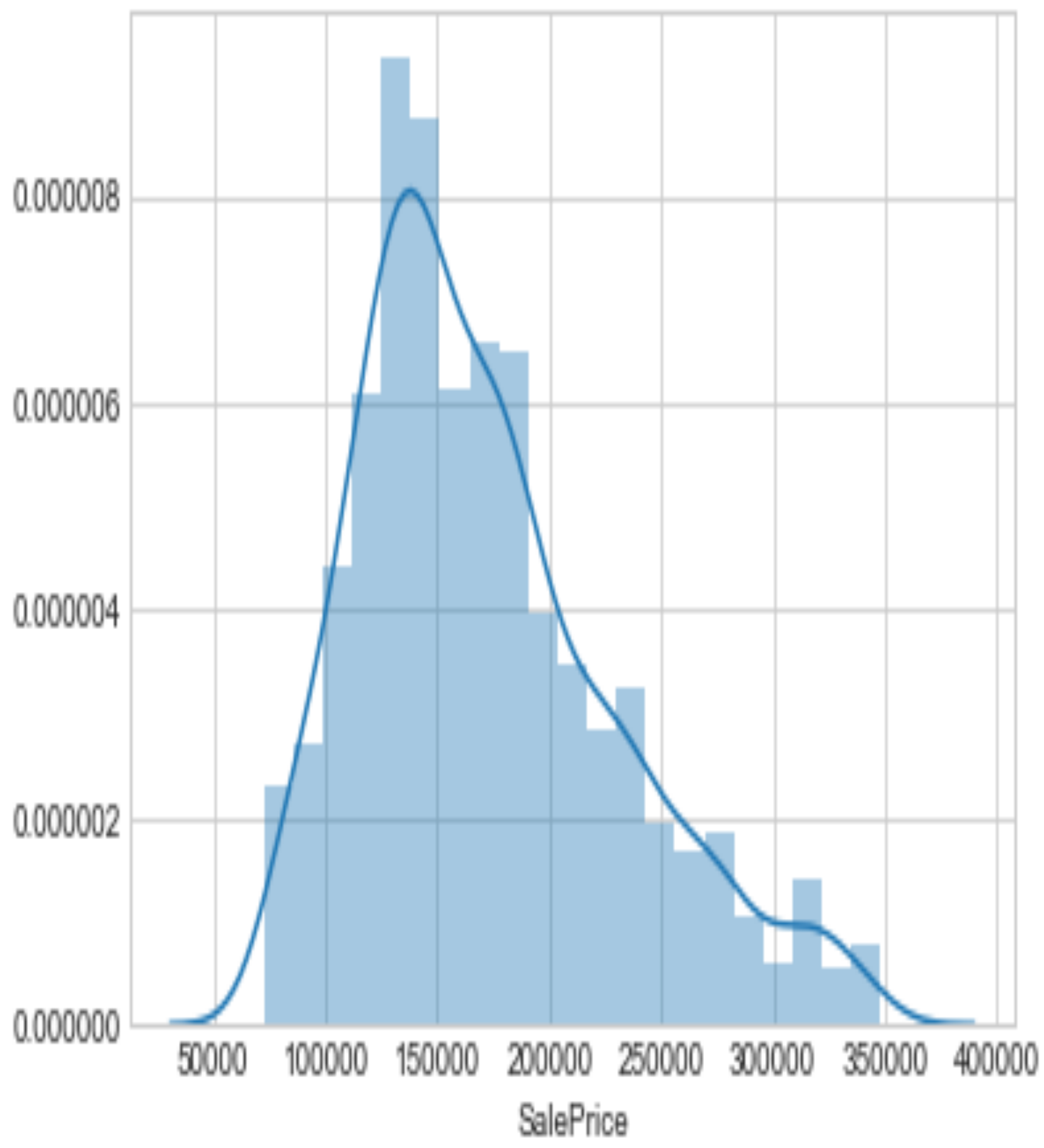
K-means Cluster

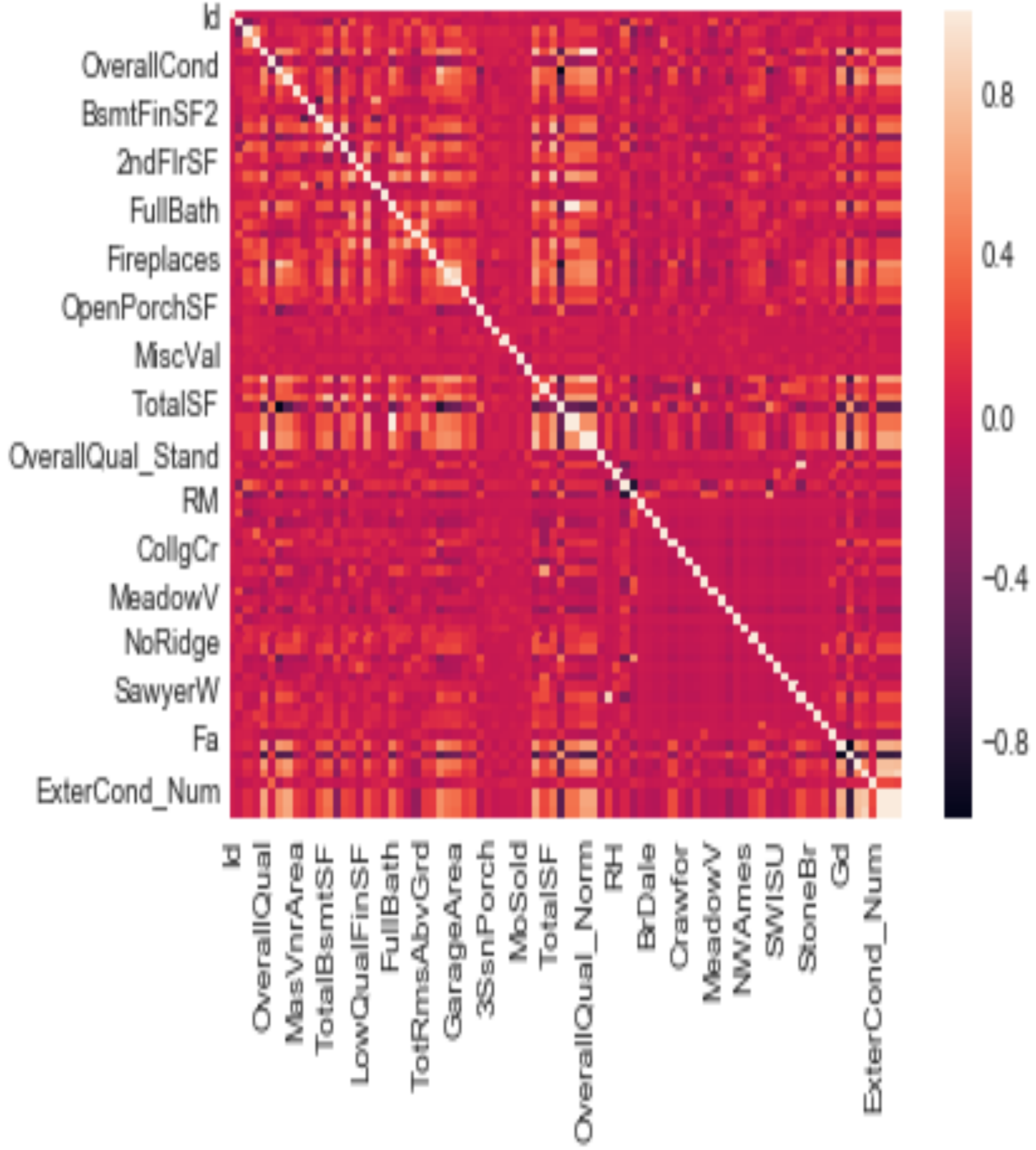


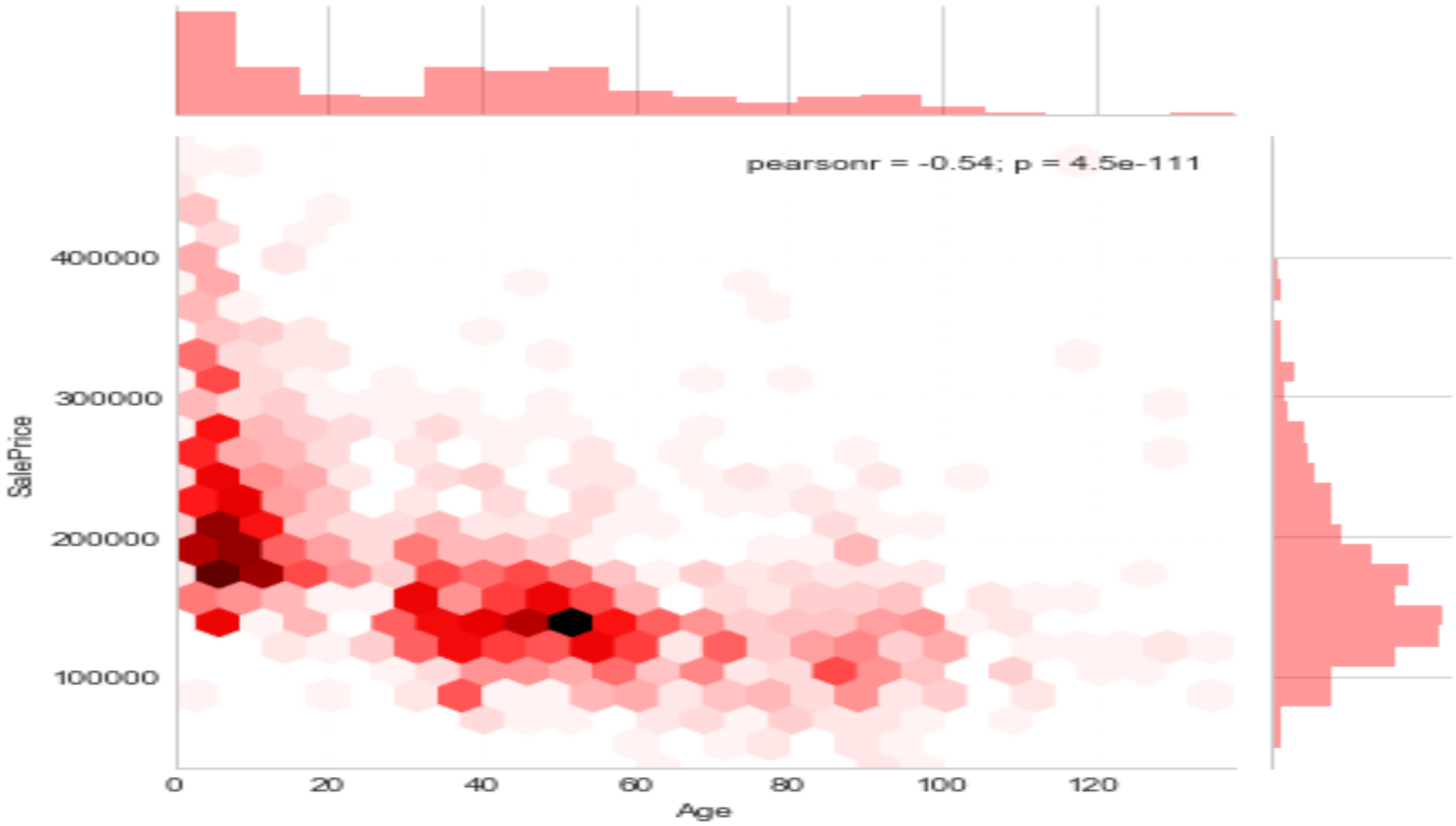
Hierarchical Cluster

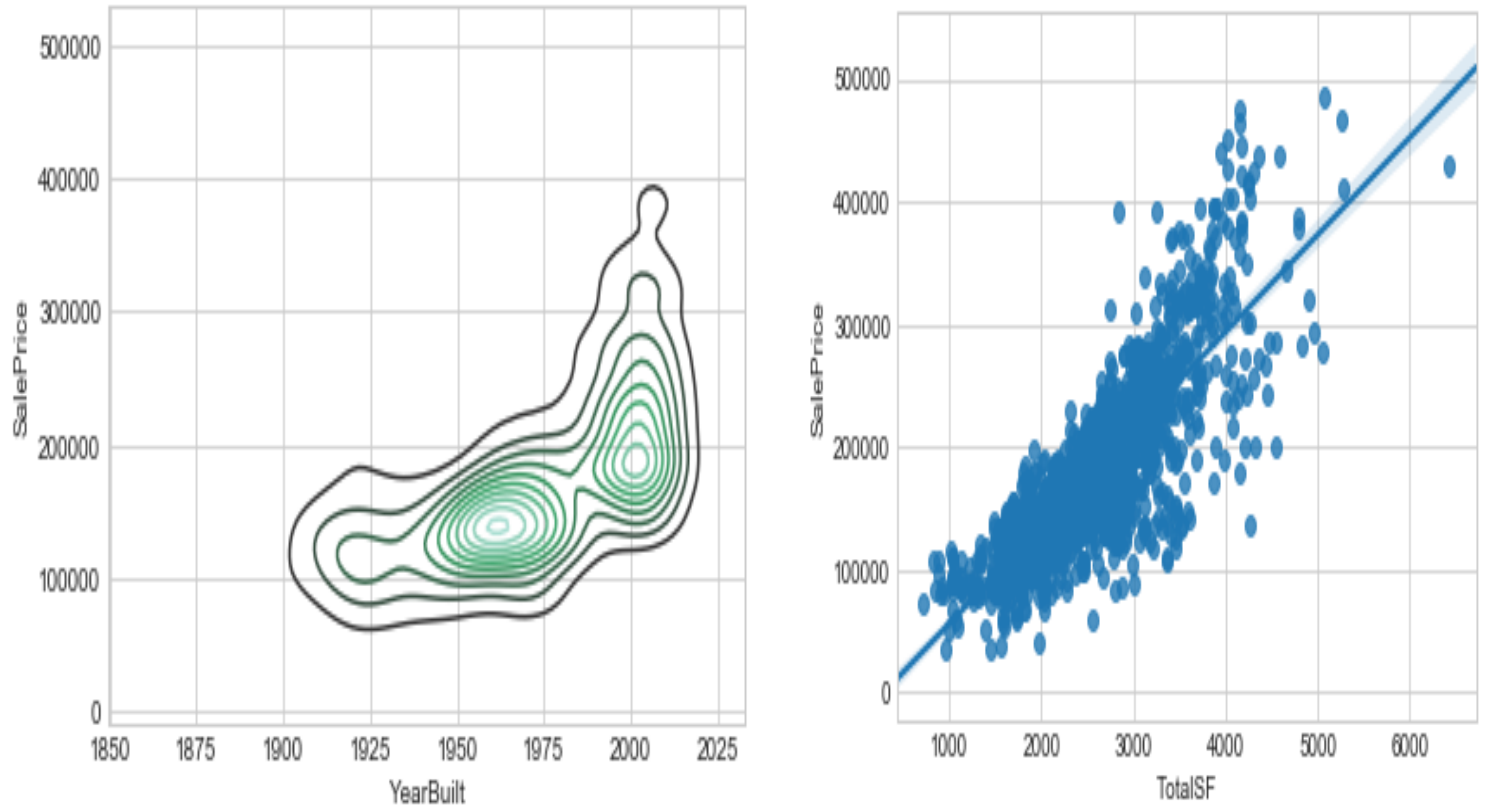


MODELLING


0.000008
0.000006
0.000004
0.000002
0.000000
50000
100000
150000
200000
250000
300000
350000
400000
SalePrice


Id
OverallCond
BsmtFinSF2
2ndFlrSF
FullBath
Fireplaces
OpenPorchSF
MiscVal
TotalSF
OverallQual_Stand
RM
CollgCr
MeadowV
NoRidge
SawyerW
Fa
ExterCond_Num
Id
OverallQual
MasVnrArea
TotalBsmtSF
LowQualFinSF
FullBath
TotRmsAbvGrd
GarageArea
3SsnPorch
MoSold
TotalSF
OverallQual_Norm
RH
BrDale
Crawfor
MeadowV
NWAmes
SWISU
StoneBr
Gd
ExterCond_Num
0.8
0.4
0.0
−0.4
−0.8

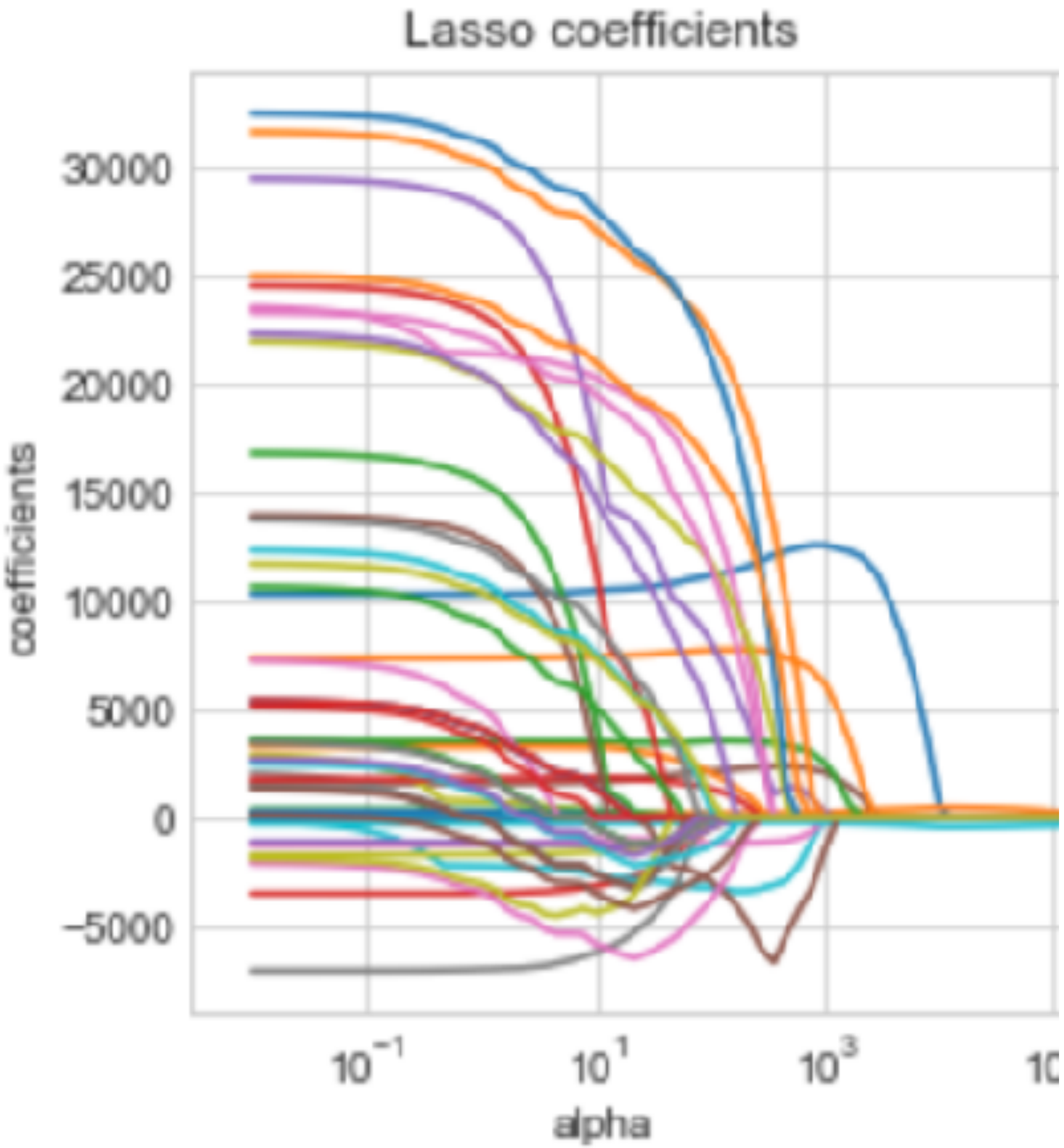

pearsonr = -0.54; p = 4.5e-111
400000
300000
200000
100000
SalePrice
0
20
40
60
80
100
120
Age

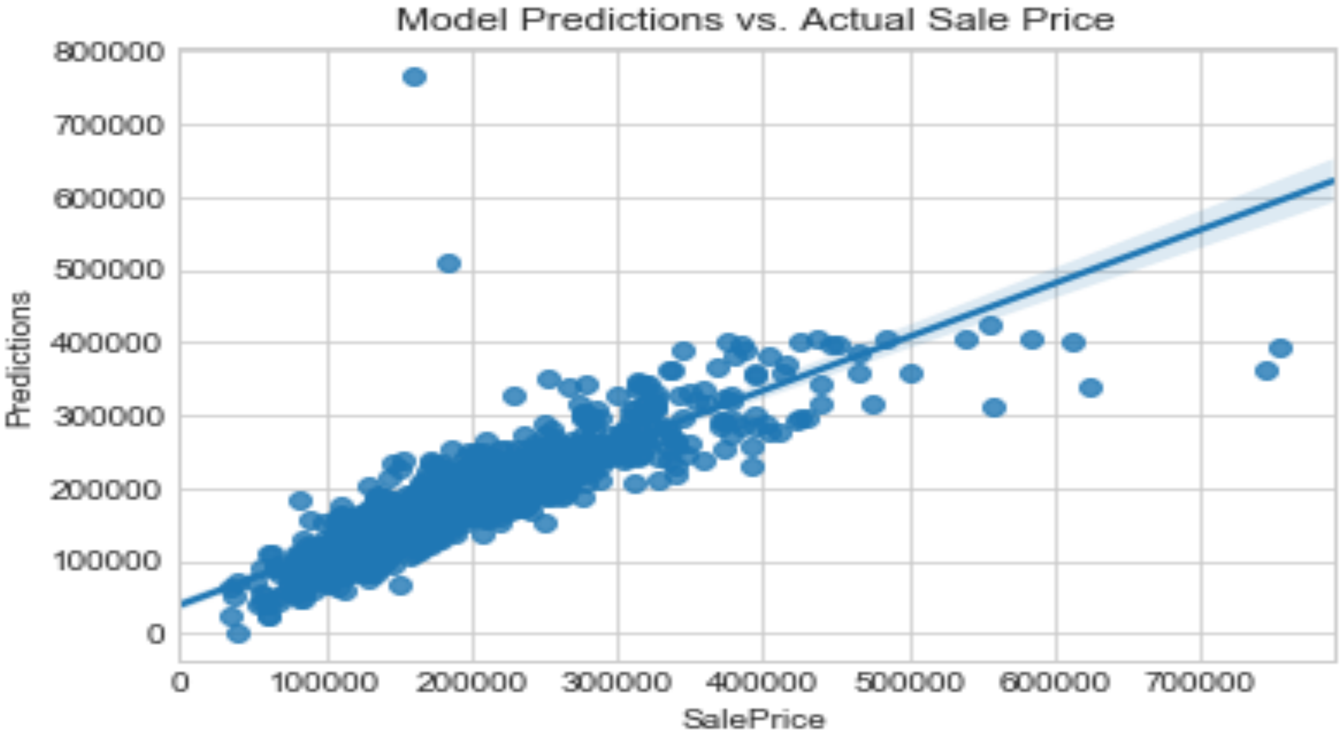


SalePrice
YearBuilt
SalePrice
TotalSF

MODEL TUNING

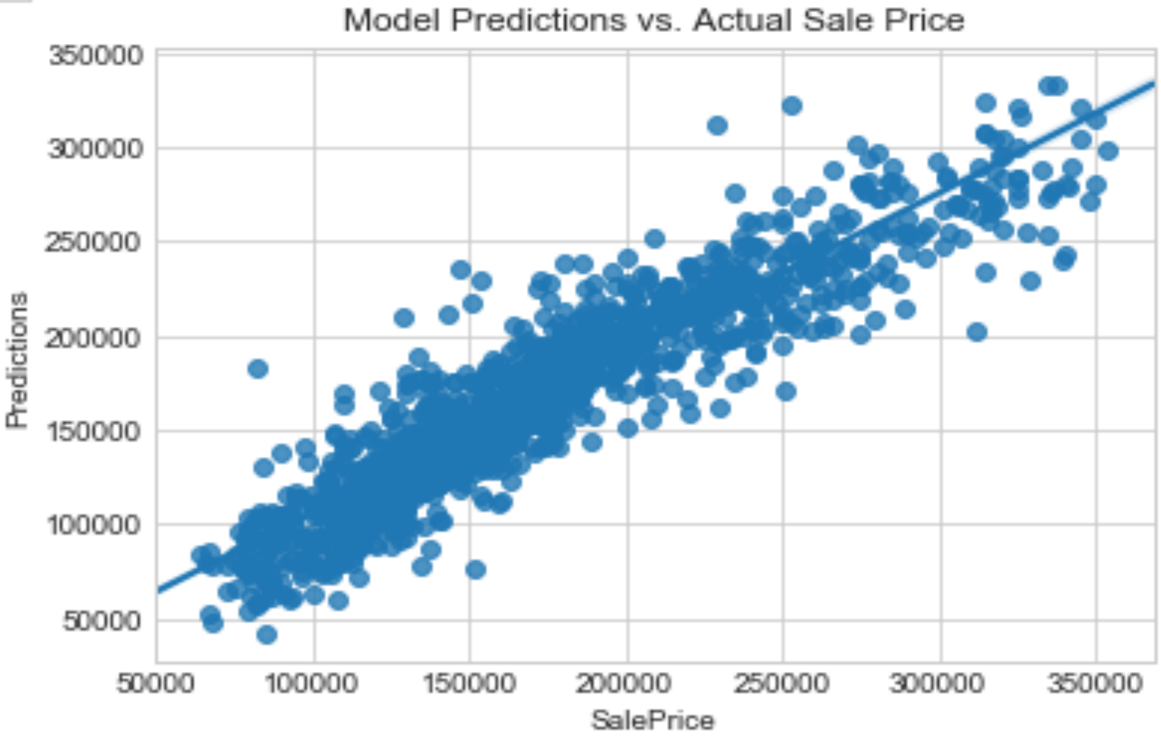
- Alpha was chosen at 10 after testing 10000 times for the best alpha.
- After 10000 tests, Lasso Regression performed better than Ridge Regression.
- For Cross-Validation we did a train-test split at 80/20
- We found cv = 10 to give the best accuracy score.





- Outliers were removed at extreme values.
- Sale Prices over $355,000 and under $63,000 were removed.

- -After removing outliers, the Lasso Score improved by 9.0%
- -The final R score improved by 10.5% after Cross-Validation.



Model Results

- Kaggle score: 0.15153

- 30% increase in error when simplifying to 6 variables:
  - OverallQ*ual*, *TotalSF*, *Condion_W_Avg*, *FullBath_Norm*, *Age*, *GarageArea*

- We use 59 variables for the Lasso final model (large number in part because of dummy variables)

```
(array([ 1.03217702e+04,  1.86218084e+02,  4.94956134e-01, -3.49969536e+03,
         1.59775171e+01,  1.60258653e+03,  2.14151325e+04, -0.00000000e+00,
         7.49617011e+02, -2.23992917e+03,  9.09276725e+01,  7.35943113e+03,
         3.35582471e+02,  4.22482620e+01,  4.54162075e+01,  1.43223335e+01,
        -1.23523988e+03, -7.02834730e+03, -1.70515523e+03, -2.86030634e+02,
         2.55455711e+02,  2.71554139e+01,  3.53765758e+03,  1.78501636e+03,
        -1.15944929e+03,  2.17796837e+01,  4.25075406e+01,  1.16162643e+01,
         1.83992485e+01,  3.15124380e+01,  1.56840772e-03,  3.27830522e+03,
         1.55241304e+04,  2.32056723e+04,  2.82651660e+04,  1.28775608e+04,
         5.10081811e+03, -0.00000000e+00,  2.04559705e+04,  1.09180459e+04,
         4.11113444e+03,  3.02725022e+04,  2.18904477e+03,  4.14364488e+03,
         2.05179189e+04, -6.25760214e+01, -3.41651463e+03,  2.06527831e+03,
        -3.04998243e+03,  1.14408468e+03,  3.12531344e+04,  2.38325510e+04,
         9.03539083e+03,  3.61459098e+03,  1.31640010e+03, -1.11486578e+03,
         2.21604032e+04,  1.24661430e+04,  1.04218727e+04]),
 -419893.1538635175)
```

CONCLUSIONS

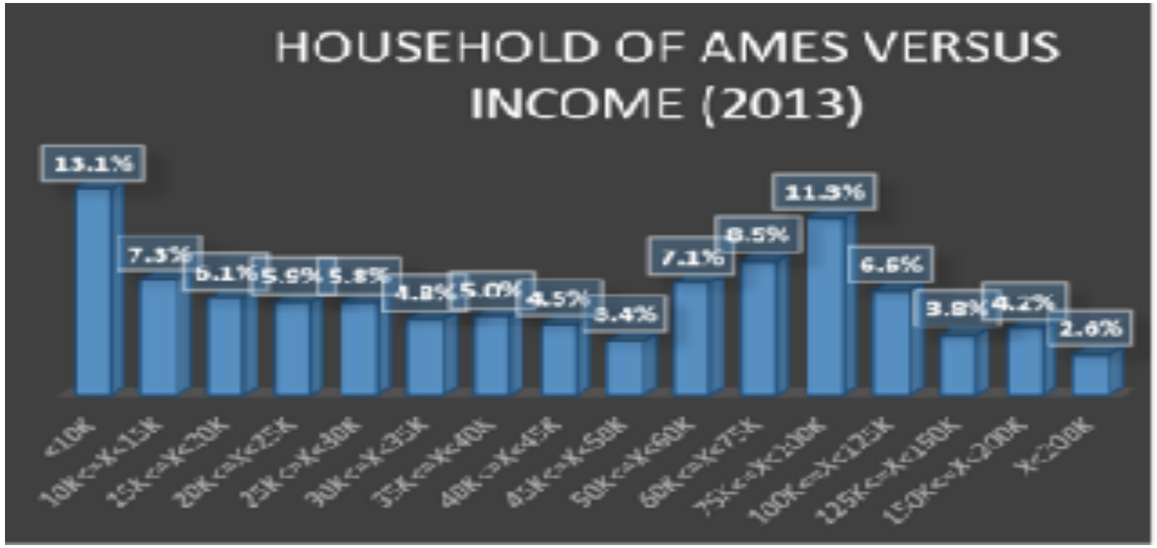
- Complexity of model translates into more accurate predictions; however a more simplistic model could be considered for practical use (trade off in lower accuracy).

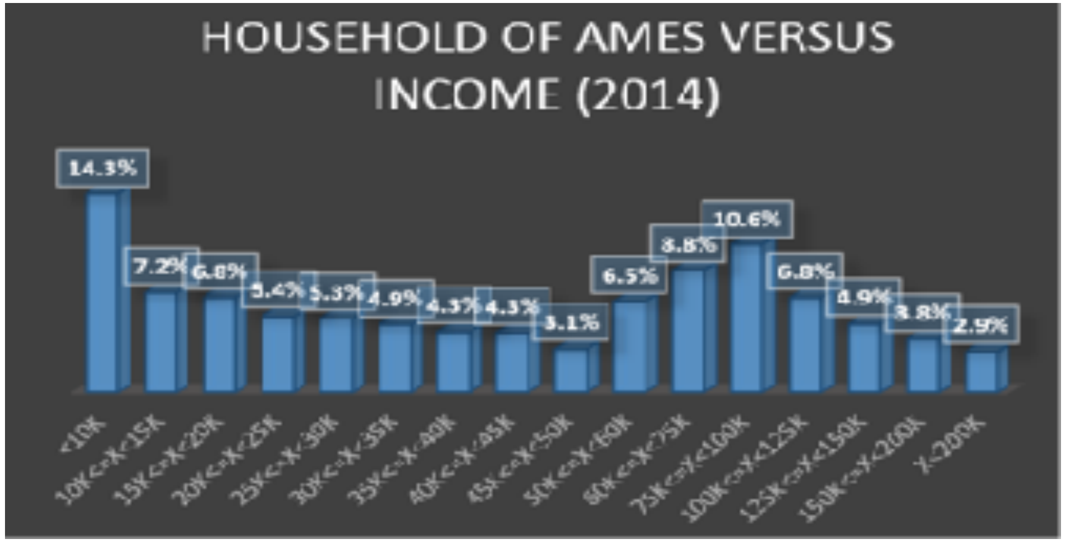
- Model could improve if neighborhoods are standardized to US census track for more demographic information such as household income.

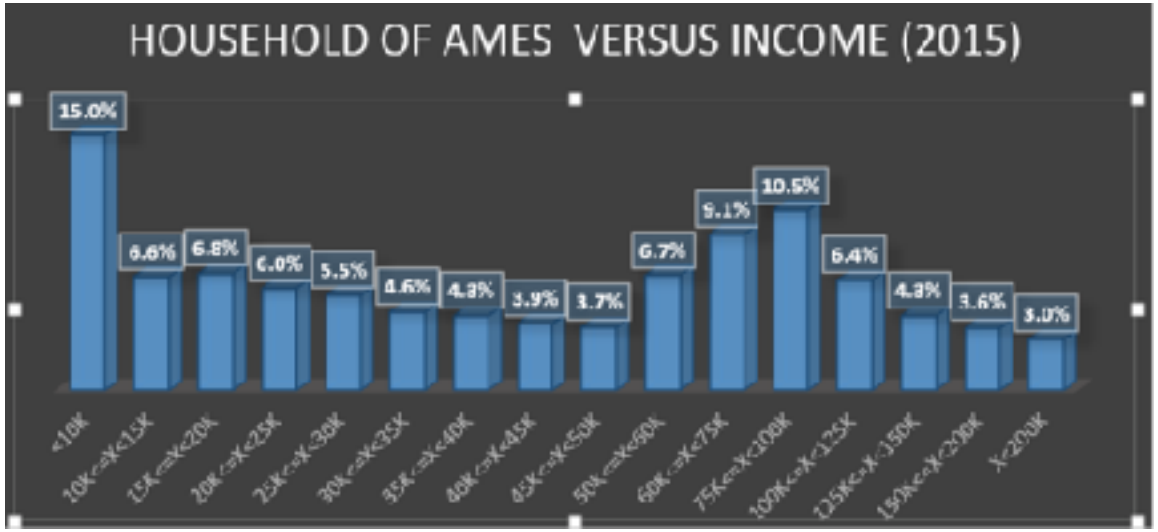
- Dataset could be complemented with other variables such as crime, school, transportation that seem to be important in house hunting/buying.

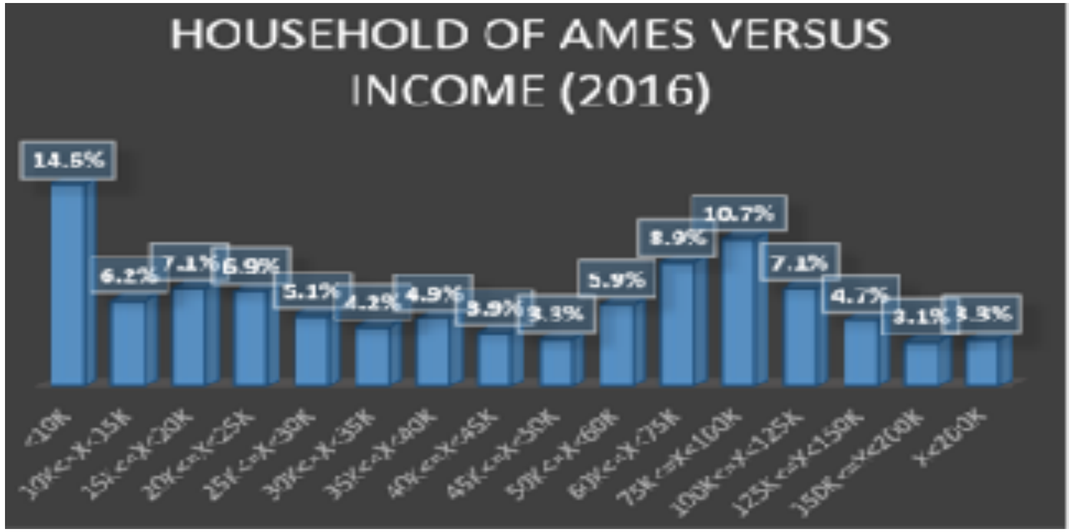
- If would be important to validate model with more recent data (years) as real estate seem cyclical with ups/downs

Household of Ames Versus Income Comparison

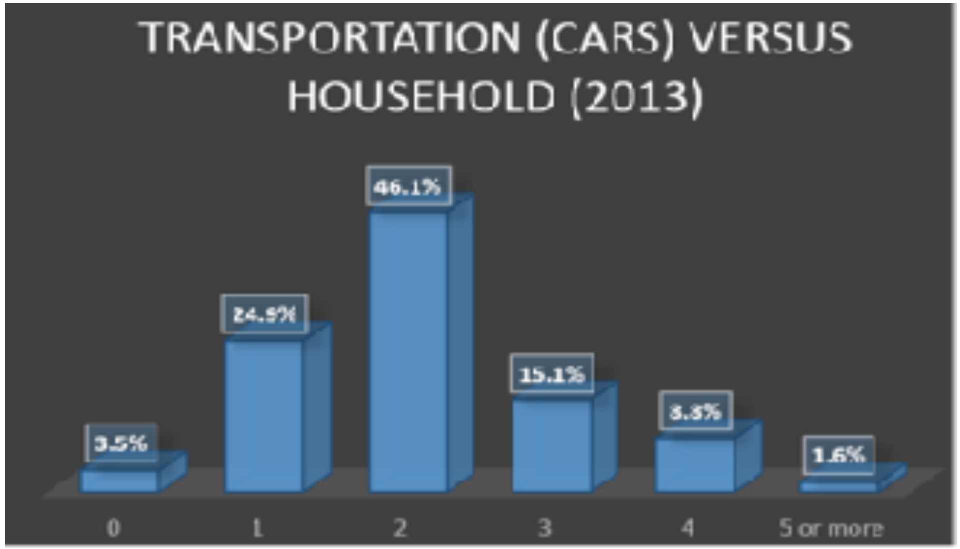


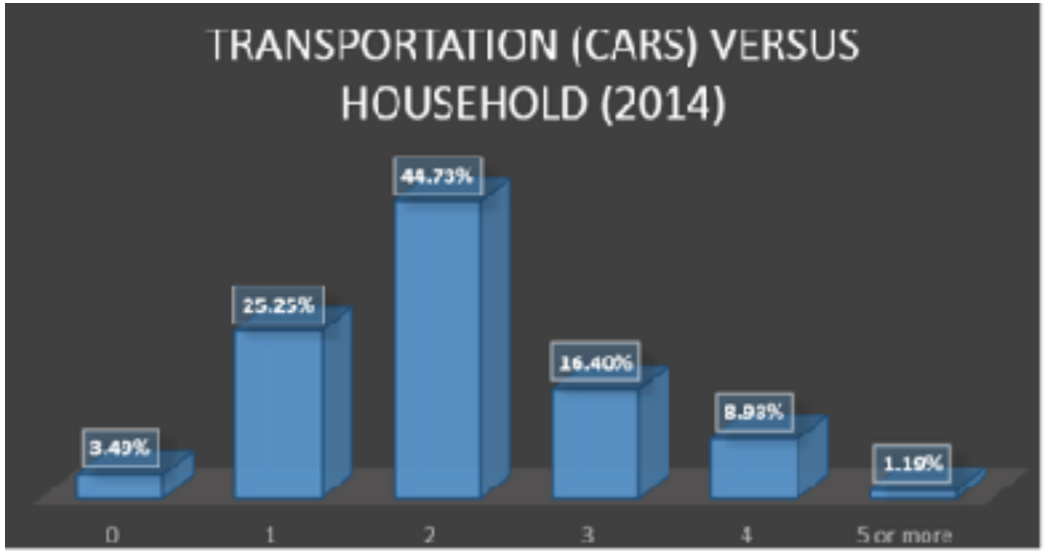


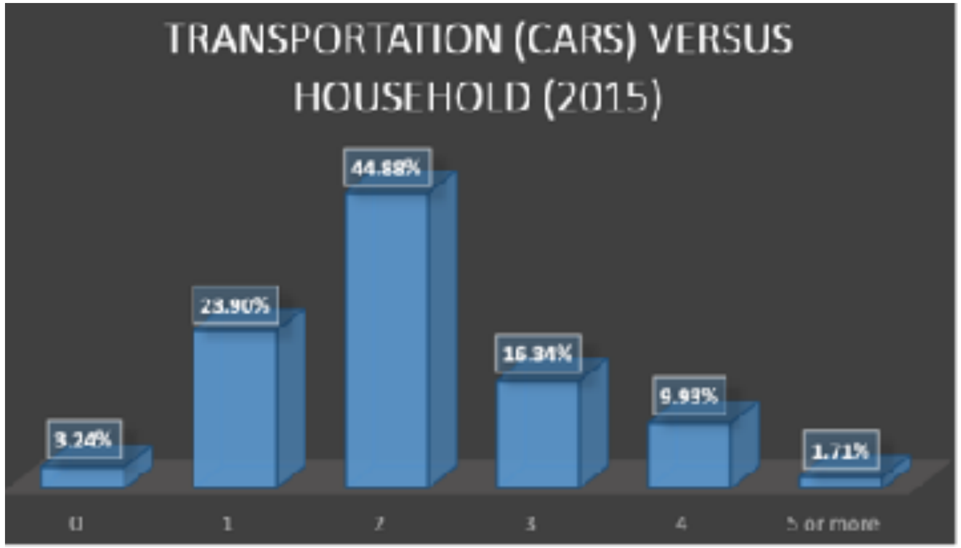


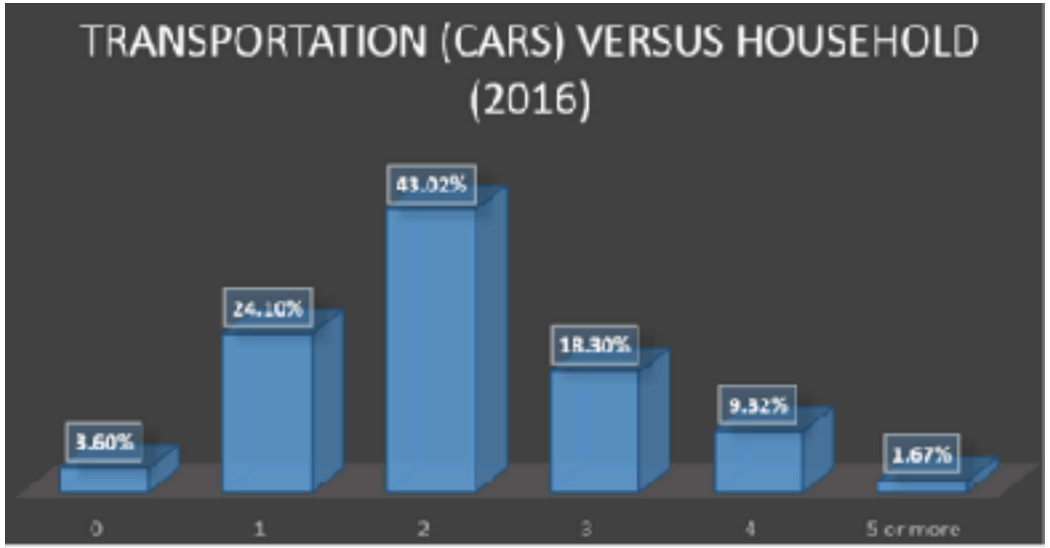


Transportation of Ames Versus Income Comparison









Transportation Types