Assignment 4A -- MSDS 410 -- Logan Strouse

INTRODUCTION

The final project for this class addresses applying the techniques learned in this class on the Ames Housing data set. The techniques were applied to the prior best performing model. This involved applying principal component analysis, factor analysis and cluster analysis to the model. The goal was to see if a better model with an improved accuracy could be modeled. I spent multiple days with Python to try and get the model to fit. I also submitted to Kaggle multiple times, but unfortunately, I was unable to beat the best score I had submitted previously to the site.

PRINCIPLE COMPONENT ANALYSIS

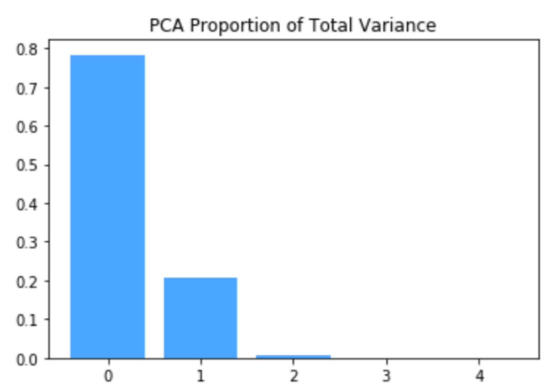
1). Can you do a dimension reduction using PCA and make the model more intuitive? Run it and show the results.

The main goal of PCA was to reduce dimensionality in the data. After multiple attempts to reduce the amount of variability through PCA, I came to the conclusion that my original model was the most accurate based upon better scores on Kaggle on and also a solid R squared score of .863. Below are some of the results I came across when I ran the PCA on my best performing model.

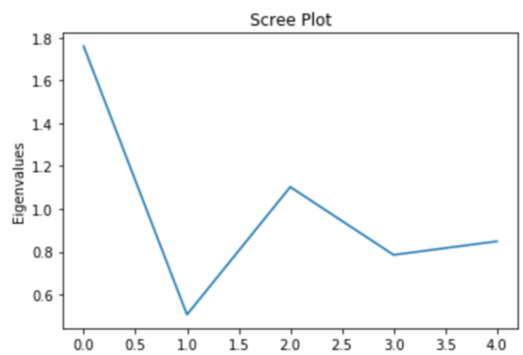
```
pca_explained_variance = pca.explained_variance_ratio_
print('Proportion of variance explained:', pca_explained_variance)
```

```
Proportion of variance explained: [0.783 0.209 0.008 0.    0.    ]
```

```
plt.bar(np.arange(len(pca_explained_variance)), pca_explained_variance,
    color = 'dodgerblue', alpha = 0.8, align = 'center')
plt.title('PCA Proportion of Total Variance')
plt.show()
```



```
   eigenvalues  0
0     1.759157  1
1     0.505695  2
2     1.102185  3
3     0.784933  4
4     0.848030  5
5          NaN  6
6          NaN  7
7          NaN  8
8          NaN  9
```



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.728e+05	1.431	1.21e+05	0.000	1.73e+05	1.73e+05
pca1	0.9997	2.17e-05	4.61e+04	0.000	1.000	1.000
pca2	-0.0139	0.000	-37.572	0.000	-0.015	-0.013

Omnibus:	48.395	**Durbin-Watson:**	2.074
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	106.260
Skew:	-0.133	**Prob(JB):**	8.43e-24
Kurtosis:	4.216	**Cond. No.**	6.59e+04

Based on the results, the number of factors to use appears to be one. If the first PCA component is chosen, the proportion of variance explained is 78%. The bar chart illustrates the percentages well in regard to giving the data in a visual form. The scree plot shows the same data as well, with where the falloff happens. I believe that suggests that one PCA component should be used. Unfortunately, this model did not score well on Kaggle and produced a score of 92,238, which is just slightly better than using the average. I moved on to the factor analysis portion next hoping for a better result.

FACTOR ANALYSIS

2). Will a PCA or FA set of variables provide a model improvement? Run it and show the results.

The factor analysis I ran did not improve the model either. Below are the results of running the regression. The model scored a poor R Squared and also did not do well on Kaggle, scoring worse than my model with PCA. In hindsight, I noticed that these two new techniques tended to help making the model more explainable. Although, this did not tend to improve their performances and accuracies when it came time to provide results.

Model:	OLS	**Adj. R-squared:**	0.144
Method:	Least Squares	**F-statistic:**	41.54
Date:	Sun, 10 Mar 2019	**Prob (F-statistic):**	8.98e-25
Time:	22:10:44	**Log-Likelihood:**	-2610.0
No. Observations:	726	**AIC:**	5228.
Df Residuals:	722	**BIC:**	5246.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	34.1033	0.328	103.988	0.000	33.459	34.747
fa1	0.8192	0.328	2.498	0.013	0.175	1.463
fa2	0.5878	0.328	1.792	0.074	-0.056	1.232
fa3	3.5197	0.328	10.732	0.000	2.876	4.164

Omnibus:	65.019	**Durbin-Watson:**	1.852
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	264.400
Skew:	0.299	**Prob(JB):**	3.86e-58
Kurtosis:	5.895	**Cond. No.**	1.00

CLUSTER ANALYSIS

3). Will a cluster analysis result in a realignment of the neighborhoods? Run it and show the results.

The cluster analysis was the most interesting part of the assignment. I used parts of the second tutorial and built upon that with information I researched online to help give me a better visual representation of the data. The neighborhood clusters still ended up being very similar to the prior groupings. I have attached a picture of that as well. This is mainly due to the neighborhoods having similar features and selling prices. Some examples include similar lot frontages as well as lot areas.

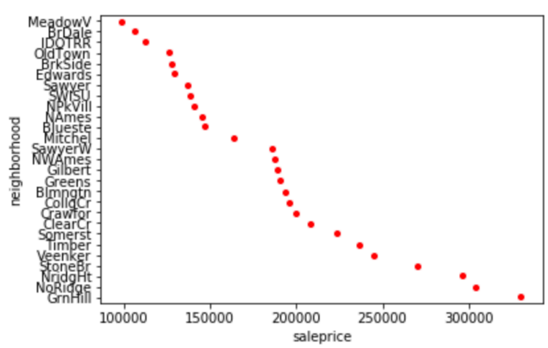
Here are the prior groupings for reference, non- cluster analysis.

	neighborhood	actual_ppsf	predicted_ppsf	Neighborhood_Group
0	GrnHill	123.318386	123.318386	1
1	Blmngtn	118.892612	117.839159	1
2	NridgHt	116.384053	119.373685	1
3	Somerst	110.542232	111.702139	1
4	StoneBr	103.866523	105.737218	1
5	Timber	103.230292	103.698200	1
6	Gilbert	101.491760	100.569597	1
7	CollgCr	99.736694	99.986461	1
8	NoRidge	97.265357	97.242276	1
9	Crawfor	96.726724	96.512449	2
10	Blueste	95.722814	95.435617	2
11	SawyerW	92.704506	92.312512	2
12	Greens	91.696116	91.476015	2
13	BrkSide	89.799890	90.008079	2
14	Veenker	88.186158	88.504960	2
15	Mitchel	86.356609	86.364205	2
16	IDOTRR	85.426961	83.905890	2
17	OldTown	85.169046	83.848181	2
18	ClearCr	85.027752	84.657259	3
19	NWAmes	83.790941	83.103876	3
20	NPkVill	82.931444	82.055189	3
21	NAmes	82.390030	81.178871	3
22	Sawyer	81.652367	80.182541	3
23	Edwards	80.688669	79.396191	3
24	BrDale	77.510648	77.060749	3
25	SWISU	76.376106	75.010812	3
26	MeadowV	68.985885	67.620480	3

This is the output I used to decide, when grouping the data into four clusters. I also used the silhouette coefficient for advice as well on this. The coefficient was high enough that I felt confident in the grouping logic.

```
import seaborn as sns
data = pd.concat(
    [
        train.groupby('neighborhood').mean()['saleprice']
    ],

    axis=1)
f, ax = plt.subplots()
sns.stripplot(data.sort_values(by='saleprice').saleprice, data.sort_values(by='saleprice').index, orient='h', color='re
```



Below are my four clusters after running the Python code. One thing that I found interesting was that each cluster seemed to have very distinct statistics. There was not much similarity between the groups in the means outputs.

```
Attribute means for segment:  0
SubClass            77.142857
LotFrontage         71.714286
LotArea           8776.714286
OverallQual          6.142857
OverallCond          5.857143
YearBuilt         1960.428571
TotalBsmtSF       1030.428571
FirstFlrSF        1022.714286
SecondFlrSF        468.857143
LowQualFinSF         0.000000
GrLivArea         1491.571429
Fireplaces           0.857143
GarageYrBlt       1966.714286
GarageCars           2.142857
GarageArea         569.000000
YrSold            2007.571429
SalePrice       188857.142857
dtype: float64

Attribute means for segment:  1
SubClass           190.0
LotFrontage         75.0
LotArea          11625.0
OverallQual          5.0
OverallCond          4.0
YearBuilt         1965.0
TotalBsmtSF       1039.0
FirstFlrSF        1039.0
SecondFlrSF          0.0
LowQualFinSF         0.0
GrLivArea         1039.0
Fireplaces           0.0
GarageYrBlt       1965.0
GarageCars           2.0
GarageArea         504.0
YrSold            2010.0
SalePrice       131500.0
dtype: float64

Attribute means for segment:  2
SubClass            20.0
LotFrontage         91.0
LotArea          11375.0
OverallQual          6.0
OverallCond          5.0
YearBuilt         1954.0
TotalBsmtSF        967.0
FirstFlrSF        1299.0
SecondFlrSF          0.0
LowQualFinSF         0.0
GrLivArea         1299.0
Fireplaces           1.0
GarageYrBlt       1954.0
GarageCars           2.0
GarageArea         494.0
YrSold            2007.0
SalePrice       150000.0
dtype: float64

Attribute means for segment:  3
SubClass            59.444444
LotFrontage         53.111111
LotArea           7931.555556
```

```
OverallQual           5.111111
OverallCond           5.222222
YearBuilt          1959.555556
TotalBsmtSF         784.777778
FirstFlrSF         1009.777778
SecondFlrSF         242.555556
LowQualFinSF          0.000000
GrLivArea          1252.333333
Fireplaces            0.333333
GarageYrBlt        1966.666667
GarageCars            1.444444
GarageArea          401.111111
YrSold             2007.222222
SalePrice        134050.000000
dtype: float64
```

The main difference I noticed was that the grouping of neighborhoods became four instead of the three original. If there was a question on whether I felt three were doable, I would answer yes. I would point to the graph which provides visual proof of three distinct clusters and also the silhouette coefficient values.

CONCLUSION

This assignment proved to be a very fun challenge and it introduced many different concepts that I hope to build on in the future to become a modeler. The concept of parsimony is very important and I learned that if too many variables or features are added, it can skew the intended results of the analysis.