Exercise 3

SDS 323

Jaeho Jang (jj36386)

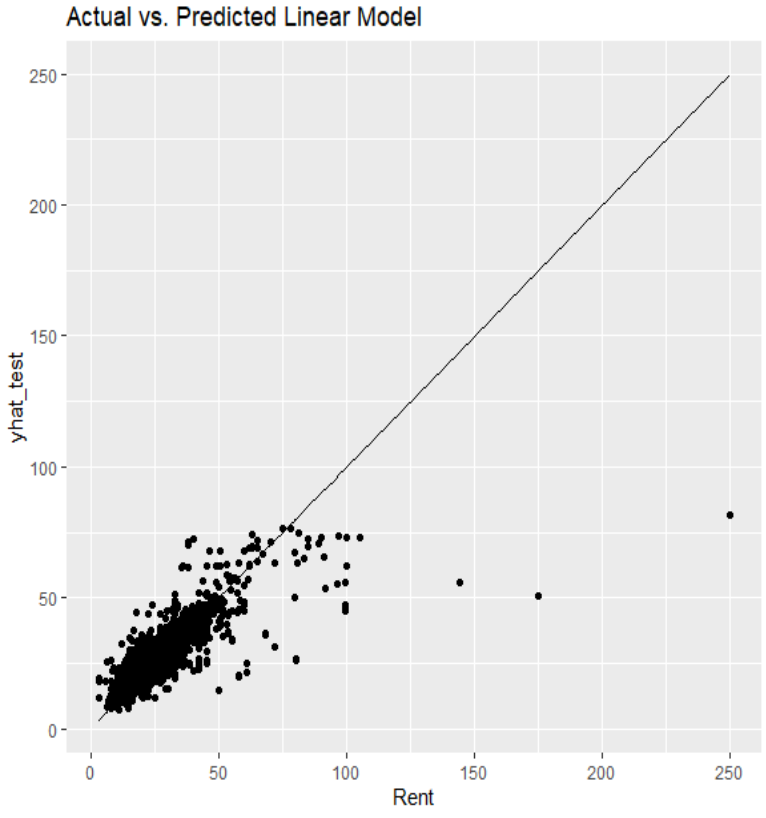
1. Predictive Model Building

1) Overview

For this analysis, I looked for a model that could best predict rent price within different rental properties in the U.S. Moreover, I quantified the average change in rental income per square foot with green certification to compare different models and tell which is most efficient.

2) Linear Model

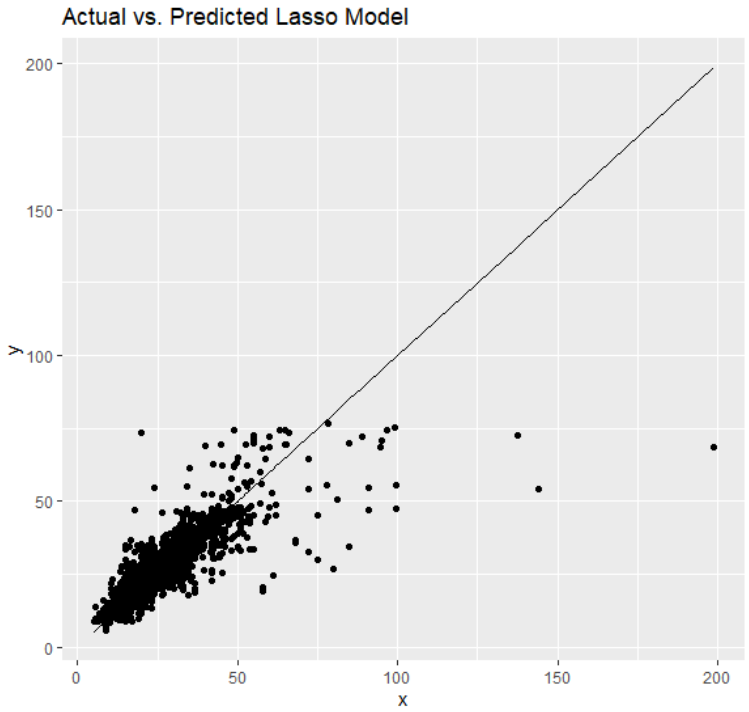
I started with the linear model to set the basic standard to compare to the other model and its RMSE to. Below is the plot of actual values and predicted value of linear model and the RMSE value.




```
Linear RMSE: 9.444652
```

3) Lasso Model

Accordingly, to capture a more detailed prediction, I implemented the lasso model. Below is the plot and RMSE value for the lasso model.




```
Lasso Regression RMSE: 8.650327
```

4) Conclusion

As seen from the two prediction models and their RMSE values, we can conclude that the lasso model depicts a more accurate version that the linear model due to lessen errors in the process of predictions.

2. What Causes What?

Q1. Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime?

A: According to the podcast, we cannot assume that there is a causation between variables. Even if there may be some correlation between variable of crime and police, it does not mean that police and the reason to crimes. There could be plenty of other variable that could be having influence

over the variable "Crime"; that being said, other variables must be controlled to see the real effect of variable police on variable crime.

Q2. How were the researchers from UPenn able to isolate this affect? Briefly describe their approach and discuss their result in the "Table 2" below, from the researchers' paper.

A: In the podcast, it is mentioned that isolation was achieved by measuring effect of police on crime when there was high number of police in area for reasons other than crime; for example, events. As seen in "Table 2", Washington D.C. was chosen to see whether if it was actually the number of police that attracted crime when there was an event non-related to crime was occurring, and when crime rate was measured, it had significantly dropped compared prior to isolation. Furthermore, metro ridership of tourist was also measured to check if number of police had effect on this as well, and this regression in fact showed there is no relationship (causation).

#Q3. Why did they have to control for Metro ridership? What was that trying to capture?

#A: As mentioned from last question, metro ridership was controlled to measure the actual effect of police number on crime rate, or if it was traffic volume of tourists (potential terrorists). It was shown that ridership was not affected.

#Q4. Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

#A: The model seems to be predicting the relationship between crime and other variables There are 2 variables and also a constant to fit the data into linear model. The table seems to tell that number of police have particularly strong effect on crime in district 1 compared to other districts. The ridership variable, as told in Q2 and Q3, states that the tourist number also has some degree of effect on crime rate. In conclusion, it seems like that police number has a strong effect on crime rate in district 1, but in other districts, there are other variable that effect crime rate.

3.Clustering and PCA

1) Overview

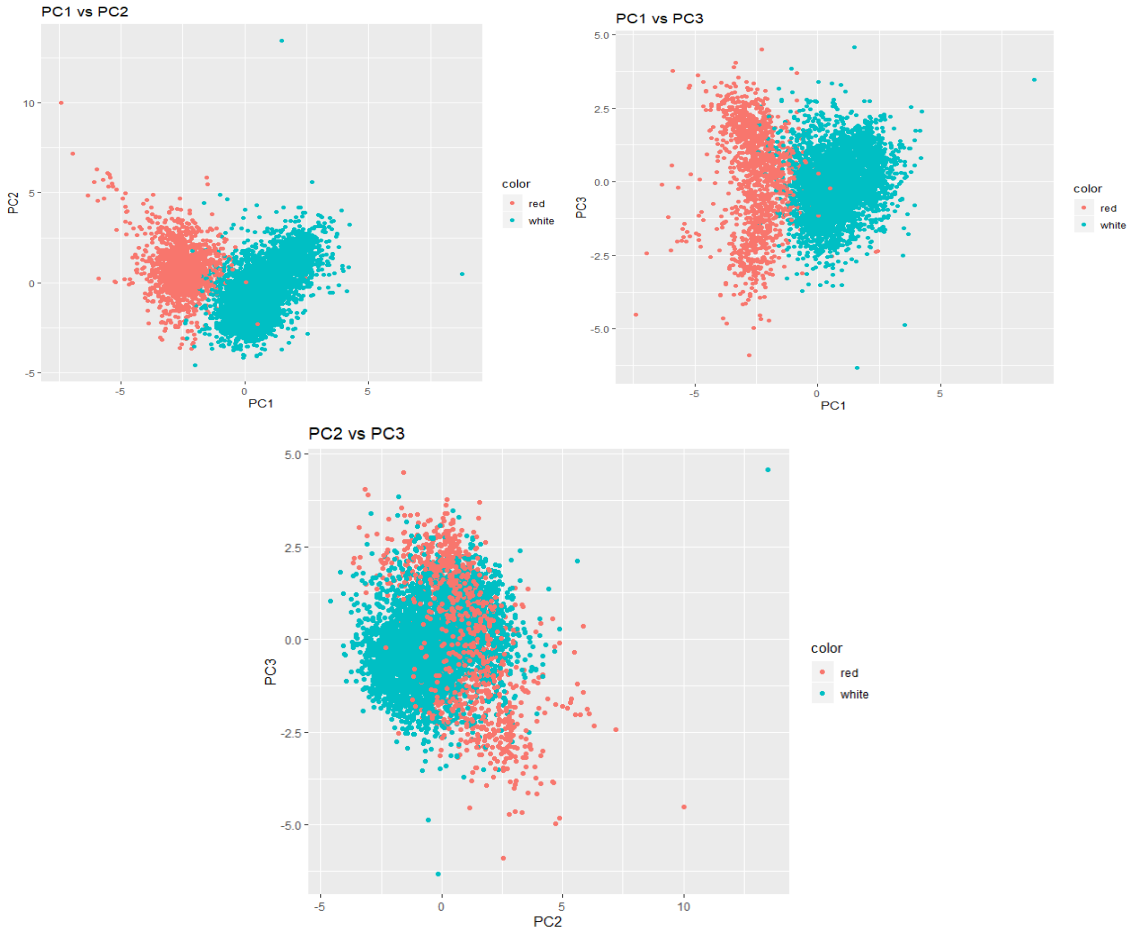
The goal of this analysis was to run both a PCA and clustering model on the 11 chemical properties of wine to summarize the results and figure out which method is more accurate for distinguishing red wine from white and using only unsupervised information, and furthermore, from high quality wine to low quality wine.

2) PCA Model

The principal component analysis (PCA) was conducted to reduce the number of data to increase adaptability but capture as much variability as possible. Below is the summary of PCA with our wine data set.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	1.7407	1.5792	1.2475	0.98517	0.84845	0.77930	0.72330	0.70817	0.58054	0.4772	0.18119
Proportion of Variance	0.2754	0.2267	0.1415	0.08823	0.06544	0.05521	0.04756	0.04559	0.03064	0.0207	0.00298
Cumulative Proportion	0.2754	0.5021	0.6436	0.73187	0.79732	0.85253	0.90009	0.94568	0.97632	0.9970	1.00000

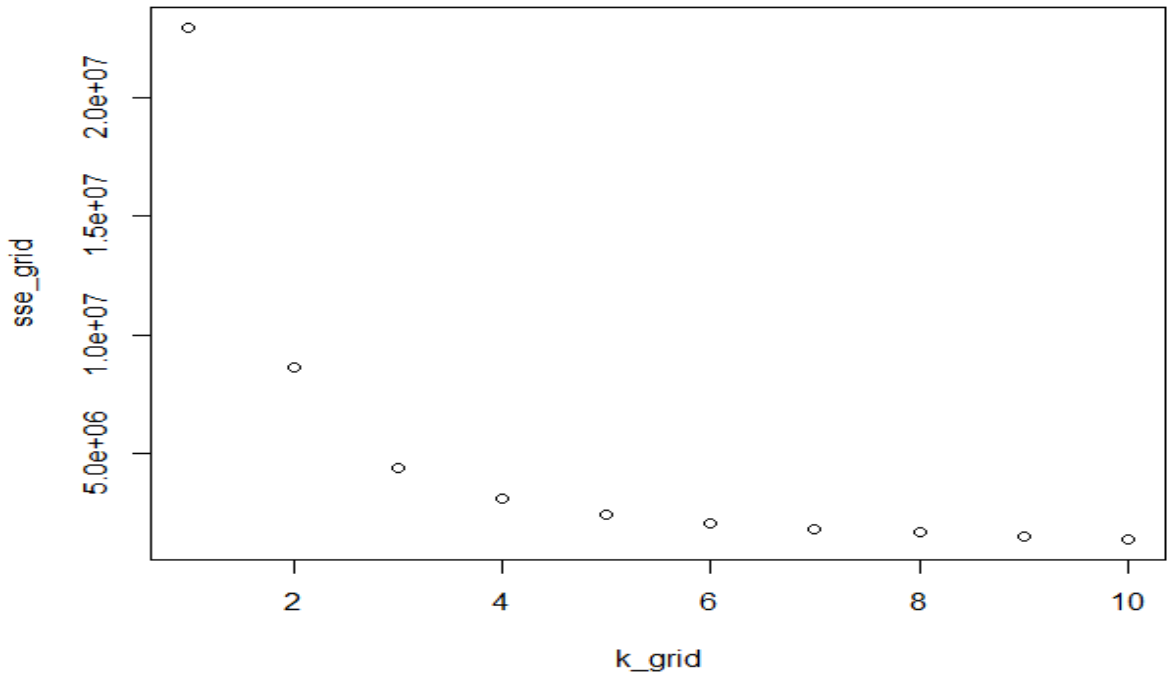
With the first 3 PCs comprising over 60% of variation, we can look at them and identify how PCA performs. The following models were created to understand how well wines were being distinguished in each PC level.



Looking at the models, we can tell that the PCA performs decently at distinguishing the wines based on color; 2 out of 3 models seem to be clustered.

3) Clustering Model; K-means++

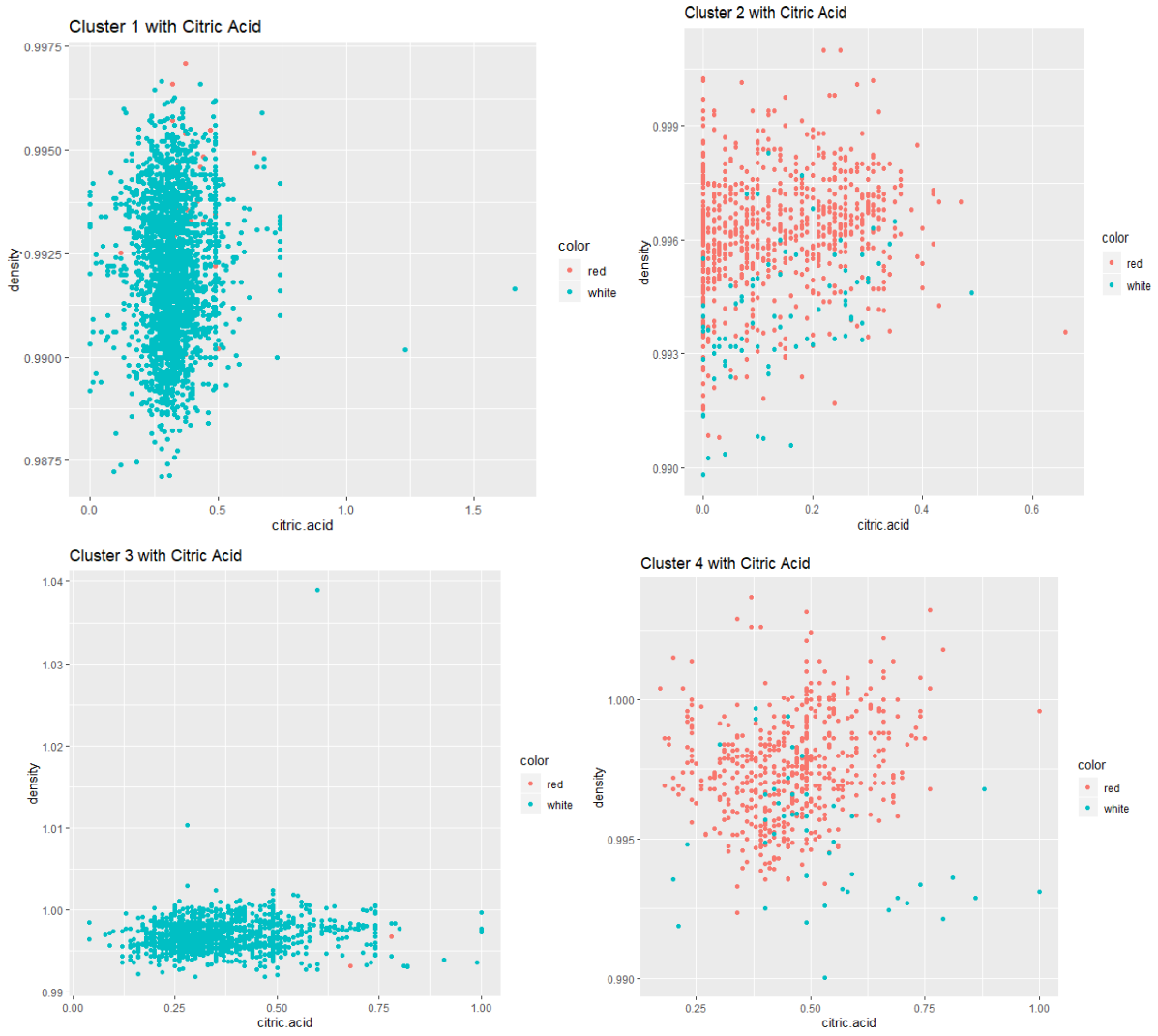
The K-means++ Clustering model was chosen because the method rids of errors in the clustering process and can tell us how the k-means fit our data. Below is the plot of k values fitted for according SSE values.



Looking at the fitted graph, our optimal k-value seems to be at k = 4. Accordingly, 4 clusters were created with average amount of each 11 chemicals that exist in a wine bottle. For each cluster, average values of chemical variables were assigned.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]
[1,]	-1.14307987	0.009927315	-0.03847891	0.71707381	-5.7854778	0.3610432	-0.25215665	-71.0768477	-1.1061091	1.39861847	0.4516172
[2,]	0.01515736	15.457796677	-0.00496425	-0.25215665	-71.0768477	-1.1061091	1.39861847	0.4516172	1.5201913	0.07055132	3.2629408
[3,]	0.15445871	62.099374227	7.25906973	1.39861847	0.4516172	1.5201913	0.07055132	3.2629408	0.7170738	-5.78547780	0.3610432
[4,]	2.39707521	-0.852639787	1.52019125	0.07055132	3.2629408	0.7170738	-5.78547780	0.3610432	-0.2521567	-71.07684772	-1.1061091

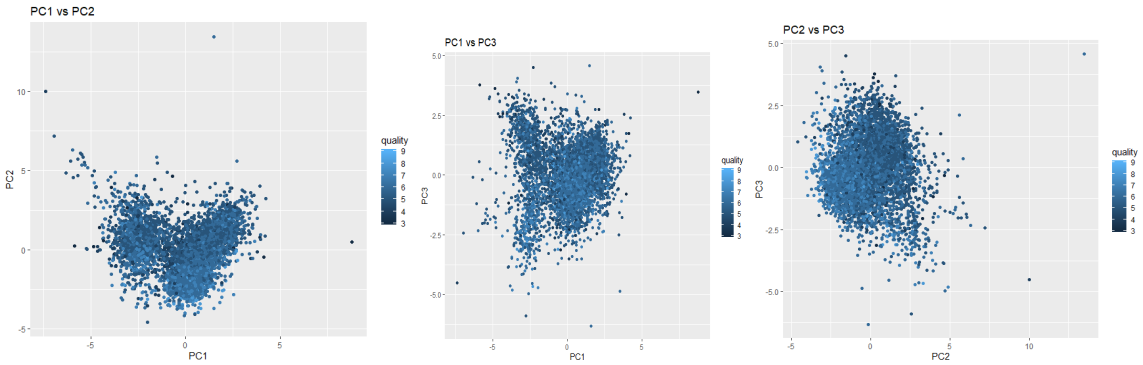
To verify that these cluster were valid, I chose one of the chemicals – citric acid – and prompted plots for each.



Looking at the single plots, the clusters seem to not only have central points but also have dominant wine color. This tells us that our cluster is reasonable.

4) Conclusion

I believe both the PCA model and K-means++ was successful at discerning the color of the wine from red to blue, based on the plots presented. However, the question of discerning quality of wine remains uncertain. Below is the PCA plots that I implemented to discern whether each PC could give an accurate cluster based on quality of wines.



As seen in the plots, there is no certain cluster that could be told. In conclusion, through on the PCA model, although color of wines could be easily discerned, the quality of wine would be hard to distinguish.

4. Market Segmentation

1) Overview

Given the social media data collected in the course of market research studying using followers of Twitter account, this data can be used to feed clients insight and direction which market they should head towards with their marketing.

2) Method

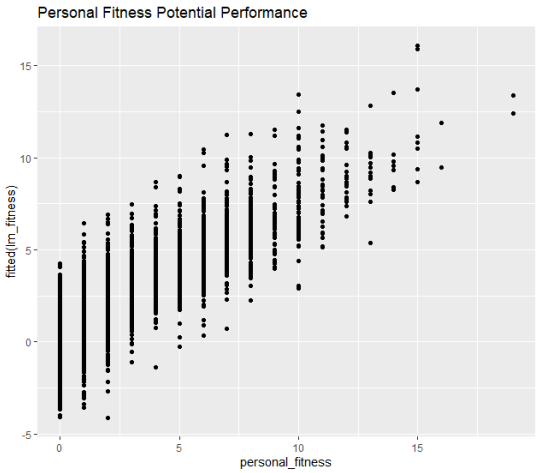
To investigate any market segment, it seems probable to approach the dataset with PCA, since it reduces ambiguity and make the data more interpretable. It would allow us to see the relationship between the more relevant variable within the dataset and leave the less relevant variables out. In conclusion, PCA would help us discover the most relevant and potential clusters within the dataset. Following is the summary of the 36 variables PCs and the first 4 PCs' – that held about 35% of data variability - coefficients.

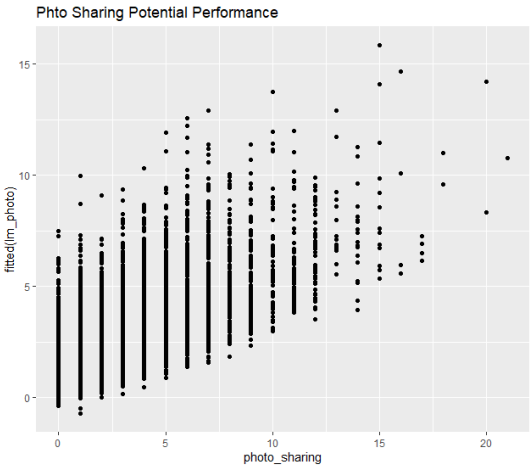
```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14    PC15
Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577 1.19277 1.15127 1.06930 1.00566 0.96785 0.96131 0.94405 0.93297
Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592 0.03952 0.03682 0.03176 0.02809 0.02602 0.02567 0.02476 0.02418
Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164 0.45369 0.49961 0.53913 0.57595 0.60771 0.63580 0.66182 0.68749 0.71225 0.73643
                          PC16    PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28    PC29    PC30
Standard deviation     0.91698 0.9020 0.85869 0.83466 0.80544 0.75311 0.69632 0.68558 0.65317 0.64881 0.63756 0.63626 0.61513 0.60167 0.59424
Proportion of Variance 0.02336 0.0226 0.02048 0.01935 0.01802 0.01575 0.01347 0.01306 0.01185 0.01169 0.01129 0.01125 0.01051 0.01006 0.00981
Cumulative Proportion  0.75979 0.7824 0.80287 0.82222 0.84024 0.85599 0.86946 0.88252 0.89437 0.90606 0.91735 0.92860 0.93911 0.94917 0.95898
                          PC31    PC32    PC33    PC34    PC35    PC36
Standard deviation     0.58683 0.5498 0.48442 0.47576 0.43757 0.42165
Proportion of Variance 0.00957 0.0084 0.00652 0.00629 0.00532 0.00494
Cumulative Proportion  0.96854 0.9769 0.98346 0.98974 0.99506 1.00000
```

	PC1	PC2	PC3	PC4
chatter	-0.1	0.2	-0.1	0.1
current_events	-0.1	0.1	-0.1	0.0
travel	-0.1	0.0	-0.4	-0.1
photo_sharing	-0.2	0.3	0.0	0.2
uncategorized	-0.1	0.1	0.0	0.0
tv_film	-0.1	0.1	-0.1	0.1
sports_fandom	-0.3	-0.3	0.1	0.1
politics	-0.1	0.0	-0.5	-0.2
food	-0.3	-0.2	0.1	-0.1
family	-0.2	-0.2	0.0	0.1
home_and_garden	-0.1	0.0	0.0	0.0
music	-0.1	0.1	0.0	0.1
news	-0.1	0.0	-0.3	-0.2
online_gaming	-0.1	0.1	-0.1	0.2
shopping	-0.1	0.2	0.0	0.1
health_nutrition	-0.1	0.1	0.2	-0.5
college_uni	-0.1	0.1	-0.1	0.3
sports_playing	-0.1	0.1	0.0	0.2
cooking	-0.2	0.3	0.2	0.0
eco	-0.1	0.1	0.0	-0.1
computers	-0.1	0.0	-0.4	-0.1
business	-0.1	0.1	-0.1	0.0
outdoors	-0.1	0.1	0.1	-0.4
crafts	-0.2	0.0	0.0	0.0
automotive	-0.1	0.0	-0.2	0.0
art	-0.1	0.1	0.0	0.1
religion	-0.3	-0.3	0.1	0.1
beauty	-0.2	0.2	0.2	0.1
parenting	-0.3	-0.3	0.1	0.0
dating	-0.1	0.1	0.0	0.0
school	-0.3	-0.2	0.1	0.1
personal_fitness	-0.1	0.1	0.2	-0.4
fashion	-0.2	0.3	0.1	0.1
small_business	-0.1	0.1	-0.1	0.1
spam	0.0	0.0	0.0	0.0
adult	0.0	0.0	0.0	0.0

3) Conclusion (Report)

In conclusion, looking at the list of coefficients for the 4 different clusters, we can tell that they each have a distinct trend. Other than cluster 1, rest of them seem to have a positive coefficient category, meaning that these positive coefficient categories could be potential marketing opportunity for the client. For example, categories such as photo sharing, personal fitness, and online gaming maintained an overall positive coefficient over different clusters. If these categories were to share an age group/cultural group, and the client had the ability to figure it the group out, it would be beneficial to them. Below are plots to verify that mentioned categories do hold potential.


Personal Fitness Potential Performance
fitted(lm_fitness)
personal_fitness
15
10
5
0
-5
0
5
10
15


Phto Sharing Potential Performance
fitted(lm_photo)
photo_sharing
15
10
5
0
0
5
10
15
20