SDS 323: Exercises 3 Report

Nikhil Ajjarapu Rithvik Saravanan

April 20, 2020

Predictive model building

Overview

In this activity, we attempt to identify the best possible model to predict rent prices for a collection of 7,894 rental properties across the United States. In addition, we want to quantify the average change in rental income per square foot associated with green certification, while holding other factors constant. To do this, we will explore a variety of different models in order to build the best predictor of price and measure their accuracies. Analyzing the nature of this best price prediction model will allow us to determine the average change in rental income associated with green certification.

Data and Model

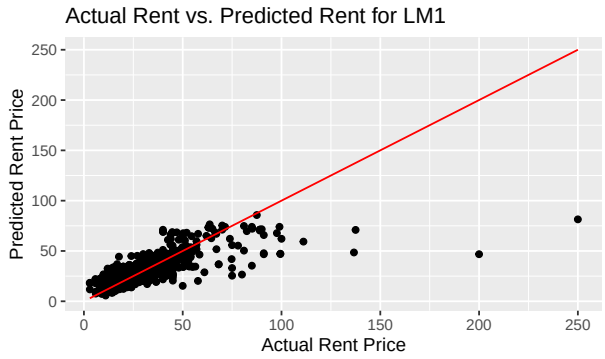
Of the 7,894 properties in the dataset, 685 properties have been awarded either LEED or EnergyStar certification as a green building. The dataset includes 22 features associated with each rental property such as size, rent, age, etc. Rent is the feature of interest in this activity as we attempt to build the best model to predict the rental price.

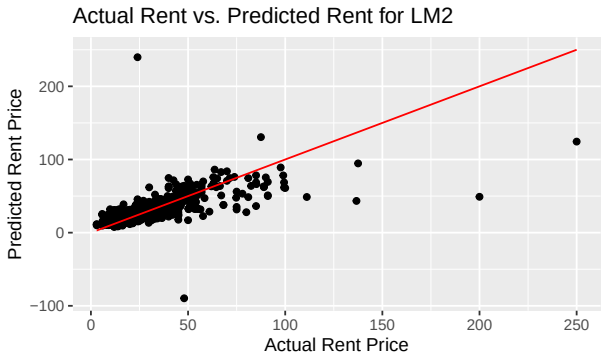
In order to test which model would be best for this activity, we built 5 different models: two linear models, a KNN regression model, a lasso regression and ridge regression model, and finally two logistic regression models. By including a variety of prediction models, we can ensure that we will have the most accurate model to describe price. We measured the performance of every model using RMSE (root mean squared error) to determine which had the least error with smaller RMSE values indicating higher accuracy. In order to reduce Monte Carlo variability, we used 200 repeated random samples of the data for each model to find the true RMSE values.

.....

Linear Model

The first model we built were two basic linear models. The first model included all of the features without interactions and the second model included all of the features with interactions. Even though lasso and ridge regression are improved versions of the linear model, we included these linear models as a benchmark. As the models involve multiple x-values, we decided to plot the actual rent price vs. predicted rent price, and include the identity function ($y = x$) as a benchmark of what perfect fit would look like.





`LM1 = LINEAR REGRESSION MODEL (without interactions) - RMSE: 9.420612`

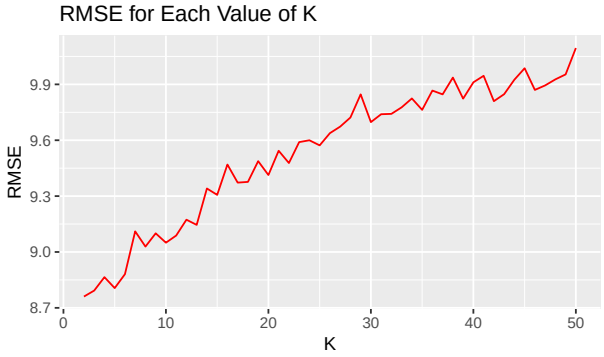
`LM2 = LINEAR REGRESSION MODEL (with interactions) - RMSE: 11.07283`

These plots show that the linear regression performed reasonably well for the most part with the model excluding interactions performing slightly better, as evidenced by the lower RMSE value.

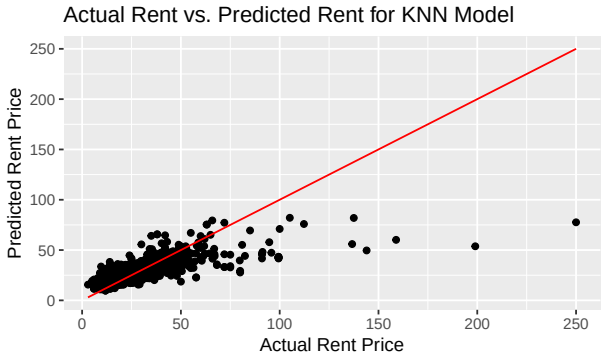
.....

KNN Model

The second model we built was a KNN model. In order to identify the optimal value of k, we tested k values from 2 - 50. From the RMSE plot below, we noted that the optimal value of k is 2 because it has the lowest RMSE value (~8.76).



Using this value of k, we again plotted actual rent price vs. predicted rent price.



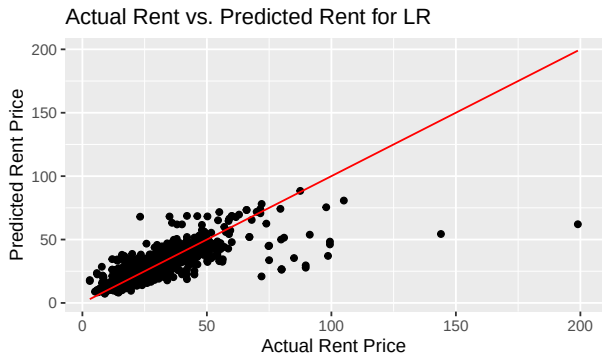
```
KNN ( k = 2  ) - RMSE: 8.760582
```

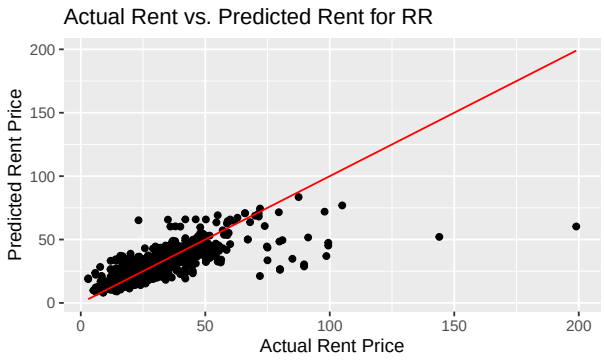
The KNN model performed comparatively better than the linear regression models because it had a lower RMSE values.

.....

Lasso and Ridge Regression Models

The third model we build were lasso and ridge regression models. Below are plots of predicted rent price vs. actual rent price for both regressions.






```
LR = LASSO REGRESSION - RMSE: 7.647866
```

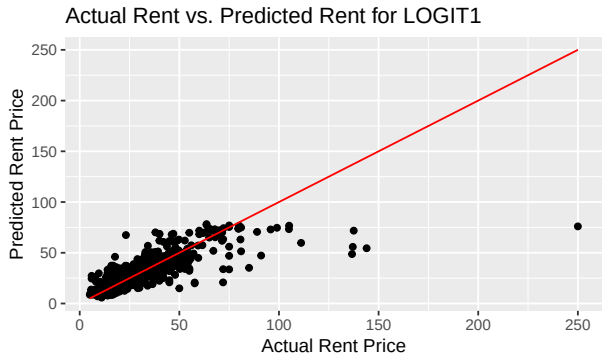
```
RR = RIDGE REGRESSION - RMSE: 7.637237
```

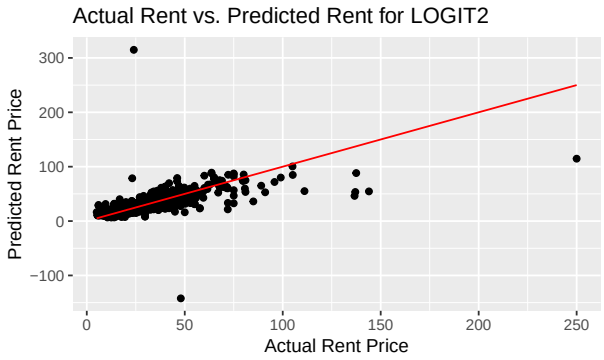
These regressions have produced the best RMSE values thus far, indicating that the predictions were more accurate than the previous models.

.....

Logistic Regression

The final model we ran was a logistic regression. Below, we have included plots of predicted rent price vs. actual rent price for logistic regression with and without interactions.






```
LOGIT1 = LOGISTIC REGRESSION (without interactions) - RMSE: 9.545176
```

```
LOGIT2 = LOGISTIC REGRESSION (with interactions) - RMSE: 7571092293
```

The RMSE values for the logistic regression are slightly higher than the previous models.

.....

Results

For each model, we ran 200 iterations of 80%/20% train/test splits, each with a random subset of data to run cross validation and get an averaged error value to prevent random splits from generating a extremely high or low error value. Although the `cv.gamlr()` has built in cross-validation, we decided to confirm it externally as it seemed the cross-validation aspect was used to determine the optimal lambda.

From the plots of each predicted model, we determined that the lasso and ridge prediction models performed the best in predicting rental price of the properties because they produced comparatively lower RMSE values. This is inline with what we expected, as we knew that lasso and ridge were regularlized linear models, which meant they would perform better than a vanilla linear regression model. However, since KNN doesn't have a predetermined model shape and logistic regression is based on a log curve, we had to test out each type of model to determine which would produce the best fit.

Conclusion

We conclude that the lasso regression prediction model functions best based on the different RMSE values that each model achieved. In order to quantify the average change in rental price, we have to look at the coefficients for the lasso regression model and determine the value for the "green rating" factor:

```
22 x 1 sparse Matrix of class "dgCMatrix"
                               1
(Intercept)       -8.490559e+00
cluster            5.667522e-04
size               7.230865e-06
empl_gr            5.753711e-02
leasing_rate       9.116945e-03
stories           -3.433687e-02
age               -1.115462e-02
renovated         -2.908905e-01
class_a            2.222371e+00
class_b            6.509654e-01
LEED               5.959820e-01
Energystar         5.687778e-02
green_rating       4.784146e-01
net               -2.522048e+00
amenities          7.103757e-01
cd_total_07       -1.630335e-04
hd_total07         5.534776e-04
total_dd_07        .
Precipitation      4.189362e-02
Gas_Costs         -2.733377e+02
Electricity_Costs  1.885409e+02
cluster_rent       1.012373e+00

Average change in rental price in regards to green rating ($): 0.4784146
```

From this, we determined that, holding other factors of the building constant, the average change in rental income per square foot associated with green certification is approximately $0.48.

What causes what?

1. **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? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)**

   We cannot do this due to the fallacy "correlation implies causation". As mentioned in the podcast, this fallacy can cause us to have irrational beliefs. In this specific example, even if there is some correlation between the variables of "Crime" and "Police", that doesn't necessarily mean that the police is the reason crime is changing. There could (and most likely are) other stronger explanations for changes in crime such as poverty, etc. Thus, all other variables must be controlled for in order to run this regression and draw any meaningful conclusions from it.

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

EFFECT OF POLICE ON CRIME

TABLE 2

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

Figure 1: The dependent variable is the daily total number of crimes in D.C. This table present the estimated coefficients and their standard errors in parenthesis. The first column refers to a model where the only variable used in the High Alert dummy whereas the model in column (2) controls form the METRO ridership. * refers to a significant coeficient at the 5% level, ** at the 1% level.

   The UPenn researchers were able to isolate this effect by measuring the effect of police on crime when there was a high number of police in an area for a reason unrelated to crime. In the example mentioned in the podcast, they said that in Washington D.C. there are often a lot of cops for events that may attract terroristic threats, which allowed them to isolate the event. When the amount of crime was measured during those times, it had significantly dropped. In addition, they also measured the number of tourists measured by metro ridership (as shown in the chart), to check if the number of police on high-alert days had any influence on the number of tourists (potential victims) out and about. The table shows that the ridership was unchanged by the number of police on high terror days, which shows that there is in fact an inverse relationship between the number of police present and the amount of crime that occurs.

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

   They controlled for Metro ridership to answer the question of whether the drop in crime was actually because of an increased police presence, or because there were just less potential victims (tourists and others who use the metro) around because they were scared by the high-alert police. As mentioned above, it was shown that ridership was not affected, which is further evidence that police themselves do have an effect on crime.

4. **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?**

TABLE 4

REDUCTION IN CRIME ON HIGH-ALERT DAYS: CONCENTRATION ON THE NATIONAL MALL

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	−2.621**	−2.621*	−2.621*
	(.044)	(1.19)	(1.225)
High Alert × Other Districts	−.571	−.571	−.571
	(.455)	(.366)	(.364)
Log(midday ridership)	2.477*	2.477**	2.477**
	(.364)	(.522)	(.527)
Constant	−11.058**	−11.058	−11.058^{+}
	(4.211)	(5.87)	(5.923)

Figure 2: The dependent variable is the daily total number of crimes in D.C. District 1 refers to a dummy variable associated with crime incidents in the first police district area. This table present the estimated coefficients and their standard errors in parenthesis.* refers to a significant coeficient at the 5% level, ** at the 1% level.

The model being estimated here is a linear model with a few variables as well as a constant to fit the data, where the dependent variable is crime. From the table, it seems to be that the theory that police influence crime holds especially strongly in District 1, but it still does hold some (albeit weak) weight in other districts as well. It seems the tourist theory mentioned earlier also holds true, as metro ridership has a positive coefficient as well. All in all, it seems that the police have a relatively strong effect on crime in District 1, and a much more moderate effect on crime in other districts after controlling for various other factors.

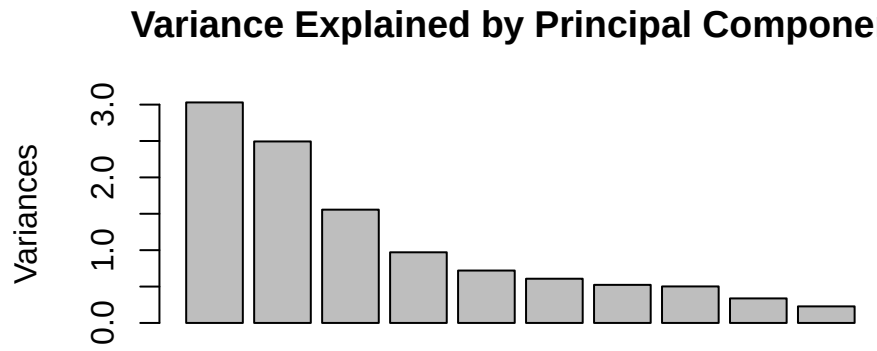
Clustering and PCA

To understand how useful PCA and clustering can be, we turn to the wine dataset. The dataset that we used for this exercise contains information on 11 chemical properties of 6500 different bottles of *vinho verde* wine from northern Portugal. This dataset also records two important features of each bottle of wine: the color (red or white) and quality (on a scale of 1-10 with 1 as low quality and 10 as high quality).

PCA

A principal components analysis would be useful in this scenario because the wine dataset includes data on 11 chemical properties of each of the wine bottles. Since these 11 properties are numerical and continuous, we attempted to explain the variation in the dataset by compressing the original 11 properties into fewer (and more manageable) values that could provide us with better insight into the dataset as a whole.

We ran a principal components analysis on the scaled data from the wine dataset and generated 11 principal components (because the number of principal components is bound by the 11 properties that we used from the dataset). The simple plot below shows the the proportion of variances for each of the 11 principal components that were generated. We have also included the characteristic information of each of the principal components to highlight their usefulness.




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

As expected, the variance plot shows a clear decrease in the variance explained for each additional principal component. This passes a basic sanity check because each additional principal component attempts to explain the remaining unexplained data after taking into consideration the previous principal components.

From the variance plot and the summary of the PCA, we noticed that the first three principal components explain about 64% of the 11 properties of wine in this dataset. Since this is a valid representation of the entire dataset, we continued our analysis focusing on the first three components. Below are the coefficients for each of the 11 properties in each of the first 3 principal components.

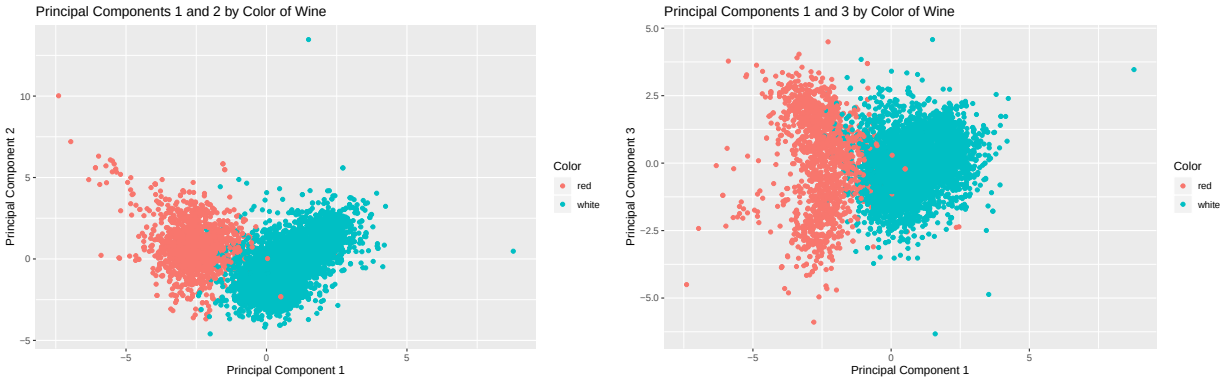
```
                       PC1   PC2   PC3
fixed.acidity        -0.24  0.34 -0.43
volatile.acidity     -0.38  0.12  0.31
citric.acid           0.15  0.18 -0.59
residual.sugar        0.35  0.33  0.16
chlorides            -0.29  0.32  0.02
free.sulfur.dioxide   0.43  0.07  0.13
total.sulfur.dioxide  0.49  0.09  0.11
density              -0.04  0.58  0.18
pH                   -0.22 -0.16  0.46
sulphates            -0.29  0.19 -0.07
alcohol              -0.11 -0.47 -0.26
```

From these coefficients, we acknowledge that the sulfur dioxide and residual sugar properties are positively weighed in PC1 (the main component) and explain a significant amount of the variance. With a little

research, we learned that sulfur dioxide is used in winemaking in order to inhibit and kill unwanted yeasts and bacteria during the fermentation process and to protect the wine from oxidation. Residual sugar is the natural grape sugar that is leftover after fermentation. We also noticed that the acidity coefficients for PC1 were strongly negative. We learned that the acidity in wine refers to the fresh, tart, and sour attributes of the wine and how well the acidity balances out the sweetness and bitterness. Since acidity is in contrast to sweetness, it is logical that sweetness (residual sugar) is strongly positive while acidity is strongly negative.

For PC2, we found that the remaing variance was mostly explained positively by density and chlorides and negatively by alcohol. For PC3, we found that the remaining variance was mostly explained positively by the pH and negatively by citric acid.

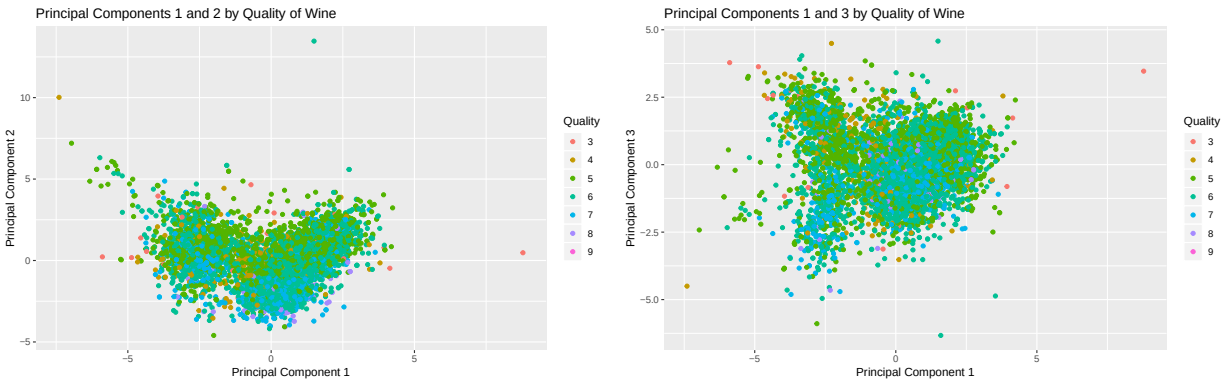
To better understand how these principal components grouped the properties of the wines in the dataset, we visualized them using the two other important features in the dataset: the color and quality of each wine. Below are plots of the principal components grouped by color and quality.



When grouping the data by the color of the wine, we observed a clear and obvious clustering of the red and white wines.

When plotting PC1 with PC2, the data points that have a lower, negative score for PC1 and a slightly positive score for PC2 are a noticeable cluster of red wines while the the data points that have a higher, positive score for PC1 and a slightly negative score for PC1 are a noticeable cluster of white wines. The distinction between the two clusters is approximately around a score of -2 for PC1.

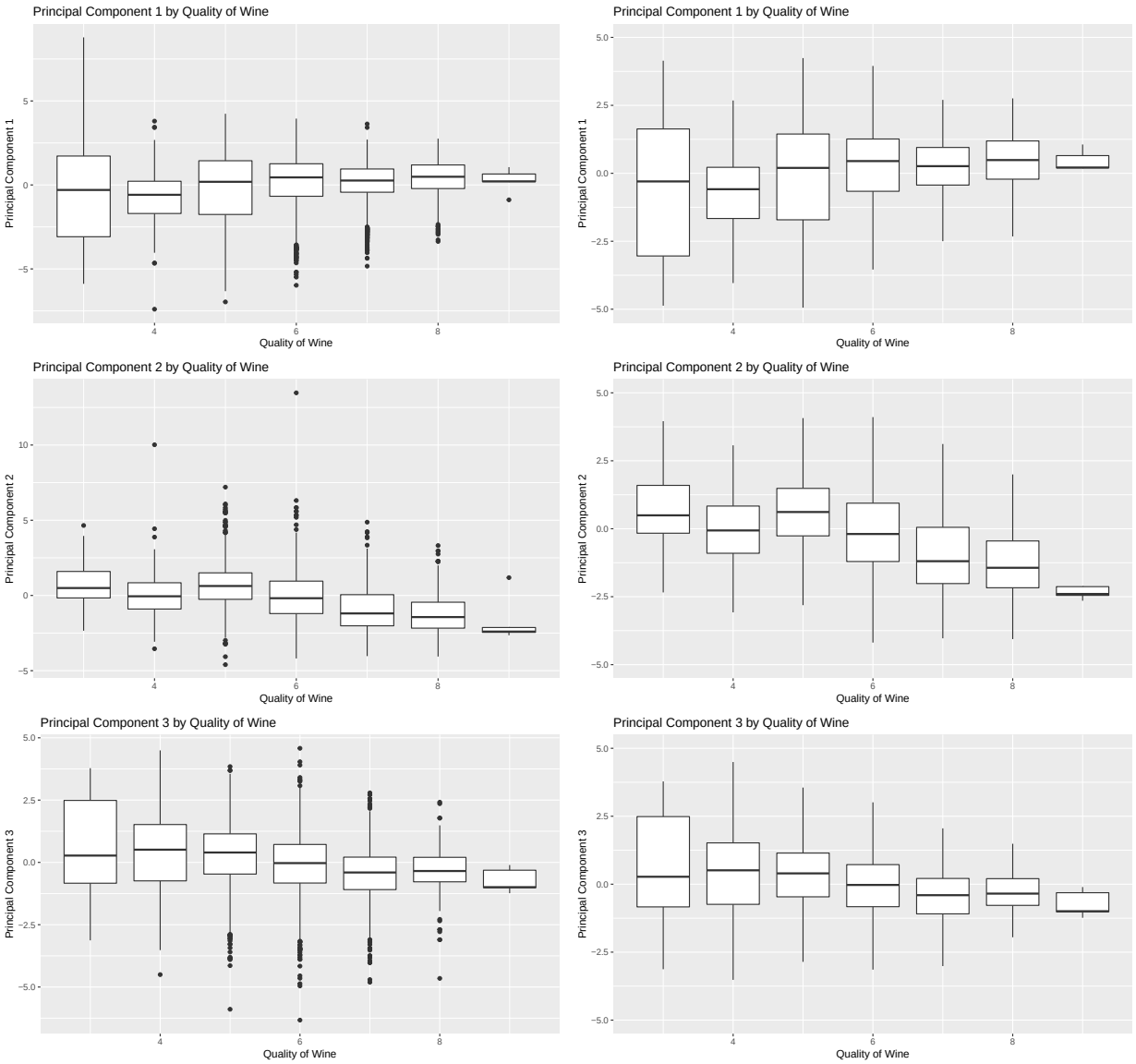
When plotting PC1 with PC3, the data points that have a lower, negative score for PC1 are a noticeable cluster of red wines while the the data points that have a higher, positive score for PC1 are a noticeable cluster of white wines. The distinction between the two clusters is approximately around a score of -1.5 for PC1.



Unlike the color of the wines, when grouping by the quality of wine, we did not observe a clear and obvious clustering to differentiate between qualities.

When plotting PC1 with PC2 and PC1 with PC3, the data points show that an overwhelming majority are categorized as mid-range qualities of 5, 6, and 7. In contrast, the colors of the wine were relatively more evenly distributed. Since there are too few wines with lower-end and higher-end qualities, using the principal components did not yield a useful grouping of the wines in the dataset.

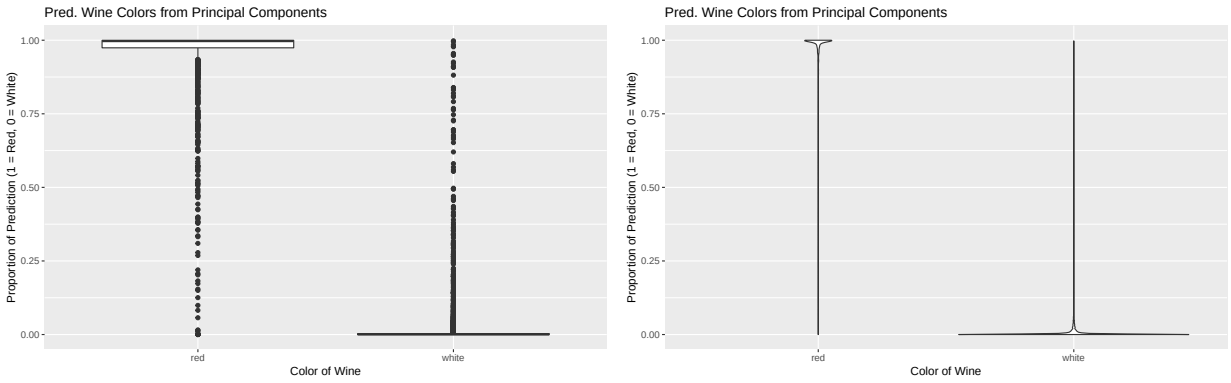
In order to get a better understanding of the principal component groupings of the qualities, we used box plots to illustrate the difference in the groupings. Below are the box plots of the wines by PC1, PC2, and PC3 in each quality of wine in the data set. We have included the plots with and without the outliers in order to provide a more holistic view of the data.



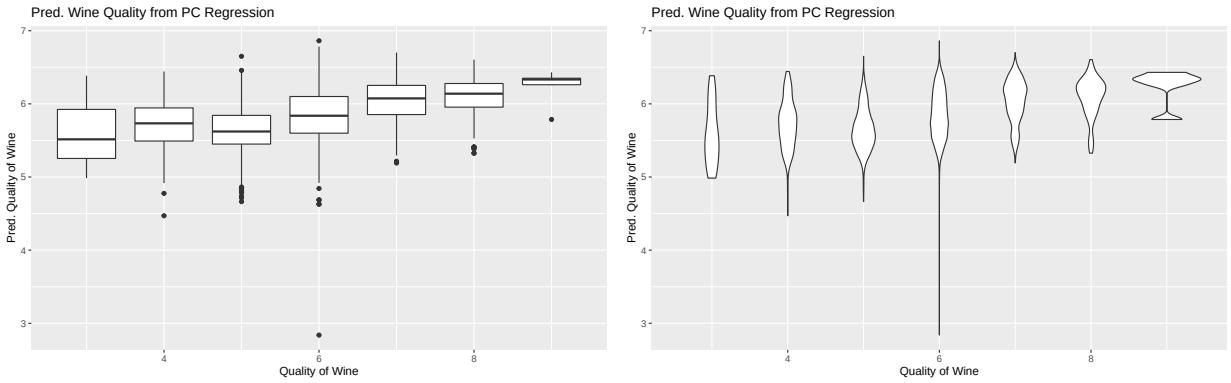
From these box plots, we notice a slight positive trend when looking at the distributions for PC1 by quality. The higher quality wines had overall slightly higher PC1 scores than lower quality wines. Looking at the distributions for PC2 and PC3 scores by quality, we identified a relatively strong negative trend for the plot

of PC2 and a slight negative trend for the plot of PC3. The higher quality wines had overall lower PC2 and PC3 scores than lower quality wines.

To further utilize our PCA, we ran a principal component regressions to predict color and quality using the principal components (PC1, PC2, and PC3). We ran a logistic regression to predict color because there were two possible outcomes: red and white, and we ran a linear regression to predict quality because quality is a continuous variable between 1 and 10. Below are the plots of both regressions in the form of box plots and violin plots.



We assigned a value of 1 to red and 0 to white when running the logistic regression. From the regression plots of predicted color, we observed that the majority of predictions were accurate because the red wine predictions were heavily populated in the value of 1 and the white wine predictions were heavily populated in the value of 0. As a result, we recognized that this regression is useful in predicting wine color.

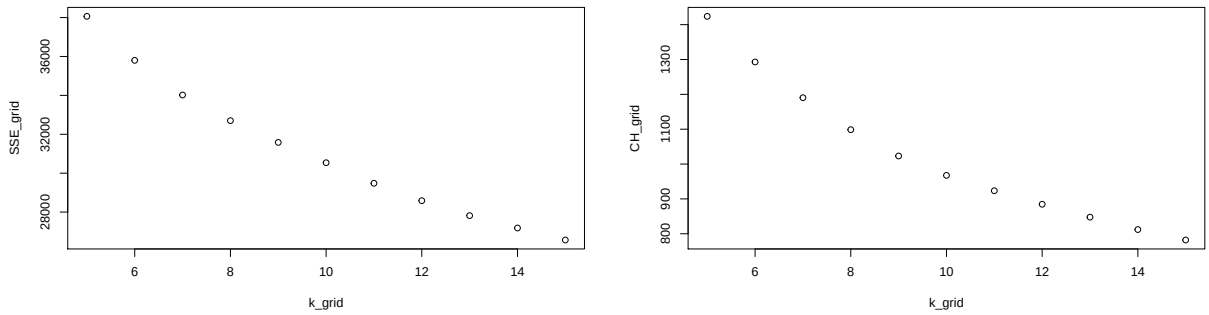


From the regression plots of predicted quality, we noticed that there is a positive, upward trend in the distributions with increasing quality. Since lower quality wines were predicted as relatively lower quality and higher quality wines were predicted as relatively higher quality, we recognized that this regression is useful in predicting general wine quality.

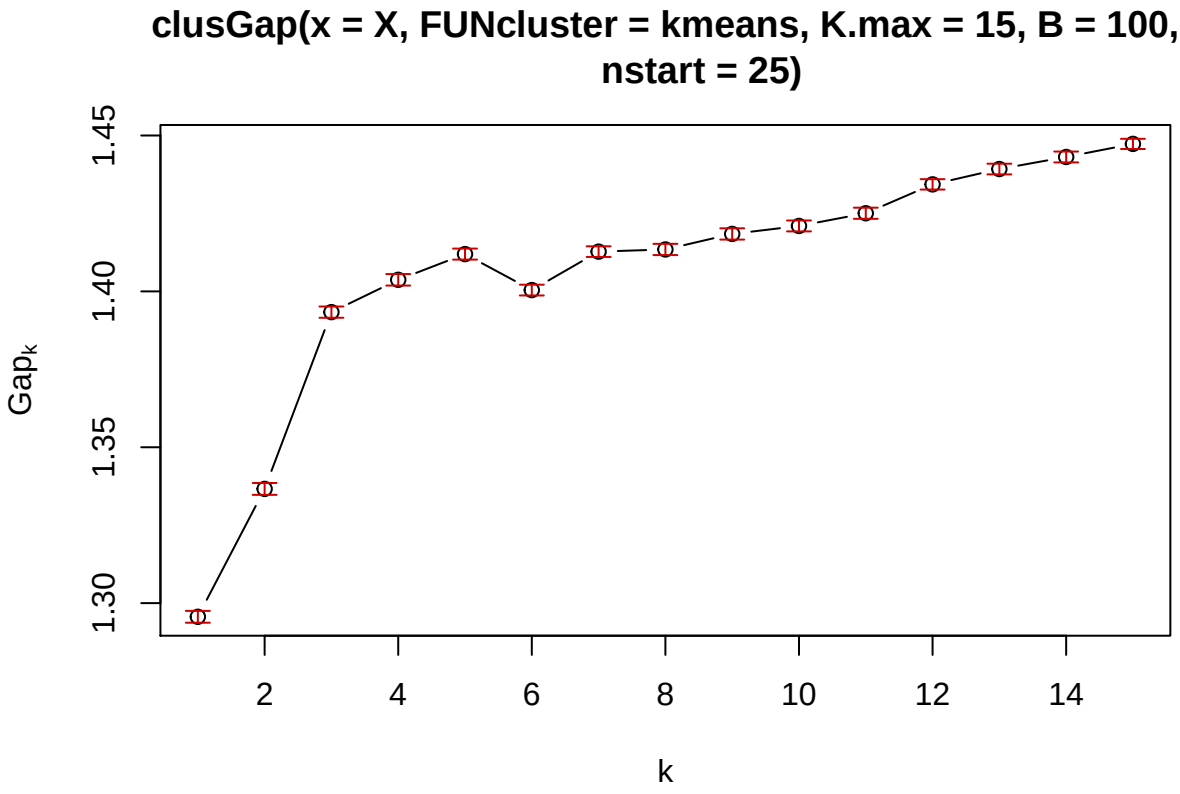
K-means Clustering

In addition to running a principal components analysis of the 11 properties in the dataset, we explored the usefulness of other clustering and grouping techniques. Specifically, we took advantage of K-means and K-means++ to identify whether we could cluster the data in a meaningful manner.

To identify the number of clusters to use in our K-means clustering algorithms, we attempted to utilize the elbow and CH plots shown below.



Since there is no significant k value identified by either of the plots above, we instead used a gap statistic plot (shown below) to help us identify the value of k.

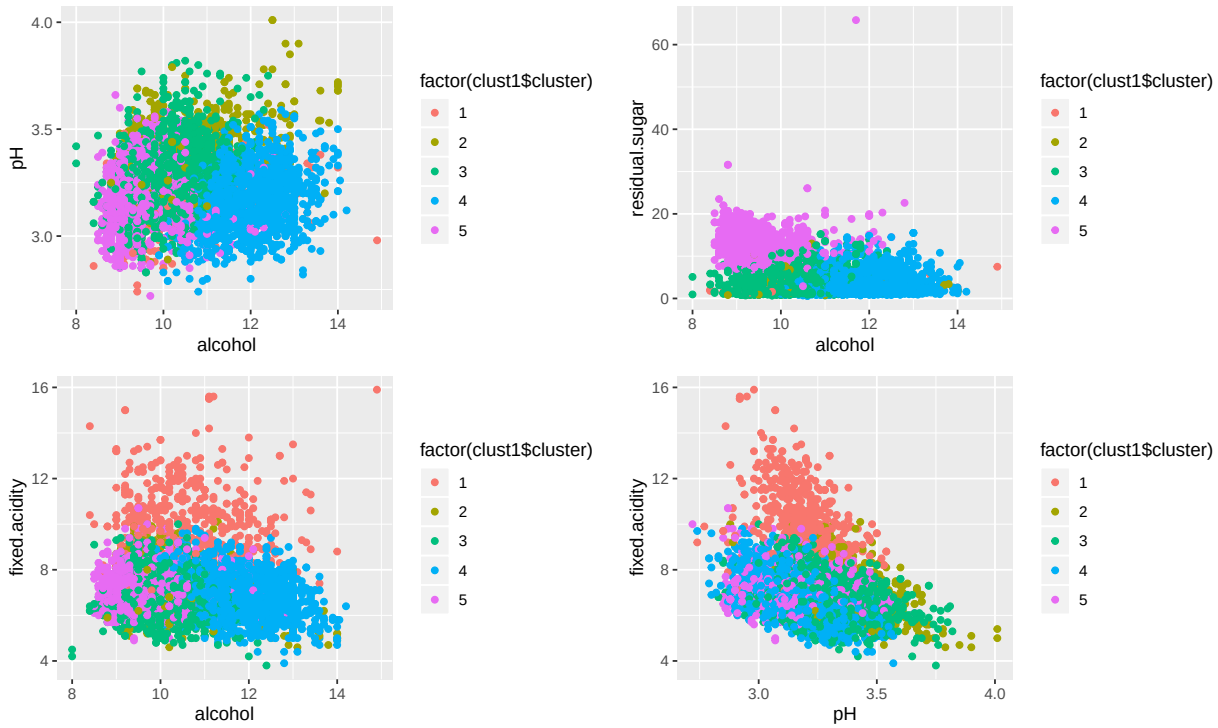


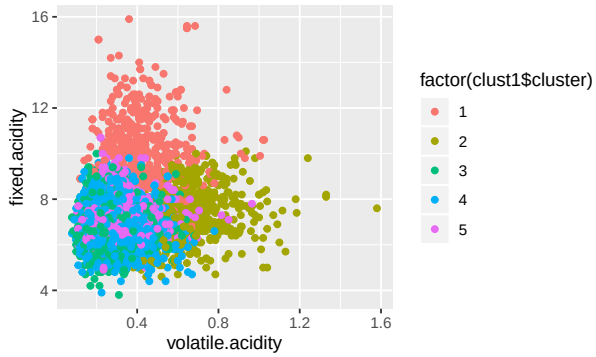

```
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = X, FUNcluster = kmeans, K.max = 15, B = 100, nstart = 25)
B=100 simulated reference sets, k = 1..15; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 5
          logW    E.logW      gap      SE.sim
```

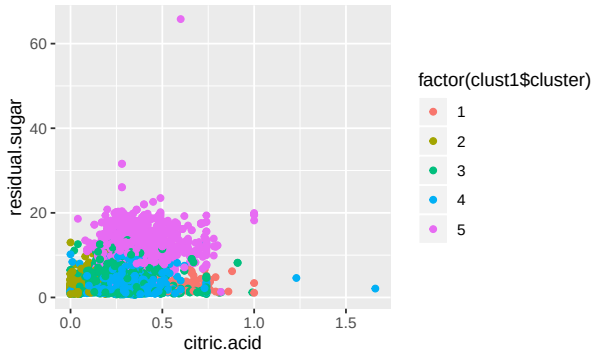
```
 [1,] 8.873237 10.168852 1.295615 0.001899947
 [2,] 8.748376 10.085008 1.336633 0.001899493
 [3,] 8.635368 10.028710 1.393342 0.001800960
 [4,] 8.584878  9.988589 1.403711 0.001844501
 [5,] 8.546191  9.958148 1.411958 0.001767827
 [6,] 8.530676  9.931093 1.400417 0.001728401
 [7,] 8.499807  9.912561 1.412754 0.001706577
 [8,] 8.481803  9.895250 1.413447 0.001799938
 [9,] 8.464968  9.883395 1.418427 0.001811486
[10,] 8.451232  9.872224 1.420992 0.001750751
[11,] 8.437011  9.862075 1.425064 0.001782774
[12,] 8.417964  9.852292 1.434328 0.001664545
[13,] 8.404414  9.843657 1.439243 0.001713981
[14,] 8.392199  9.835311 1.443113 0.001741035
[15,] 8.379712  9.827018 1.447306 0.001635170
```

From this gap statistic plot, we observed that the function is non-increasing from $k = 5$ to $k = 6$. Thus, we initialized both of our algorithms with generating 5 clusters.

To get a rough idea of how this basic clustering algorithm performs, we attempted to observe the accuracy of the clustering on some of the properties from the dataset.

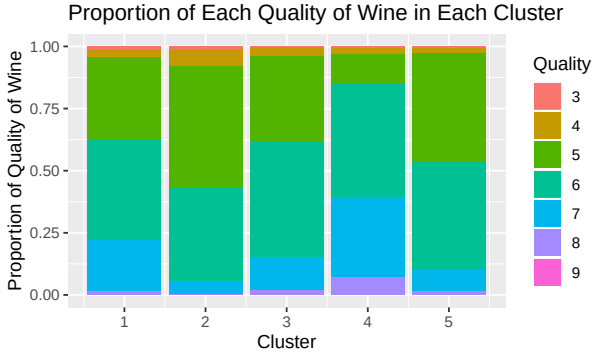






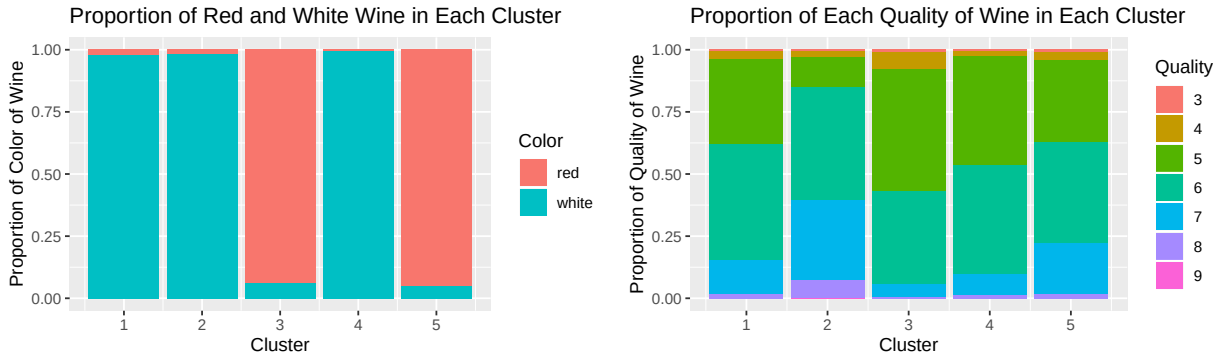
It is important to note that K-means generates reasonable clusterings because each cluster has a relatively well-defined region in each plot. For example, the wines in cluster 4 have noticeably less alcohol and more residual sugar than wines in cluster 5 (from the plot at the top right). However, the issue persists that there is a significant amount of overlap between the individual clusters in each of the plots. This issue may be attributed to the fact that we only plotted a few pairs of the 11 total properties.

To get a better visualization of the clusters, we observed how well they group the two additional features: color and quality.



These plots show a more descriptive explanation of the clusters. From the plot of red vs. white by cluster above, we understood that clusters 1 and 1 are predominantly red wines while clusters 3, 4, and 5 are predominantly white wines. Additionally, the plot of quality by cluster shows that clusters 2, 3, and 5 are roughly mid-quality wines while clusters 1 and 4 contain comparatively higher-quality wines.

To determine whether our clustering could be further improved, we ran a K-means++ clustering algorithm with $k = 5$. Below are the same plots that were generated to observe a meaningful description of each cluster.



From the plot of red vs. white by cluster above, we acknowledge that clusters 3 and 5 are predominantly red wines while clusters 1, 2, and 4 are predominantly white wines. Additionally, the plot of quality by cluster shows that clusters 1, 3, and 4 have mid-quality wines while clusters 2 and 5 are roughly higher-quality wines. This clusteirng algorithm shows slightly better groupings from the basic K-means apprach because each cluster is more distinct and can be grouped together more clearly, especially in terms of the quality groupings. To see if this improvement is reflected in the data, we observed the within-cluster and between-cluster average distances for the two clustering algorithms.

```
K-means total within-cluster distances: 38063.17
```

```
K-means++ total within-cluster distances: 38063.17
```

```
K-means between-cluster distances: 33392.83
```

```
K-means++ between-cluster distances: 33392.83
```

Since our value $k = 5$ is relatively small, the distance within and between clusters between the K-means and K-means++ clustering algorithms is not very distinguishable. So, these measures are not the best way to convey that K-means++ is preferred over K-means for the wine dataset. Given this, we acknowledge that the clusters from K-means++ show more understandable groupings than the clusters from K-means. Since the main goal of clustering is to cluster the data in a manner that makes it easy to interpret, we acknowledge that K-means++ accomplishes this goal comparatively better for the wine dataset.

Hierarchical Clustering

To see if we could further build better clusterings, we tried hierarchical clustering of the wine data with 15 clusters.

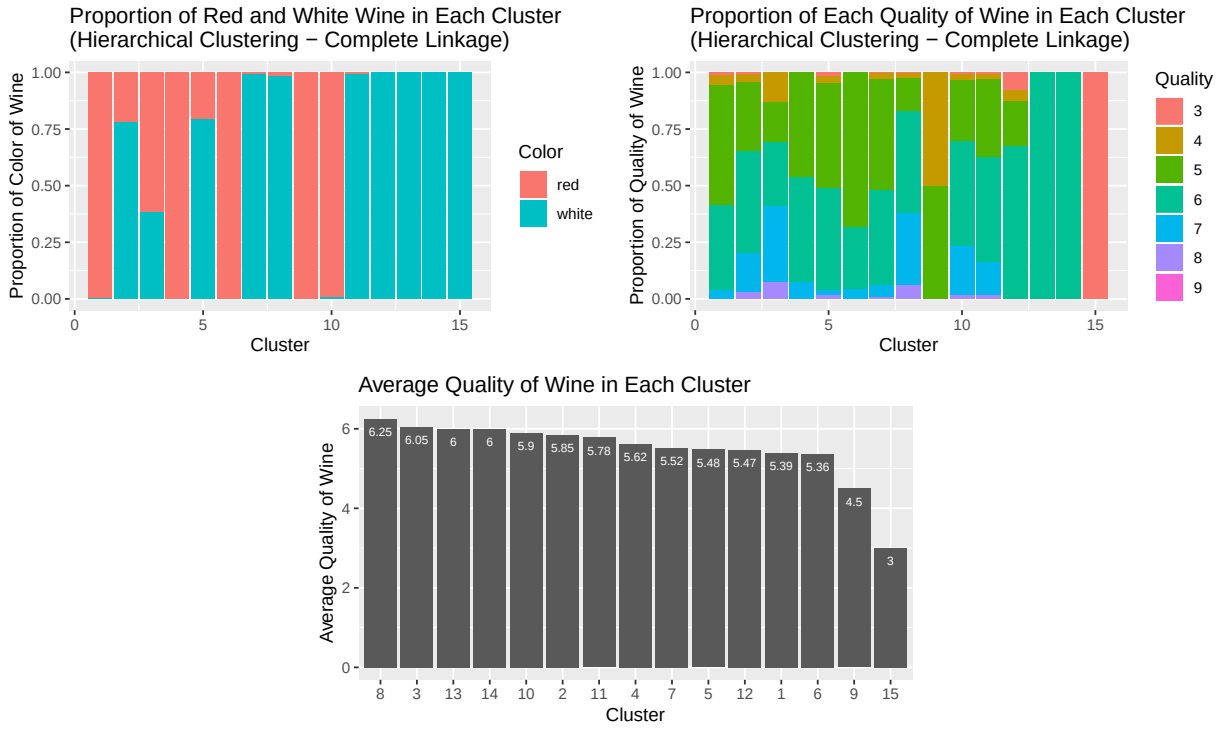
.....

Below is the number of wine bottles in each cluster for hierarchical clustering with complete linkage. Since the clusters are relatively well balanced, we determined that complete linkage was a reasonable method. Looking at the plot of the color of wine by cluster, we observed that clusters 1, 3, 4, 6, 9, and 10 are predominantly red wines while clusters 2, 5, 7, 8, 11, 12, 13, 14, and 15 are predominantly white wines. From the plot of quality of wine by cluster, we observed that clusters 9, 12, and 15 are lower-quality wines, clusters 1, 4, 5, 6, 7, 13, and 14 are mid-quality wines, and clusters 2, 3, 8, 10, and 11 are higher-quality wines. This shows that complete linkage is a useful and insightful method for clustering the wine data because we were able to find relatively well delineated groupings for both red and white wine as well as different qualities of wine.

Although, it is important to not that clusters 9, 13, 14, and 15 are not as useful as the other clusters because they have comparatively fewer data points.

Complete linkage

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
866	2344	39	13	128	22	962	1516	2	115	446	40	2	1	1



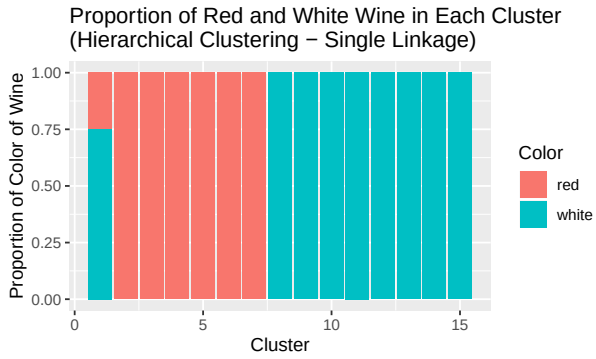
The 6 possible groupings for the wine from this method are shown below.

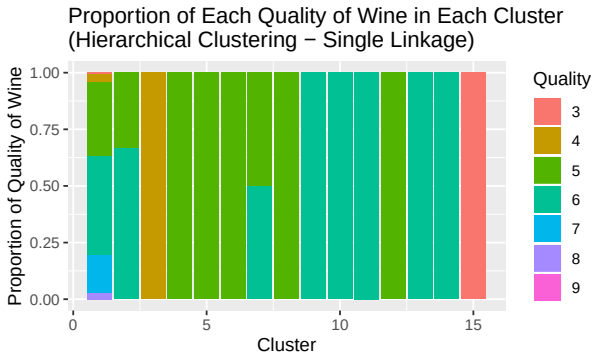
Grouping of Wine	Cluster
Low-quality red wine	9
Mid-quality red wine	1, 4, 6
High-quality red wine	3, 10
Low-quality white wine	12, 15
Mid-quality white wine	5, 7, 13, 14
High-quality white wine	2, 8, 11

Below are the number of wines in each cluster for hierarchical clustering with single, average, and centroid linkage. Since the clusters are very unbalanced (the number of wines in each cluster is very unevenly distributed), we determined that these types of linkage were not viable approaches. For reference, we have also included the plots of the color of wine by cluster and the quality of wine by cluster for each type of linkage. Since we have determined that these types of linkage are not feasible methods, these plots are not very useful for identifying clusters. We have included them in order to further emphasize the usefulness of the complete linkage clustering.

Single linkage

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6477	3	1	1	1	1	4	1	1	1	2	1	1	1	1



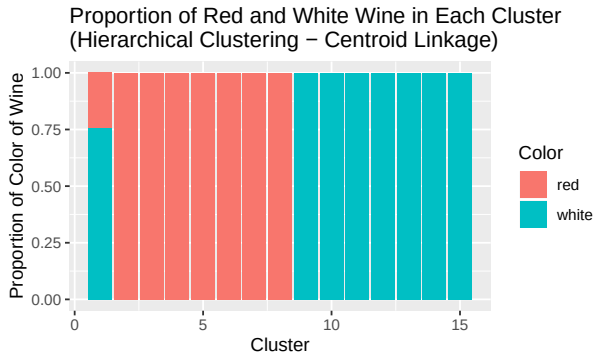


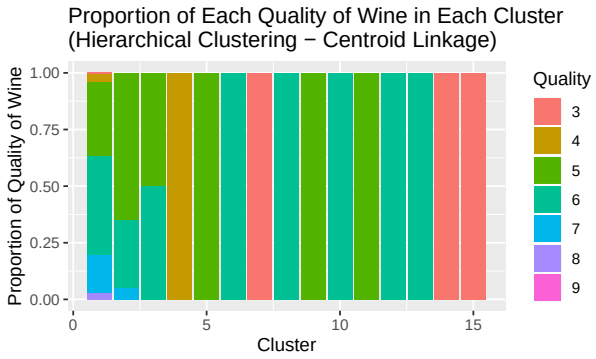
Average linkage

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6373	6	22	29	1	1	25	5	27	1	2	1	2	1	1

Centroid linkage

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6461	20	4	1	1	1	1	1	1	1	1	1	1	1	1





Conclusion

Which dimensionality reduction technique makes more sense for this data? Can your dimensionality reduction help sort the red wines from the white wines and the higher quality wines from the lower quality wines?

After conducting a principal components analysis, K-means clustering, K-means++ clustering, and hierarchical clustering, we have found that each method had advantages and drawbacks.

PCA performed remarkably well in using the principal components to group the wines by color because the clusterings were clear and obvious and the regression was very accurate. For quality, the principal components regression performed reasonably well because it was able to extract the general pattern of predicting the increasing wine qualities.

K-means and K-means++ clustering both also performed well in separating the red wines from the white wines. K-means++ performed slightly better than K-means in identifying clusters for quality because the clusters were slightly more distinct. However, a drawback to both of the methods was that the quality clusterings were not clear enough to make defined delineations between the different clusters.

Hierarchical clustering with complete linkage performed reasonably well in generating clusters with majorities of red and white wines. The quality groupings of complete linkage were also useful because some clusters were clearly higher-quality wines while others were clearly lower-quality wines. Using these clusters, we were able to identify each cluster as belonging to one of the 6 possible types of wines in terms of general quality and color.

We have determined that, for the wine dataset, using PCA to compress the original 11 chemical properties into fewer variables proved to be the most useful in grouping the wines because the principal components allowed us to produce reasonable and understandable groupings. Reducing the 11 properties to 3 principal components provided us the opportunity to gain insight into the dataset with fewer variables. As illustrated in the plots from the PCA above, the first three principal components and the regression were capable of usefully identifying the data by color and quality using regression.

Market segmentation

Overview

In this activity, we analyzed data from a market-research study using followers of the Twitter account of a large consumer brand NutrientH20. In order to allow NutrientH20 to hone its messaging a little more sharply,

we examined the dataset and highlighted any noteworthy market segments in NutrientH20's social-media audience.

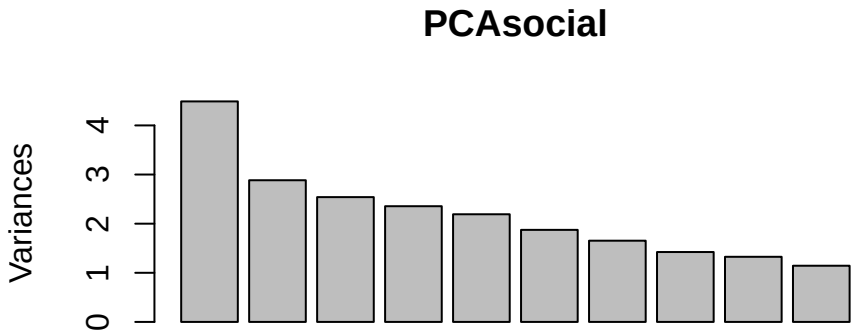
Data and Model

This social marketing dataset contains counts of Tweets from a random anonyomous sample of the Twitter account's followers over a seven-day period in June 2014. Each Tweet was categorized based on its content using a pre-specified scheme of 36 different categories, each representing a broad area of interest (e.g. politics, sports, family, etc.) and the number of Tweets per anonymous user per category was stored in the social marketing dataset. Our goal in this activity was to identify any notable market segments that we observed in NutrientH20's social-media audience using a variety of dimensionality reduction and clustering techniques.

PCA

A principal components analysis would be helpful for the dimensionality reduction of this data set because the 36 categories are not useful for interpretation as it is. Since the social marketing dataset contains only numerical, continous data for each category, performing a dimensionality reduction could provide us with better insight into the dataset as a whole.

We scaled the Tweet data and ran a principal components analysis and generated 36 principal components (because the number of principal components is bound by the 36 categories in the dataset). The variance plot below for this PCA is shown below.




```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577
Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592
Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164 0.45369 0.49961
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     1.19277 1.15127 1.06930 1.00566 0.96785 0.96131 0.94405
Proportion of Variance 0.03952 0.03682 0.03176 0.02809 0.02602 0.02567 0.02476
Cumulative Proportion  0.53913 0.57595 0.60771 0.63580 0.66182 0.68749 0.71225
                          PC15    PC16   PC17    PC18    PC19    PC20    PC21
```

```
Standard deviation     0.93297 0.91698 0.9020 0.85869 0.83466 0.80544 0.75311
Proportion of Variance 0.02418 0.02336 0.0226 0.02048 0.01935 0.01802 0.01575
Cumulative Proportion  0.73643 0.75979 0.7824 0.80287 0.82222 0.84024 0.85599
                          PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.69632 0.68558 0.65317 0.64881 0.63756 0.63626 0.61513
Proportion of Variance 0.01347 0.01306 0.01185 0.01169 0.01129 0.01125 0.01051
Cumulative Proportion  0.86946 0.88252 0.89437 0.90606 0.91735 0.92860 0.93911
                          PC29    PC30    PC31   PC32    PC33    PC34    PC35
Standard deviation     0.60167 0.59424 0.58683 0.5498 0.48442 0.47576 0.43757
Proportion of Variance 0.01006 0.00981 0.00957 0.0084 0.00652 0.00629 0.00532
Cumulative Proportion  0.94917 0.95898 0.96854 0.9769 0.98346 0.98974 0.99506
                          PC36
Standard deviation     0.42165
Proportion of Variance 0.00494
Cumulative Proportion  1.00000
```

Since the variance plot shows a clear and obvious decrease in the variance explained for each additional principal component, we can continue with the rest of the PCA. This is the baseline sanity check that each additional principal component attempts to explain the remaining unexplained data after taking into consideration the previous principal components.

For this scenario, since there are a relatively large number of categories, we used the first five principal components which explain about 40% of the variance of the 36 categories in the dataset. We continued our analysis using these five components. Below are the coefficients for each of the 36 properties in each of the first 5 principal components.

```
                   PC1   PC2   PC3   PC4   PC5
chatter          -0.13  0.20 -0.07  0.11 -0.19
current_events   -0.10  0.06 -0.05  0.03 -0.06
travel           -0.12  0.04 -0.42 -0.15 -0.01
photo_sharing    -0.18  0.30  0.01  0.15 -0.23
uncategorized    -0.09  0.15  0.03  0.02  0.06
tv_film          -0.10  0.08 -0.09  0.09  0.21
sports_fandom    -0.29 -0.32  0.05  0.06 -0.03
politics         -0.13  0.01 -0.49 -0.20 -0.06
food             -0.30 -0.24  0.11 -0.07  0.07
family           -0.24 -0.20  0.05  0.07 -0.01
home_and_garden  -0.12  0.05 -0.02 -0.01  0.04
music            -0.12  0.14  0.01  0.08  0.07
news             -0.13 -0.04 -0.34 -0.18 -0.03
online_gaming    -0.07  0.08 -0.06  0.22  0.48
shopping         -0.13  0.21 -0.05  0.10 -0.20
health_nutrition -0.12  0.15  0.23 -0.46  0.17
college_uni      -0.09  0.12 -0.09  0.26  0.49
sports_playing   -0.13  0.11 -0.04  0.18  0.37
cooking          -0.19  0.31  0.19  0.01 -0.12
eco              -0.15  0.09  0.03 -0.12  0.02
computers        -0.14  0.04 -0.37 -0.14 -0.06
business         -0.14  0.10 -0.11  0.01 -0.05
outdoors         -0.14  0.11  0.14 -0.41  0.15
crafts           -0.19 -0.02  0.00  0.02  0.04
automotive       -0.13 -0.03 -0.19 -0.04 -0.06
art              -0.10  0.06 -0.05  0.06  0.16
religion         -0.30 -0.32  0.09  0.07 -0.02
```

```
beauty           -0.20  0.21  0.15  0.15 -0.19
parenting        -0.29 -0.30  0.09  0.05 -0.04
dating           -0.11  0.07 -0.03 -0.03 -0.01
school           -0.28 -0.20  0.08  0.09 -0.09
personal_fitness -0.14  0.14  0.22 -0.44  0.16
fashion          -0.18  0.28  0.14  0.14 -0.17
small_business   -0.12  0.09 -0.10  0.08  0.03
spam             -0.01  0.00 -0.01 -0.02  0.02
adult            -0.03 -0.01  0.00 -0.02  0.01
```

From these coefficients, we understood that all categories are weighed negatively in PC1 which can be attributed to the large dataset and the variety of data points sampled. Of the coefficients for PC1, the most strongly negative categories included parenting, school, family, and religion. This market segment is relatively broad, but is likely to include millenials and young adults without children/families (no interest in parenting or school or family or religion).

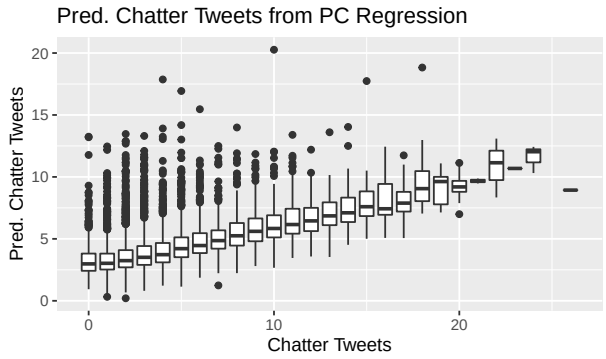
For PC2, we noticed an interesting association of the strongly positively weighed coefficents because they included the photo sharing, shopping, cooking, beauty, and fashion categories. These categories share the common underlying theme of activites typically attributed to females and models in our society. This is an noteworthy finding, and is an important starting point for understanding one potential market segment of NutrientH20's social-media audience. We also noticed that the sports fandom, family, and religion categories negatively weighed the PC2 score. This is logical because this market segment seems to include young females and models, who are not likely to post or be interested in sports, families, or religion.

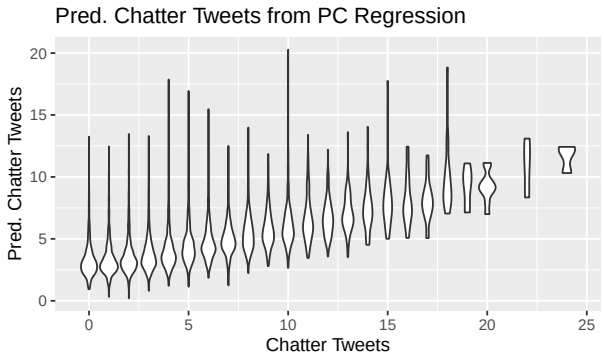
For PC3, we noticed the strongly positively weighed categories were health and nutrition and personal fitness. These categories share the common underlying theme of activities typically attributed to people interested in exercising and working out and taking care of their physical health. This is another potential market segment that we could understand further with more analyses. The negatively weighed categories for PC3 are mainly politics, news, and computers. This is reasonable because these categories represent different interests than mainly exercising and physical health.

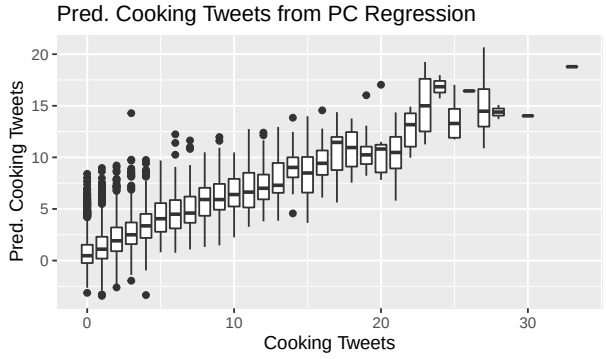
For PC4, we noticed the strongly positively weighed categories were online gaming and college/university. These categories share the common underlying theme of interests shared by teenage and young adult college students who play video games and are not interested in outdoor activites. This is yet another potential market segment that warrants further analysis. The negatively weighed categories for PC4 are primarily travel, health and nutrition, outdoors, and personal fitness. This follows from our initial understanding because these categories are very unlikely to be of interest to the gaming, indoorsy types.

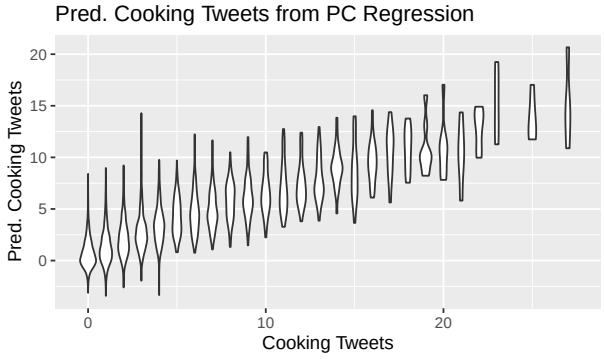
For PC5, we noticed the strongly positively weighed categories were TV/film, online gaming, college/university, sports playing and personal fitness. These categories seem to share some common ground with the market segment from PC4, but this segment seems to instead include college students who are interested in playing sports and video games (perhaps sports related video games such as NBA 2k or Madden). This is one more market segment that is worth further exploring. The negatively weighed categories are photo sharing, beauty, and fashion. This follows from our understanding because college students interested in sports are not likely to have interests in beauty or fashion.

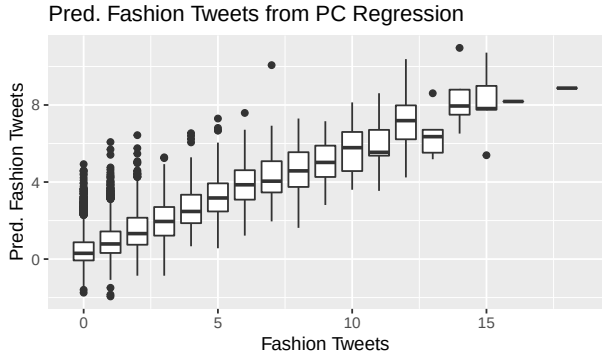
To further verify the validity of our PCA, we ran a principal component regression to predict a few key categories. Below are the plots of predicted vs. actual distributions for a few categories from the social marketing dataset. We have included both boxplots and violin plots in order to provide a holistic visualization of the distribtions.

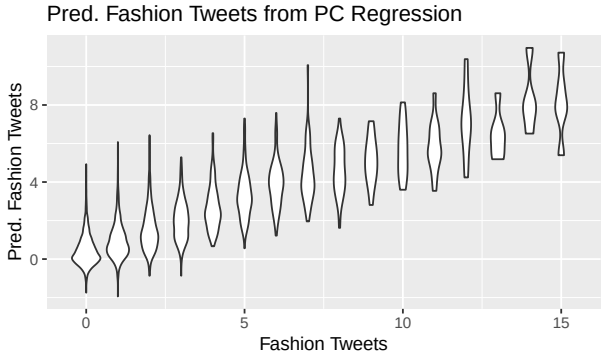

Pred. Chatter Tweets from PC Regression
Pred. Chatter Tweets
Chatter Tweets
0
5
10
15
20

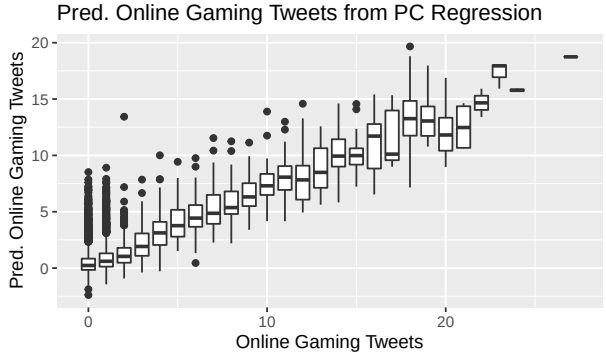

Pred. Chatter Tweets from PC Regression
Pred. Chatter Tweets
Chatter Tweets
0
5
10
15
20
25

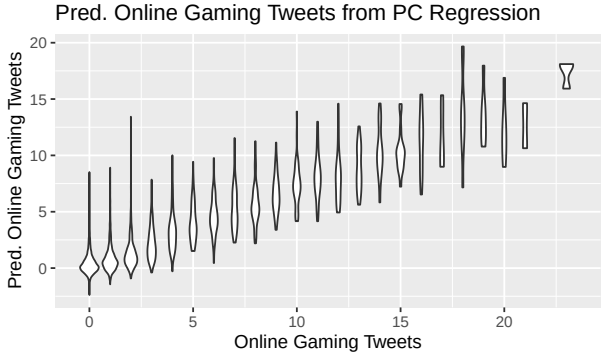

Pred. Cooking Tweets from PC Regression
Pred. Cooking Tweets
Cooking Tweets
0
5
10
15
20
30

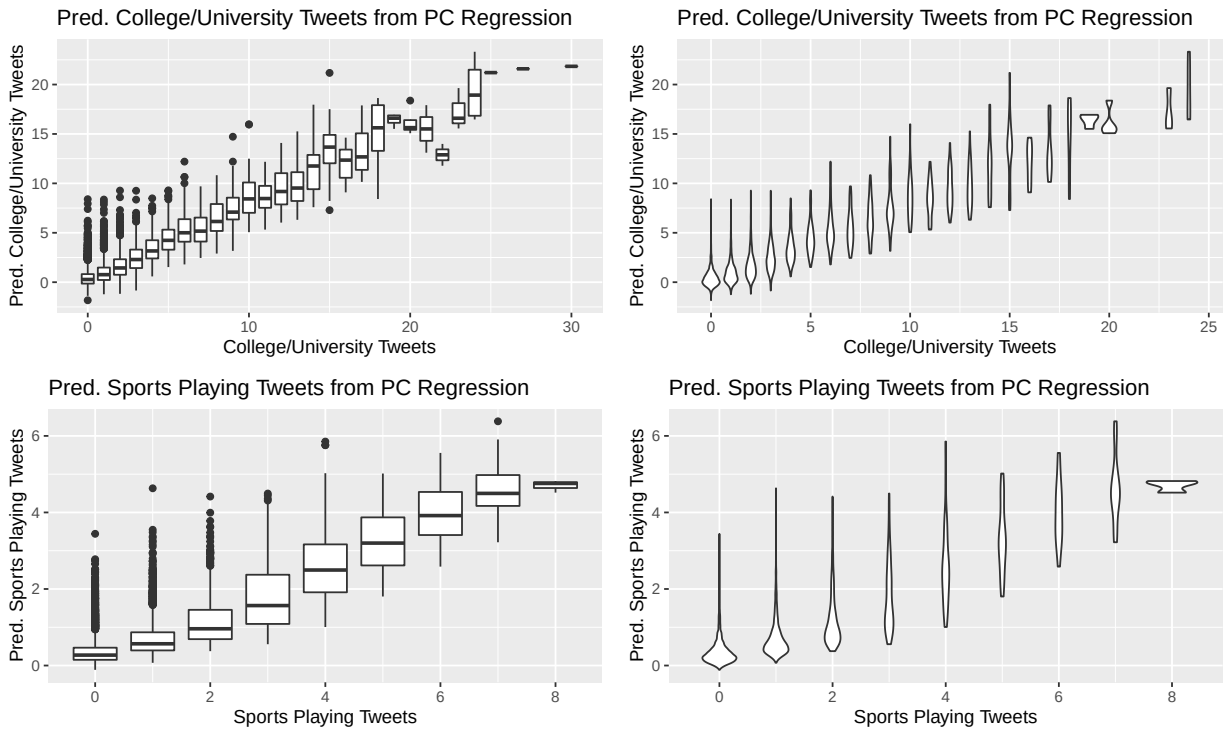

Pred. Cooking Tweets from PC Regression
Pred. Cooking Tweets
Cooking Tweets
0
5
10
15
20


Pred. Fashion Tweets from PC Regression
Pred. Fashion Tweets
Fashion Tweets
0
4
8
5
10
15


Pred. Fashion Tweets from PC Regression
Pred. Fashion Tweets
Fashion Tweets
0
4
8
5
10
15


Pred. Online Gaming Tweets from PC Regression
Pred. Online Gaming Tweets
Online Gaming Tweets
0
5
10
15
20


Pred. Online Gaming Tweets from PC Regression
Pred. Online Gaming Tweets
Online Gaming Tweets
0
5
10
15
20



Since each of these plots is close to the identity function ($y = x$), our PCA components are a reliable compression of the original 36 categories in the dataset.

K-Means Clustering

In addition to running the PCA on the social marketing dataset, we explored the utility of K-means and K-means++ clustering. In order to determine whether the market segments we discovered from the PCA were rational, we attempted to find any outstanding cluster that associated with one of the 5 principal components.

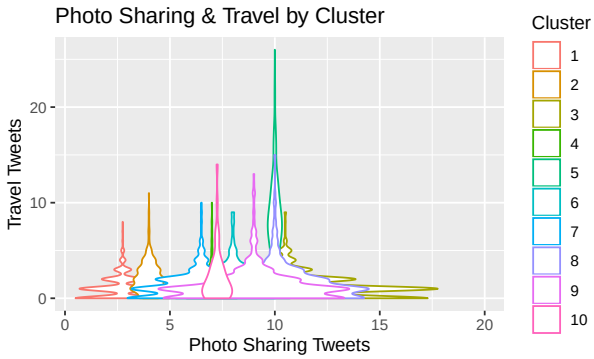
The sizes of the clusters in each clustering method are shown below.

```
Size of clusters with K-means clustering: 412 350 349 1049 675 475 766 3328 429 49
```

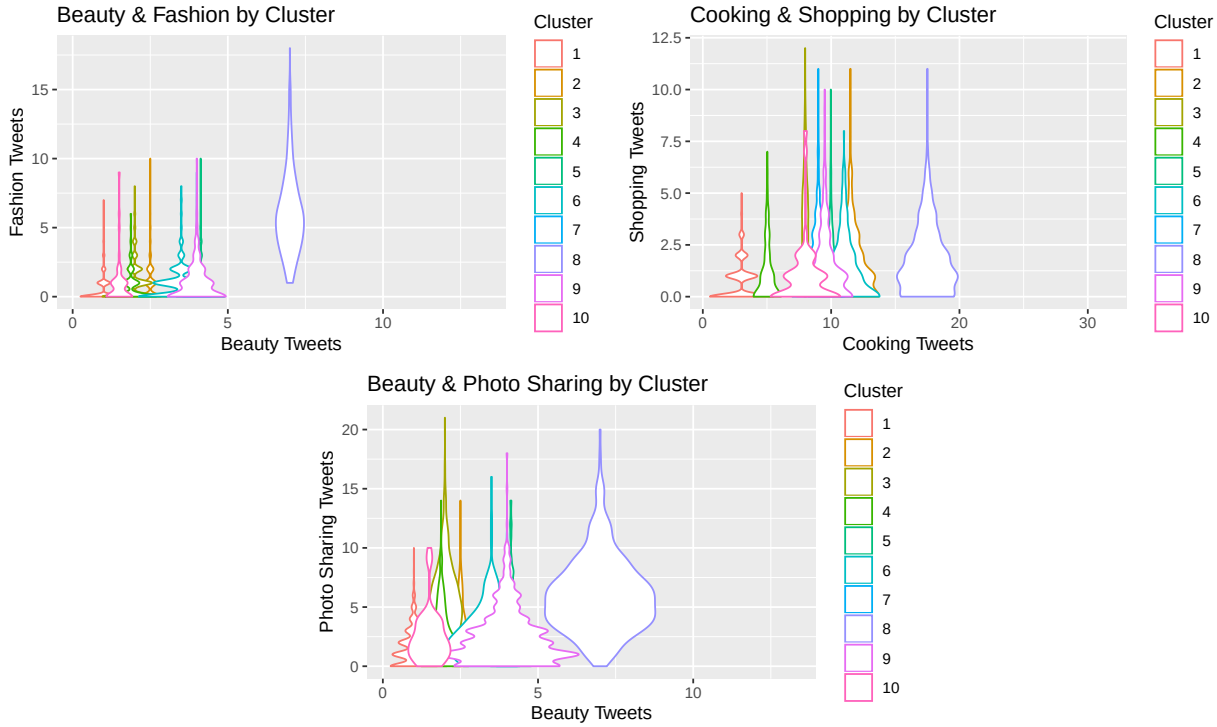
```
Size of clusters with K-means++ clustering: 3311 412 1065 429 349 350 768 475 674 49
```

From this, we decided to move forward in our analysis with the K-means++ clustering because it is an optimization of the K-means clustering technique by using better initial centroids. Also, K-means++ provided slightly more even cluster sizes.

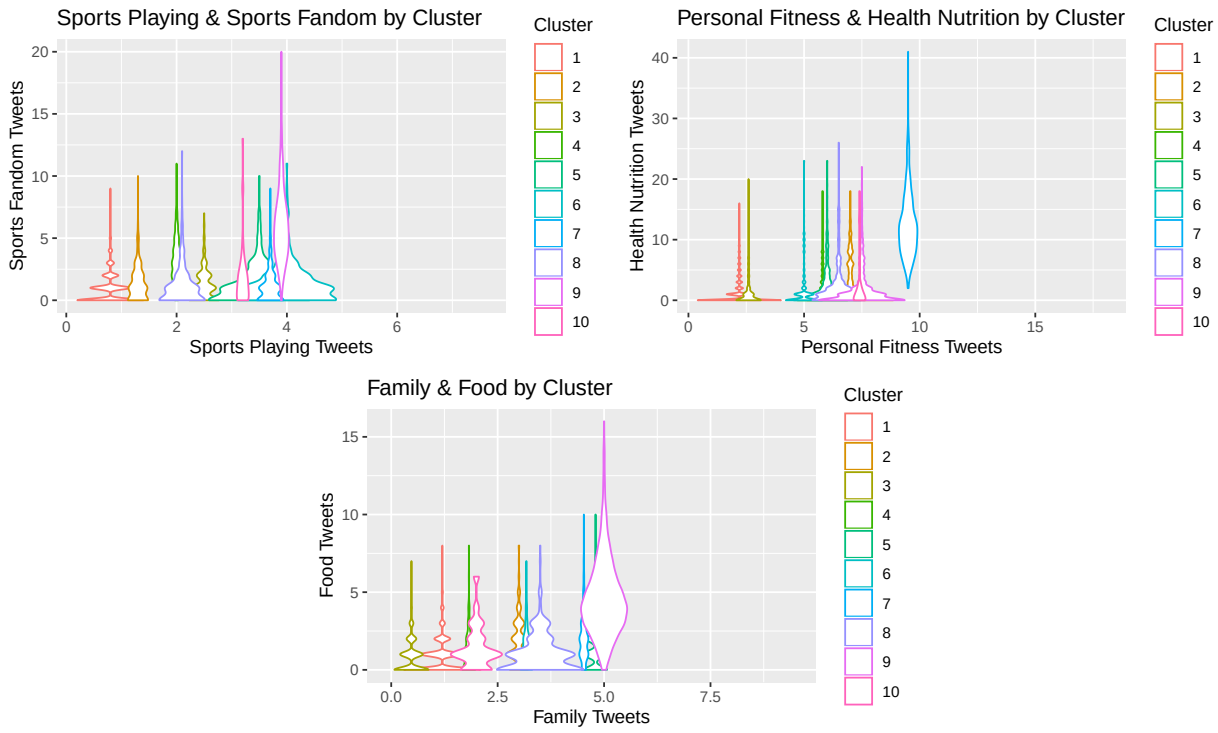
In order to determine whether any clusters stood out, we plotted pairs of categories from the dataset as violin plots to emphasize the distribution.



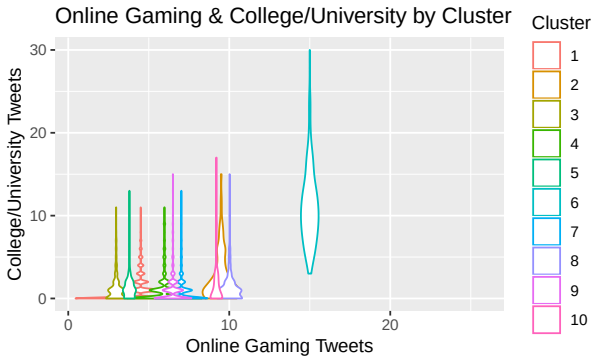
From the plot of photo sharing by travel, we noticed that cluster 5 (green) had a noticeably high volume of travel Tweets. This could represent a potential market segment that is interested in traveling and sharing photos of their travels.



From the plots of beauty, fashion, cooking, and shopping, we noticed that cluster 8 (purple) stood out in each plot. This further bolsters our understanding of the market segment discovered in PC2 above. The underlying theme of females and models who post a high volume of photos begins to strengthen from this clustering.



From the plots of sports and family and food, we noted that clusters 7 (blue) and 9 (pink) were distinct. High volumes of Tweets in these categories indicates that this cluster is associated with the market segment discovered in PC3, which consisted of people interested in exercising, working out, and physical health.



From the plot of online gaming and college/university, we observed that cluster 6 (teal) was very unique. High volumes of Tweets in these categories follow closely with the market segment discovered in PC4 and PC5, which included college students interested in different types of video games.

Hierarchical Clustering

To further flush out our understanding of these market segments we ran hierarchical clustering on the social marketing dataset with 15 clusters.

We explored each type of linkage for hierarchical clustering: single, complete, average, and centroid. Below are the cluster sizes for each type of linkage.

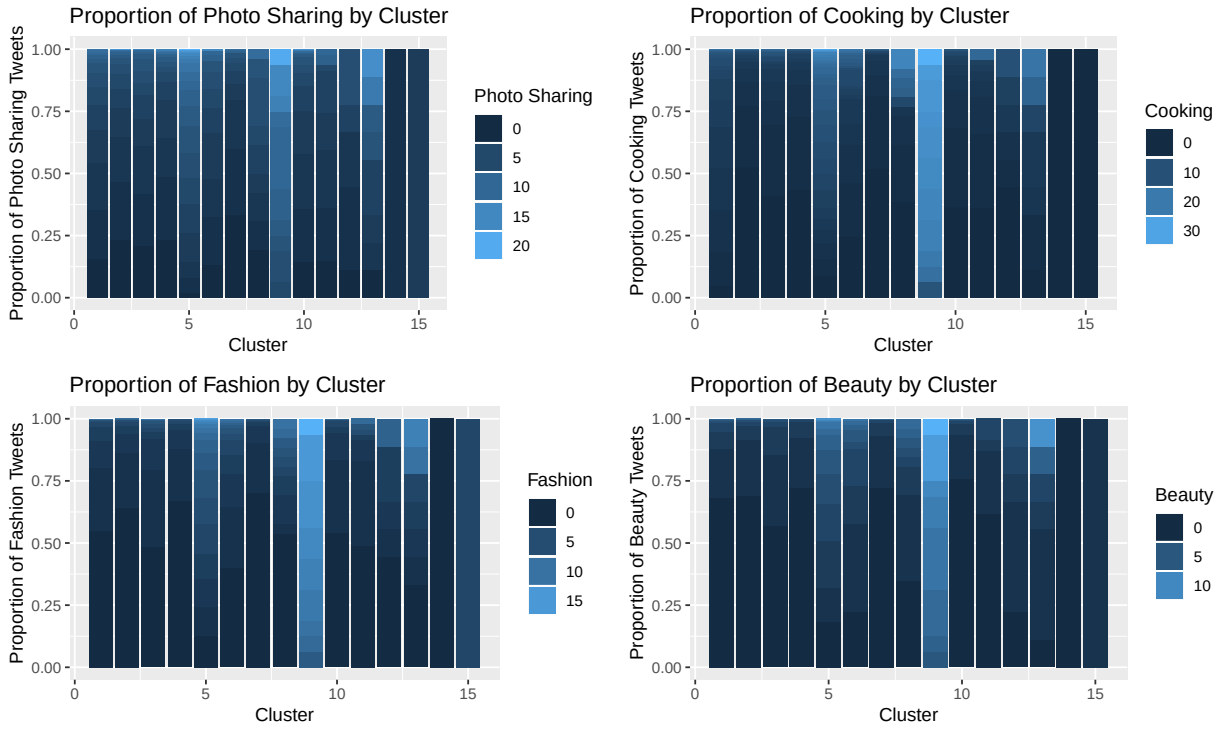
```
Singe linkage cluster sizes: 7823 46 1 1 1 1 1 1 1 1 1 1 1 1

Complete linkage cluster sizes: 487 4868 284 408 734 130 451 26 16 410 47 9 9 2 1

Average linkage cluster sizes: 7796 8 46 6 8 2 2 2 5 1 2 1 1 1 1

Centroid linkage cluster sizes: 7866 1 1 1 1 2 1 1 1 1 1 1 2 1 1
```

From this, we determined that complete linkage provided us the most balanced sizes for the 15 clusters.

Complete linakge

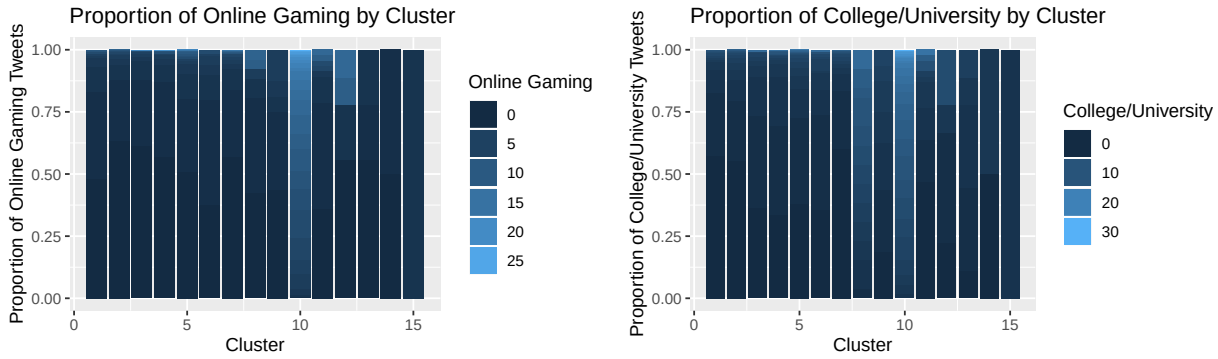
Using hierarchical clustering with complete linkage, we plotted the proportion of the number of Tweets in each cluster for a variety of useful categories from the social marketing dataset. We intended to observe whether any of the clusters in this method provided us further insight to verify the market segments that we have found in the above methods.

The following plots show the proportion of the number of Tweets in each cluster for the photo sharing, cooking, fashion, and beauty categories.



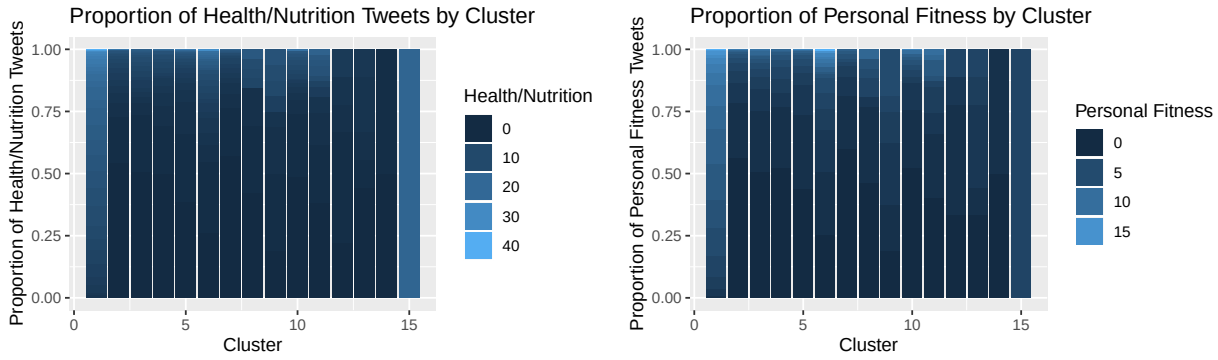
One commonality that we found was in cluster 9. These plots show a noticeably high proportion of high numbers of photo sharing, cooking, fashion, and beauty Tweets in cluster 9 (the cluster whose bar is a significantly lighter shade). This follows our understanding of the distinct grouping associated with PC2 because we had established that PC2 had a strong association with young females and models interested in beauty, fashion, and posting photos.

The following plots show the proportion of the number of Tweets in each cluster for the online gaming and college/university categories.



Another commonality that we found was in cluster 10. These plots show a noticeably high proportion of high numbers of online gaming and college/university Tweets in cluster 10 (the cluster whose bar is a significantly lighter shade). This follows our understanding of the distinct groupings associated with PC4 and PC5 because we had established that these components had a strong association with college and playing video games. Since this grouping involves activites associated with teenagers and young adults, we have solidified our insight into this market segment.

The following plots show the proportion of the number of Tweets in each cluster for the health and nutrition and personal fitness categories.



A third commonality that we found was in cluster 1. These plots show a noticeably high proportion of high numbers of health and nutrition and personal fitness Tweets in cluster 1 (the cluster whose bar is a significantly lighter shade). This follows our understanding of the distinct groupings associated with PC3 because we had established that PC3 had a strong association with exercising, fitness, and physical health. Since this grouping involves activities that require exercise and physical health, we solidified our insight into this grouping.

Conclusion

After conducting a PCA, K-means and K-means++ clustering, and hierarchical clustering with complete linakge, we have determined that there exist a few unique market segments in NutrientH20's social-media audience.

One market segment is young females and models who are interested in beauty, cooking, fashion, and shopping. These individuals are not particularly interested in sports, religion, or school. The best way to appeal to this group would be to include advertisements and promotions that express beauty and fashion.

Another market segment is individuals who are interested in personal fitness and health and nutrition. These individuals are not particularly interested in politics, news, or computers. The best way to appeal to this group would be to include advertisements and promotions that involve exercising and personal/physical health through working out and proper nutrition.

Another market segment is individuals who are college students (teenagers and young adults) who are interested in video games. The best way to appeal to this group would be to include advertisements and promotions that involve cheap deals (because college students run on a tight budget) and possibly games.

Using these dimensionality reduction and clustering techniques, we were able to reduce the original 36 categories of Tweets into a more manageable set of variables that allowed us to extract distinctive market segments in NutrientH20's social-media audience.