SDS Exercise 3

by Cheng Peng, Zhiyuan Wei, Erich Schwartz

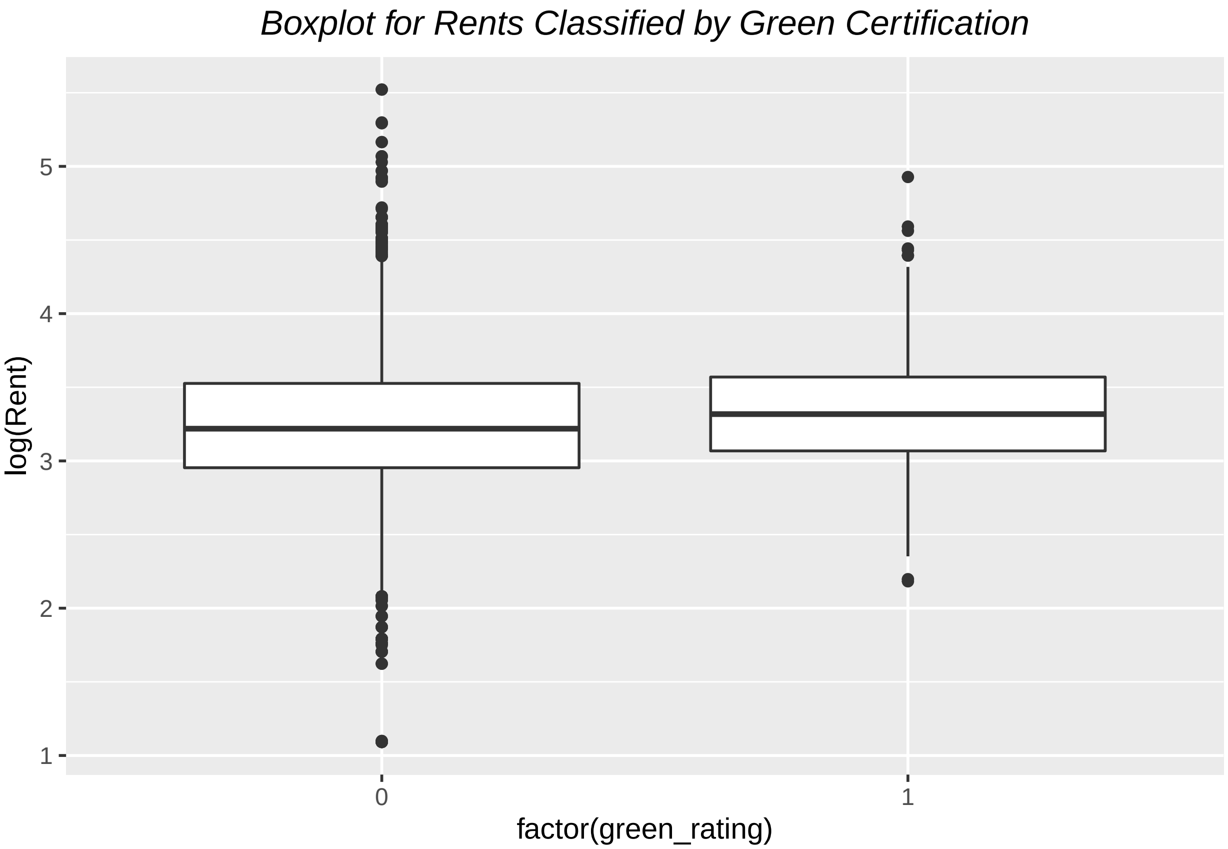
1. Predictive Model Buidling

1.1 Introduction

The greenbuildings.csv data contains information on 7894 commercial properties across the United States. All of 7894 observations carry 21 distinct features, one of which indicates whether the property is designated as a green building. Our goal is to build the best predictive model for rent and quantify the average change in rental income per square foot associated with green certification. To achieve these objectives, we will apply the knowledge of model selection and regularization in our following analysis.

1.2 Data Cleaning

In this analysis, we focus on the variable "green_rating," which indicates a building either has LEED or EnergyStar certification. To build a useful predictive model, we first clean the data by removing all N/As from the dataset and delete the "LEED" and "EnergyStar" columns. We also remove the variable CS_propertyID as the buildings' identifiers can't be predictive of their rents. The distribution of rents classified by green certification is as follows:



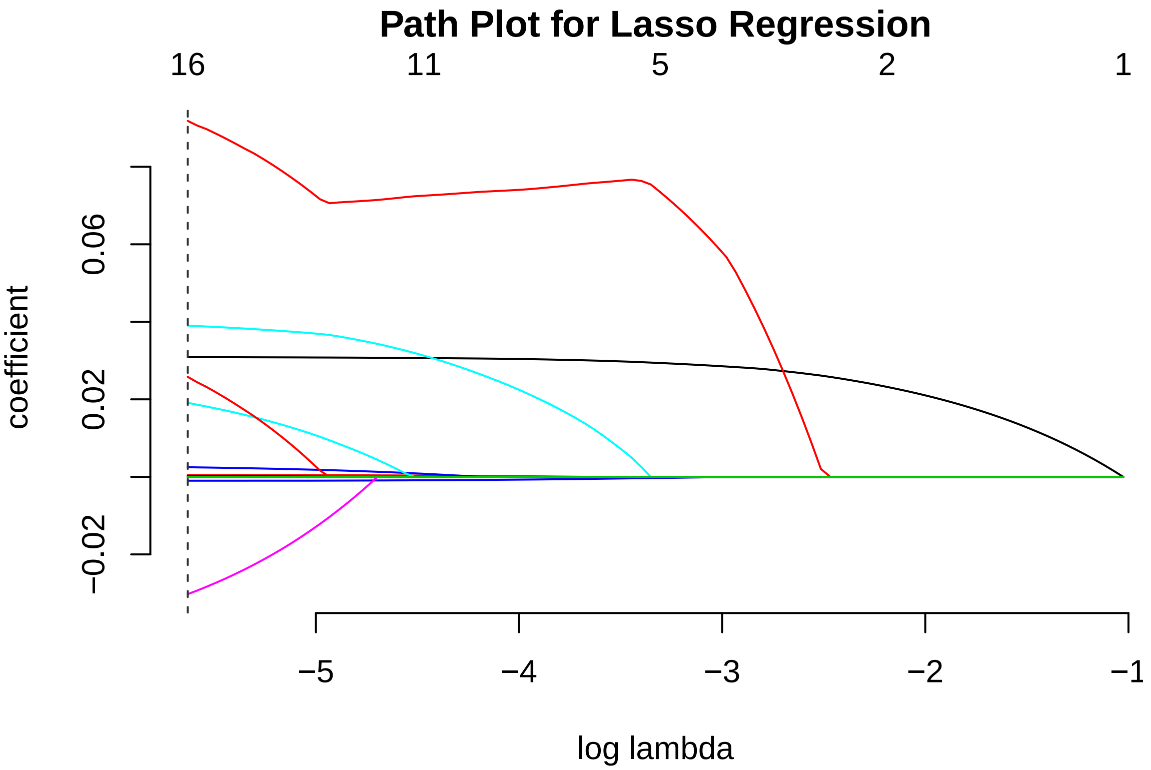
1.3 Lasso Regression

Next, we employ Lasso Regression in our analysis. We first generate a sparse matrix with log(Rent) and all other variables of green buildings. Next, we use "gamlr" function and attempt to find the best lambda.

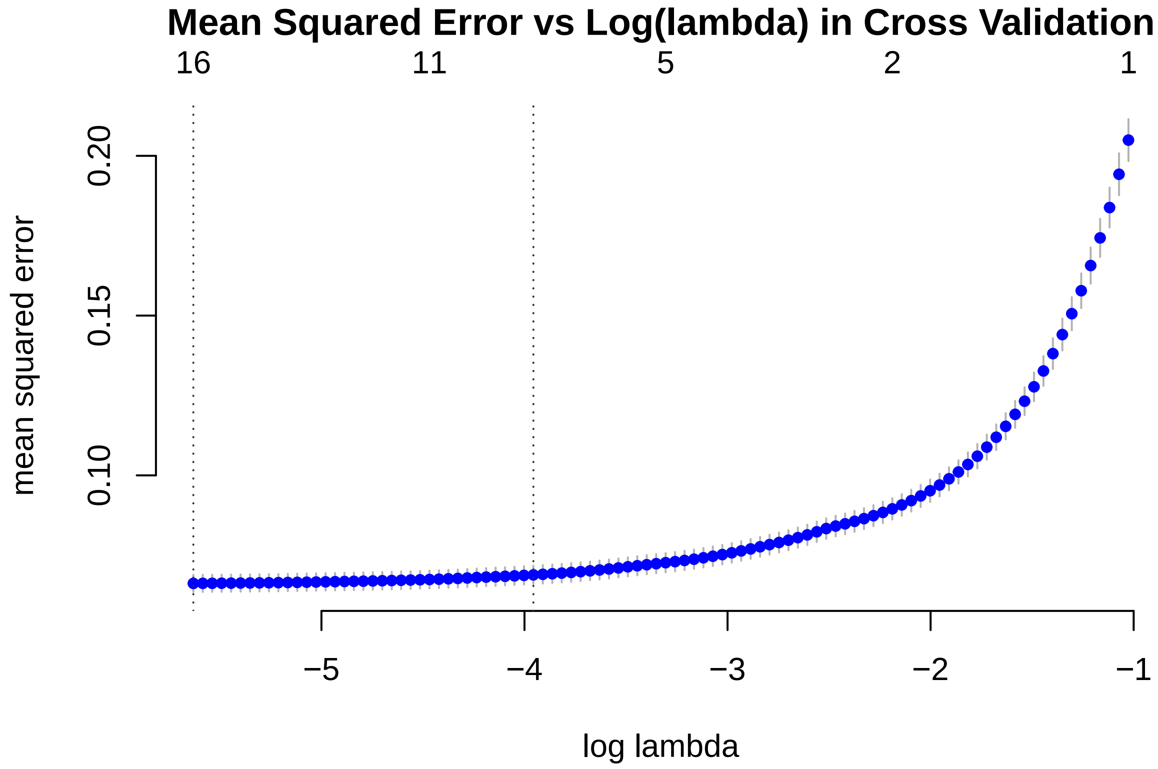
We return a path plot for Lasso Regression, featuring the change in regression coefficients for each variable as log(lambda) increases. We then use Cross Validation to obtain the best-performing lambda.

We generate the plot showing how the mean squared error varies as log(lambda) increases. It shows that the mean square error strictly increases as lambda increases. Finally, we select the lambda that returns the smallest out of sample error.

The result shows that the best log(lambda) is -5.630747. 16 variables are kept in this model. Given that 16 variables are still reasonable to interpret, we choose this model as our best model instead of using the "1-standard-error" model, which is less accurate in comparison. The exact coefficients of our best model are as follows:



```
## fold 1,2,3,4,5,6,7,8,9,10,done.
```



```
## [1] -5.630747
```

1.4 Conclusions

By using regularization, more specifically, Lasso regression, we find the best model with log(lambda) at -5.630747 and with 16 variables. Renovated (whether the building has undergone substantial renovations), hd_total07 (number of heating degree days), Gas_Costs, and Electricity_Costs are the four variables eliminated from the model. Namely, the four variables excluded are irrelevant with rents. class_a, class_b, green_rating, net, amenities, and cluster_rent are among the more robust drivers of rental income. The final regression coefficients also show that when a building is awarded the green certification, its rental income per square foot increases by 1.91%.

2. What causes what?

1. Cities with higher crime rates tend to hire more police. This inherent correlation between the crime rates and the number of cops may return a regression result that says more policing causes higher crime rates.

2. The researchers looked for examples where the number of cops is not correlated with the crime rates. They used the terrorism alert system in Washington, D.C. When D.C. is on terrorism alert, more police officers will be deployed to the streets. Such deployments are unrelated to the crime rates of D.C. So the researchers could examine whether the crime rates decrease due to the increased policing. The researchers discovered that on high alert days, with more cops on the streets, crimes in D.C. decrease by 7.316 cases, and the result is at 5% significance level. This proves that an increased number of police officers to result in lower crime rates.

3. Researchers were worried that the lower crime rates were the result of fewer tourists on high alert days. So they tried to use midday ridership to measure the number of tourists visiting D.C. and see whether increased policing contributes to lower crime rates when the number of tourists is the same.

4. The model estimates when D.C. is on high alerts, how the crime rates in Police District 1, and all other districts in D.C. change. Police District 1 is where the White House, Congress, and Smithsonian Institution locate, which will have most of the increased police attention when D.C. is on terrorism alert. The result shows that on high alert days, District 1 will experience a 2.621-case reduction in crimes while in other Districts, the reduction in crimes is not statistically significant. This proves that an increased number of police officers do reduce crime rates.

3. Clustering and PCA

3.1 Introduction

The wine.csv data contains information on 11 chemical properties of 6497 different bottles of vinho verde wine from northern Portugal. All of the 6497 obsevations carry 13 distinct features including 11 chemical properties; Two other variables that are the quality rate of the wine and the color of the wine. Our goal is to run both clustering and PCA algorithms and determine which dimensionality reduction technique is easily capable of predicting the color of the wine.

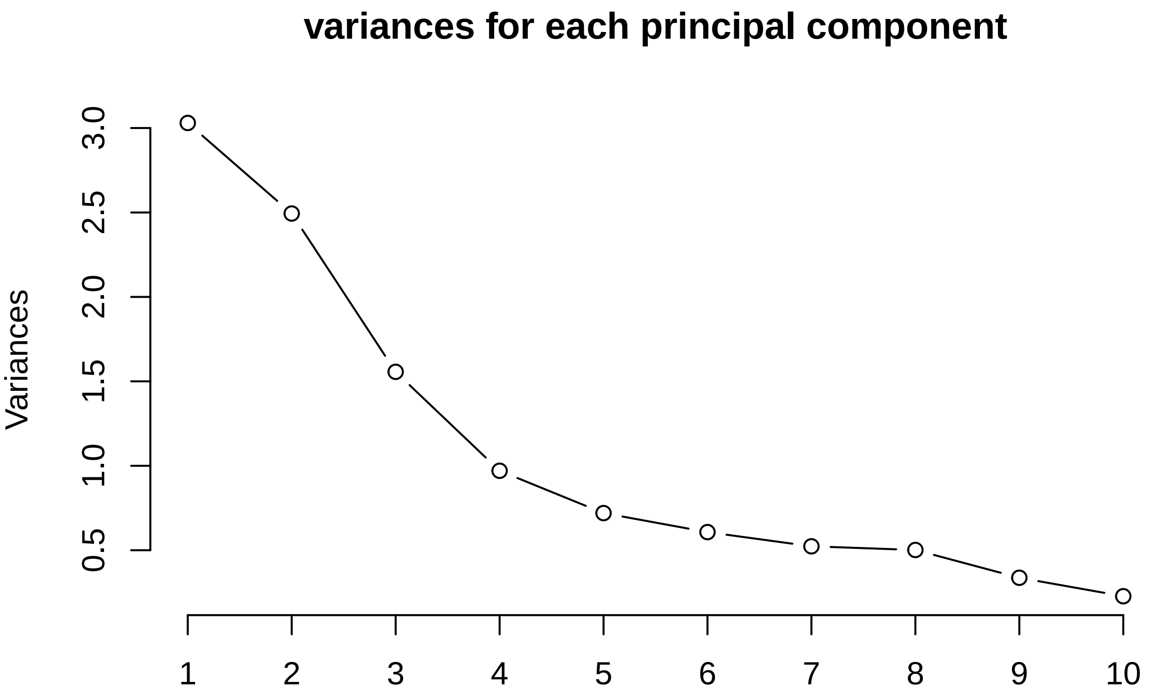
we start by creating a new dataset from wine.csv except we remove the “quality” and “color” columns. Our new dataset new_wine contains information on 11 chemical properties.

3.2 Color of the wine

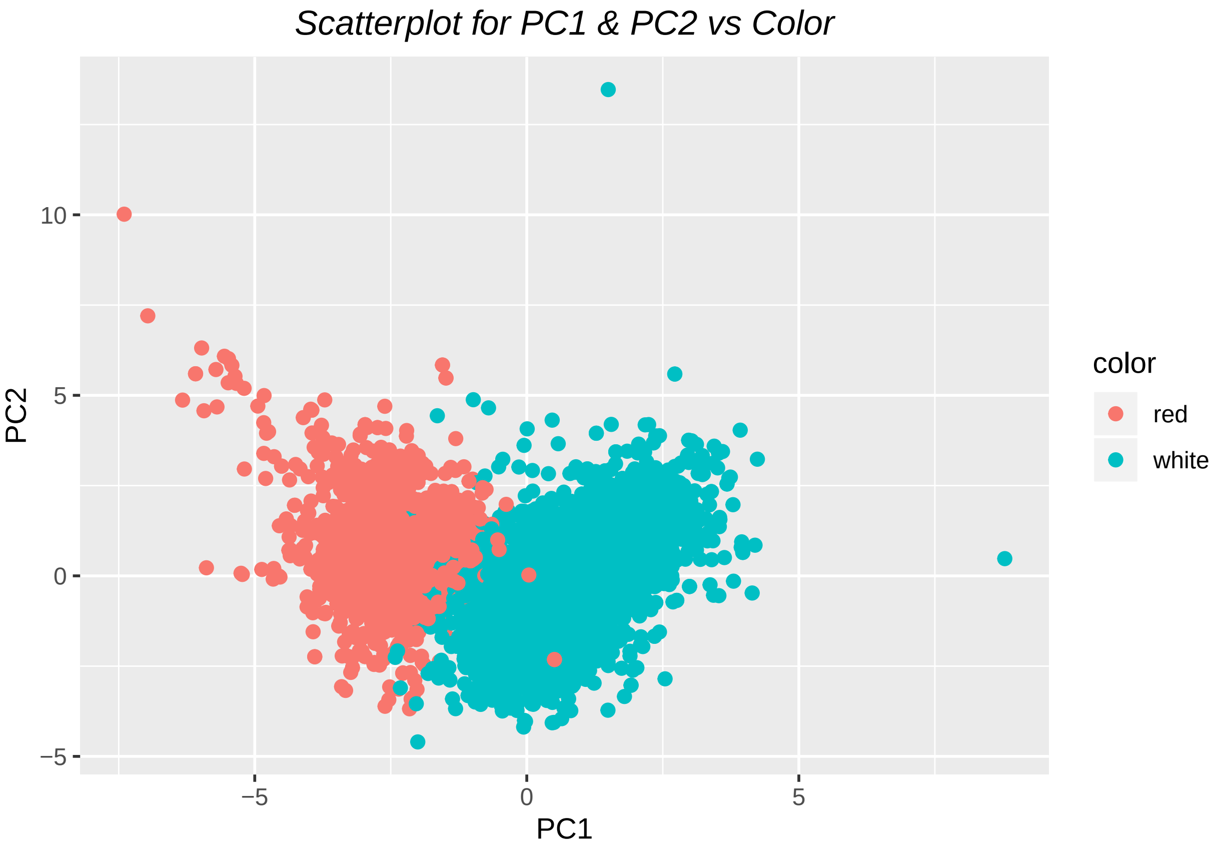
3.2.1 PCA Approach

Our first approach is to use PCA. We first construct a PCA model for the dataset. From the summaries, we can derive components that incorporate most information about the dataset. More specifically, PC1, PC2, and PC3 obtained from the dimensionality reduction algorithm can give us information about the dataset with only three variables. Next, we examine the direction of the top three principal components, namely how PCx relates to the 11 chemical properties in detail.

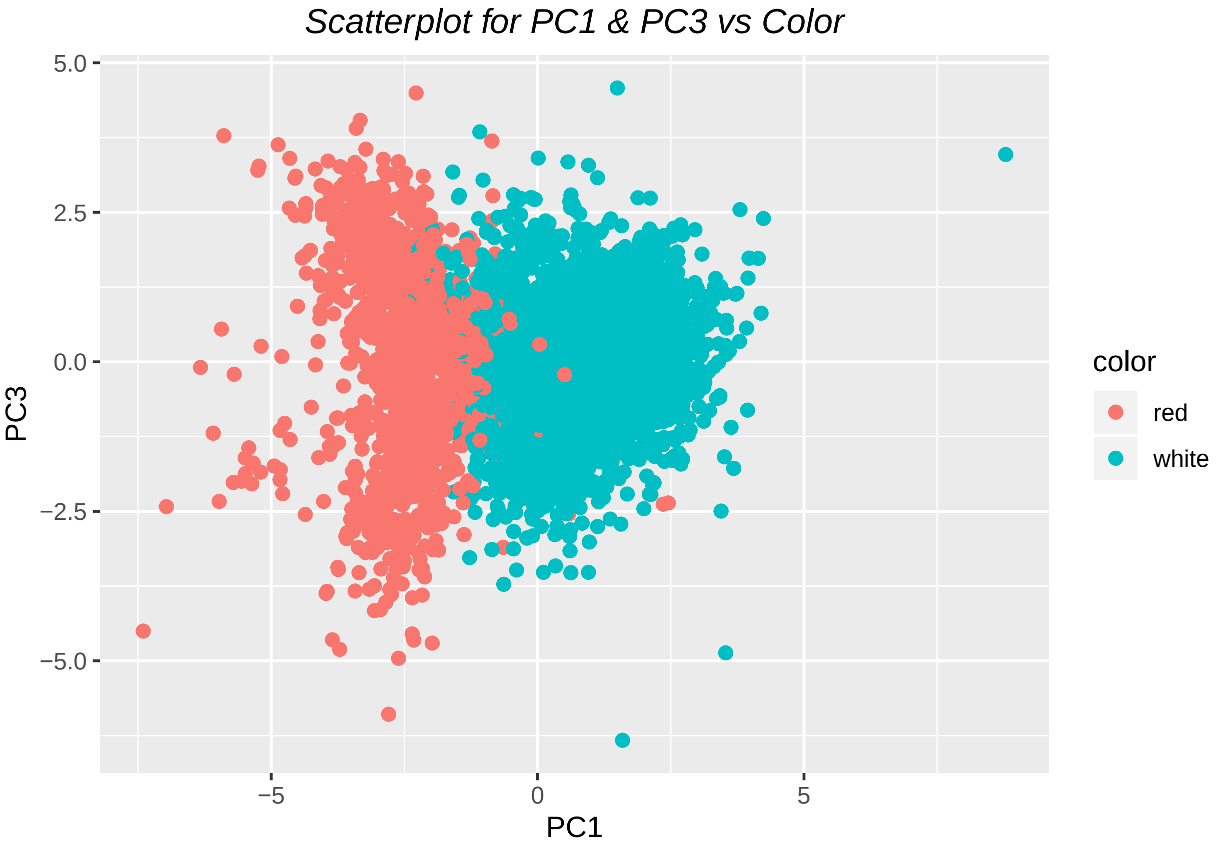
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 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
```



From the summaries and plots above, we know that PC1, PC2, and PC3 collectively contain 64.36% of the information in the dataset. Thus, we take PC1, PC2, and PC3 as variables of interest for further analysis.



Scatterplot for PC1 & PC2 vs Color



Scatterplot for PC1 & PC3 vs Color

We plot the graphs showing the relationship between PC1, PC2 & PC3, and the wines' color. The plots above demonstrate that PC1 appears to be a good indicator of the color of the wine. However, to get this information, we use the “supervised” information within the dataset, namely, the variable “color.” Without variable “color,” we cannot obtain the fact that PC1 relates to the color of wines (higher PC1 indicates white wine), nor can we know the cutoffs for white/red wine in PC1. Though experienced chemists might infer from the PC1 compositions (which is presented below) that PC1 is related to the color of the wine, this complicates the analysis and fails to simplify the issue for data analyst without chemistry backgrounds. Thus, PCA is not the best dimensionality reduction approach to be employed here.

```
##        fixed.acidity     volatile.acidity          citric.acid
##                -0.24                -0.38                 0.15
##       residual.sugar            chlorides  free.sulfur.dioxide
##                 0.35                -0.29                 0.43
## total.sulfur.dioxide              density                   pH
##                 0.49                -0.04                -0.22
##            sulphates              alcohol
##                -0.29                -0.11
```

(In PC1, citric acid, residual sugar, free sulfur dioxide, and total sulfur dioxide have positive projections while sulphates, fixed and volatile acidity have high negative projections; Experienced chemists may be able to infer from the coefficients above that PC1 is strongly correlated with colors of wines.)

3.2.2 Clustering Approach

Now, we attempt the clustering approach. We start by scaling the data, and pick k=2 since the wine we are studying has only two colors: red and white. We return the average chemical composition in both clusters. Details are as follows:

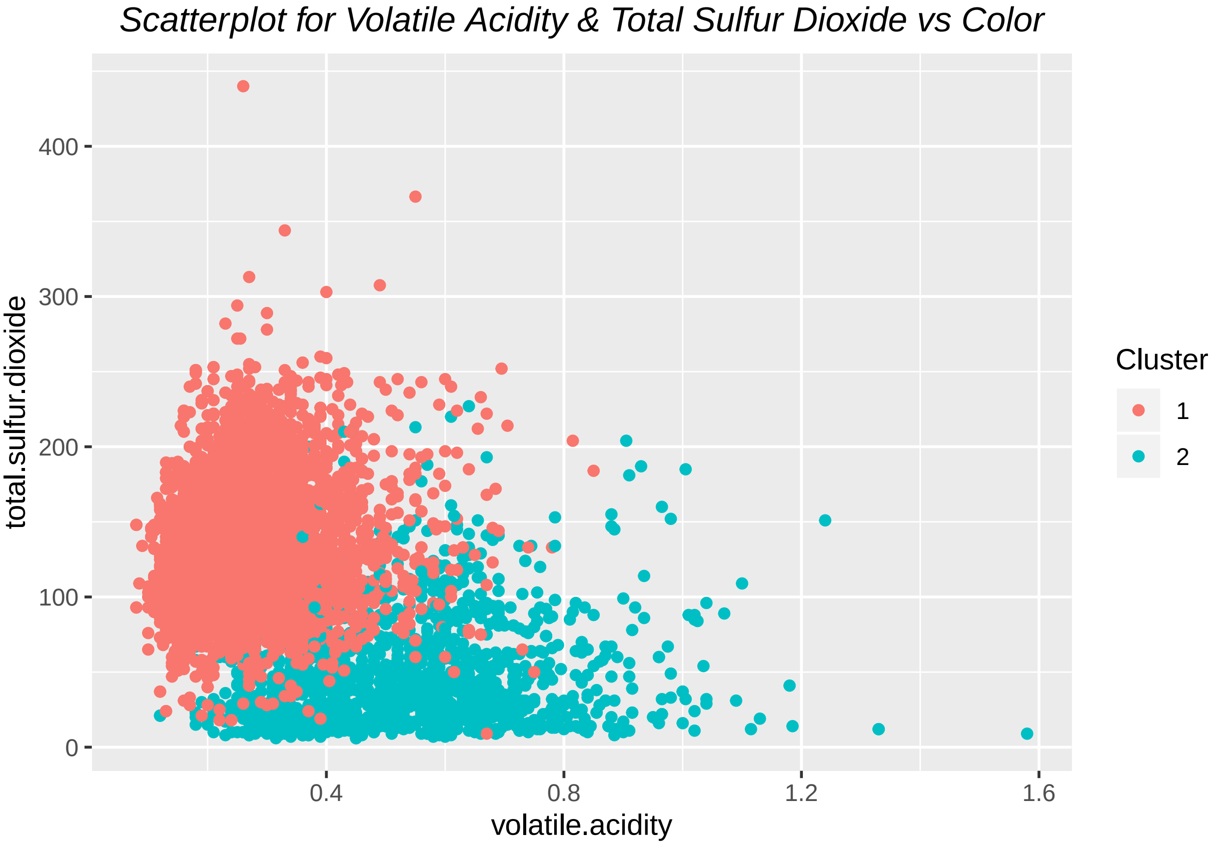
Cluster1:

```
##        fixed.acidity     volatile.acidity          citric.acid
##                 6.85                 0.27                 0.34
##       residual.sugar            chlorides  free.sulfur.dioxide
##                 6.39                 0.05                35.52
## total.sulfur.dioxide              density                   pH
##               138.46                 0.99                 3.19
##            sulphates              alcohol
##                 0.49                10.52
```

Cluster2:

```
##        fixed.acidity     volatile.acidity          citric.acid
##                 8.29                 0.53                 0.27
##       residual.sugar            chlorides  free.sulfur.dioxide
##                 2.63                 0.09                15.76
## total.sulfur.dioxide              density                   pH
##                48.64                 1.00                 3.31
##            sulphates              alcohol
##                 0.66                10.40
```

From the plot demonstrating how clustering works with volatile acidity and total sulfur dioxide on the x, y-axis, and the confusion matrix, we can conclude that clustering successfully distinguishes the white wine from the red wine. 98.6% of the observations fall in the clusters of their respective colors. Cluster 1 contains the white wines predominately, whereas Cluster 2 includes mostly red wines.



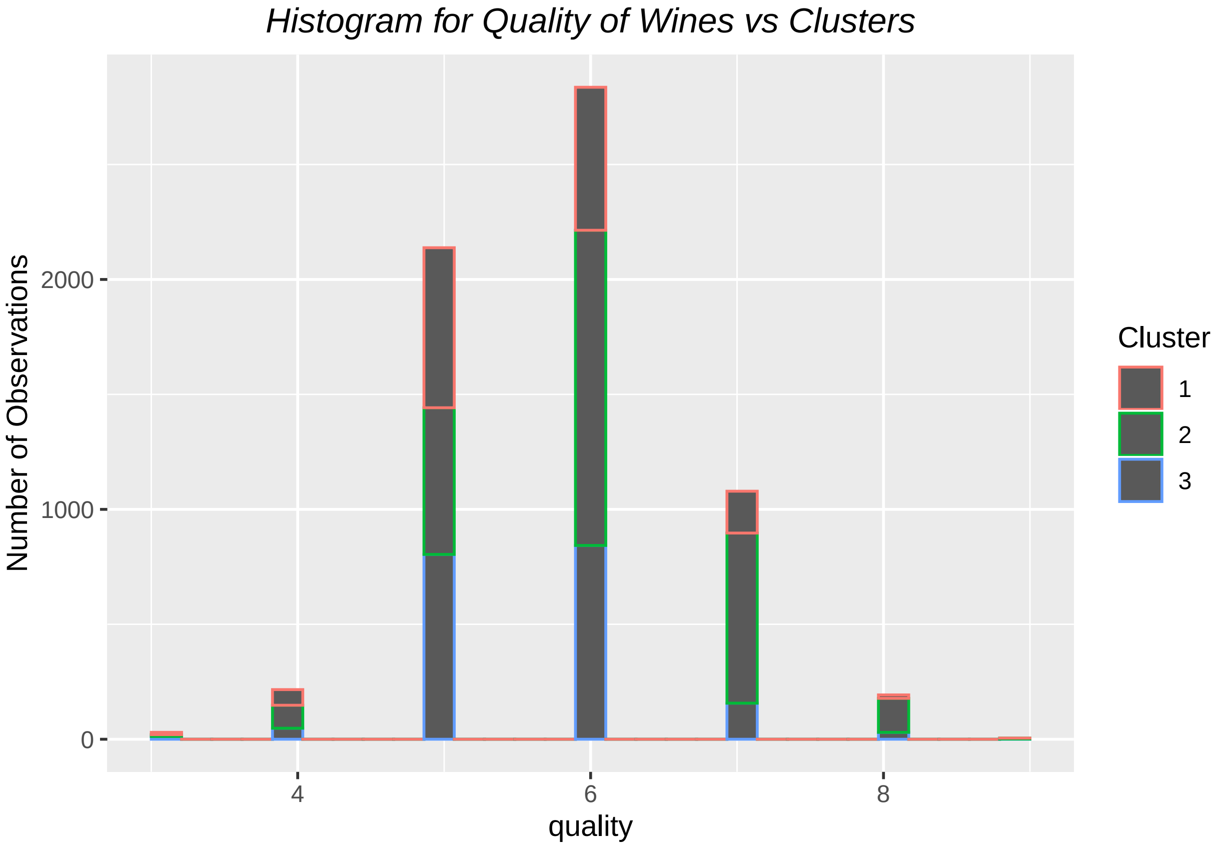
```
##        cluster
## color      1    2
##   red     24 1575
##   white 4830   68
```

In the PCA method, data analysts have to read from the technicalities of PC1 to infer that PC1 correlates with wines' color, which can be a challenge for anyone without chemistry backgrounds. In clustering, however, the dimensionality reduction technique automatically separates the observations into two groups, each one of which has predominantly one color of wines. Plus, unlike in PCA, we don't need to access the "supervised" information to determine the cutoffs for red/white wines in the clustering approach. Thus, we conclude that clustering is a better methodology than the PCA to distinguish different colors of wines with their chemical properties.

3.3 Quality of the wine

Next, we use clustering to distinguish between wines of different qualities. We degsinate that there are three types of wine, namely, the wines of high, average, and low qualities.

```
##        cluster
## quality    1    2    3
##       3   10    8   12
##       4   68  100   48
##        5  696  638  804
##        6  622 1371  843
##        7  182  740  157
##        8   15  148   30
##        9    0    4    1
```



From both the confusion matrix and the plot featuring the distribution of each cluster across different qualities of wine, we find that the clustering can't successfully distinguish between wines of different qualities. All clusters fail to show a concentration over certain levels of quality. Instead, their distributions across varying levels of quality are relatively spread out. Thus, we conclude that clustering is NOT capable of sorting the higher from the lower quality wines using their chemical compositions.

3.4 Conclusion

We run both PCA and Clustering algorithm on wine.csv, which contains the chemical composition of wines, and we try to find the dimensionality reduction methodology that is most effective at sorting different colors of wines. Via the analysis of 3 principal components returned from the PCA algorithm, we find it hard to sort different colors of wines without sufficient chemistry backgrounds or accessing "supervised" information. Also, though PCA maximizes the overall variance of the data along a small set of directions to concentrate information, it can well pick directions that make it hard to separate classes. Clustering, however, is proven to have distinguished the red wines from the white ones, with an accuracy of 0.986. Based on the success of the cluetring approach, we apply it to sort wines of different qualities. However, the algorithm fails to successfully sort the higher from the lower quality wines, showing that clustering is also not universal when trying to distinguish between different properties.

4. Market Segmentation

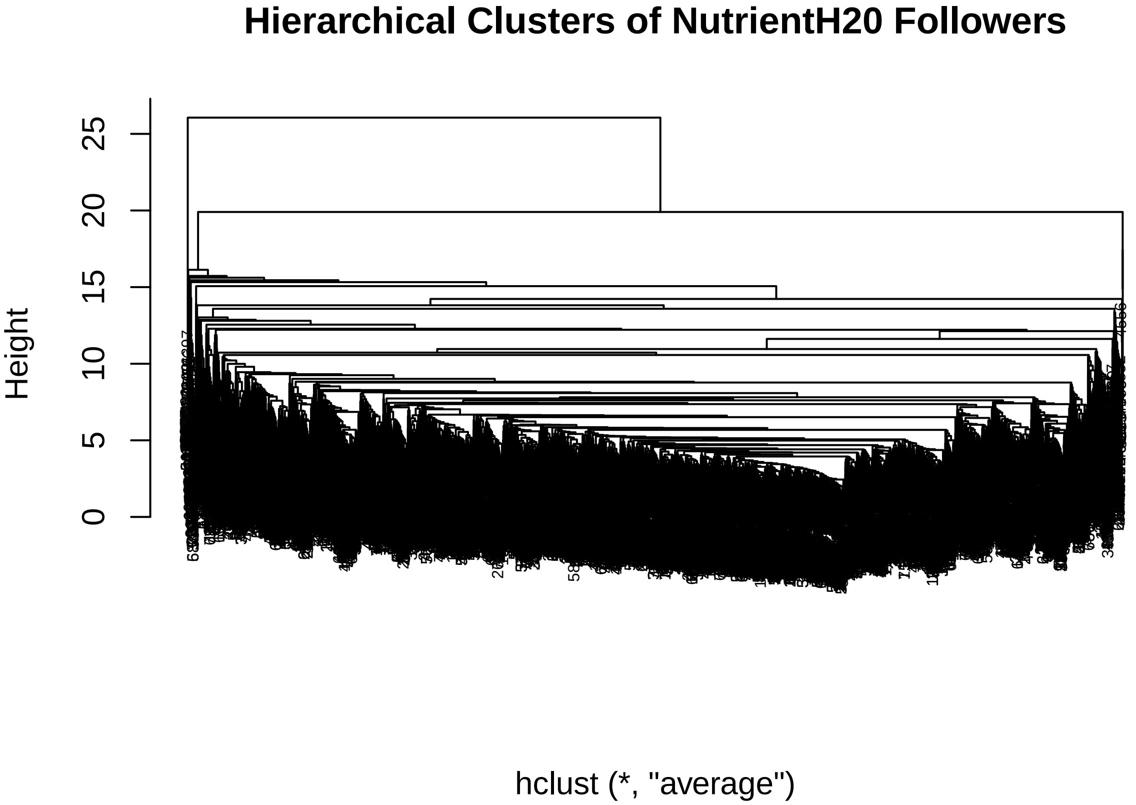
4.1 Introduction

The social_marketing.csv contains 7882 entries with 36 different categories. Each category represents a broad area of interest, like politics and sports. The entries denote the number of tweets a follower of NutrientH20 has on a given subject. The categories of interests are not mutually exclusive. This report will identify market segments that stand out among the audiences of NutrientH20 and thus provide insights into how the owner of NutrientH20 can better target its followers.

4.2 Clustering

Before the analysis, we first remove the variable "X" from the dataset, which is a random identification number unrelated to our study.

We first apply the clustering methodology. After scaling the data, we use hierarchical clustering with average linkage to obtain the dendrogram below:



Next, we cut the tree at k = 20 to obtain 20 distinct clusters and tabulate the number of accounts in each cluster.

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16 
## 7724   64    6   46    2    6    1    8    3    4    2    2    2    5    1    2 
##   17   18   19   20 
##    1    1    1    1
```

From the tabulation of 20 clusters above, we find that cluster 1 contains significantly more observations than the remaining clusters. Thus, we hypothesize that cluster 1 represents the typical audience of NutrientH20 and is, therefore, the market segment that the owner of NutrientH20 should target.

4.3 PCA

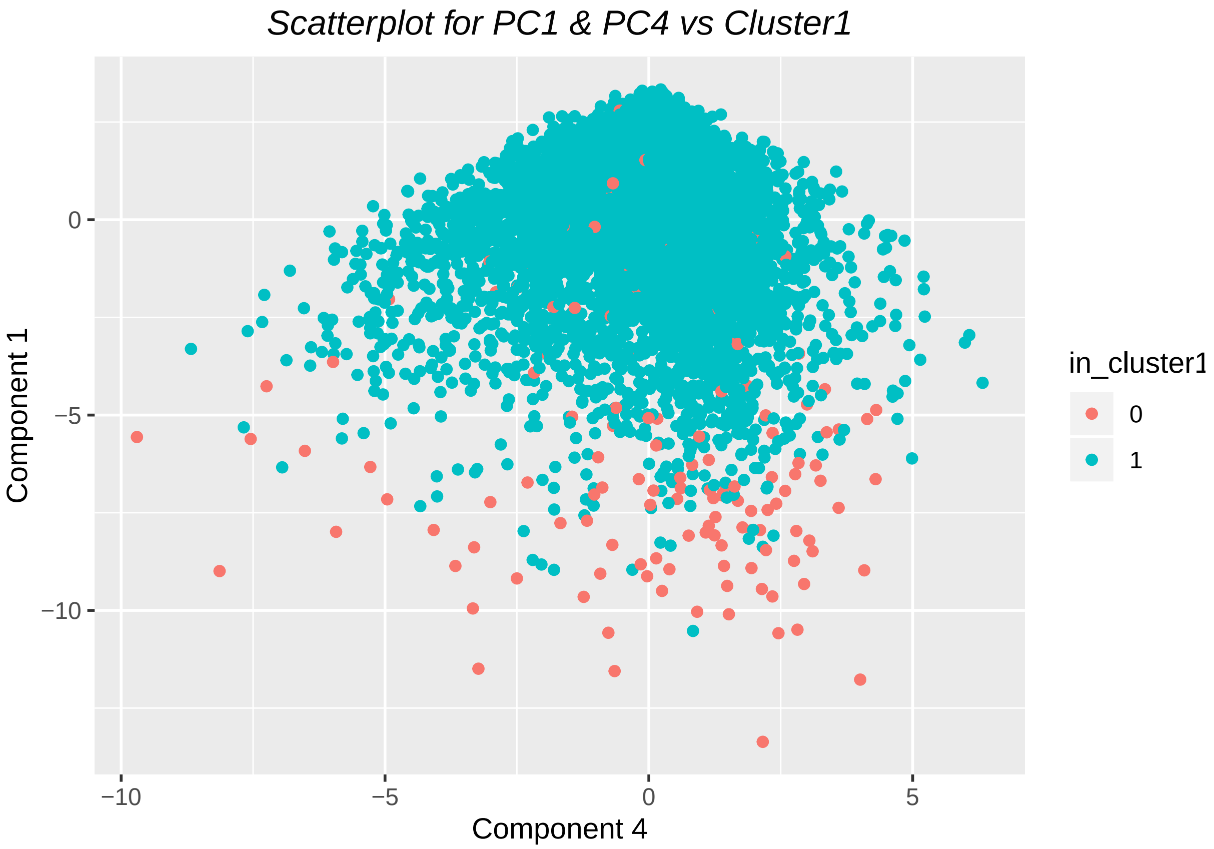
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027 1.36885 1.28577
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087 0.05205 0.04592
## Cumulative Proportion  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
```

The principal component analysis summary demonstrates that the first four principal components contain 34.08% of the information within the dataset. Next, we use the four components to predict whether an account will fall on cluster 1 using linear regression.

We generate a new variable cluster_1 denoting whether the observation belongs to cluster1. The new variable is used as the dependent variable in the linear regression, with the four principal components being the explanatory variables.

```
##
## Call:
## glm(formula = cluster_1 ~ PC1 + PC2 + PC3 + PC4, data = social_marketing)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.98705   0.01814   0.01962   0.02162   0.03599
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.800e-01  1.579e-03 620.683   <2e-16 ***
## PC1          7.891e-04  7.453e-04   1.059    0.290
## PC2         -8.761e-05  9.298e-04  -0.094    0.925
## PC3          1.120e-04  9.906e-04   0.113    0.910
## PC4         -1.587e-03  1.029e-03  -1.543    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.01964753)
##
##     Null deviance: 154.83  on 7881  degrees of freedom
## Residual deviance: 154.76  on 7877  degrees of freedom
## AIC: -8599.6
##
## Number of Fisher Scoring iterations: 2
```



The regression result shows that PC1 and PC4 are closest to being statistically significant. Whereas a higher PC1 makes it more likely to be in cluster 1, a higher PC4 makes it less likely to be cluster 1. Thus, the variables in PC1 with positive coefficients and PC4 with negative coefficients are the topics more prevalent among people in cluster 1. The details of CPC1 and PC4 is as follows:

PC1

```
##          chatter   current_events           travel    photo_sharing
##            -0.13            -0.10            -0.12            -0.18
##    uncategorized          tv_film   sports_fandom         politics
##            -0.09            -0.10            -0.29            -0.13
##             food           family  home_and_garden            music
##            -0.30            -0.24            -0.12            -0.12
##             news    online_gaming         shopping health_nutrition
##            -0.13            -0.07            -0.13            -0.12
##      college_uni   sports_playing          cooking              eco
##            -0.09            -0.13            -0.19            -0.15
##        computers         business         outdoors           crafts
##            -0.14            -0.14            -0.14            -0.19
##       automotive              art         religion           beauty
##            -0.13            -0.10            -0.30            -0.20
##        parenting           dating           school personal_fitness
##            -0.29            -0.11            -0.28            -0.14
##          fashion   small_business             spam            adult
##            -0.18            -0.12            -0.01            -0.03
```

All variables in PC1 have negative coefficients. In particular, sports_fandom, religion, beauty, food, parenting, and family have the most significant absolute negative values. Namely, sports_fandom, religion, beauty, food, parenting, and family may not be welcomed by the followers of NutrientH20. Thus, the account owner should avoid posting topics of the above categories since they're unlikely to attract the largest group of the account's potential audiences.

PC4

```
##          chatter   current_events           travel    photo_sharing
##             0.11             0.03            -0.15             0.15
##    uncategorized          tv_film    sports_fandom         politics
##             0.02             0.09             0.06            -0.20
##             food           family  home_and_garden            music
##            -0.07             0.07            -0.01             0.08
##             news    online_gaming         shopping health_nutrition
##            -0.18             0.22             0.10            -0.46
##      college_uni   sports_playing          cooking              eco
##             0.26             0.18             0.01            -0.12
##        computers         business         outdoors           crafts
##            -0.14             0.01            -0.41             0.02
##       automotive              art         religion           beauty
##            -0.04             0.06             0.07             0.15
##        parenting           dating           school personal_fitness
##             0.05            -0.03             0.09            -0.44
##          fashion   small_business             spam            adult
##             0.14             0.08            -0.02            -0.02
```

The negative coefficients within PC4 generally feature health-related topics such as personal_fitness and more serious topics like politics and news. Among the negative coefficients, health_nutrition, personal_fitness, and outdoors have the largest absolute values. This means that the typical audiences of the twitter account NutrientH20 prefer information about health_nutrition, personal_fitness, and outdoors better than others. Also, eco, travel, politics, and computers may be topics that can attract more followers to NutrientH20.

4.4 Conclusion

By using both clustering and PCA, we identify the cluster that is the characterization of typical followers of the account NutrientH20 and believe users in cluster1 is the market segment the owner of NutrientH20 should target. Running regression of a dummy variable denoting cluster1 and four principal components, we find that PC1 and PC4 are more statistically significant, among others. Then, from the breakdown of PC1 and PC4, we infer that NutrientH20 should focus more on ealth_nutrition, personal_fitness, and outdoors as typical users appear to be a population with an emphasis on personal health. To attract more users, the brand should accentuate its ability to share information regarding personal fitness, nutrition, and outdoor activities. Meanwhile, the brand can also share a limited amount of eco, travel, politics, and computers but avoid topics such as sports_fandom, religion, beauty, food, parenting, and family.

In short, NutrientH20 should primarily target audiences similar to those within cluster1, which we believe represents the typical image of the account's followers by sharing information more on personal health and fitness topics.