Exercise 3

Predictive Model Building

```
data <- read.csv("~/Documents/SDS323Assignments/greenbuildings.csv")
dataFiltered = filter(data, leasing_rate >= 10)
green = filter(dataFiltered, green_rating == 1)
nonGreen = filter(dataFiltered, green_rating == 0)
dataFiltered = dataFiltered[!is.na(dataFiltered$Rent), ] #drop those whose rent is NA
dataFiltered = replace_na(dataFiltered)
n = nrow(dataFiltered)
n_train = round(0.8*n)
n_test = n - n_train
rmse = function(y, yhat) {
  sqrt(mean((y - yhat)^2, na.rm = TRUE))
}
```

We first find the mean rent and build a null model:

```
avg = mean(dataFiltered$Rent)
xData = model.matrix(~ . - Rent, data = dataFiltered)
yData = model.matrix(~Rent - 1, data = dataFiltered)
totalErr = 0
for(i in 1:100)
{
train_cases = sample.int(n, n_train, replace=FALSE)
test_cases = setdiff(1:n, train_cases)
y_train = yData[train_cases,]
y_test = yData[test_cases,]
totalErr = totalErr + rmse(avg, y_test)
}
totalErr / 100
```

```
## [1] 15.13563
```

Our benchmark is around 15. We now run a linear regression, removing total__dd__07, because it depends entirely on cd.total.07 and hd.total07, as well as the Property ID and cluster number which don't make sense to use as a quantitative variable.

```
data_train = dataFiltered[train_cases,]
data_test = dataFiltered[test_cases,]
totalErr = 0
for(i in 1:100)
{
train_cases = sample.int(n, n_train, replace=FALSE)
test_cases = setdiff(1:n, train_cases)
data_train = dataFiltered[train_cases,]
data_test = dataFiltered[test_cases,]
lm2 = lm(Rent ~ . - CS_PropertyID - total_dd_07 - cluster, data = data_train)
yhat_test = predict(lm2, data_test)
totalErr = totalErr + rmse(yhat_test, data_test$Rent)
}
totalErr / 100
```

```
## [1] 9.576231
```

```
lm2
```

```
##
## Call:
## lm(formula = Rent ~ . - CS_PropertyID - total_dd_07 - cluster,
##     data = data_train)
##
## Coefficients:
##       (Intercept)               size            empl_gr       leasing_rate
##        -8.427e+00          7.288e-06          6.612e-02          1.181e-02
##           stories                age          renovated            class_a
##        -5.239e-02         -1.050e-02         -3.306e-01          3.092e+00
##           class_b               LEED         Energystar       green_rating
##         8.745e-01         -1.547e+00         -4.310e+00          4.858e+00
##               net          amenities        cd_total_07         hd_total07
##        -2.772e+00          5.343e-01         -7.976e-05          6.062e-04
##     Precipitation          Gas_Costs  Electricity_Costs       cluster_rent
##         4.796e-02         -3.880e+02          2.207e+02          9.987e-01
```

We observe that the coefficients of LEED, Energystar, and green_rating have different signs, which suggests there may be noise in the data. Similarly, it is unclear why Gas_Costs and Electricity_Costs have different signs.

```
data_train = dataFiltered[train_cases,]
data_test = dataFiltered[test_cases,]
totalErr = 0
recip = function(x){1 / x}
for(i in 1:100)
{
train_cases = sample.int(n, n_train, replace=FALSE)
test_cases = setdiff(1:n, train_cases)
data_train = dataFiltered[train_cases,]
data_test = dataFiltered[test_cases,]
lm2 = lm(Rent ~ . + recip(size) - CS_PropertyID - total_dd_07 - cluster - Energystar - Electricity_Cost
yhat_test = predict(lm2, data_test)
totalErr = totalErr + rmse(yhat_test, data_test$Rent)
}
totalErr / 100
```

```
## [1] 9.302295
```

```
lm2
```

```
##
## Call:
## lm(formula = Rent ~ . + recip(size) - CS_PropertyID - total_dd_07 -
##     cluster - Energystar - Electricity_Costs, data = data_train)
##
## Coefficients:
##   (Intercept)           size        empl_gr   leasing_rate        stories
##    -5.003e+00      7.733e-06      1.659e-02      1.033e-02     -2.890e-02
##           age     renovated       class_a       class_b          LEED
##    -8.865e-03    -3.106e-01     2.789e+00     1.413e+00     1.740e+00
##  green_rating           net     amenities   cd_total_07     hd_total07
##     7.003e-01    -2.429e+00     4.795e-01    -2.613e-04     1.796e-04
## Precipitation     Gas_Costs  cluster_rent    recip(size)
##    -1.531e-02    -2.853e+01     1.096e+00     1.514e+03
```

To summarize, I ran two regressions and a null model (outputting the mean each time) on the rent per square foot of each house versus several characteristics including the size, the age, and the green rating. For the first one I used all variables except ones that were clearly redundant and ones that were not suitable for using numerically (such as the numerical property ID). I then removed a few variables which I think affected the output variable in unnatural ways, likely due to noise. The mean squared error of my model was about 40% better than the null model of choosing the mean each time. The coefficients of the regression are listed above. After removing the EnergyStar variable, the rent per square foot seemed to increase slightly with green certification, all other factors being equal. In conclusion, it seems that green rating appears to be very slightly correlated with an increased price per square foot.

What causes what?

1. Firstly, correlation does not imply causality. In fact, the causation may be the other way around, in that, as a response to higher crime, more cops are deployed in certain cities. This may lead to the wrongful conclusion from regression that more police causes more crime. Secondly, choosing different cities does not control for other confounding variables that may also strongly affect crime rates, such as socioeconomic status. In other words, we are comparing apples to oranges when we compare different cities.

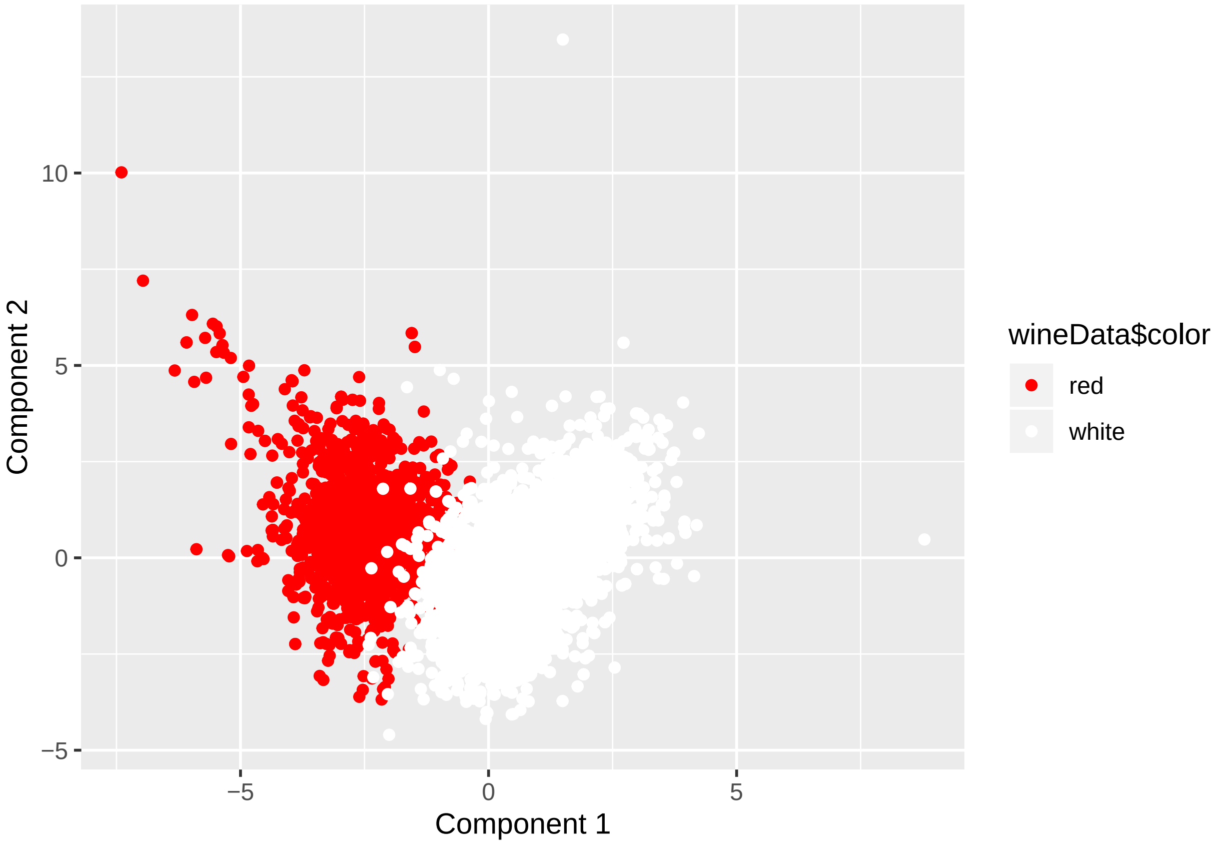
2. The researchers set a control with a baseline cop alertness and computed the baseline amount of crime, and, in the same city, the amount of crime when the cop alertness is increased due to an orange alert for terrorism. These increases in alertness was not caused by more crime but by an unrelated reason. However, on orange alert days, street crimes tended to decrease. Using a control in the same city helps control for other confounding factors.

3. The researchers thought of another possible confounding factor: that there may have been fewer potential victims when the terrorism alert is increased, which may make the would-be criminals less likely to feel the incentive to go out and commit a crime. By checking the metro ridership, the researchers discovered that this was not the case; the ridership was essentially the same.

4. The high alert in district 1 was significantly correlated with a decrease in the amount of crime, after controlling for mid-day ridership (i.e. potential number of victims). Also, as predicted, the logarithm of the mid-day ridership was significantly, at the 5% level, correlated with an increase in crime, again all else being equal.

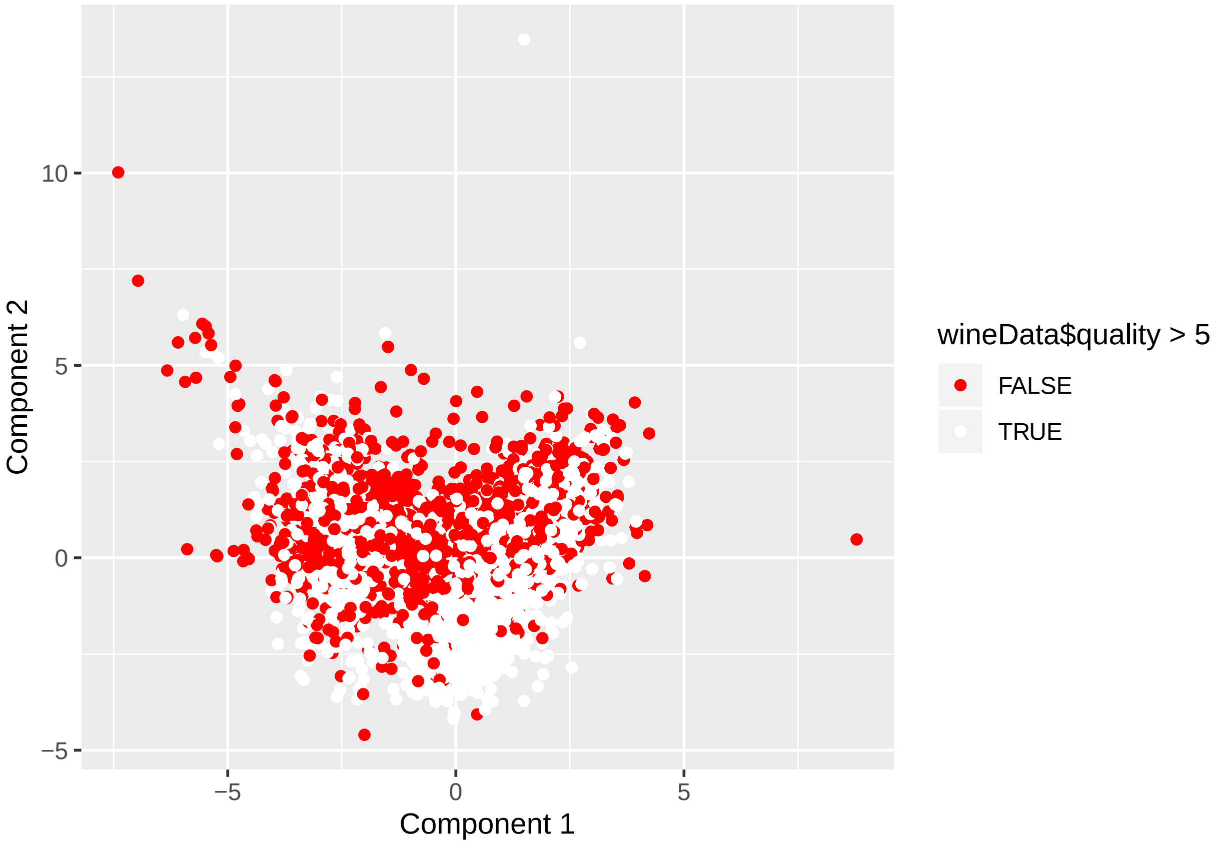
Clustering and PCA

```r
wineData <- read.csv("~/Documents/SDS323Assignments/wine.csv")
pc2 = prcomp(wineData[ ,!(colnames(wineData) == c("color", "quality"))], scale=TRUE, rank=2)
scores = pc2$x
qplot(scores[,1], scores[,2], color=wineData$color, xlab='Component 1', ylab='Component 2')+ scale_colo
```



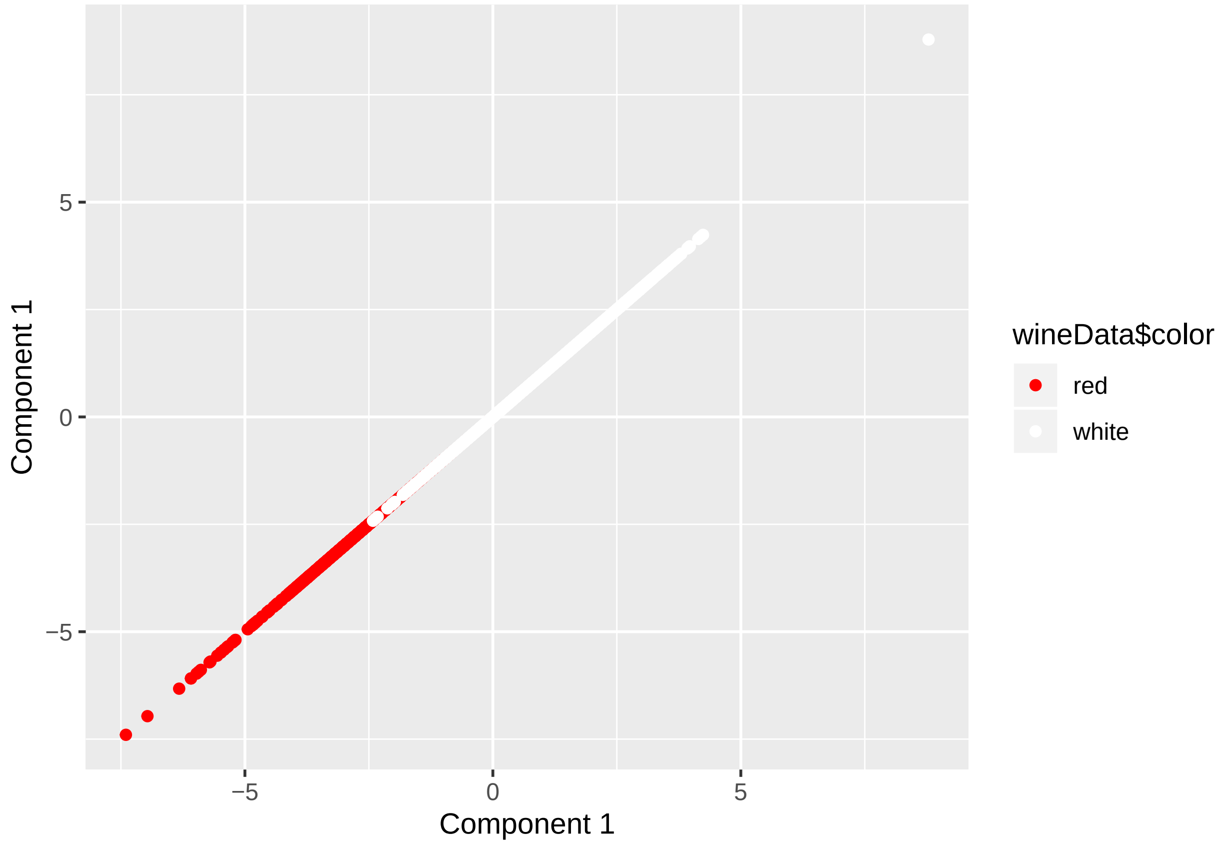
Wine quality (greater than 5 is good but less than 5 is bad)

```
qplot(scores[,1], scores[,2], color=wineData$quality>5, xlab='Component 1', ylab='Component 2')+ scale_
```



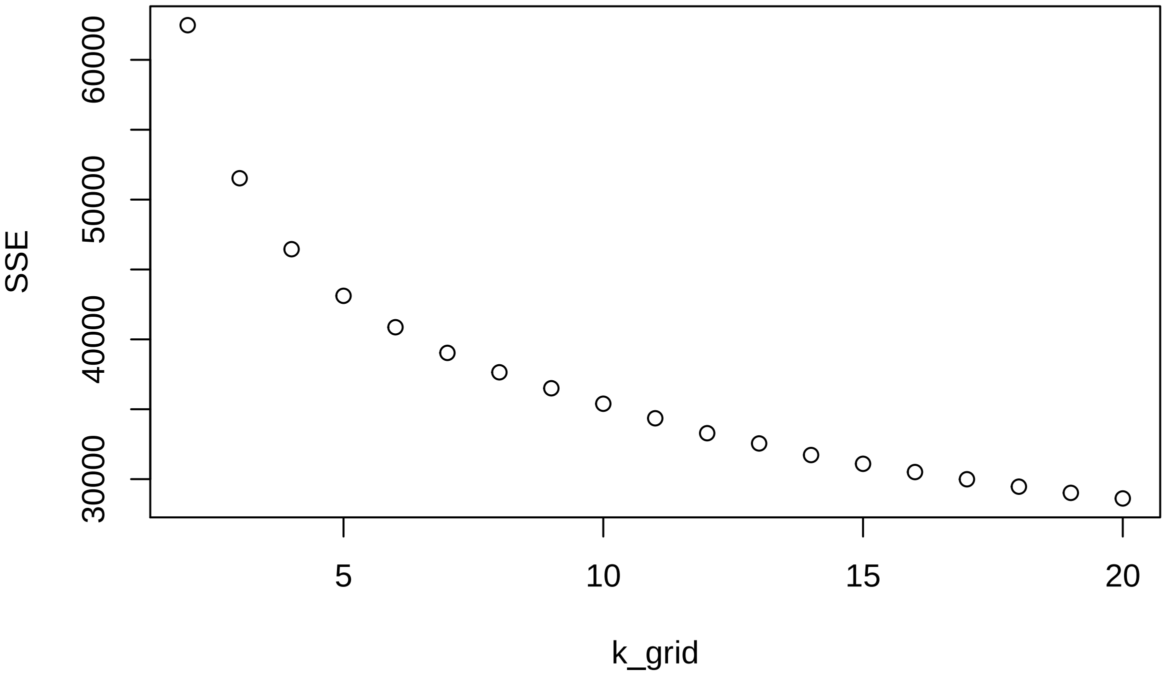
Interestingly, a higher component 1 seems to be associated with white wine, while a lower component 1 seems to be associated with red wine. There does not seem to be a correlation between quality and the PCA values. Component 2 seems not to matter much. We display another graph with only 1 component.

```
wineData <- read.csv("~/Documents/SDS323Assignments/wine.csv")
pc1 = prcomp(wineData[ ,!(colnames(wineData) == c("color", "quality"))], scale=TRUE, rank=1)
scores = pc1$x
qplot(scores[,1], scores[,1], color=wineData$color, xlab='Component 1', ylab='Component 1')+ scale_colo
```



We now use k-means clustering. We first create an elbow plot to help us decide which value of k to choose.

```
X = scale(wineData[ ,!(colnames(wineData) == "color")], center=TRUE, scale=TRUE)
k_grid = seq(2, 20, by = 1)
SSE = c()
for(i in k_grid){
  cluster_k = kmeans(X, i, nstart = 50)
  SSE = append(SSE, cluster_k$tot.withinss)
}
plot(k_grid, SSE)
```



We choose k = 5

```
clust1 = kmeans(X, 5, nstart=50)
table(wineData[which(clust1$cluster == 1),]$color)
```

```
##
##   red white
##    29  1652
```

```
table(wineData[which(clust1$cluster == 2),]$color)
```

```
##
##   red white
##    56  1578
```

```
table(wineData[which(clust1$cluster == 3),]$color)
```

```
##
##   red white
##     2  1584
```

```
table(wineData[which(clust1$cluster == 4),]$color)
```

```
##
##   red white
##   597    19
```

```
table(wineData[which(clust1$cluster == 5),]$color)
```

```
##
##   red white
##   915    65
```

It appears that red wines seem to cluster in clusters 2 and 4, and white wines in 1, 3, and 5; each cluster contained more than 90 percent of one color wine.

```
table(wineData[which(clust1$cluster == 1),]$quality)
```

```
##
##   3   4   5   6   7   8
##   9 100 731 804  36   1
```

```
table(wineData[which(clust1$cluster == 2),]$quality)
```

```
##
##   4   5   6   7   8   9
##   1  24 697 745 162   5
```

```
table(wineData[which(clust1$cluster == 3),]$quality)
```

```
##
##   3   4   5   6   7   8
##   9  30 693 710 127  17
```

```
table(wineData[which(clust1$cluster == 4),]$quality)
```

```
##
##   3   4   5   6   7   8
##   5   9 189 271 131  11
```

```
table(wineData[which(clust1$cluster == 5),]$quality)
```

```
##
##   3   4   5   6   7   8
##   7  76 501 354  40   2
```

Cluster 3 seems to be notably associated with better quality, with 6 and 7 the most frequent values, compared to 5 and 6 for the other clusters. While both k-means and PCA predicted color very well, it seems that k-means was better at predicting wine quality. I think k-means may have made more sense because certain variables such as acidity may cause degradation if it is either too high or too low, which PCA doesn't account for since it is a linear dimensionality reduction technique. K-means, on the other hand, only compares wines "close" to each other, which makes it more robust, at least in terms of predicting quality.

Market segmentation

```
mktData <- read.csv("~/Documents/SDS323Assignments/social_marketing.csv")
mktData = mktData[,!(colnames(mktData) == 'X')]
length(mktData$spam)
```

```
## [1] 7882
```

```
sum(mktData$adult)
```

```
## [1] 3179
```

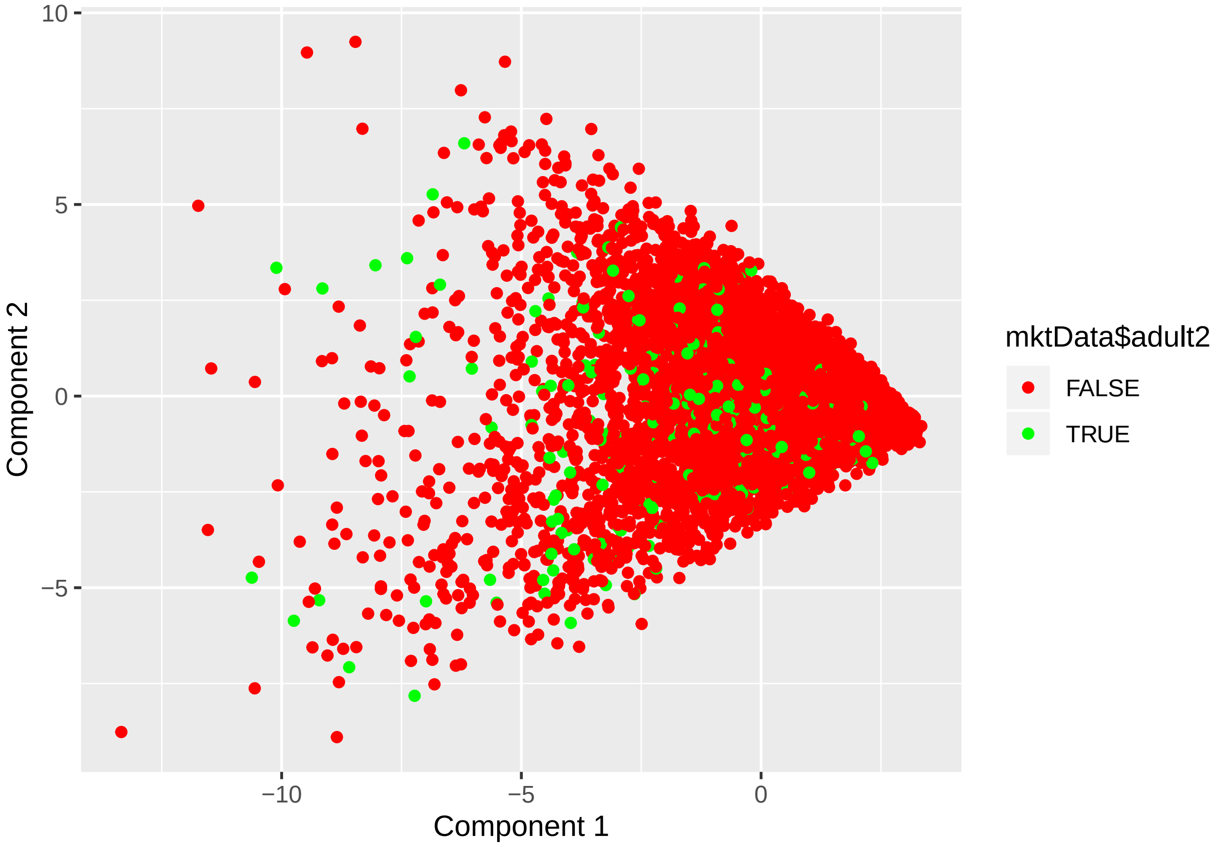
```
sum(mktData$spam)
```

```
## [1] 51
```

#The amount of spam may be too small for meaningful analysis, but the adult content is quite prevalent. We use PCA in an attempt to gain insights from the data, as well as taking means based on adult content.

```
mktData$adult2 = mktData$adult >= 1
mktData$isSpam = mktData$spam >= 1
pc2 = prcomp(mktData, scale=TRUE, rank=2)
```

```
scores = pc2$x
qplot(scores[,1], scores[,2], color=mktData$adult2, xlab='Component 1', ylab='Component 2')+ scale_colo
```



```
df1 = colMeans(filter(mktData, adult2 == 0))
df2 = colMeans(filter(mktData, adult == 1))
df = rbind(df1, df2)
df
```

```
##      chatter current_events   travel photo_sharing uncategorized   tv_film
## df1 4.394283       1.520925 1.574672      2.714305     0.8032002 1.0812363
## df2 4.753623       1.565217 1.449275      2.449275     0.7391304 0.7826087
##     sports_fandom politics     food    family home_and_garden     music
## df1      1.594502 1.813047 1.385941 0.8546225       0.5140864 0.6865427
## df2      1.507246 1.144928 1.333333 0.8260870       0.4492754 0.4492754
##         news online_gaming shopping health_nutrition college_uni sports_playing
## df1 1.215399      1.201176 1.402626         2.580826    1.565372      0.6442834
## df2 0.884058      1.362319 1.333333         2.826087    1.231884      0.4927536
##      cooking       eco computers  business  outdoors    crafts automotive
## df1 2.009163 0.5005470 0.6446937 0.4246444 0.7679158 0.5106674  0.8176969
## df2 1.840580 0.6376812 0.4347826 0.3768116 0.8260870 0.4492754  0.9420290
##           art  religion    beauty parenting    dating    school
## df1 0.7138950 1.0932713 0.7050055 0.9100109 0.7112965 0.7579322
## df2 0.5942029 0.9565217 0.4637681 0.7681159 0.6666667 0.8260870
##     personal_fitness   fashion small_business         spam adult adult2
## df1         1.463074 1.0006838      0.3263129 0.0004102845     0      0
## df2         1.463768 0.7826087      0.3333333 0.0434782609     1      1
##           isSpam
## df1 0.0004102845
## df2 0.0434782609
```

While PCA didn't lead to any interesting observations, except the fact that a higher component 1 leads to a lower variance in component 2 (which may be just how the PCA is structured), we do notice that adult content seems to be correlated with spam.

```
df1 = colMeans(filter(mktData, isSpam == 0))
df2 = colMeans(filter(mktData, isSpam == 1))
df = rbind(df1, df2)
df
```

```
##      chatter current_events   travel photo_sharing uncategorized  tv_film
## df1 4.397166       1.524065 1.580876      2.698328     0.8123324 1.071492
## df2 4.653061       1.877551 2.244898      2.448980     0.9183673 0.877551
##     sports_fandom politics     food    family home_and_garden     music
## df1      1.592110 1.785778 1.397038 0.8642921       0.5195966 0.6791778
## df2      1.897959 2.244898 1.469388 0.7959184       0.6938776 0.6938776
##         news online_gaming  shopping health_nutrition college_uni
## df1 1.205541      1.207328 1.3920592         2.565811    1.547172
## df2 1.204082      1.448980 0.9591837         2.795918    1.918367
##     sports_playing  cooking       eco computers  business  outdoors    crafts
## df1      0.6398570 1.999489 0.5101494 0.6468786 0.4247415 0.7804162 0.5147453
## df2      0.5306122 1.795918 0.8571429 1.0000000 0.1836735 1.1428571 0.6938776
##     automotive      art religion    beauty parenting    dating    school
## df1  0.8288012 0.721435 1.093961 0.7059875  0.919571 0.7109664 0.7670114
## df2  1.0000000 1.265306 1.326531 0.5714286  1.204082 0.6938776 0.8775510
##     personal_fitness   fashion small_business     spam     adult     adult2
## df1         1.460232 0.9968084      0.3351206 0.000000 0.3607813 0.06689646
## df2         1.755102 0.9591837      0.5306122 1.040816 7.2040816 0.93877551
##     isSpam
## df1      0
## df2      1
```

Note that nearly 94% of spam-marked messages contain adult content, with an average of 7.2 instances, but only less than 7% of nonspam-marked messages do. We can try running a logistic regression on how spam can be predicted. Because the proportion of spam is so low we don't expect a robust model; we are only doing to to gain insights on how to predict spam.

```
model = glm(isSpam ~ . - isSpam - spam, data = mktData, family = 'binomial')
model
```

```
## 
## Call:  glm(formula = isSpam ~ . - isSpam - spam, family = "binomial", 
##     data = mktData)
## 
## Coefficients:
##      (Intercept)           chatter    current_events            travel  
##        -8.404270          0.049485          0.161775          0.009303  
##    photo_sharing     uncategorized           tv_film     sports_fandom  
##         0.099098         -0.133784         -0.264181          0.197102  
##         politics              food            family   home_and_garden  
##         0.117113         -0.143135         -0.387700          0.153225  
##            music              news     online_gaming          shopping  
##         0.148637         -0.040129         -0.172998         -0.369030  
## health_nutrition       college_uni    sports_playing           cooking  
##         0.025860          0.280729         -0.229407         -0.026496  
##              eco        computers         business         outdoors
##         0.087411         0.010147        -0.833738        -0.012857
##           crafts       automotive              art         religion
##         0.152993        -0.151531         0.250588         0.131983
##           beauty        parenting           dating           school
##        -0.292148        -0.037358        -0.065598        -0.071168
## personal_fitness          fashion   small_business            adult
##         0.062329         0.077303         0.017598         0.136229
##       adult2TRUE
##         4.795617
##
## Degrees of Freedom: 7881 Total (i.e. Null);  7845 Residual
## Null Deviance:       595.6
## Residual Deviance: 316.9     AIC: 390.9
```

Clearly, the fact that adult content exists at all greatly increases the likelihood of spam, with greater adult content further leading to more likely spam. Also, all else being equal, the presence of less business content seem to be associated with less spam. We also analyze the correlations between different variables to see if any clusters stand out.

```
cor(mktData)
```

```
##                       chatter current_events       travel photo_sharing
## chatter           1.000000000    0.156211163  0.014666814   0.536266603
## current_events    0.156211163    1.000000000  0.050946163   0.145952862
## travel            0.014666814    0.050946163  1.000000000   0.024067950
## photo_sharing     0.536266603    0.145952862  0.024067950   1.000000000
## uncategorized     0.066993440    0.029674226  0.030865562   0.096398099
## tv_film           0.012898633    0.077673464  0.096989954   0.021170912
## sports_fandom     0.014392631    0.061780370 -0.008709190   0.019921002
## politics          0.051448378    0.068282733  0.660210000   0.039766998
## food             -0.004704080    0.059767526  0.075142216   0.006802181
## family            0.079170381    0.063367006  0.017534763   0.098587939
## home_and_garden   0.070157596    0.053511465  0.040931872   0.083955114
## music             0.091317767    0.072366139  0.038640317   0.146188166
## news             -0.001163561    0.060284226  0.250616947  -0.011980028
## online_gaming     0.004784789   -0.001645689  0.013222873   0.037234010
## shopping          0.583373219    0.150143619  0.019907780   0.535621020
## health_nutrition  0.003727738    0.019863015 -0.011922499   0.034804791
## college_uni       0.033991439    0.030952662  0.053835136   0.061494841
## sports_playing    0.061073951    0.030720781  0.054961948   0.098696993
## cooking           0.002081642    0.046716479  0.017597488   0.360590994
## eco               0.155117735    0.076903468  0.061434287   0.173411683
## computers         0.072489571    0.054824859  0.602934879   0.093000186
## business          0.166106904    0.074580868  0.161845668   0.178636351
## outdoors         -0.006597319    0.017898948  0.027133231   0.032783874
## crafts            0.111741452    0.073723629  0.086953794   0.110945709
## automotive        0.133767354    0.072444806 -0.002784110   0.115249028
## art              -0.005076772    0.053477246  0.086394257   0.024752344
## religion         -0.023111699    0.067694509  0.063933064   0.003299888
## beauty            0.026364231    0.070112215  0.012565524   0.317965902
## parenting         0.020742516    0.050237162  0.042341130   0.041169648
## dating            0.184452453    0.031045863  0.086756030   0.028511721
## school            0.116678625    0.068047127  0.022199544   0.106209238
## personal_fitness  0.032747707    0.038388654 -0.005299245   0.062523682
## fashion           0.063314293    0.055906219  0.026015648   0.347248795
## small_business    0.116730623    0.065497930  0.116950707   0.138112387
## spam              0.004603242    0.019402989  0.022773349  -0.008664858
## adult             0.015372712    0.016764279  0.020244743  -0.012707084
## adult2            0.004540130    0.015067856  0.016192250  -0.022984361
## isSpam            0.005699655    0.021897852  0.022837522  -0.007175568
##                  uncategorized       tv_film sports_fandom     politics
## chatter           0.0669934396  0.0128986328  0.0143926306  0.051448378
## current_events    0.0296742259  0.0776734640  0.0617803698  0.068282733
## travel            0.0308655618  0.0969899543 -0.0087091896  0.660210000
## photo_sharing     0.0963980989  0.0211709116  0.0199210022  0.039766998
## uncategorized     1.0000000000  0.1632789878 -0.0005287314 -0.001143143
## tv_film           0.1632789878  1.0000000000  0.0307587998  0.032254541
## sports_fandom    -0.0005287314  0.0307587998  1.0000000000  0.067097905
## politics         -0.0011431434  0.0322545409  0.0670979049  1.000000000
## food              0.0353476665  0.0806833248  0.5326383665  0.059017379
## family           -0.0046281840  0.0217763915  0.4378103823  0.045470650
## home_and_garden   0.0742591369  0.1065910522  0.0848219920  0.057986105
## music             0.1439077049  0.2748322309  0.0545375138  0.007545661
## news              0.0037403342  0.0674405581  0.2002896757  0.561842213
## online_gaming     0.0236016809  0.0353318018  0.0247608769 -0.006917410
## shopping          0.0553125510  0.0416246434  0.0262612699  0.046393973
## health_nutrition  0.0798559719 -0.0017906843 -0.0112292550 -0.016851900
## college_uni       0.0947186290  0.2042182554  0.0264564138  0.008964787
## sports_playing    0.0838301172  0.1032631811  0.0710499117  0.032736511
## cooking           0.1612570088  0.0006017955  0.0076921174 -0.007322368
## eco               0.0472320334  0.0628254703  0.0858508297  0.064256293
## computers         0.0266392164 -0.0054859445  0.0506332951  0.572150641
## business          0.0659654221  0.1011473763  0.0681327694  0.150675278
## outdoors          0.0939022767  0.0289285413  0.0622176473  0.073373236
## crafts            0.0891965709  0.1846894463  0.2011885972  0.058597842
## automotive        0.0139164228  0.0205075271  0.2396464725  0.285508994
## art               0.1064475118  0.4987718266  0.0223195336  0.025625587
## religion          0.0170384882  0.0450275641  0.6379748428  0.032529266
## beauty            0.1373705347  0.0167831053  0.1228632403 -0.011292710
## parenting         0.0062680638 -0.0017880905  0.6077181198  0.044281427
## dating            0.1270238084  0.0041708935  0.0169255299  0.078288457
## school            0.0582599351  0.0250222744  0.4931061924  0.027795499
## personal_fitness  0.0847295855 -0.0004447435  0.0142720768 -0.008096476
## fashion           0.1414087167  0.0176088683  0.0307672871 -0.006793131
## small_business    0.0848320319  0.1887982408  0.0486622355  0.105261525
## spam              0.0138980529 -0.0042106462  0.0089574643  0.009438922
## adult             0.0451966858 -0.0217002490  0.0079861641 -0.027102370
## adult2            0.0374753359 -0.0236437628 -0.0008130612 -0.028850793
## isSpam            0.0089062375 -0.0091903923  0.0111255614  0.011906307
##                          food       family home_and_garden         music
## chatter          -0.004704080  0.079170381      0.07015760  0.0913177669
## current_events    0.059767526  0.063367006      0.05351146  0.0723661388
## travel            0.075142216  0.017534763      0.04093187  0.0386403172
## photo_sharing     0.006802181  0.098587939      0.08395511  0.1461881658
## uncategorized     0.035347667 -0.004628184      0.07425914  0.1439077049
## tv_film           0.080683325  0.021776392      0.10659105  0.2748322309
## sports_fandom     0.532638366  0.437810382      0.08482199  0.0545375138
## politics          0.059017379  0.045470650      0.05798610  0.0075456607
## food              1.000000000  0.375533627      0.08921731  0.0721457958
## family            0.375533627  1.000000000      0.06504292  0.0351173712
## home_and_garden   0.089217308  0.065042916      1.00000000  0.0632560502
## music             0.072145796  0.035117371      0.06325605  1.0000000000
## news              0.064499690  0.094370314      0.08137338  0.0155152209
## online_gaming     0.045756746  0.082462535      0.03037615  0.0234658858
## shopping          0.020982946  0.084098539      0.07322790  0.1059920572
## health_nutrition  0.223043246  0.026603921      0.06612165  0.0512084191
## college_uni       0.048260040  0.068586525      0.06022031  0.1881590868
## sports_playing    0.089333676  0.099209710      0.06847339  0.1080771385
## cooking           0.067684449  0.055368653      0.07302935  0.1699371547
## eco               0.148171958  0.085396673      0.07238756  0.0699570243
## computers         0.112482997  0.062071539      0.05303950  0.0425994678
## business          0.076995867  0.063453202      0.04751696  0.1024759787
## outdoors          0.190841529  0.057691067      0.07788630  0.0871445544
## crafts            0.238647962  0.171104760      0.07652389  0.0778324671
## automotive        0.059687427  0.149865803      0.07189290  0.0297480759
## art               0.101791832  0.030163186      0.10488798  0.0202095405
## religion          0.591318063  0.452768549      0.09439831  0.0753480995
## beauty            0.102454336  0.118087188      0.07316316  0.1769706252
## parenting         0.544948125  0.420578018      0.08011189  0.0440734267
## dating            0.033434336  0.017710590      0.10761722  0.0098874584
## school            0.432403928  0.334689575      0.10006929  0.0518095525
## personal_fitness  0.223856925  0.038185942      0.07695813  0.0513248587
## fashion           0.036963917  0.062704099      0.07612052  0.1604436447
## small_business    0.063499437  0.062871670      0.08015803  0.1220162157
## spam              0.001482766 -0.006802752      0.02160085  0.0005284301
## adult             0.017630551  0.034583067      0.03939379 -0.0042387481
## adult2            0.023293983  0.029236794      0.03205864 -0.0252932792
## isSpam            0.003202982 -0.004745485      0.01859592  0.0011218083
##                           news online_gaming     shopping health_nutrition
## chatter          -1.163561e-03  0.0047847889  0.583373219     0.0037277381
## current_events    6.028423e-02 -0.0016456887  0.150143619     0.0198630146
## travel            2.506169e-01  0.0132228733  0.019907780    -0.0119224993
## photo_sharing    -1.198003e-02  0.0372340099  0.535621020     0.0348047911
## uncategorized     3.740334e-03  0.0236016809  0.055312551     0.0798559719
## tv_film           6.744056e-02  0.0353318018  0.041624643    -0.0017906843
## sports_fandom     2.002897e-01  0.0247608769  0.026261270    -0.0112292550
## politics          5.618422e-01 -0.0069174099  0.046393973    -0.0168519001
## food              6.449969e-02  0.0457567463  0.020982946     0.2230432459
## family            9.437031e-02  0.0824625354  0.084098539     0.0266039206
## home_and_garden   8.137338e-02  0.0303761455  0.073227895     0.0661216523
## music             1.551522e-02  0.0234658858  0.105992057     0.0512084191
## news              1.000000e+00 -0.0023217561 -0.011813142     0.0184723484
## online_gaming    -2.321756e-03  1.0000000000 -0.007932408    -0.0000176422
## shopping         -1.181314e-02 -0.0079324080  1.000000000     0.0314086858
## health_nutrition  1.847235e-02 -0.0000176422  0.031408686     1.0000000000
## college_uni      -6.632975e-03  0.7728392923  0.029046828    -0.0277788561
## sports_playing    3.755406e-02  0.4912993420  0.039791700     0.0447899666
## cooking           1.316988e-02  0.0353209993  0.086908714     0.2489527151
## eco               3.916141e-02  0.0275257081  0.165475294     0.2208463589
## computers         2.117802e-01  0.0142759909  0.069347702     0.0235923180
## business          5.202271e-02  0.0032965473  0.160388318     0.0330434015
## outdoors          1.407314e-01  0.0056496732  0.018558832     0.6082253668
## crafts            5.326177e-02  0.0417262775  0.107245497     0.0797254552
## automotive        5.554175e-01  0.0483166936  0.115744356    -0.0238249985
## art               4.135701e-02  0.0828918943  0.032911113     0.0319408677
## religion          2.294591e-02  0.0057441594  0.007956775     0.0106622133
## beauty            1.722328e-02  0.0045270085  0.068034649     0.0069553398
## parenting         7.889741e-02  0.0226973412  0.020248306     0.0292332391
## dating            5.206573e-02  0.0228324583  0.004068363     0.0751908487
## school            5.272822e-02 -0.0008595235  0.076918012     0.0066410255
## personal_fitness  2.873937e-02  0.0119043235  0.054419081     0.8099023568
## fashion           2.132337e-03  0.0360133351  0.096854665     0.0297479557
## small_business    4.759885e-02  0.0419629288  0.112662828    -0.0093025415
## spam             -1.800012e-03  0.0058670252 -0.019251616     0.0023958243
## adult            -1.120403e-02  0.0125787755 -0.017128756    -0.0090925481
## adult2           -1.682464e-02  0.0102013912 -0.026252633    -0.0108218821
## isSpam           -5.459293e-05  0.0070680586 -0.018811123     0.0040229344
##                    college_uni sports_playing       cooking        eco
## chatter           0.0339914391    0.061073951  0.0020816423 0.15511774
## current_events    0.0309526621    0.030720781  0.0467164788 0.07690347
## travel            0.0538351360    0.054961948  0.0175974884 0.06143429
## photo_sharing     0.0614948410    0.098696993  0.3605909943 0.17341168
## uncategorized     0.0947186290    0.083830117  0.1612570088 0.04723203
## tv_film           0.2042182554    0.103263181  0.0006017955 0.06282547
## sports_fandom     0.0264564138    0.071049912  0.0076921174 0.08585083
## politics          0.0089647874    0.032736511 -0.0073223677 0.06425629
## food              0.0482600403    0.089333676  0.0676844485 0.14817196
## family            0.0685865247    0.099209710  0.0553686528 0.08539667
## home_and_garden   0.0602203114    0.068473391  0.0730293453 0.07238756
## music             0.1881590868    0.108077138  0.1699371547 0.06995702
## news             -0.0066329747    0.037554057  0.0131698817 0.03916141
## online_gaming     0.7728392923    0.491299342  0.0353209993 0.02752571
## shopping          0.0290468283    0.039791700  0.0869087138 0.16547529
## health_nutrition -0.0277788561    0.044789967  0.2489527151 0.22084636
## college_uni       1.0000000000    0.506374768  0.0326211357 0.02999765
## sports_playing    0.5063747684    1.000000000  0.1134243174 0.04883664
## cooking           0.0326211357    0.113424317  1.0000000000 0.09468926
## eco               0.0299976539    0.048836643  0.0946892641 1.00000000
## computers         0.0357173059    0.055137058  0.0625767717 0.07170567
## business          0.0568632385    0.058557218  0.0904536300 0.06903281
## outdoors          0.0002266462    0.058386618  0.1926335927 0.17481079
## crafts            0.0445461695    0.073583667  0.0683961801 0.08596322
## automotive        0.0392353381    0.046380642  0.0195136512 0.05997131
## art               0.0903808391    0.056542328  0.0555088964 0.06668392
## religion          0.0278312665    0.076103199  0.0339513209 0.09915385
## beauty            0.0196241668    0.093295239  0.6642389459 0.04762071
## parenting         0.0103665861    0.068607722  0.0506783916 0.11429246
## dating            0.0240141334    0.094629504  0.0290986583 0.06553289
## school           -0.0055988726    0.067138279  0.0843002915 0.09004071
## personal_fitness -0.0215268678    0.051434835  0.2336229315 0.21261355
## fashion           0.0520235384    0.108221244  0.7214026744 0.06381744
## small_business    0.1138833315    0.079653227  0.0727561881 0.07650412
## spam              0.0094529848   -0.008742556 -0.0061781291 0.03141528
## adult            -0.0117086302   -0.011257322 -0.0045564273 0.06021275
## adult2           -0.0196480567   -0.018746119 -0.0114237291 0.05472183
## isSpam            0.0100712609   -0.008802764 -0.0046653309 0.03543554
##                     computers     business      outdoors     crafts
## chatter           0.072489571  0.166106904 -0.0065973195 0.11174145
## current_events    0.054824859  0.074580868  0.0178989481 0.07372363
## travel            0.602934879  0.161845668  0.0271332314 0.08695379
## photo_sharing     0.093000186  0.178636351  0.0327838735 0.11094571
## uncategorized     0.026639216  0.065965422  0.0939022767 0.08919657
## tv_film          -0.005485945  0.101147376  0.0289285413 0.18468945
## sports_fandom     0.050633295  0.068132769  0.0622176473 0.20118860
## politics          0.572150641  0.150675278  0.0733732362 0.05859784
## food              0.112482997  0.076995867  0.1908415291 0.23864796
## family            0.062071539  0.063453202  0.0576910670 0.17110476
## home_and_garden   0.053039498  0.047516956  0.0778863049 0.07652389
## music             0.042599468  0.102475979  0.0871445544 0.07783247
## news              0.211780215  0.052022707  0.1407313846 0.05326177
## online_gaming     0.014275991  0.003296547  0.0056496732 0.04172628
## shopping          0.069347702  0.160388318  0.0185588318 0.10724550
## health_nutrition  0.023592318  0.033043402  0.6082253668 0.07972546
## college_uni       0.035717306  0.056863238  0.0002266462 0.04454617
## sports_playing    0.055137058  0.058557218  0.0583866177 0.07358367
## cooking           0.062576772  0.090453630  0.1926335927 0.06839618
## eco               0.071705671  0.069032814  0.1748107912 0.08596322
## computers         1.000000000  0.145387107  0.0430388875 0.09283321
## business          0.145387107  1.000000000  0.0509135858 0.10165359
## outdoors          0.043038887  0.050913586  1.0000000000 0.07239357
## crafts            0.092833206  0.101653592  0.0723935740 1.00000000
## automotive        0.013813021  0.040319930  0.0887386055 0.03397123
## art              -0.003641485  0.085921419  0.0069278190 0.23353456
## religion          0.107529163  0.075390623  0.0426490984 0.23640037
## beauty            0.050833394  0.087306971  0.0414714063 0.09871424
## parenting         0.092961564  0.083133411  0.0613567343 0.20838376
## dating            0.094839229  0.111825266  0.0953960198 0.09845243
## school            0.076726075  0.104868599  0.0428262486 0.21799280
## personal_fitness  0.021830496  0.058177210  0.5677902744 0.09075479
## fashion           0.063047239  0.109494748  0.0557789123 0.08554645
## small_business    0.104650495  0.107021298  0.0081708055 0.09108434
## spam              0.023117451 -0.025491850  0.0215189501 0.01993976
## adult             0.030290997 -0.010963774  0.0710361684 0.03101611
## adult2            0.013302091 -0.007250782  0.0436925122 0.02276470
## isSpam            0.023534042 -0.027368370  0.0235554143 0.01723796
##                    automotive          art     religion       beauty
## chatter           0.133767354 -0.005076772 -0.023111699  0.026364231
## current_events    0.072444806  0.053477246  0.067694509  0.070112215
## travel           -0.002784110  0.086394257  0.063933064  0.012565524
## photo_sharing     0.115249028  0.024752344  0.003299888  0.317965902
## uncategorized     0.013916423  0.106447512  0.017038488  0.137370535
## tv_film           0.020507527  0.498771827  0.045027564  0.016783105
## sports_fandom     0.239646473  0.022319534  0.637974843  0.122863240
## politics          0.285508994  0.025625587  0.032529266 -0.011292710
## food              0.059687427  0.101791832  0.591318063  0.102454336
## family            0.149865803  0.030163186  0.452768549  0.118087188
## home_and_garden   0.071892898  0.104887976  0.094398311  0.073163163
## music             0.029748076  0.020209541  0.075348099  0.176970625
## news              0.555417450  0.041357011  0.022945910  0.017223284
## online_gaming     0.048316694  0.082891894  0.005744159  0.004527009
## shopping          0.115744356  0.032911113  0.007956775  0.068034649
## health_nutrition -0.023824999  0.031940868  0.010662213  0.006955340
## college_uni       0.039235338  0.090380839  0.027831267  0.019624167
## sports_playing    0.046380642  0.056542328  0.076103199  0.093295239
## cooking           0.019513651  0.055508896  0.033951321  0.664238946
## eco               0.059971307  0.066683922  0.099153848  0.047620711
## computers         0.013813021 -0.003641485  0.107529163  0.050833394
## business          0.040319930  0.085921419  0.075390623  0.087306971
## outdoors          0.088738605  0.006927819  0.042649098  0.041471406
## crafts            0.033971234  0.233534563  0.236400370  0.098714237
## automotive        1.000000000 -0.001711121  0.070620972  0.048793819
## art              -0.001711121  1.000000000  0.040783857  0.055029988
## religion          0.070620972  0.040783857  1.000000000  0.144553885
## beauty            0.048793819  0.055029988  0.144553885  1.000000000
## parenting         0.119122238  0.029719842  0.655597304  0.153620364
## dating            0.003453045  0.021181878  0.037641132  0.085459256
## school            0.102745787  0.071435572  0.516217988  0.188665957
## personal_fitness -0.009861229  0.030535005  0.031587379  0.024227535
## fashion           0.019730231  0.061987356  0.065365540  0.634973942
## small_business    0.033961568  0.154498608  0.061433981  0.091303432
## spam              0.008560904  0.024339269  0.008062977 -0.009135427
## adult             0.037330372  0.021976194  0.006347298  0.011082636
## adult2            0.031904315  0.024004843  0.003995430  0.000392486
## isSpam            0.009850430  0.026234619  0.009547183 -0.007965250
##                     parenting       dating        school personal_fitness
## chatter           0.020742516 0.1844524533  0.1166786249     0.0327477067
## current_events    0.050237162 0.0310458633  0.0680471275     0.0383886537
## travel            0.042341130 0.0867560304  0.0221995438    -0.0052992453
## photo_sharing     0.041169648 0.0285117208  0.1062092377     0.0625236821
## uncategorized     0.006268064 0.1270238084  0.0582599351     0.0847295855
## tv_film          -0.001788090 0.0041708935  0.0250222744    -0.0004447435
## sports_fandom     0.607718120 0.0169255299  0.4931061924     0.0142720768
## politics          0.044281427 0.0782884565  0.0277954990    -0.0080964757
## food              0.544948125 0.0334343361  0.4324039280     0.2238569249
## family            0.420578018 0.0177105901  0.3346895745     0.0381859420
## home_and_garden   0.080111889 0.1076172151  0.1000692908     0.0769581279
## music             0.044073427 0.0098874584  0.0518095525     0.0513248587
## news              0.078897406 0.0520657288  0.0527282215     0.0287393678
## online_gaming     0.022697341 0.0228324583 -0.0008595235     0.0119043235
## shopping          0.020248306 0.0040683633  0.0769180117     0.0544190810
## health_nutrition  0.029233239 0.0751908487  0.0066410255     0.8099023568
## college_uni       0.010366586 0.0240141334 -0.0055988726    -0.0215268678
## sports_playing    0.068607722 0.0946295042  0.0671382792     0.0514348348
## cooking           0.050678392 0.0290986583  0.0843002915     0.2336229315
## eco               0.114292458 0.0655328878  0.0900407104     0.2126135451
## computers         0.092961564 0.0948392286  0.0767260752     0.0218304964
## business          0.083133411 0.1118252662  0.1048685990     0.0581772096
## outdoors          0.061356734 0.0953960198  0.0428262486     0.5677902744
## crafts            0.208383762 0.0984524297  0.2179927962     0.0907547913
## automotive        0.119122238 0.0034530446  0.1027457872    -0.0098612290
## art               0.029719842 0.0211818783  0.0714355719     0.0305350048
## religion          0.655597304 0.0376411317  0.5162179885     0.0315873791
## beauty            0.153620364 0.0854592558  0.1886659575     0.0242275347
## parenting         1.000000000 0.0480947718  0.4996163935     0.0556830580
## dating            0.048094772  1.0000000000  0.2324943729     0.0751219669
## school            0.499616394  0.2324943729  1.0000000000     0.0289927455
## personal_fitness  0.055683058  0.0751219669  0.0289927455     1.0000000000
## fashion           0.070474628  0.1649797106  0.1465750199     0.0419365922
## small_business    0.051275509  0.0811375200  0.0892821248     0.0092225435
## spam              0.016097412 -0.0019265102  0.0062148420     0.0078760698
## adult             0.043453094  0.0040897011  0.0400706672     0.0081773126
## adult2            0.026777962 -0.0008768237  0.0294395372    -0.0015024887
## isSpam            0.014758297 -0.0007536582  0.0073124037     0.0096366079
##                      fashion small_business          spam        adult
## chatter           0.063314293    0.116730623  0.0046032416  0.015372712
## current_events    0.055906219    0.065497930  0.0194029891  0.016764279
## travel            0.026015648    0.116950707  0.0227733485  0.020244743
## photo_sharing     0.347248795    0.138112387 -0.0086648576 -0.012707084
## uncategorized     0.141408717    0.084832032  0.0138980529  0.045196686
## tv_film           0.017608868    0.188798241 -0.0042106462 -0.021700249
## sports_fandom     0.030767287    0.048662236  0.0089574643  0.007986164
## politics         -0.006793131    0.105261525  0.0094389223 -0.027102370
## food              0.036963917    0.063499437  0.0014827661  0.017630551
## family            0.062704099    0.062871670 -0.0068027516  0.034583067
## home_and_garden   0.076120516    0.080158027  0.0216008547  0.039393795
## music             0.160443645    0.122016216  0.0005284301 -0.004238748
## news              0.002132337    0.047598846 -0.0018000119 -0.011204031
## online_gaming     0.036013335    0.041962929  0.0058670252  0.012578776
## shopping          0.096854665    0.112662828 -0.0192516162 -0.017128756
## health_nutrition  0.029747956   -0.009302541  0.0023958243 -0.009092548
## college_uni       0.052023538    0.113883331  0.0094529848 -0.011708630
## sports_playing    0.108221244    0.079653227 -0.0087425556 -0.011257322
## cooking           0.721402674    0.072756188 -0.0061781291 -0.004556427
## eco               0.063817437    0.076504123  0.0314152819  0.060212751
## computers         0.063047239    0.104650495  0.0231174511  0.030290997
## business          0.109494748    0.107021298 -0.0254918495 -0.010963774
## outdoors          0.055778912    0.008170806  0.0215189501  0.071036168
## crafts            0.085546453    0.091084342  0.0199397644  0.031016108
## automotive        0.019730231    0.033961568  0.0085609039  0.037330372
## art               0.061987356    0.154498608  0.0243392694  0.021976194
## religion          0.065365540    0.061433981  0.0080629769  0.006347298
## beauty            0.634973942    0.091303432 -0.0091354265  0.011082636
## parenting         0.070474628    0.051275509  0.0160974121  0.043453094
## dating            0.164979711    0.081137520 -0.0019265102  0.004089701
## school            0.146575020    0.089282125  0.0062148420  0.040070667
## personal_fitness  0.041936592    0.009222543  0.0078760698  0.008177313
## fashion           1.000000000    0.092180469 -0.0031873217  0.003439968
## small_business    0.092180469    1.000000000  0.0316621449  0.082381865
## spam             -0.003187322    0.031662145  1.0000000000  0.294399003
## adult             0.003439968    0.082381865  0.2943990032  1.000000000
## adult2           -0.008050178    0.058078616  0.2606201610  0.796639822
## isSpam           -0.001617529    0.024859345  0.9822983553  0.296632358
##                         adult2        isSpam
## chatter           0.0045401295  5.699655e-03
## current_events    0.0150678558  2.189785e-02
## travel            0.0161922498  2.283752e-02
## photo_sharing    -0.0229843610 -7.175568e-03
## uncategorized     0.0374753359  8.906237e-03
## tv_film          -0.0236437628 -9.190392e-03
## sports_fandom    -0.0008130612  1.112556e-02
## politics         -0.0288507927  1.190631e-02
## food              0.0232939826  3.202982e-03
## family            0.0292367942 -4.745485e-03
## home_and_garden   0.0320586410  1.859592e-02
## music            -0.0252932792  1.121808e-03
## news             -0.0168246429 -5.459293e-05
## online_gaming     0.0102013912  7.068059e-03
## shopping         -0.0262526328 -1.881112e-02
## health_nutrition -0.0108218821  4.022934e-03
## college_uni      -0.0196480567  1.007126e-02
## sports_playing   -0.0187461185 -8.802764e-03
## cooking          -0.0114237291 -4.665331e-03
## eco               0.0547218335  3.543554e-02
## computers         0.0133020909  2.353404e-02
## business         -0.0072507816 -2.736837e-02
## outdoors          0.0436925122  2.355541e-02
## crafts            0.0227647032  1.723796e-02
## automotive        0.0319043146  9.850430e-03
## art               0.0240048432  2.623462e-02
## religion          0.0039954302  9.547183e-03
## beauty            0.0003924860 -7.965250e-03
## parenting         0.0267779617  1.475830e-02
## dating           -0.0008768237 -7.536582e-04
## school            0.0294395372  7.312404e-03
## personal_fitness -0.0015024887  9.636608e-03
## fashion          -0.0080501782 -1.617529e-03
## small_business    0.0580786159  2.485935e-02
## spam              0.2606201610  9.822984e-01
## adult             0.7966398218  2.966324e-01
## adult2            1.0000000000  2.645835e-01
## isSpam            0.2645835411  1.000000e+00
```

The most obvous clusters of highly correlated variables ($r > 0.5$) include (chatter, photo_sharing, shopping) and (online gaming, sports playing, university). The cluster between chatter and photo_sharing seems obvious, but between those two and shopping not immediately so. The cluster between online gaming, sports playing, and university may be due to the fact that playing sports and online gaming are more popular with university students, i.e. young people.

To summarize, we observed the difference in means of other variables in the Social Marketing dataset by whether or not the datapoint is marked isSpam (spam >= 1) and as adult content (adult >= 1). We saw no obvious correlation other than the fact that spam tends to strongly imply adult content is present (94% vs 7% for non-spam data). We then ran a logistic regression with isSpam and observed that the presence of business content may be a significant indicator of a lower likelihood of spam, as can be seen in the table of coefficients above. Furthermore, just the presence of adult content itself was a far better indicator of spam than the actual amount of adult content. Finally, when analyzing clusters of highly correlated variables, we discovered that online gaming and sports may be more popular among young people, including those in universities. This should be taken into account when young people are part of the target audience.