Final project

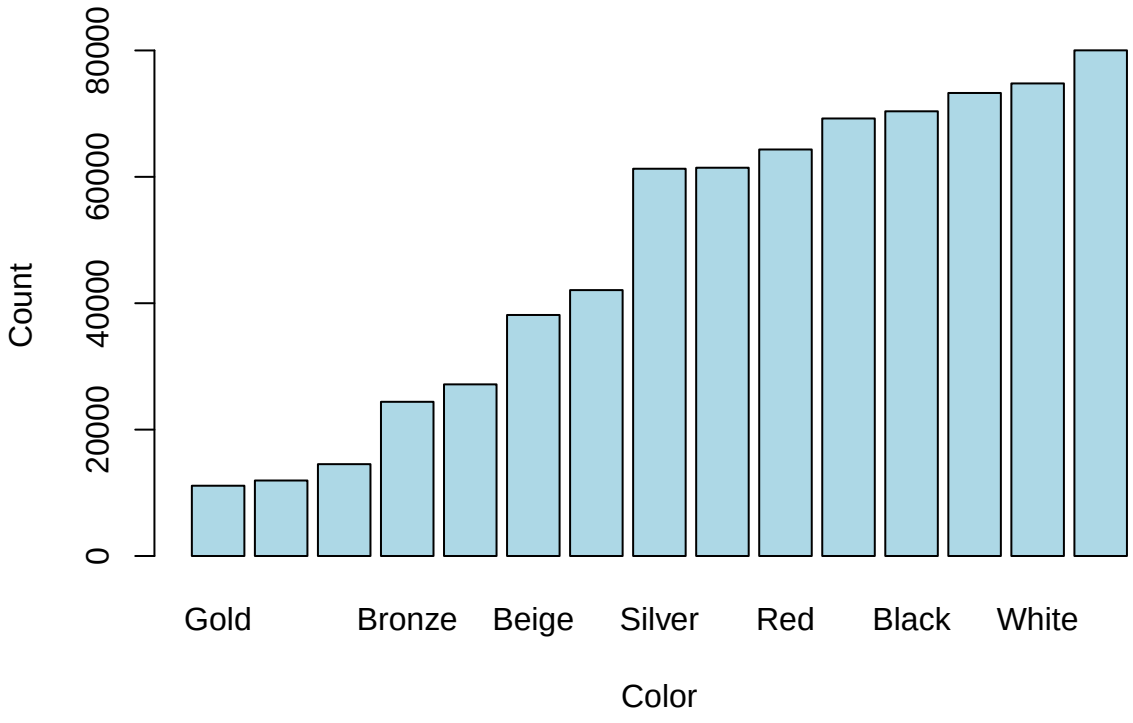
2023-11-23

Question 1: Use the cars_big.csv to an external site. dataset to build a predictive model pricing cars based on its features

Step 1:Data Collection

1. Using cars_big data set that includes observations for the dependent and independent variables.
2. Identify the dependent variable (price) and the independent variables (all other numeric variables).

Step 2: Data Exploration & Visualization



Step 3: Data cleaning & feature selection

Explore and clean the dataset.Check for missing values, outliers, and other anomalies.

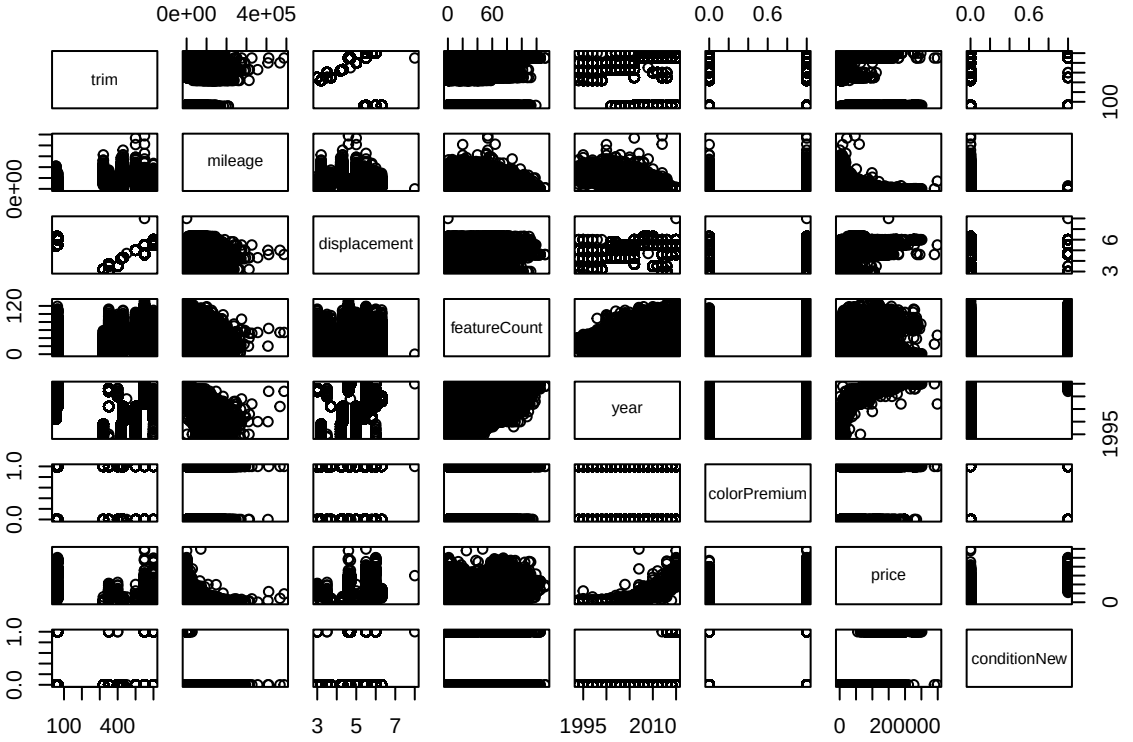
Assumptions Check:Verify that the assumptions of multiple linear regression are met, including linearity, independence, homoscedasticity, and normality of residuals.

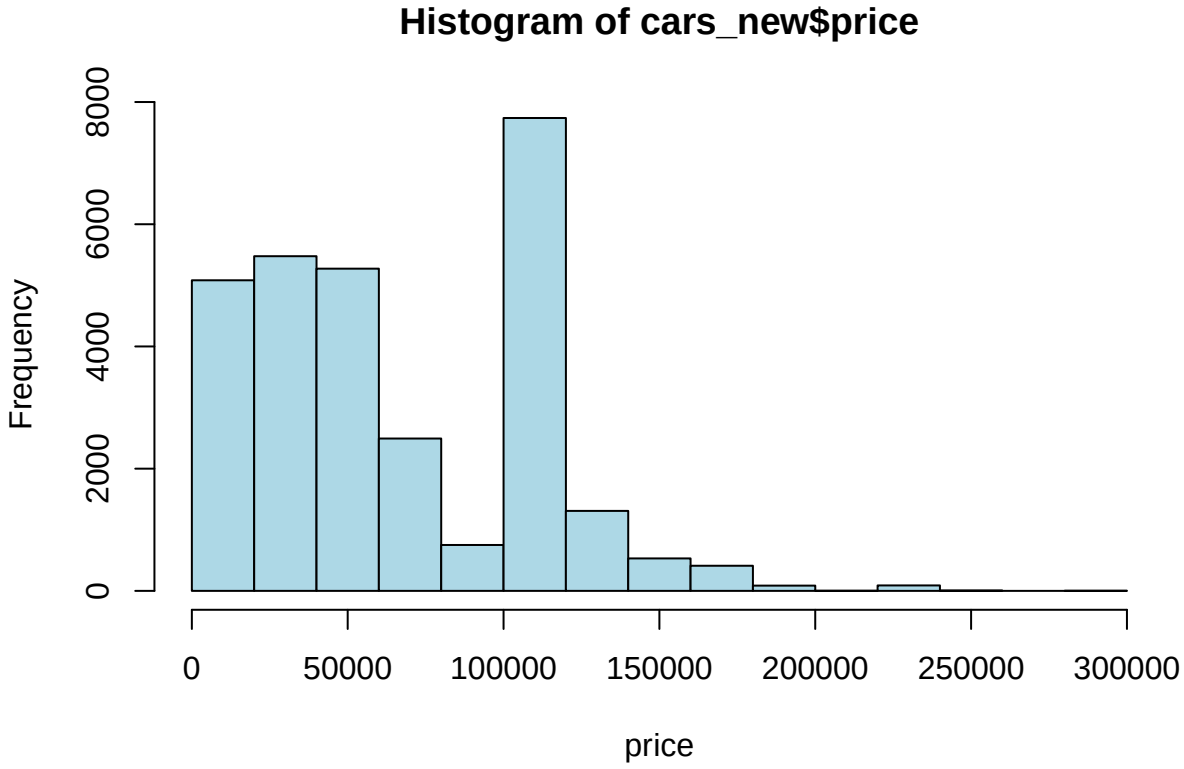
```
## Warning: NAs introduced by coercion
```

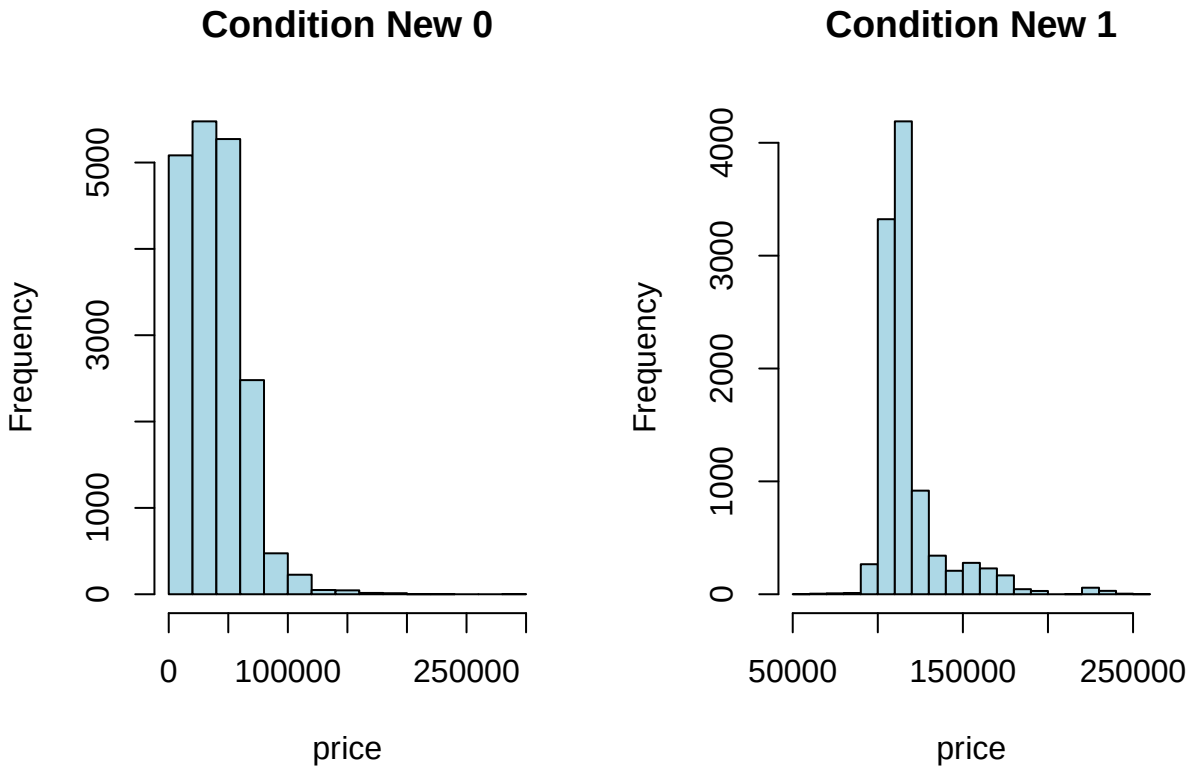
```
## Warning: NAs introduced by coercion
```

```
##  [1] "trim"                          "isOneOwner"
##  [3] "mileage"                       "year"
##  [5] "displacement"                  "featureCount"
##  [7] "price"                         "colorPremium"
##  [9] "conditionNew"                  "conditionUsed"
## [11] "regionESC"                     "regionMid"
## [13] "regionMtn"                     "regionNew"
## [15] "regionPac"                     "regionSoA"
## [17] "regionunsp"                    "regionWNC"
## [19] "regionWSC"                     "soundSystemBang Olufsen"
## [21] "soundSystemBose"               "soundSystemBoston Acoustic"
## [23] "soundSystemHarman Kardon"      "soundSystemPremium"
## [25] "soundSystemunsp"               "wheelTypeChrome"
## [27] "wheelTypePremium"              "wheelTypeSteel"
## [29] "wheelTypeunsp"
```

Step 4: Data visualization



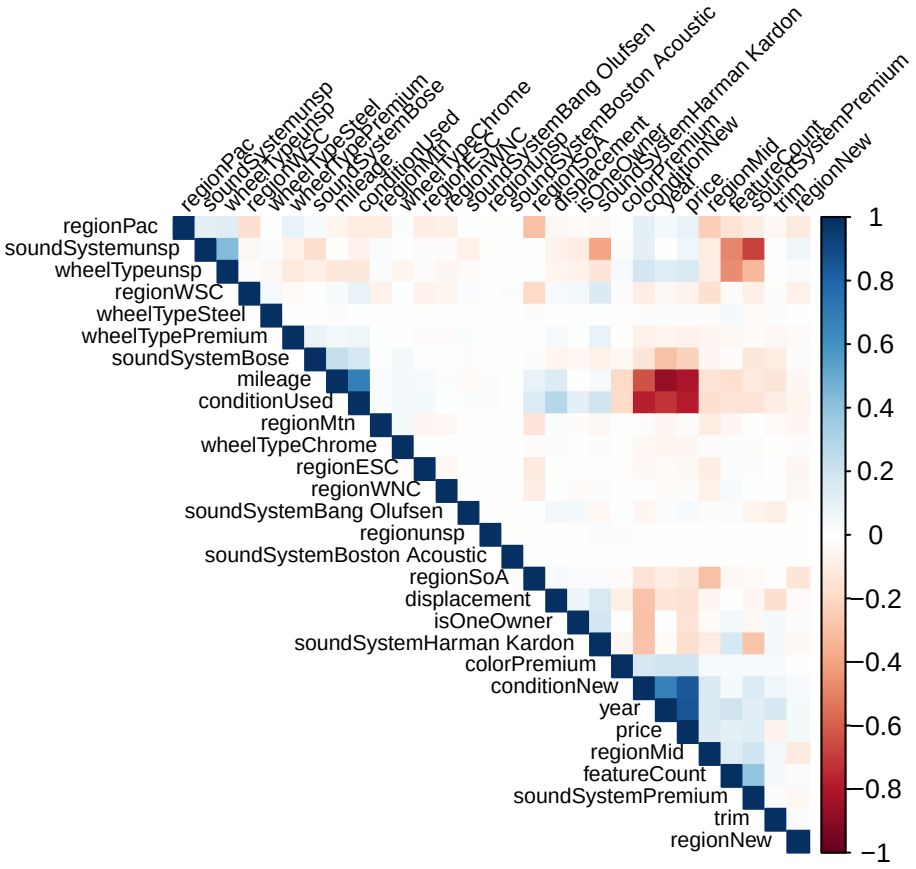




Interpretation: Price of new cars is higher compared to the used cars.

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```



Interpretation:

1. Price is strongly positively co-related with year, positively correlated with cars that are new in condition and negatively correlated with cars that are used. There is negative correlation with mileage & region_MTN.

2. There is no clear linear relationship observed between features & price, rather a more exponential relationship is observed. Linear regression model may not be the best fit model but is simple to interpret. Hence we create a simple linear model with the numeric features above.

Step 5: Multiple linear regression model

```
##
## Call:
## lm(formula = price ~ ., data = cars_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54778  -6348  -1859   3748 269116
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 -7.603e+06  7.535e+04 -100.908  < 2e-16 ***
## trim                        -6.372e+01  6.206e-01 -102.672  < 2e-16 ***
## isOneOwner                   3.826e+01  2.361e+02    0.162 0.871294
```

```
## mileage                       -1.568e-01  3.401e-03  -46.122  < 2e-16 ***
## year                           3.825e+03  3.730e+01  102.535  < 2e-16 ***
## displacement                   2.711e+03  1.465e+02   18.501  < 2e-16 ***
## featureCount                  -9.937e+00  3.304e+00   -3.007 0.002637 **
## colorPremium                   9.013e+02  2.749e+02    3.279 0.001044 **
## conditionNew                   4.019e+04  2.935e+02  136.962  < 2e-16 ***
## conditionUsed                 -6.750e+03  2.702e+02  -24.977  < 2e-16 ***
## regionESC                      1.005e+03  4.654e+02    2.159 0.030855 *
## regionMid                     -6.423e+02  3.110e+02   -2.065 0.038908 *
## regionMtn                      1.692e+03  4.330e+02    3.908 9.33e-05 ***
## regionNew                      2.978e+02  4.282e+02    0.695 0.486790
## regionPac                      1.742e+02  3.122e+02    0.558 0.576968
## regionSoA                     -8.711e+01  2.957e+02   -0.295 0.768337
## regionunsp                     3.736e+03  5.726e+03    0.653 0.514062
## regionWNC                      1.972e+03  5.355e+02    3.683 0.000231 ***
## regionWSC                      8.510e+02  3.542e+02    2.403 0.016277 *
## `soundSystemBang Olufsen`      7.598e+03  9.097e+03    0.835 0.403609
## soundSystemBose               -7.541e+03  9.053e+03   -0.833 0.404812
## `soundSystemBoston Acoustic` -1.111e+04  1.566e+04   -0.709 0.478242
## `soundSystemHarman Kardon`    -7.216e+03  9.046e+03   -0.798 0.425065
## soundSystemPremium            -4.559e+03  9.044e+03   -0.504 0.614238
## soundSystemunsp               -4.219e+03  9.044e+03   -0.466 0.640877
## wheelTypeChrome                8.101e+02  1.439e+03    0.563 0.573486
## wheelTypePremium              -7.565e+02  6.432e+02   -1.176 0.239537
## wheelTypeSteel                 1.816e+04  1.837e+03    9.887  < 2e-16 ***
## wheelTypeunsp                 -1.649e+02  1.847e+02   -0.892 0.372141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12780 on 29223 degrees of freedom
## Multiple R-squared:  0.9167, Adjusted R-squared:  0.9167
## F-statistic: 1.149e+04 on 28 and 29223 DF,  p-value: < 2.2e-16
```

Interpretation:

1. R-square is 91% which means our model is able to explain 91% of the variance in the data. We may be adding noise and we can check with the residual error from the test set.
2. Relationship of price is statistically significant with trim, year, displacement, milage, condition new & used, color Grsy/Silver, region MTN,WNC,Wheel type steel. This would be used in feature selection.

Step 5: multiple linear regression with selected features based on the significance as well as using a few interaction terms

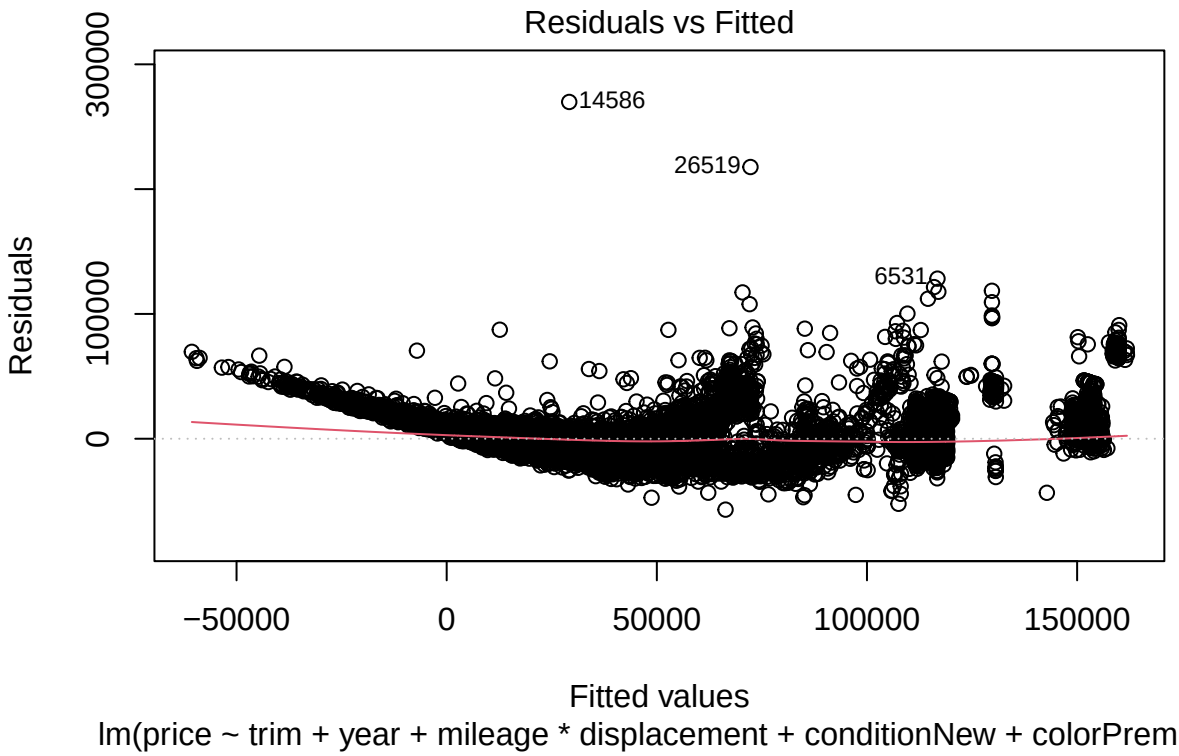
```
##
## Call:
## lm(formula = price ~ trim + year + mileage * displacement + conditionNew +
##     colorPremium + regionMtn + regionWNC + wheelTypeSteel + featureCount,
##     data = cars_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56797  -6986  -1821   4077 269839
```

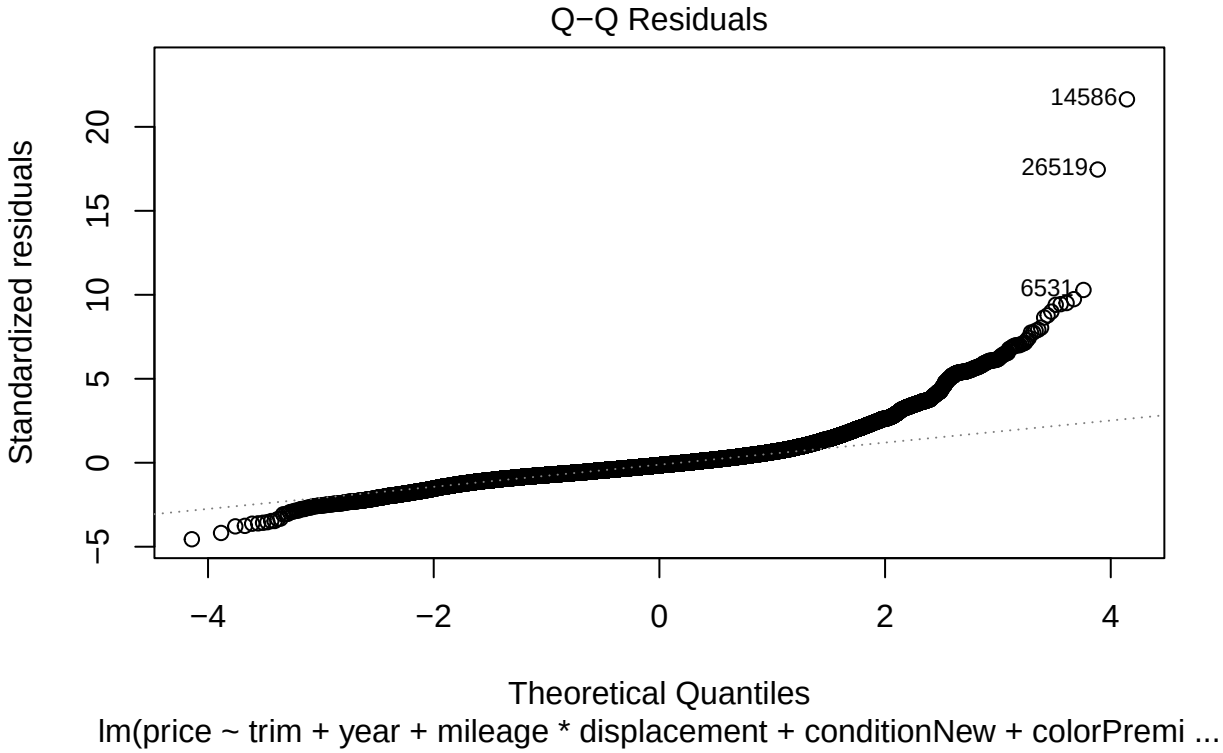
```
##
## Coefficients:
##                             Estimate      Std. Error  t value
## (Intercept)          -8694774.104752    70740.817696 -122.910
## trim                       -54.621780        0.633463  -86.227
## year                      4338.151073       35.114781  123.542
## mileage                      0.650932        0.016446   39.581
## displacement             11891.218263      238.757811   49.805
## conditionNew             44153.860853      219.825690  200.859
## colorPremium              1365.087811      267.797345    5.097
## regionMtn                 1944.684094      347.368744    5.598
## regionWNC                 2283.151176      464.981489    4.910
## wheelTypeSteel           19131.248042     1786.408423   10.709
## featureCount               -12.787415        2.573534   -4.969
## mileage:displacement        -0.166511        0.003282  -50.737
##                                  Pr(>|t|)
## (Intercept)          < 0.0000000000000002 ***
## trim                 < 0.0000000000000002 ***
## year                 < 0.0000000000000002 ***
## mileage              < 0.0000000000000002 ***
## displacement         < 0.0000000000000002 ***
## conditionNew         < 0.0000000000000002 ***
## colorPremium                 0.0000003464 ***
## regionMtn                    0.0000000218 ***
## regionWNC                    0.0000009147 ***
## wheelTypeSteel       < 0.0000000000000002 ***
## featureCount                 0.0000006774 ***
## mileage:displacement < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12470 on 29240 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9207
## F-statistic: 3.086e+04 on 11 and 29240 DF,  p-value: < 0.00000000000000022
```

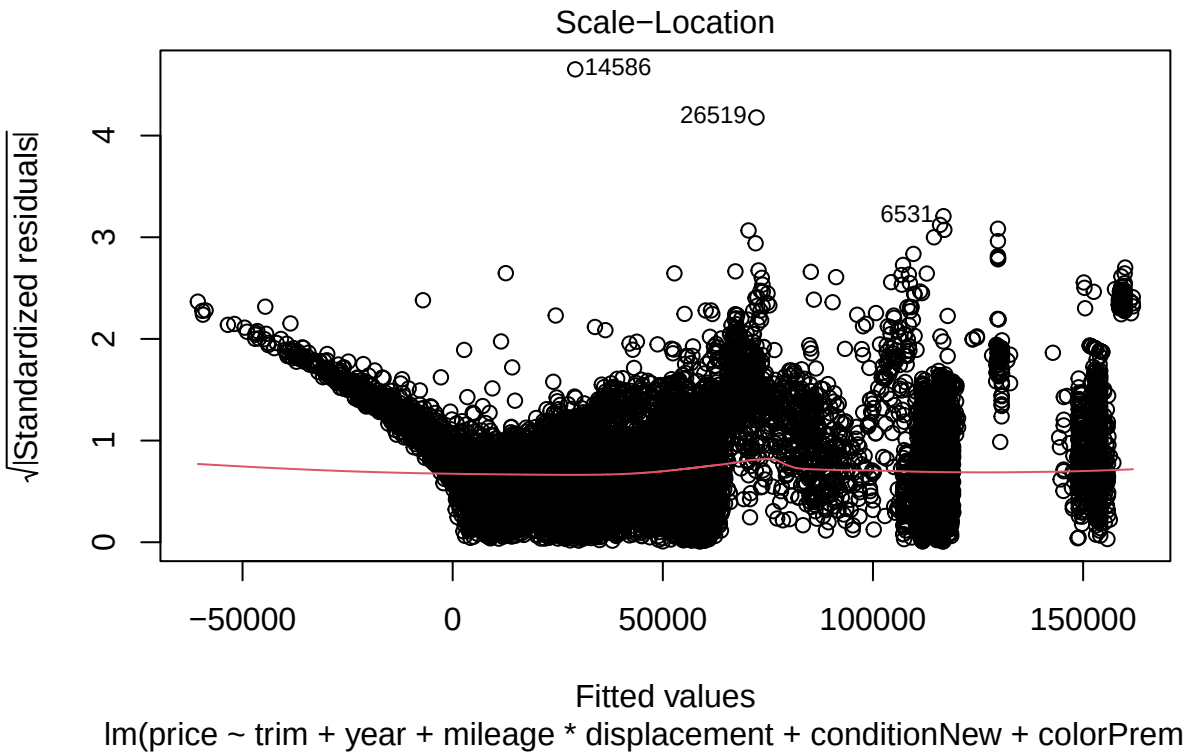
Interpretation

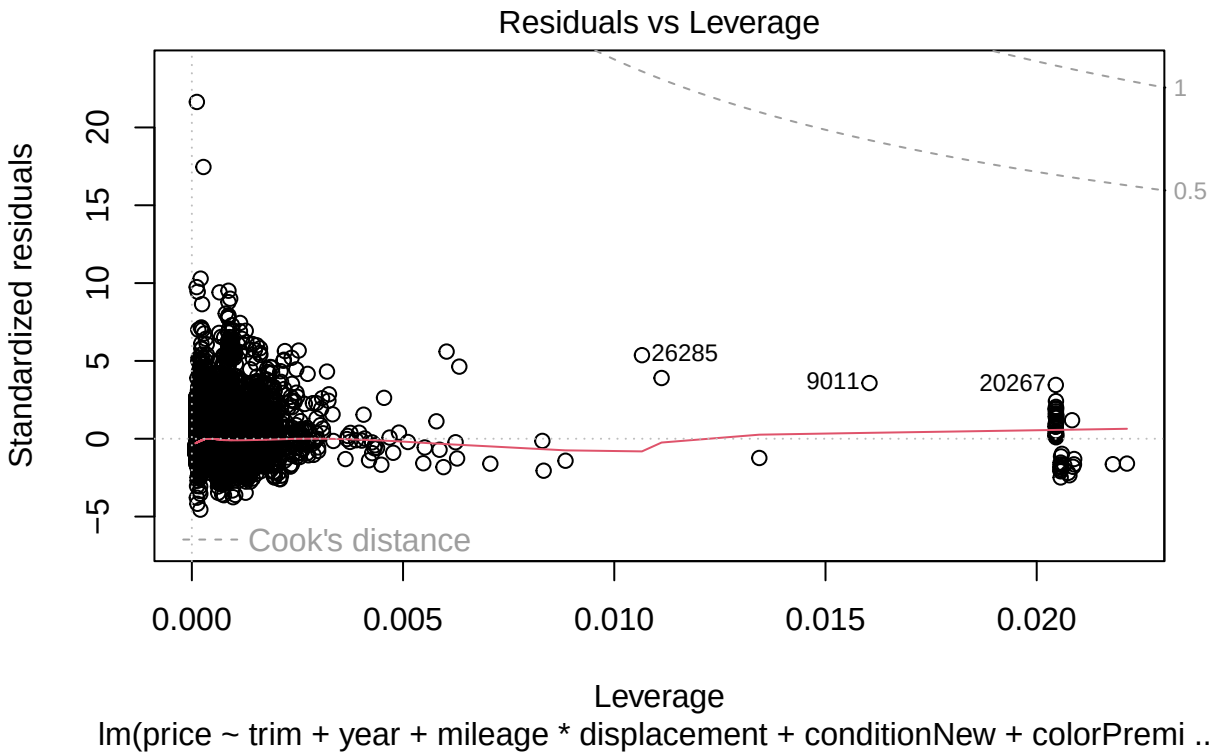
1. We are able to get 92% of the variance with feature selection and keeping only the important ones.
2. Increase of 1 year leads to an increase in price by 4320 unit
3. A new car has an increase of 44213 unit in price compared to the used car
4. A premium color car has a increase of an avg 1365 unit compared to other colors
5. A car with wheel has a price increase of 18696 compared to the other wheel types.

#Step 5: Residual plot to see if linear model is a good fit.


Residuals vs Fitted
Residuals
300000
100000
0
14586
26519
6531
−50000
0
50000
100000
150000
Fitted values
lm(price ~ trim + year + mileage * displacement + conditionNew + colorPremi ...


Q−Q Residuals
Standardized residuals
−5
0
5
10
15
20
14586
26519
6531
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ trim + year + mileage * displacement + conditionNew + colorPremi ...


Scale−Location
14586
26519
6531
√|Standardized residuals|
0
1
2
3
4
−50000
0
50000
100000
150000
Fitted values
lm(price ~ trim + year + mileage * displacement + conditionNew + colorPremi ...



Interpretation The residual error is not normally distributed around the horizontal line. However there is no a curve/pattern in the distribution suggestion linear regression. There are also large outliers in the residual plot. While linear model is not the best fit model, it is able to explain the variance and is simple to interpret.

Step 6: Testing the model on the test data to evaluate the performance of the model

Split Data:Divide the dataset into two parts: a training set and a testing (or validation) set. The training set is used to train the model, and the testing set is used to evaluate its performance.Use the training set to fit the multiple linear regression model.Use the testing set to make predictions and evaluate the model's performance on unseen data.

```
## [1] 12652.56
```

Interpretation

The square root mean error is 12652 units, which is a reduction in the test error compared to the model with all the features. We will further try to reduce the test error by penalizing the complexity.

Further optimizing the model to introduce polynominal interaction as the relationship between price & year is not linear.

```
## [1] 10895.15
```

Interpretation

The square root mean error is reduced to 10895 units, which is a reduction in the test error compared to the model without polynomial features. .

Question 2: Your task is to analyze this data as you see fit and to prepare a report for NutrientH20. Identify market segments that appear to stand out in their social-media audience.

We do this by using association graph and K-Mean clsutering

Model 1: Using Graph Network for the tweets and segmenting customers based on the similar tweets pattern

##we are categorizing the users by taking the top 3 tweets based on the number of tweets for each theme. We then find association between tweets and communities classes to draw interesting insights for e.g.

Interesting patterns:

1. Which tweets are most influential - we can segment the most influential users and incentivise them for social media marketing
2. Segmenting the users based on the similar tweets/associated tweets based on the degree. This could tell us if a person is tweeting on outdoor, they are most likely to tweet on personal_fitness. So, we can send them targeted ads based on the close association.

Step 1: Reading the dataset.

Dataset is clean with no missing values.

Step 2: Applying Apriori model for networks

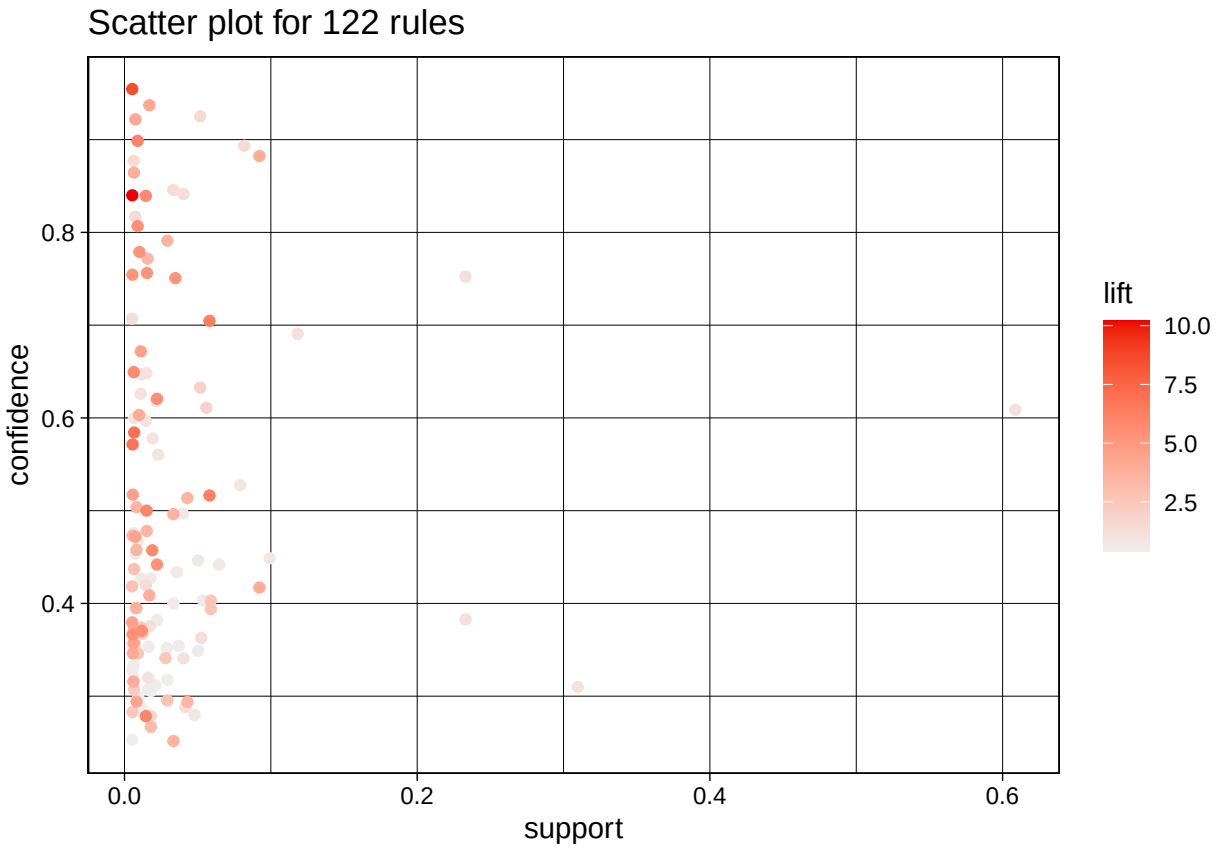
```
## transactions as itemMatrix in sparse format with
##  7882 rows (elements/itemsets/transactions) and
##  35 columns (items) and a density of 0.08571429
##
## most frequent items:
##          chatter    photo_sharing health_nutrition   current_events
##             4798             2442             1743             1351
##           travel          (Other)
##             1181            12131
##
## element (itemset/transaction) length distribution:
## sizes
##    3
## 7882
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3       3       3       3       3       3
```

```
##
## includes extended item information - examples:
##       labels
## 1      adult
## 2        art
## 3 automotive
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.25    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       3  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 39
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[35 item(s), 7882 transaction(s)] done [0.00s].
## sorting and recoding items ... [31 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [122 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
## set of 122 rules
##
## rule length distribution (lhs + rhs):sizes
##  1  2  3
##  2 62 58
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.459   3.000   3.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.005202   Min.   :0.2517   Min.   :0.005582   Min.   : 0.4158
##  1st Qu.:0.007390   1st Qu.:0.3537   1st Qu.:0.017286   1st Qu.: 0.9075
##  Median :0.015288   Median :0.4419   Median :0.033431   Median : 2.0066
##  Mean   :0.034308   Mean   :0.5043   Mean   :0.074204   Mean   : 2.6710
##  3rd Qu.:0.035524   3rd Qu.:0.6246   3rd Qu.:0.082403   3rd Qu.: 4.0160
##  Max.   :0.608729   Max.   :0.9545   Max.   :1.000000   Max.   :10.1860
##      count
##  Min.   :  41.00
##  1st Qu.:  58.25
##  Median : 120.50
##  Mean   : 270.42
##  3rd Qu.: 280.00
##  Max.   :4798.00
##
## mining info:
```

```
##          data ntransactions support confidence
##  social_trans          7882   0.005       0.25
##                                                                                call
##  apriori(data = social_trans, parameter = list(support = 0.005, confidence = 0.25, maxlen = 3))
```



Interpreation: Please see the graph in Gephi (attahed image)

color: Degree Size: Betweenness Centrality

1. Tweets with most edges/degree are: Cooking, photo_sharing, food, religion, sports_phandon & politics.
2. The most influential tweet theme is sports_fandom.
3. Detecting community using modularity class, we can clearly see communities of

a) Fitness based coomunities outdoor, personal_fitness, health_nutition
b) Culture based communities like sport_fandom, food, religion, parenting
c) Global communities like politics, travel, news, automative, computers
d) Generation based communities - college_uni, sports_playing, online-gaming, music, tv, art
e) Product based community - beauty, cooking, shopping, photo_sharing

3. We can see chatter is not influential node and is associated with all the themes.

Based on the classes, we can extract the user details to target them for marketing promotions/recommendations.

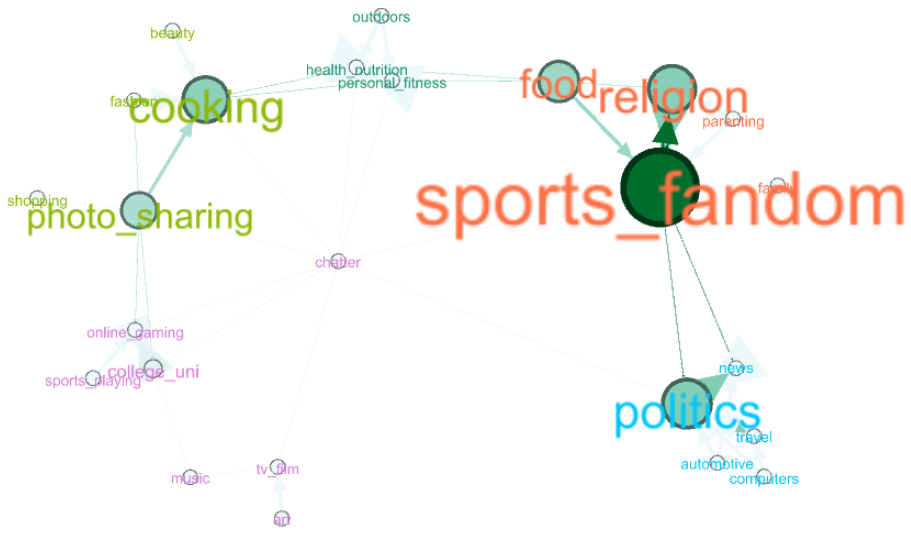


Figure 1: Social Tweets graph

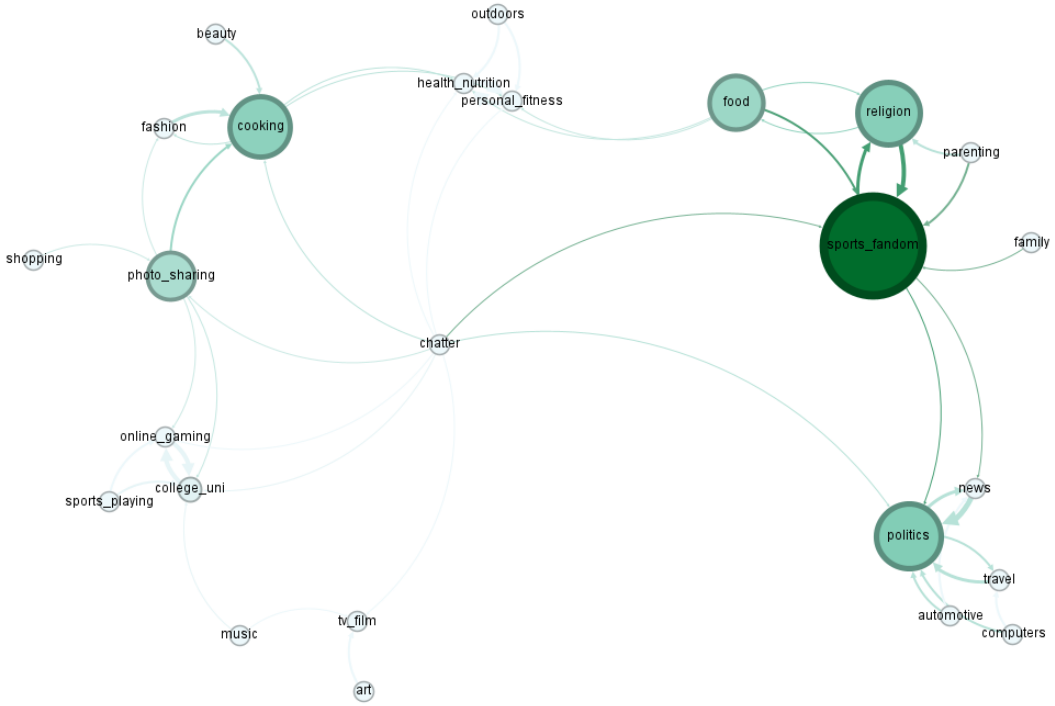
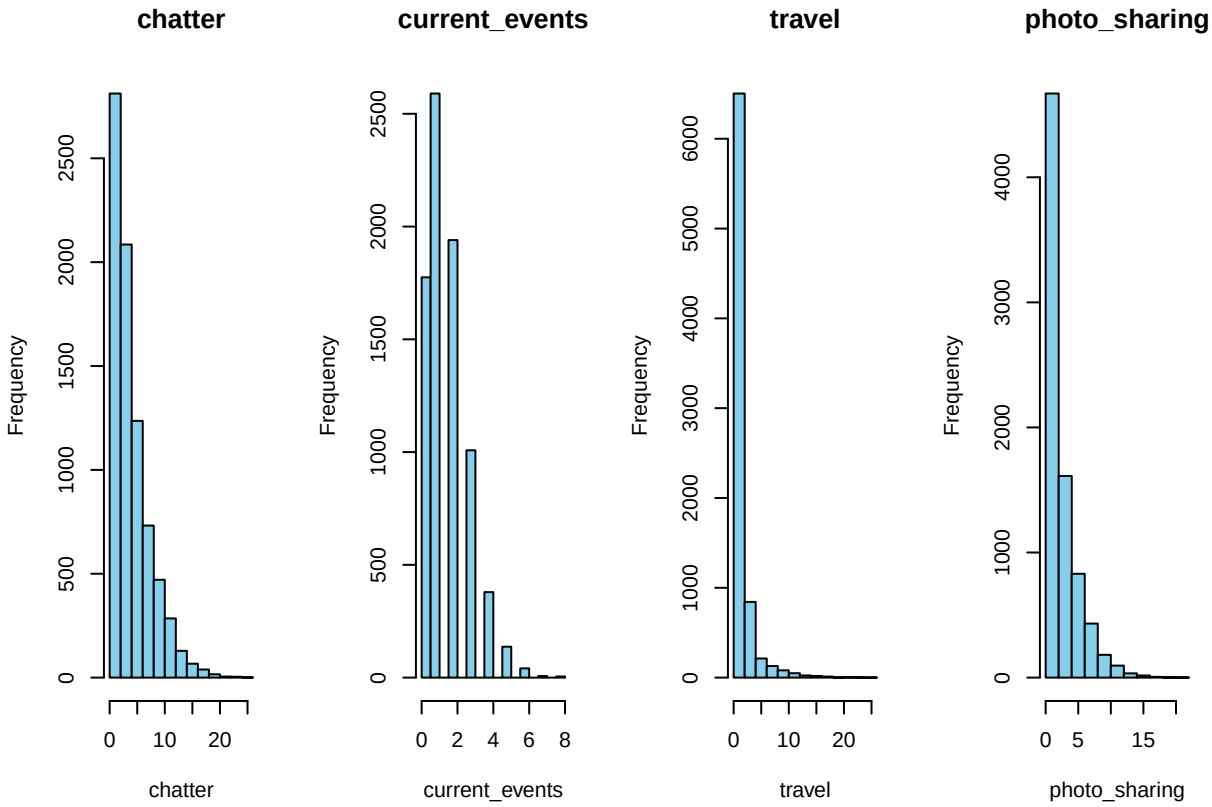
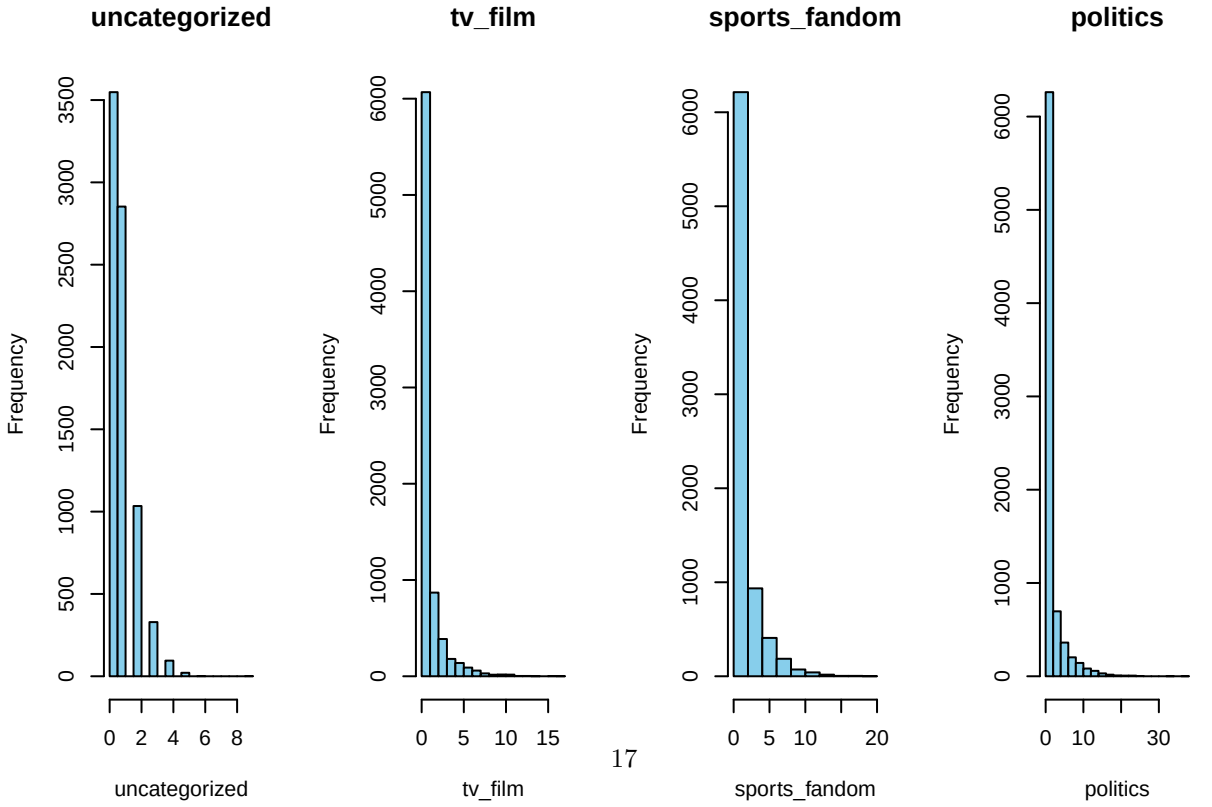


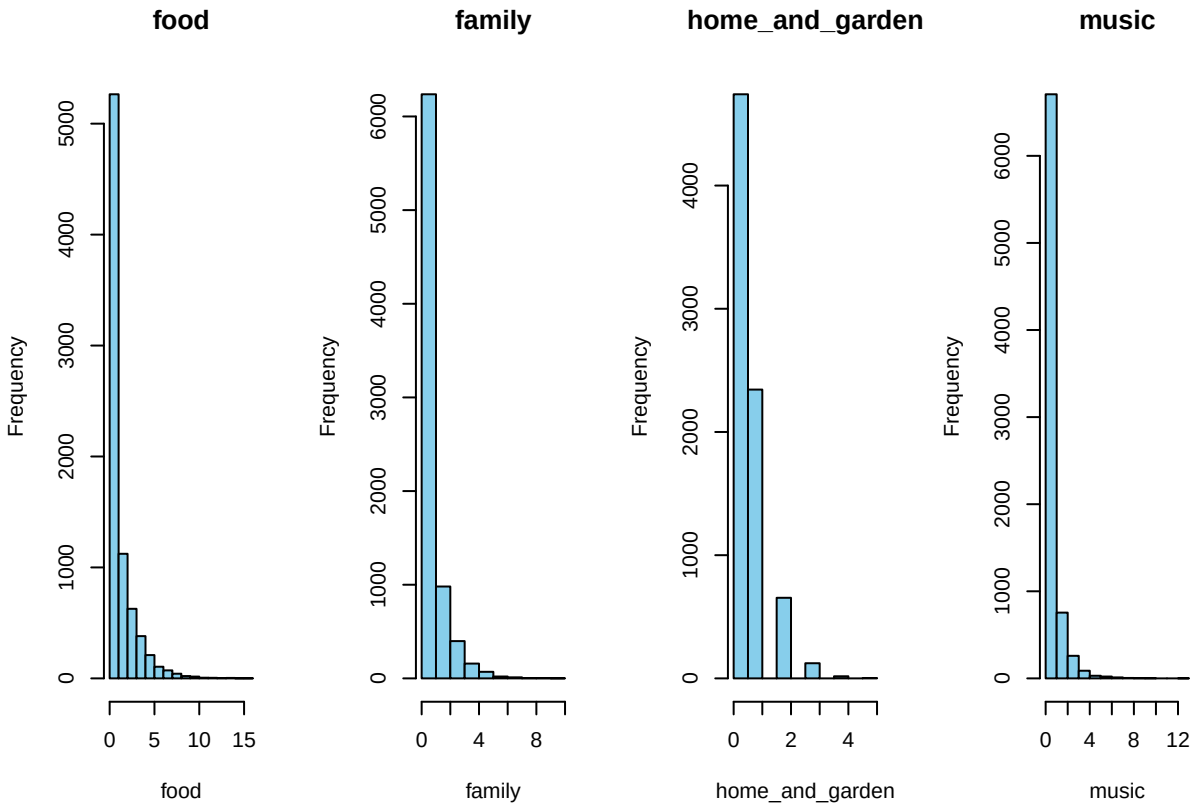
Figure 2: Social Tweets graph

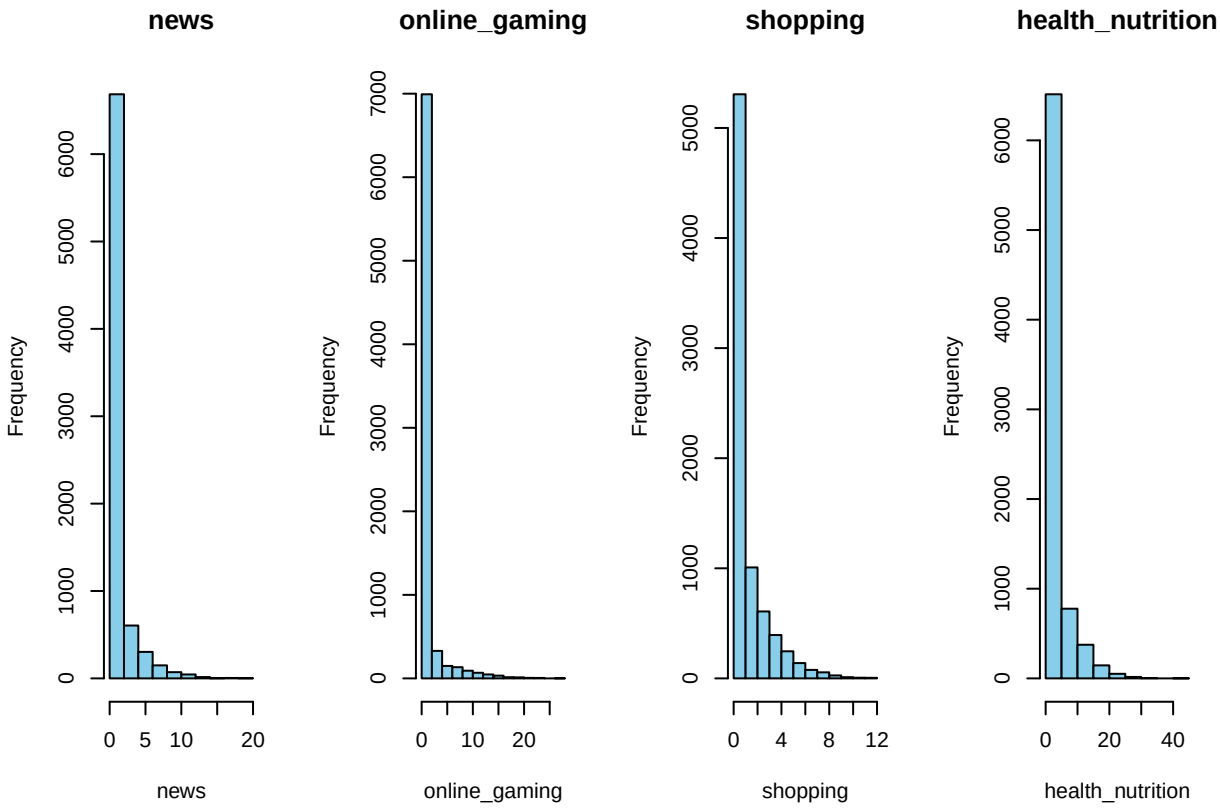
Model 2: K- Means Clustering

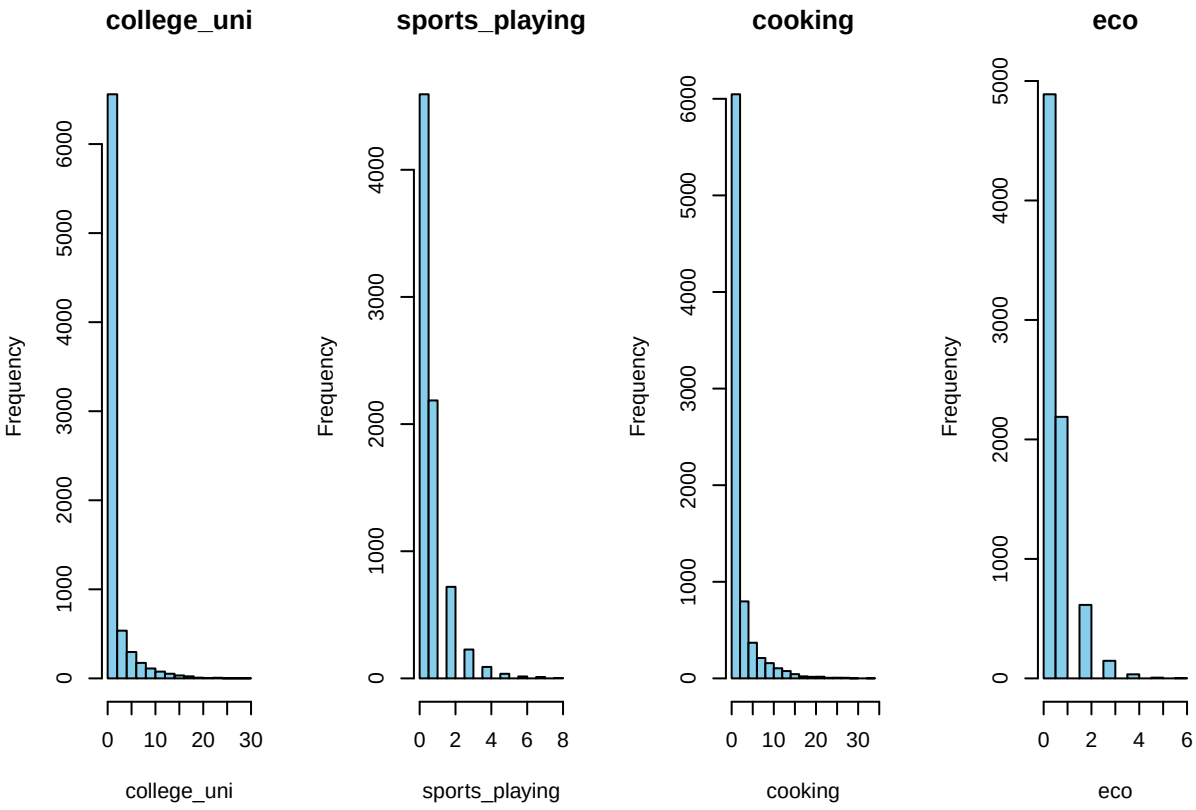
Step 1: Data visualization/distribution

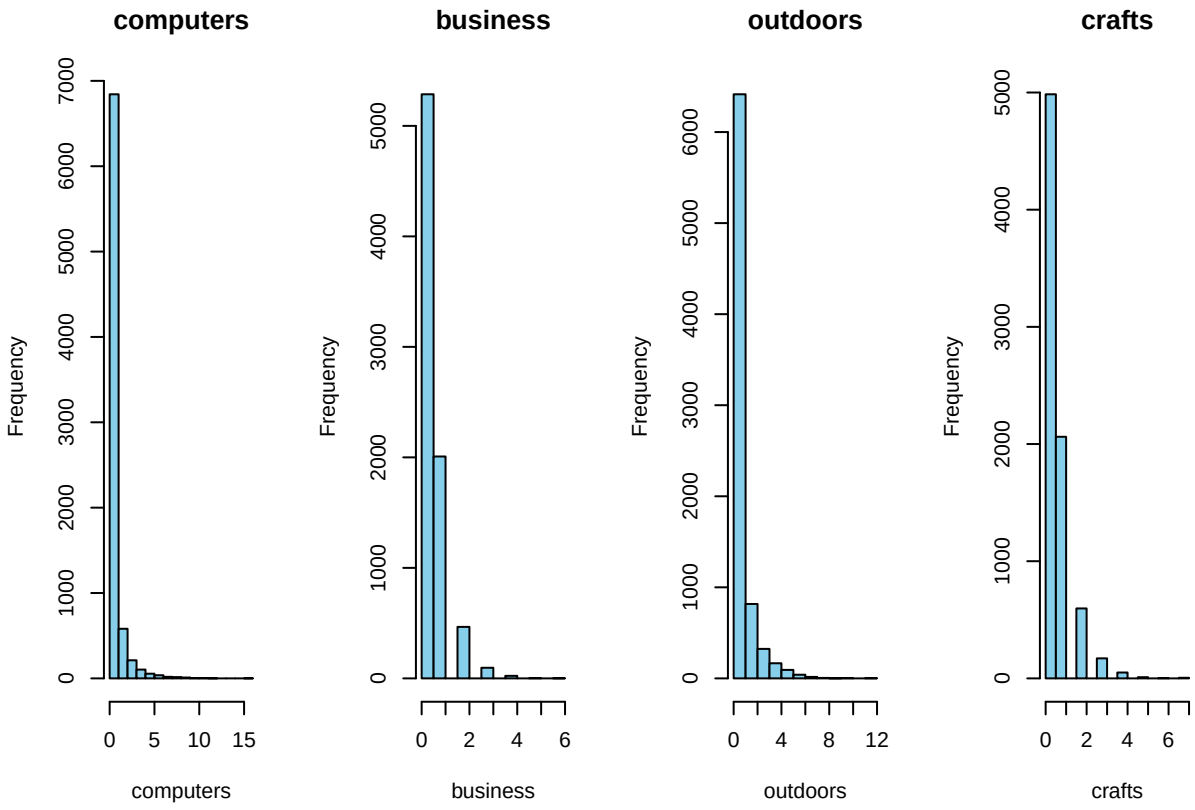


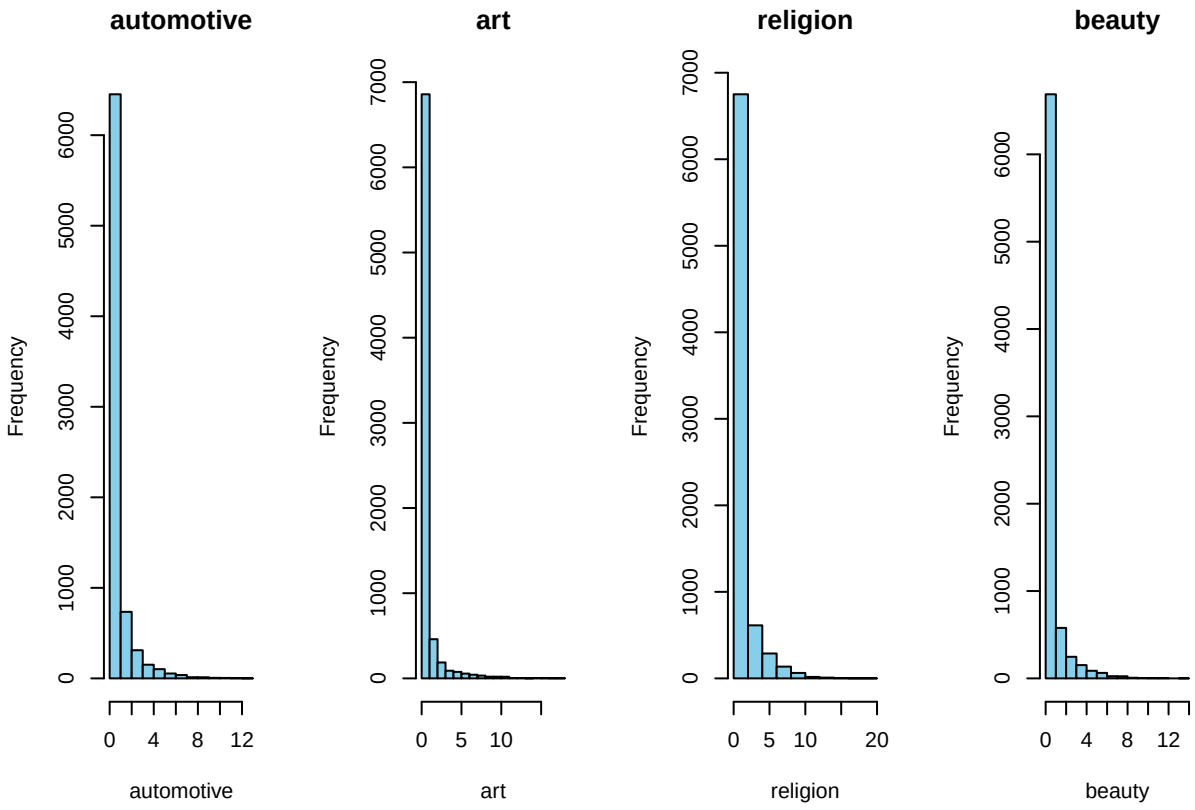


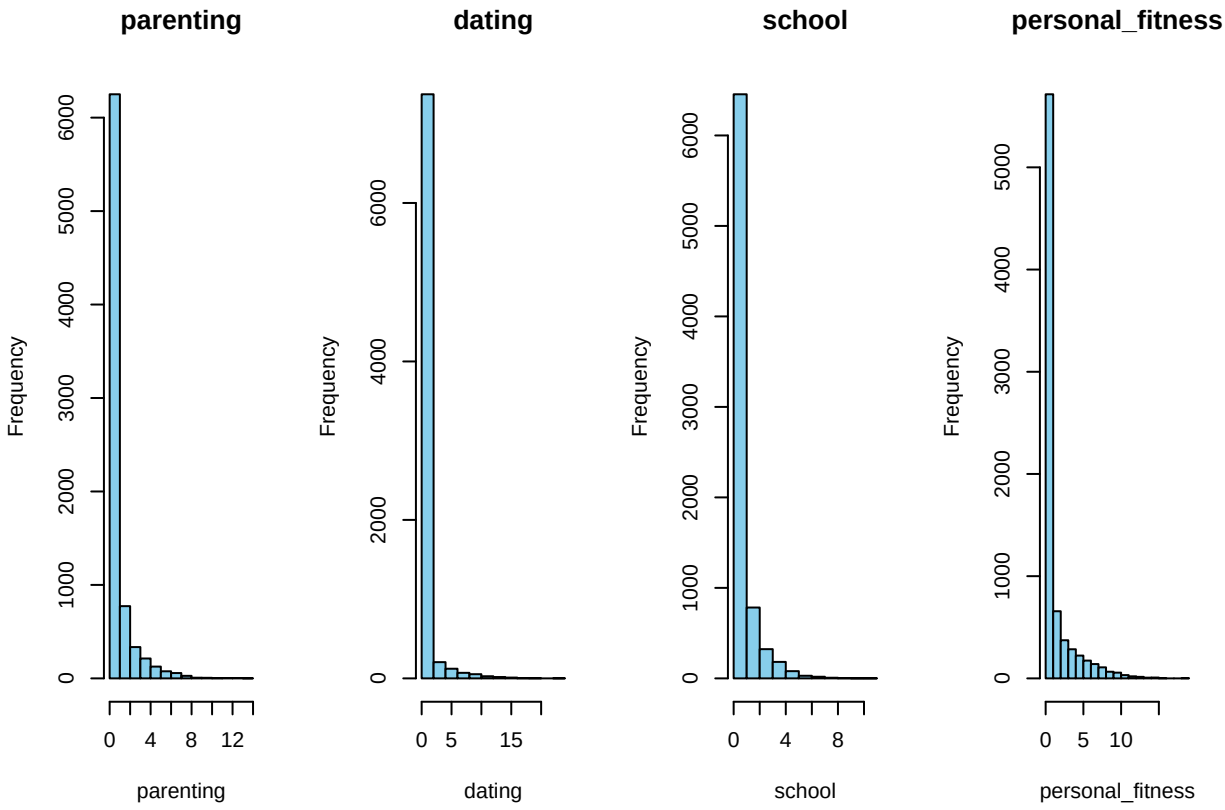

food
family
home_and_garden
music
Frequency
0
1000
2000
3000
4000
5000
6000
0
5
10
15
food
0
4
8
family
0
2
4
home_and_garden
0
4
8
12
music

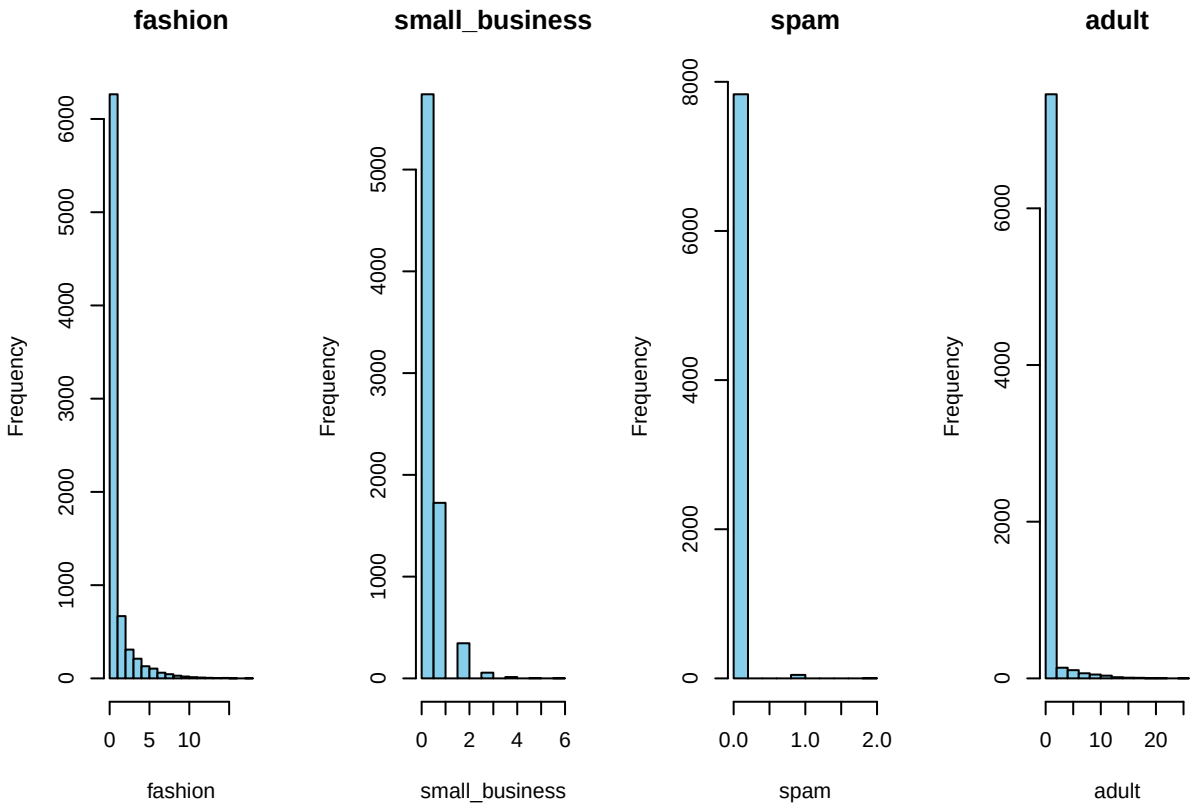

news
Frequency
0
1000
2000
3000
4000
5000
6000
0
5
10
20
news
online_gaming
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
10
20
online_gaming
shopping
Frequency
0
1000
2000
3000
4000
5000
0
4
8
12
shopping
health_nutrition
Frequency
0
1000
2000
3000
4000
5000
6000
0
20
40
health_nutrition


college_uni
Frequency
0
1000
2000
3000
4000
5000
6000
0
10
20
30
college_uni
sports_playing
Frequency
0
1000
2000
3000
4000
0
2
4
6
8
sports_playing
cooking
Frequency
0
1000
2000
3000
4000
5000
6000
0
10
20
30
cooking
eco
Frequency
0
1000
2000
3000
4000
5000
0
2
4
6
eco


computers
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
5
10
15
computers
business
Frequency
0
1000
2000
3000
4000
5000
0
2
4
6
business
outdoors
Frequency
0
1000
2000
3000
4000
5000
6000
0
4
8
12
outdoors
crafts
Frequency
0
1000
2000
3000
4000
5000
0
2
4
6
crafts


automotive
Frequency
0
1000
2000
3000
4000
5000
6000
0
4
8
12
automotive
art
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
5
10
art
religion
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
5
10
20
religion
beauty
Frequency
0
1000
2000
3000
4000
5000
6000
0
4
8
12
beauty


parenting
Frequency
0
1000
2000
3000
4000
5000
6000
0
4
8
12
parenting
dating
Frequency
0
2000
4000
6000
0
5
15
dating
school
Frequency
0
1000
2000
3000
4000
5000
6000
0
4
8
school
personal_fitness
Frequency
0
1000
2000
3000
4000
5000
0
5
10
personal_fitness

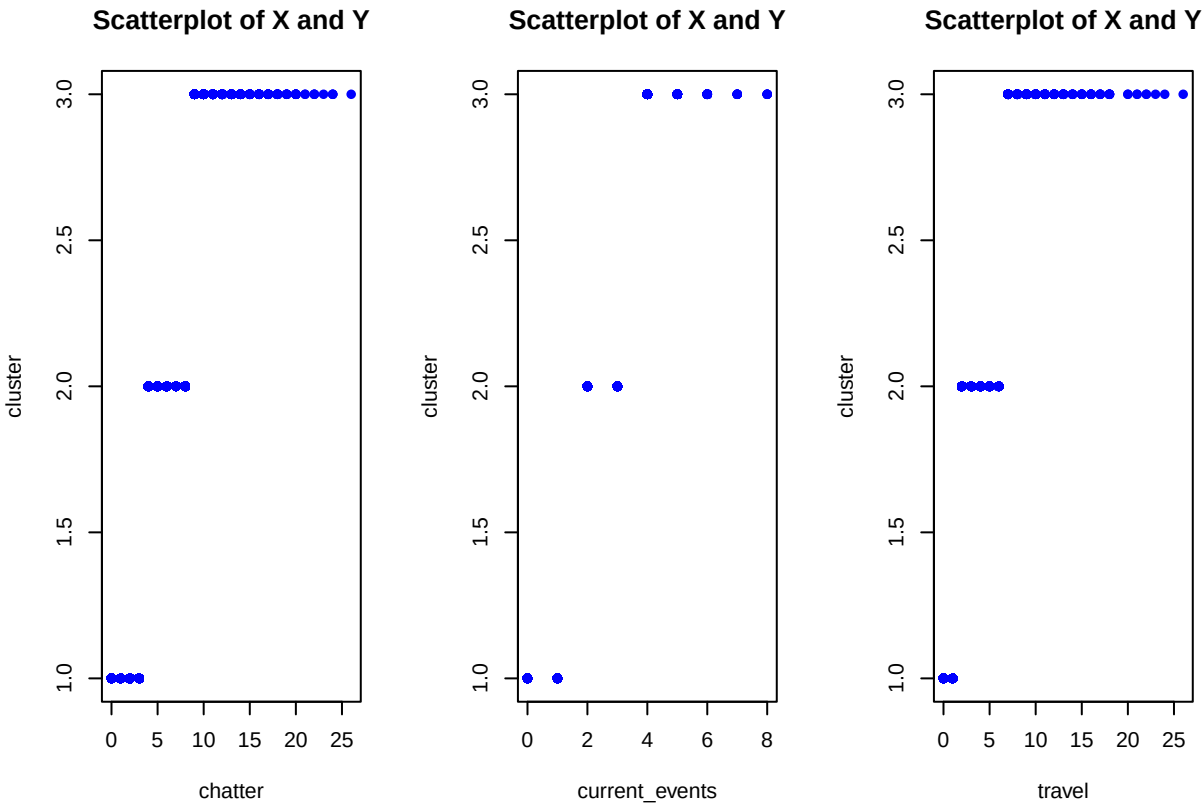


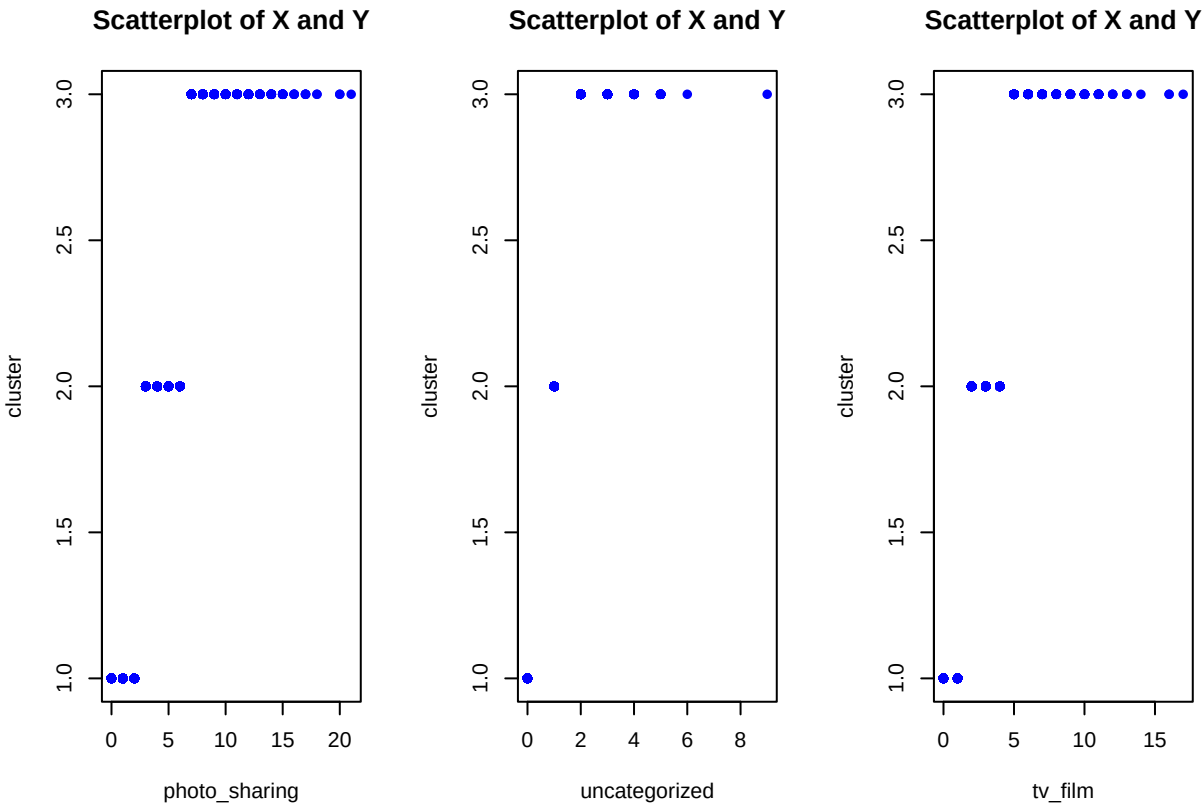
Interpretation:

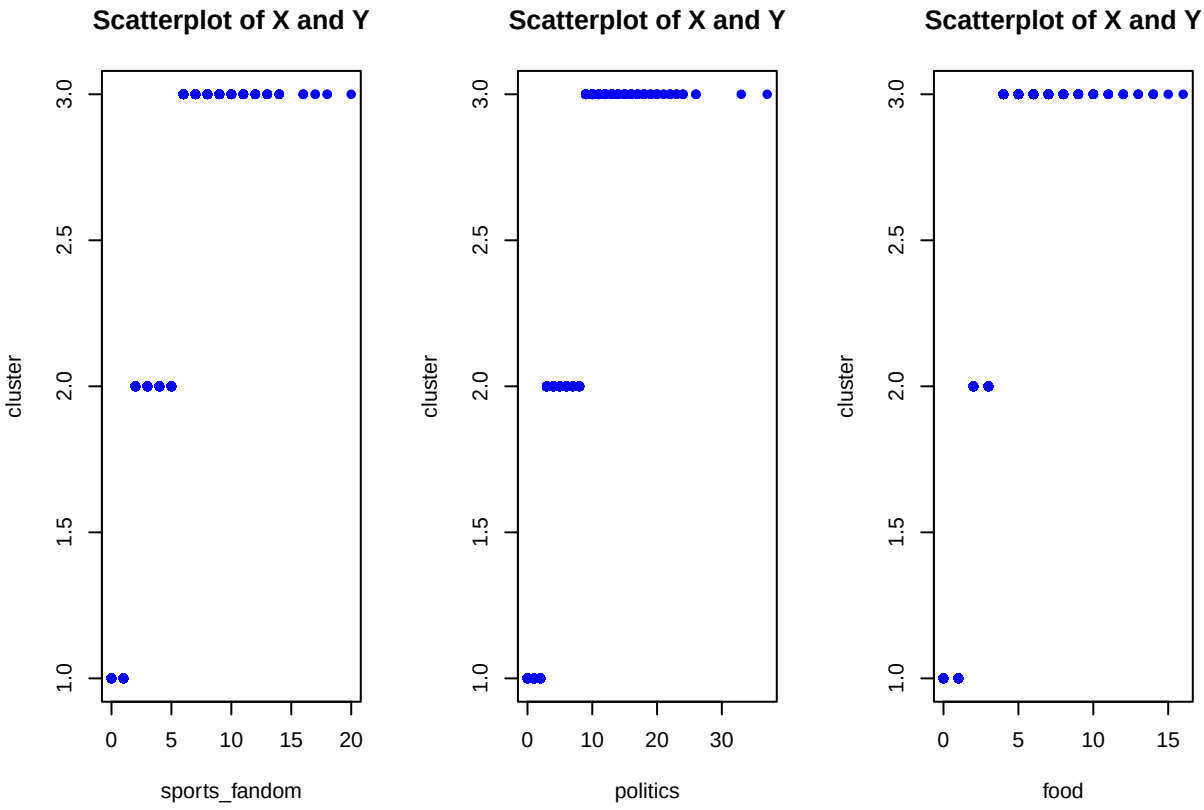
We can see that the distribution of tweets is similar for all tweet themes with a high number of users with who do not tweet on the theme, and low segment of users with high tweets.

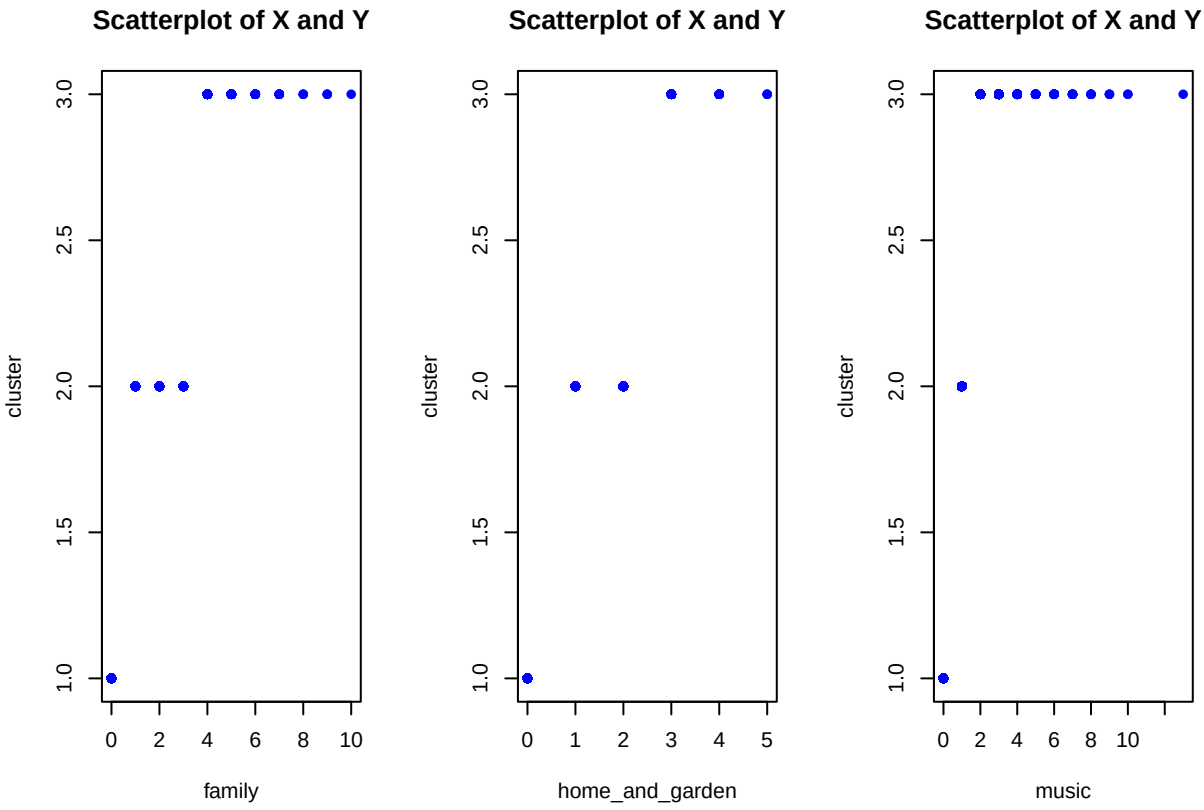
Step 2: Applying K-Means clustering model

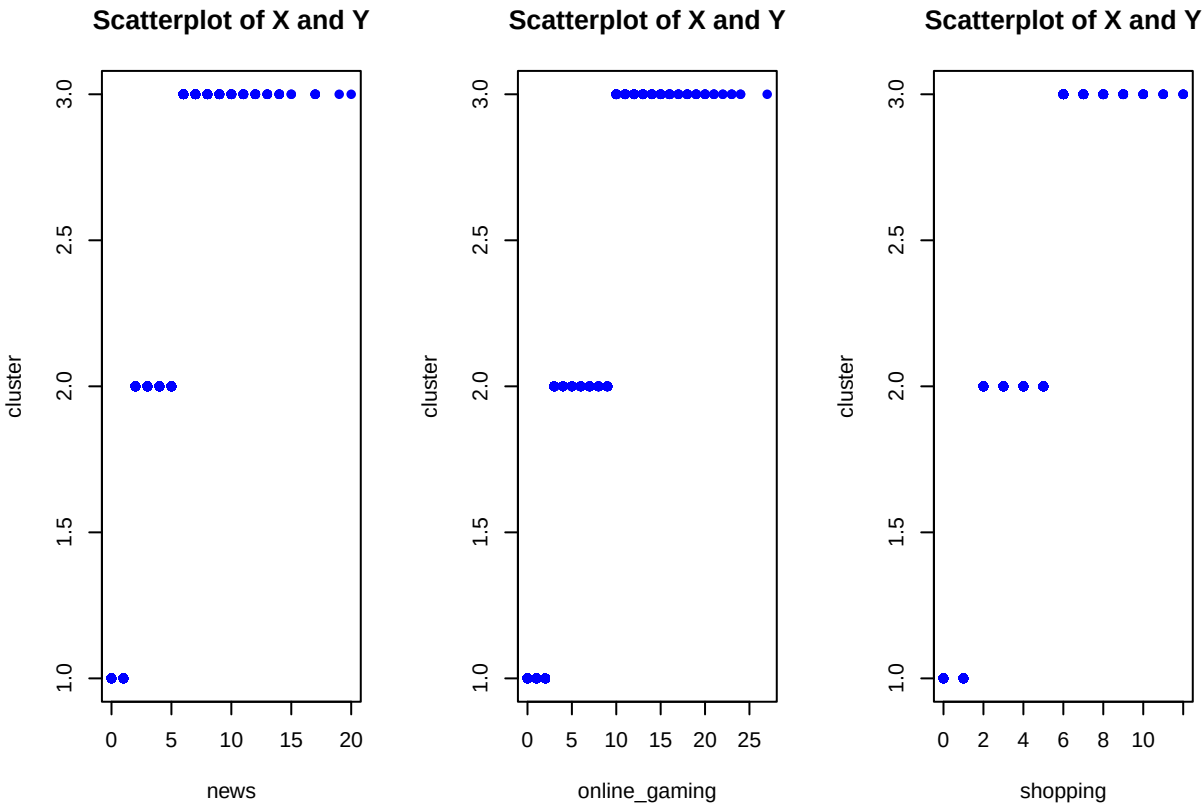
Using K-Means clustering to create clustering for each column tweet into 3 clusters based on the number of tweets. For e,g, cluster 1: users with lowest number of tweets on cooking , cluster 2: users with medium number of tweets on cooking 3. cluster 3: users with highest number of tweets on cooking. Not standardizing the data since all the units are in the same scale.

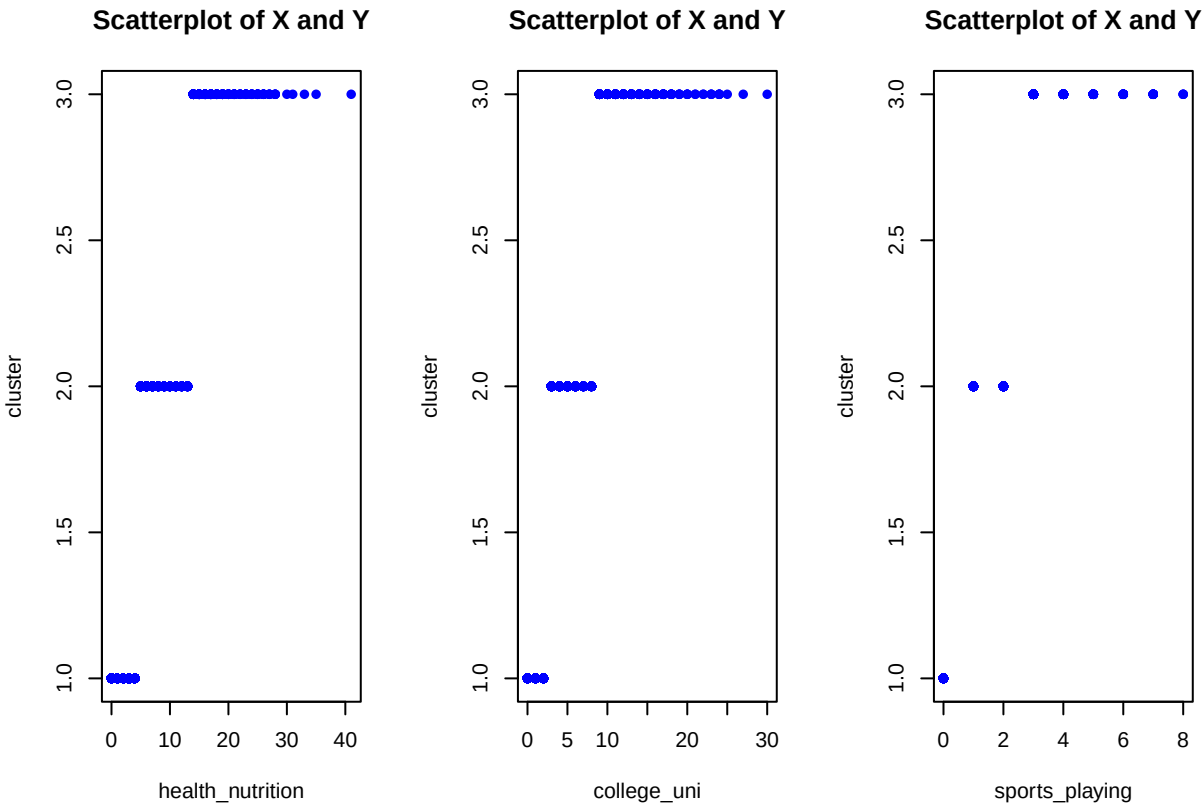

Scatterplot of X and Y
cluster
chatter
1.0
1.5
2.0
2.5
3.0
0
5
10
15
20
25
Scatterplot of X and Y
cluster
current_events
0
2
4
6
8
Scatterplot of X and Y
cluster
travel

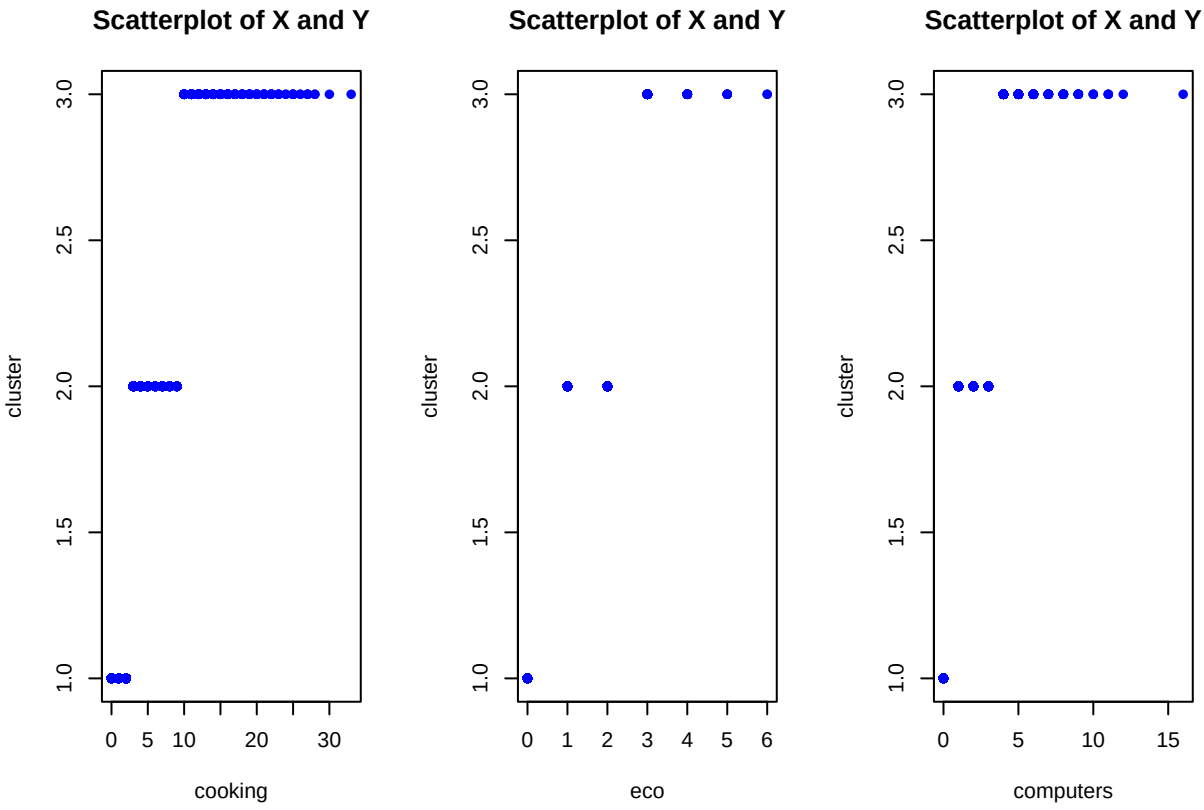

Scatterplot of X and Y
cluster
photo_sharing
Scatterplot of X and Y
cluster
uncategorized
Scatterplot of X and Y
cluster
tv_film

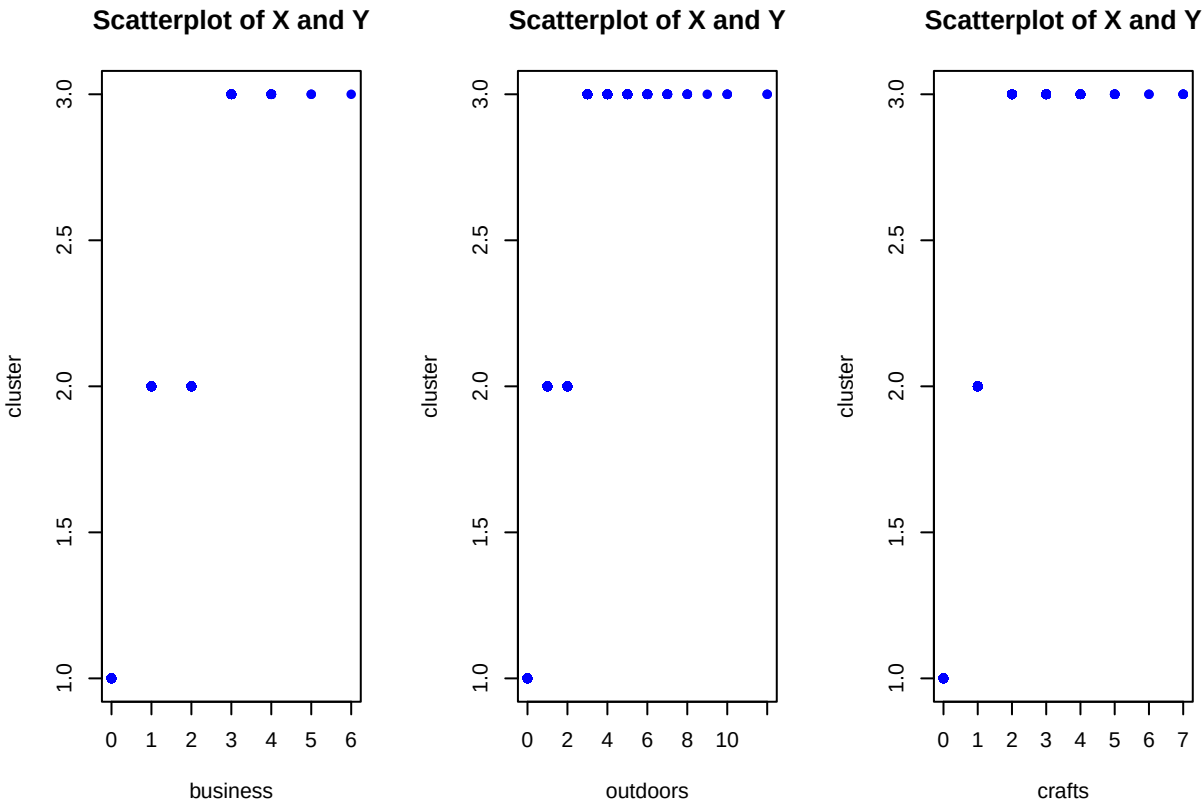

Scatterplot of X and Y
cluster
sports_fandom
Scatterplot of X and Y
cluster
politics
Scatterplot of X and Y
cluster
food

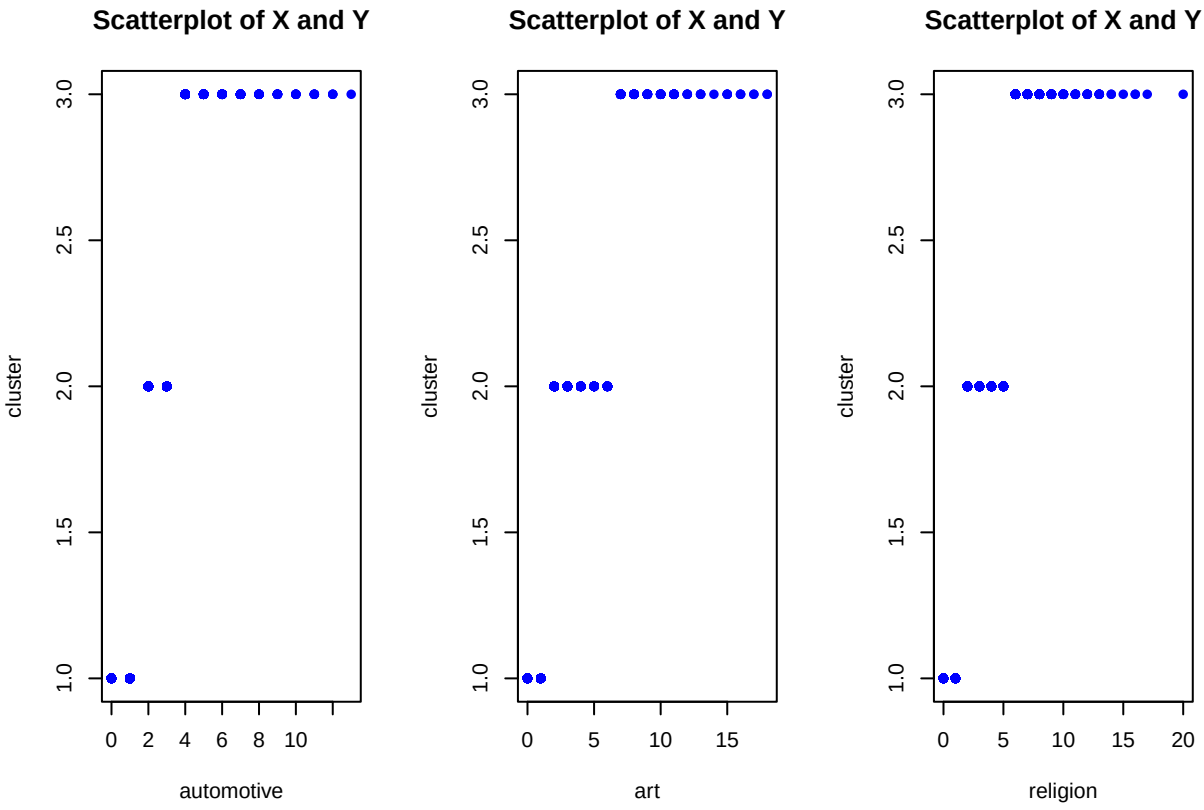

Scatterplot of X and Y
cluster
family
Scatterplot of X and Y
cluster
home_and_garden
Scatterplot of X and Y
cluster
music

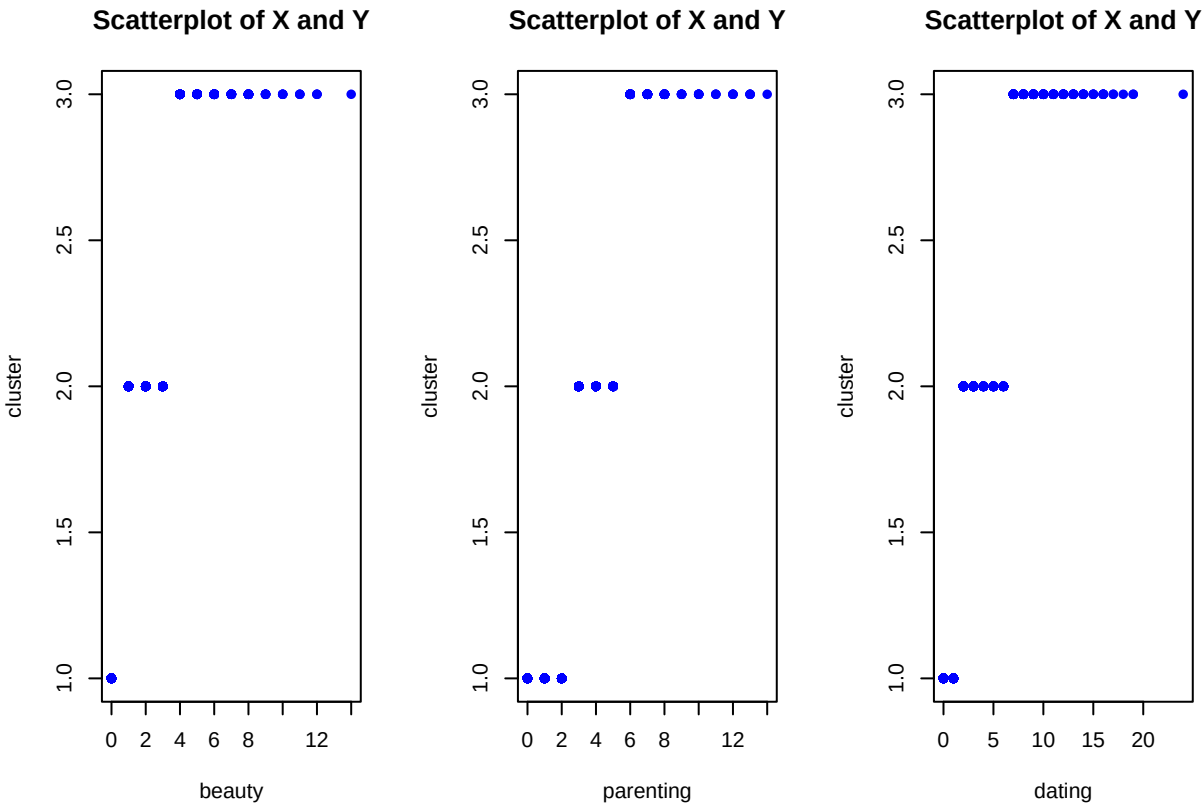

Scatterplot of X and Y
cluster
news
Scatterplot of X and Y
cluster
online_gaming
Scatterplot of X and Y
cluster
shopping

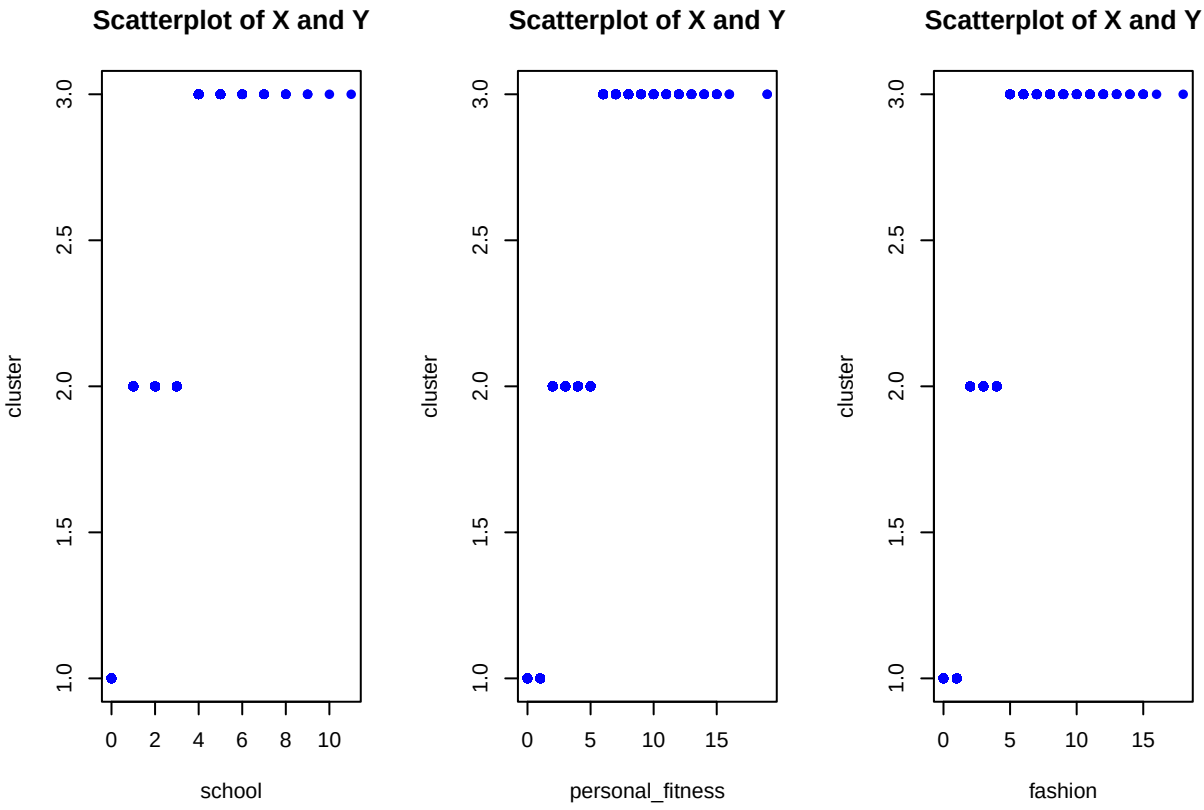

Scatterplot of X and Y
cluster
health_nutrition
Scatterplot of X and Y
cluster
college_uni
Scatterplot of X and Y
cluster
sports_playing

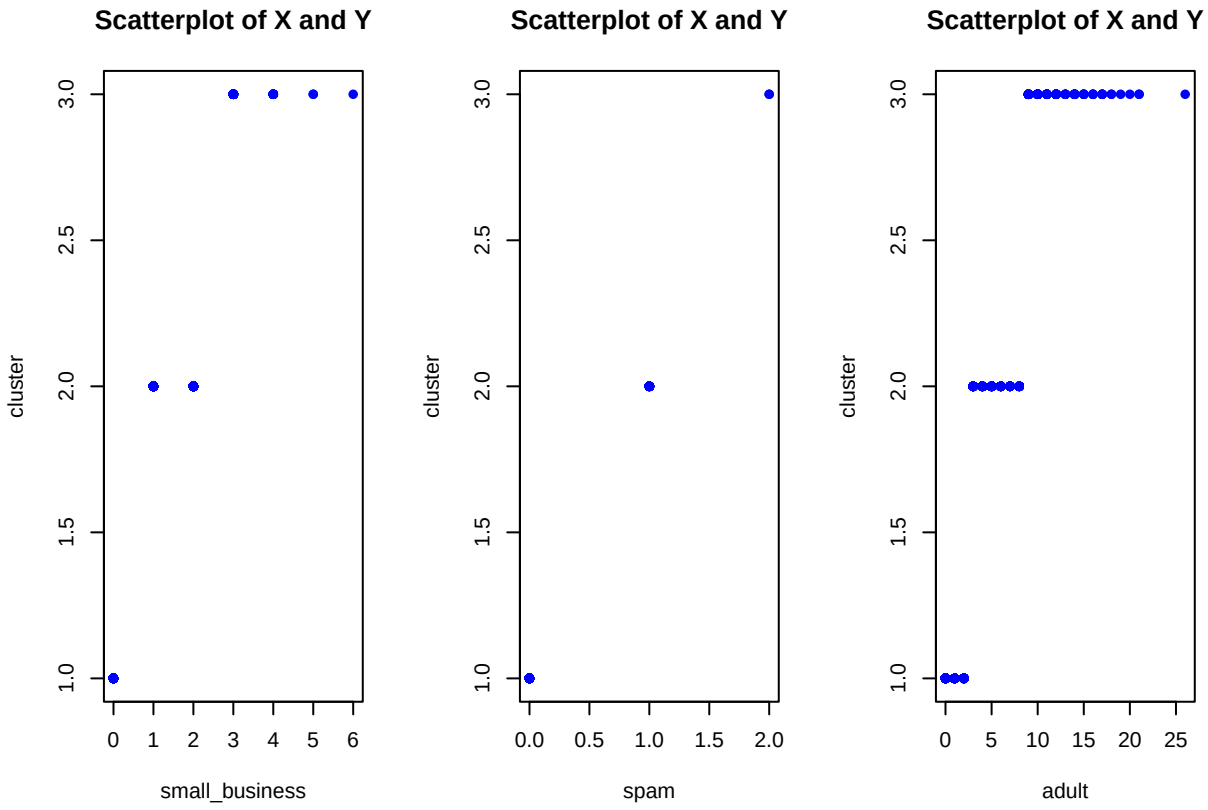

Scatterplot of X and Y
cluster
cooking
1.0
1.5
2.0
2.5
3.0
0
5
10
20
30
Scatterplot of X and Y
cluster
eco
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
6
Scatterplot of X and Y
cluster
computers
1.0
1.5
2.0
2.5
3.0
0
5
10
15


Scatterplot of X and Y
cluster
business
Scatterplot of X and Y
cluster
outdoors
Scatterplot of X and Y
cluster
crafts


Scatterplot of X and Y
cluster
automotive
Scatterplot of X and Y
cluster
art
Scatterplot of X and Y
cluster
religion


Scatterplot of X and Y
cluster
beauty
Scatterplot of X and Y
cluster
parenting
Scatterplot of X and Y
cluster
dating


Scatterplot of X and Y
cluster
school
Scatterplot of X and Y
cluster
personal_fitness
Scatterplot of X and Y
cluster
fashion



Interpretation:

Based on the clusters, we will targer the users for ads. For example. Above we the extracting customer segments who have lowest travel tweets in results for travel discounts & promotions.