Solo 1: Segmentation Scheme for The App Happy Company

Christina Macholan | Predict 450, Section 55

INTRODUCTION

The following report presents a data-driven segmentation framework developed for The App Happy Company (AHC). This framework will guide the product development and marketing strategy for their new B2C social entertainment app. To date, AHC has only developed B2B apps, and the launch of this new app will be their first foray into the consumer app market. Because of their inexperience, AHC lacks direct insight into current consumer needs and demands. They will depend on this analysis to understand consumers' needs and to build and market a successful product that will delight their customers.

AHC hired the Consumer Spy Corporation (CSC) to collect detailed demographic, behavioral, and attitudinal data from 1,800 respondents via a research survey and in-person focus groups. The segmentation scheme that follows was derived after building a series of clustering models with the survey data (including ones using k-means clustering, hierarchical clustering, and PAM clustering techniques), and selecting the result with the best combination of overall model fit, cluster distinctiveness, and interpretability.

SEGMENTATION FRAMEWORK

The following framework identifies five different consumer market segments from the CSC survey (see the Methodology section for complete details on the cluster modeling and profiling processes).

Our recommendation is that three of the segments (1, 2, and 3) should be considered high priority targets for product design and marketing since they demonstrate the best profile fit with the proposed social entertainment app. Segment 4, on the other hand, has very low affinity for apps and is generally not socially connected aside from relatively infrequent use of Facebook. As the largest proportion of the surveyed consumers, segment 5 does present some opportunity for growth, however the the clear app value-proposition for this segment has yet to be clear (see the Additional Research & Recommendations section for more details).

The **My Phone is** **My Entertainer** Segment

Segment 1: *17% of survey respondents*

Psychographics: Consumers in this segment, though young, tend to either aspire to influence or already have influence within their peer groups. They are socially-motivated, and place a strong emphasis on others' perceptions (e.g. what's "hot or not" and how "cool" something is). They have strong grasp of technology, having grown up with it, and show a strong willingness to pay for apps or app features.

Behavioral: These consumers are super-users of technology and are very entertainment-oriented. Music is important to them and they stay in tune with what is on TV even when they are away.

Demographics: Users in this segment are the youngest (mostly under the age of 30) and do not have children.

Tech Profile: Always connected. Likely to have a second device (iPod or Tablet) in addition to a phone. Has an app for everything (typically more than 30 apps) with Social, Communication, and Entertainment apps dominating their phones.

Product Opportunities: Focus on social features that keep the user engaged and visible within his/her social circle. Music and TV will both be important.

Where to Target Them Online: Facebook, YouTube, Pandora, and social, music, gaming, and shopping apps, in general.

Segment 2: *22% of survey respondents*

The **“My Phone is My Personal Assistant”** Segment

Psychographics: Consumers in this segment have a positive view of technology and have a self-concept of being an active and driven opinion leader and advisor. They are motivated to make choices that reflect their personal style rather than allowing the influence of external expectations (e.g. they care less about what is “hot or not”).

Behavioral: These consumers tend to use their phones for organization and communication purposes and limit the number of apps they download. They log onto sites less frequently than segments 1 and 3.

Demographics: Consumers categorized in this segment are more likely to be above 30, well-educated, and are less likely to have children.

Tech Profile: Less likely to have a second device and more likely to own a Blackberry (+xx%) instead of an iPhone or Android, though.

Product Opportunities: Build app features that help these consumers keep their entertainment libraries organized. This group will gravitate towards elements that feel customizable and allow them to display their personal sense of style (e.g. customizable profile page).

Where to Target Them Online: Facebook, YouTube, social media and communication apps

The **“My Phone is My Lifeline”** Segment

Segment 3: *21% of survey respondents*

Psychographics: Consumers in this segment are doers and are focused on balancing efficiency with fun (or at least maintaining the image that they can -- they are very “image” and “style” oriented). They are heavily tech dependent, pride themselves on being tech savvy, and enjoy using technology for entertainment and social connection, but this dependency can sometimes be stress-inducing given everything else they have going on in their lives, including likely having a partner and children at home (e.g. this group scored highest on the negative sentiments towards technology question, despite indicating on other questions that they are glad that mobile keeps them connected and are delighted to be entertained through apps). They are looking to be delighted, but not overwhelmed. These consumers are well-educated, are willing to spend, and have more income to spend with than segment 1, but will likely be conscientious about getting good value for their money and time.

Behavioral: These consumers are time sensitive because their lives are busy, and seek out tech solutions to help stay organized and in control. They are very active on social media and keep up with TV even when they are not watching.

Demographics: Mostly between the ages 25 and 40, likely with 1 or more children and a partner at home.

Tech Profile: Always connected. Very likely to have a second device (iPod or Tablet) in addition to a phone. Has an app for everything.

Where to Target Them Online: social and entertainment apps.

The **“My Phone is Sometimes a Mystery to Me”** Segment

Segment 4: *12% of survey respondents*

Psychographics: These consumers see likely to see their phone as entirely utilitarian or even mildly perplexing and are unlikely to see the benefits of paying for a mobile app. They don’t view their phone as an entertainment device.

Behavioral: These consumers are much less likely to be frequent visitors of the same site.

Demographics: Mostly 50+ with grown children (18+).

Tech Profile: Have few to little apps on their device and are less likely to be using Android or iOS. May not be aware of what an “app” is, exactly (10% answered that they do not know if they have an app on their device). Rarely have a second device. If they are using apps, it is mostly for occasional check-ins on social media or gaming apps.

Product Opportunities: N/A since this group would be very unlikely to use the app.

Where to Target Them Online: This segment will be difficult to reach since they are much less frequently online or using their phone for entertainment purposes.

The **"My Phone is Nice to Have"** Segment ***Segment 5:*** *28% of survey respondents*	**Psychographics:** These consumers are in a similar demographic to segment 5, but are more tech savvy and open to the influence of tech on their lives. They aren't looking to be leaders or influencers in tech, but are content to use their devices for entertainment occasionally. **Behavioral:** These users are not influencers or looking to be influenced. They are likely to follow the same patterns of using their phone and may be difficult to sign on to a new app without significant effort. **Demographics:** Mostly 50+ with grown children (18+) and married. **Tech Profile:** Have an average number of apps on their device and use it for basic entertainment (social media, music, apps). Even on iPhone and Android split. **Product Opportunities:** Integrate the app with social networks or other popular platforms to keep the brand present in the user's daily flow without disruption. This may earn more staying power than trying to dramatically change the user's' basic routine habits. **Where to Target Them Online:** Social apps, gaming apps, music apps.

METHODOLOGY

We tested several clustering techniques and formulations for the basis variables before selecting the final segmentation scheme above. The following is a comparison of each of the clustering methods tested based on their goodness-of-fit, cluster distinctiveness, and interpretability.

EXPLORATORY ANALYSIS & OUTLIERS

The survey data contains 40 attitudinal variables measuring responders' agreement with statements about their technology preferences, personality, and purchasing behaviors on the six-point Likert scale (1 = Strongly Agree, 2 = Agree, 3 = Agree Somewhat, 4 = Disagree Somewhat, 5 = Disagree, 6 = Disagree Strongly). Only these 40 variables were considered for use in building the clustering algorithms, while the other demographic and behavioral variables were used later on for preparing the segmentation profiles.

After careful consideration of the data, we chose to remove 46 records where the respondent marked the same numeric answer for each of the 40 attitudinal statements. We chose do so knowing that one negative statement in particular -- *"Responsibility is overrated; I'd rather be told what to do."* -- serves as a control against other positive questions, which we would expect to receive an extreme opposite, or at least slightly different, number on the Likert scale. These are suspected of being lazily marked questionnaires. All other records were left as is and the stability of the cluster analysis was confirmed with and without these 46 cases included to ensure important information was not lost with their exclusion.

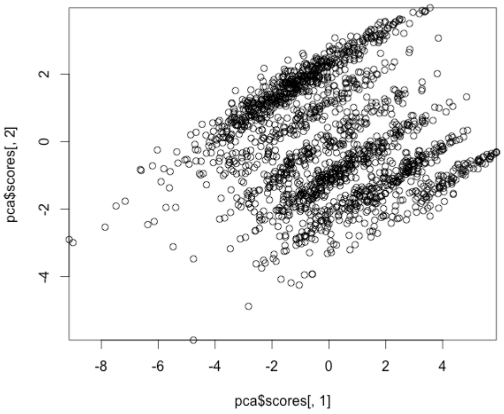
BASIS VARIABLES

In order to reduce the complexity of the clustering models that follow and to produce a more interpretable segmentation schema, we went through several iterations of variable reduction for the 40 attitudinal variables. Visual analysis of the correlation plots and bivariate plots (see the *Appendix A*) as well as results from a principal components analysis and a factor analysis run in R provided initial guidance on which variables share information and could possibly be combined (view the other variable reduction schemes tested in *Appendix B*).

However, in the end, the best performing variable reduction was largely driven by a combination of both the statistical relationships between the variables and my own intuitions about what statements would be meaningful to combine.

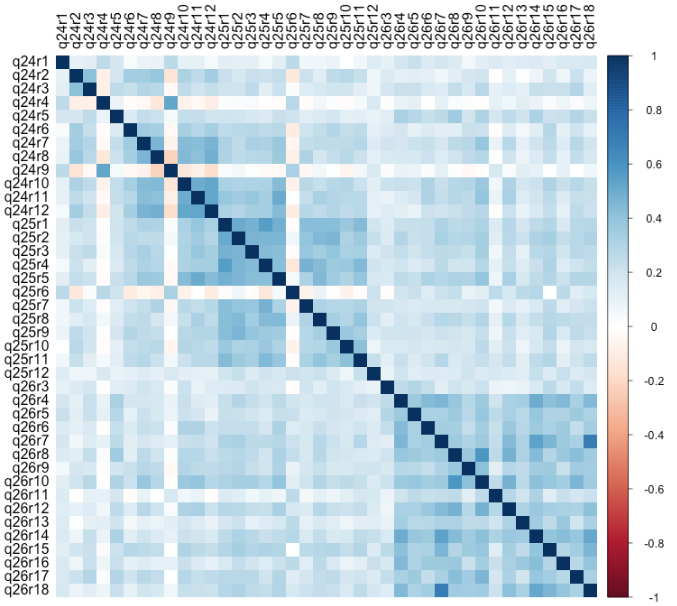
- **Tech Positive:** Combines responses from q24.1, q24.2, q24.3, q24.5, q24.6
- **Media Positive:** Combines responses from q24.7, q24.8, and q26.17
- **Internet Communicator:** Combines responses from q24.10, q24.11
- **Opinion Leader:** Combines responses from q25.1, q25.2, q25r3, q25r4, q25.5
- **Control Oriented:** Combines responses from q25.7, q25.8
- **Driven:** Combines responses from q25.9, q25.10, q25.11
- **Bargain / Cost Oriented:** Combines responses from q26.3 and q26.5
- **App Fan:** Combines responses from q26.8, q26.9, q26.10, and q26.12
- **Has Children:** Combines responses from q26.11
- **Comfortable Spending:** Combines responses from q26.4, q26.6, q26.13, and q26.16
- **Brand / Image Oriented:** Combines responses from q26.7, q26.14, q26.15, and q26.18

Plot of First Two Components of PCA after Variable Reduction

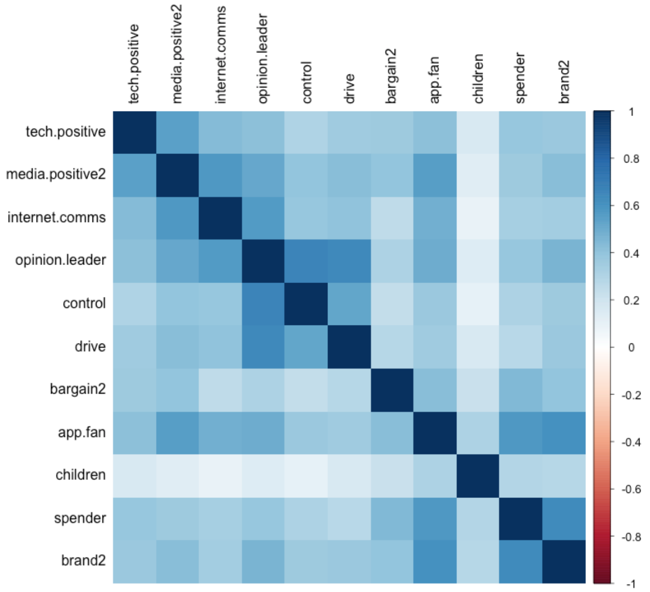


The plots below demonstrate the consolidation of information into a more dense set of variables from 40 original variables to 11 reduced variables. Note that all negative statements were removed from the final set of variables since they hindered the goodness-of-fit for the clustering algorithms. The remaining 11 variables were used to build all subsequent clustering models.

Correlation Plot for all 40 Variables



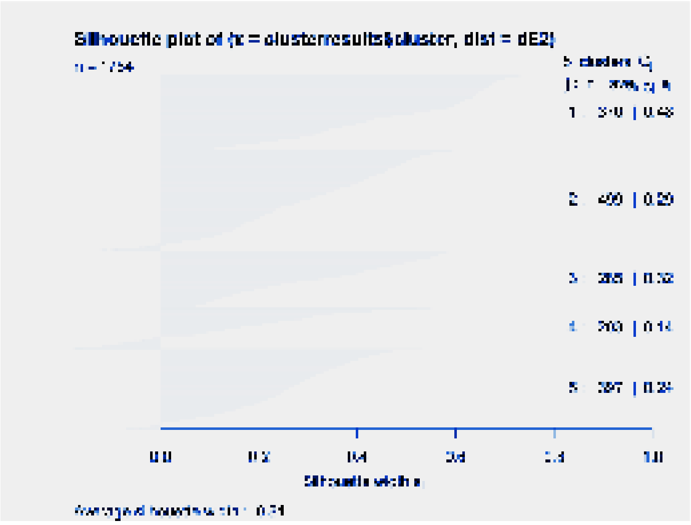
Correlation Plot after Variable Reduction

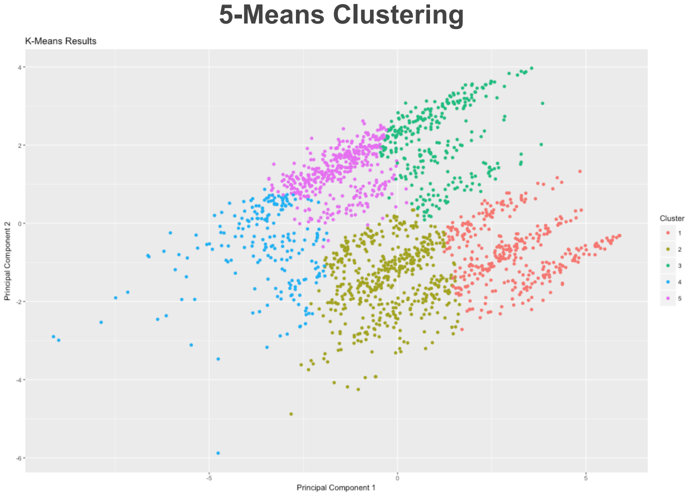


K-MEANS CLUSTERING

K-means clustering proved to produce the best clustering model for the survey data. Guided by the soft "elbow" in the Scree plot near 4 and 5, we selected to run the k-means algorithm with 5 clusters prescribed using the 11 basis variables mentioned in the previous section.

The resulting model has an R^2 of 0.4857047 and a Silhouette width of 0.31. The segments have a reasonably even distribution, and high enough silhouette widths for each segment to make them distinguishable for profiling purposes. The plot on the right below shows the structure of the first two principal components for the five cluster solution, with reasonable division between the clusters (and distinct striation because of the removal of the uniform value responses).



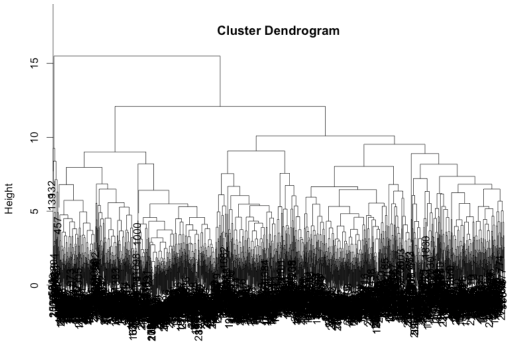


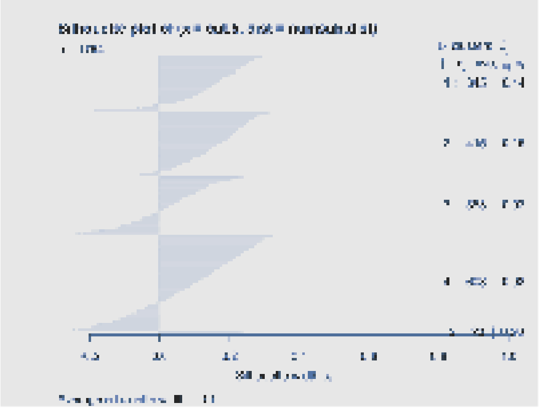
HIERARCHICAL CLUSTERING

Using the hierarchical clustering algorithm in R, we attempted to outperform the previous k-means clustering model.

Single linkage, average linkage, and complete linkage algorithms were all run with various combinations of the basis variables, however with the best five segment (i.e. fit "cut") model had an R^2 of only 0.3856964 and an average silhouette width of only 0.1. wo of the identified clusters were very near the wall with 0.02 and 0.08 silhouette widths. Also, the most distinctive segment, with a width of 0.20, only had 23 records assigned to it. The small size of which would likely make the segment operationally useless for targeting, even if the model performed better overall.

The dendrogram for this complete linkage model is shown to the right. Because of the lack of discrimination between the segments and due to a both lower R^2 and silhouette width, we did not pursue this approach any further.





PAM CLUSTERING

The highest silhouette width achieved with partitioning around medoids (PAM) clustering for the five cluster model was only 0.1508228, which also underperformed compared to the k-means clustering model.

Average silhouette width per cluster:

- 0.08827990
- 0.12073922
- 0.20831277
- 0.24973538
- 0.09591348

Average silhouette width of total data set: 0.1508228

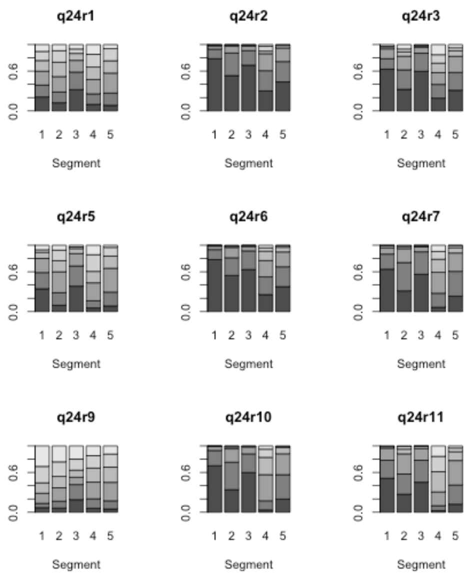
PROFILING

After building the k-means clustering model and assigning each of the respondents to one of the five clusters, we returned to the full data set -- demographic and behavioral variables included -- to conduct an comparison of each profile.

In order to quickly identify differences between the five groups, we reviewed group means for each variable and created two-way proportional frequency plots to examine the distribution of each response by cluster.

For example, in the panel to the right you can see that for question 24 part 6 ("*I look for web tools and Apps that help me save time."*) respondents in clusters 1 and 3 were more likely to respond with a 1 (*Strongly Agree*) for this statement.

I systematically reviewed plots for each of the original variables as well as the derived variables to develop the profiles outlined in the *Profiling* section above. The profiles are deeply tied to a visual analysis of the plots, which can be examined in *Appendix C*.



TYPING TOOLS TO OPERATIONALIZE SEGMENTATION

Because we cannot conduct the complete consumer survey for every new potential customer, we need to define a plan to operationalize our user segment assignment and monitor the stability of segmentation scheme over time.

Based on the results of our profile, we recommend designing and testing the following two typing tools in tandem to see which one performs the best:

- **Mobile Analytics Based Typing Tool:** Mobile app tracking technology has advanced features that allow you to track information about your app viewers' technology (device and operating system), assumed demographic (gender and age), social media activity related to the app (sharing links to the app or visiting the app from a social channel), and the referring site or app that brought the user to the site. Because each of the five profiles was highly differentiated on most of these traits, we could build a simple logistic regression, k-nearest neighbors, or decision tree classification model using the data from our mobile app analytics platform to quickly categorize each new visitor in alignment with our current profiles. The added benefit of using this approach is that mobile app remarketing technology will allow us to use this same data to easily retarget user profiles with online ads.

- **App Profile:** Alternatively, users could fill out a short profile when they first enter the app to share key demographic data to assign them to the assumed segment (age, gender). A short web survey could also periodically ask users to answer some of the most polarizing questions (e.g. about attitudes towards apps and social media) to draw clear distinction between the segments. A similar classification model as described for the mobile-based solution would need to be built.

ADDITIONAL RESEARCH RECOMMENDATIONS & CONCLUSION

This initial analysis of the CSC survey data resulted in a well-defined five segment framework for AHC to begin using for guidance in designing and promoting their new social engagement app.

In additional to the segmentation scheme presented above, we recommend following this study with additional focus groups, surveys, or research to answer at the following:

- **Was the sample for the survey representative?** The assumptions of this analysis are only as good as the data that CSC shared. Follow up with CSC to ensure that the sampling approach used makes the assumptions of this analysis viable.
- **What type of entertainment?** With only vague music, TV, and entertainment questions asked on the survey, it is not yet clear whether each segment has different interests or desires for what type of entertainment info would live within the AHC app. Bring together a focus group representing each segment to conduct questioning or user studies to guide the app product development before it gets out of line with user's interests.
- **Start with Android or iOS?** What is the market share of Android vs. iOS in the markets where you plan to launch? The survey indicates that more of the respondents have iPhones or Apple products than iOS devices, but this may not be the case for the entire market you target (especially since Android leads in market share worldwide).

Since the data for the App Happy segmentation is based on a survey questionnaire, we should question whether the sampling approach and / or the delivery method could have possibly biased the data collected and, therefore, the end result of the analysis.

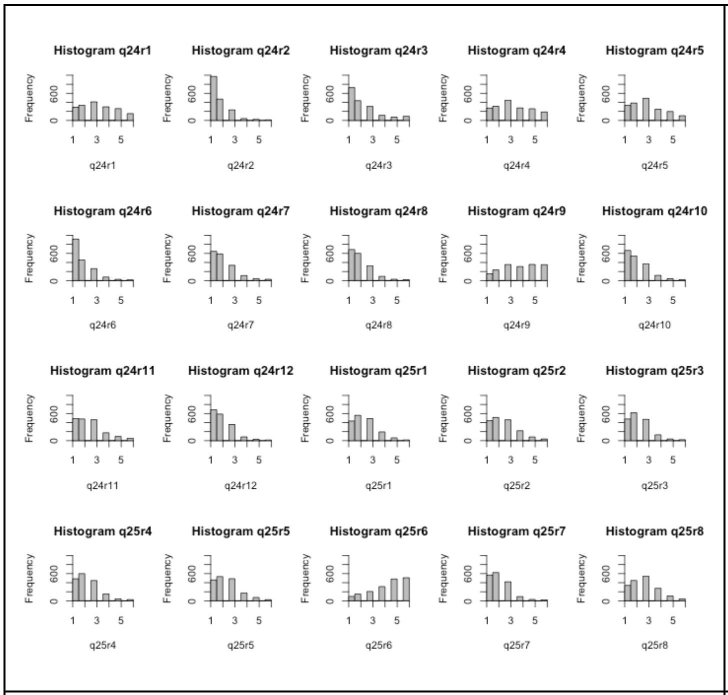
For example, if the survey was delivered on the web, there may be a response bias towards people who are more enthusiastic about apps and technology than the actual target market. In this case the sample of consumers surveyed may not accurately represent the demographics, attitudes, or behaviors of the whole target market population.

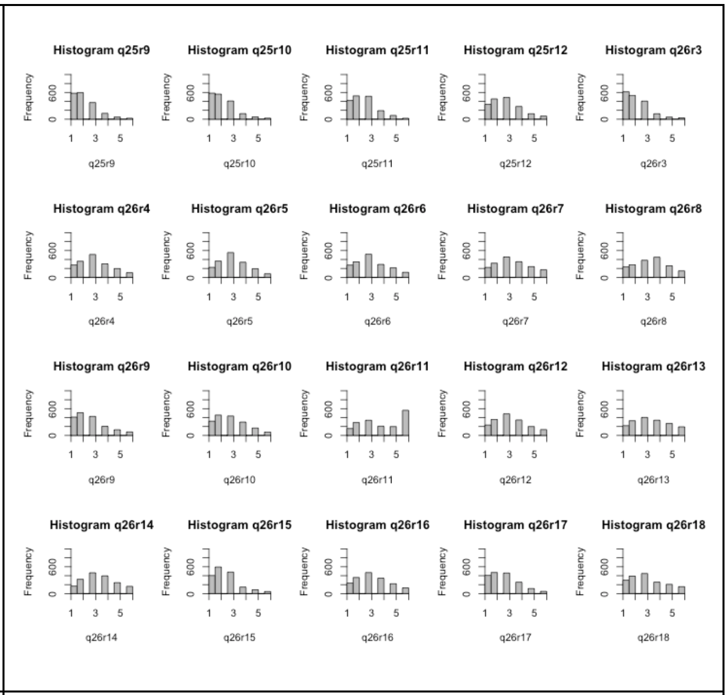
The structure of the questionnaire could possibly bias responses as well. For example, for most of the questions on the Likert scale, 1 is associated with a positive sentiment and 6 with a negative. However, for a few of the questions (e.g. q24r4 and q24r9) this is reversed. Users who are quickly running through the survey may accidentally answer this question the opposite of what they intended.

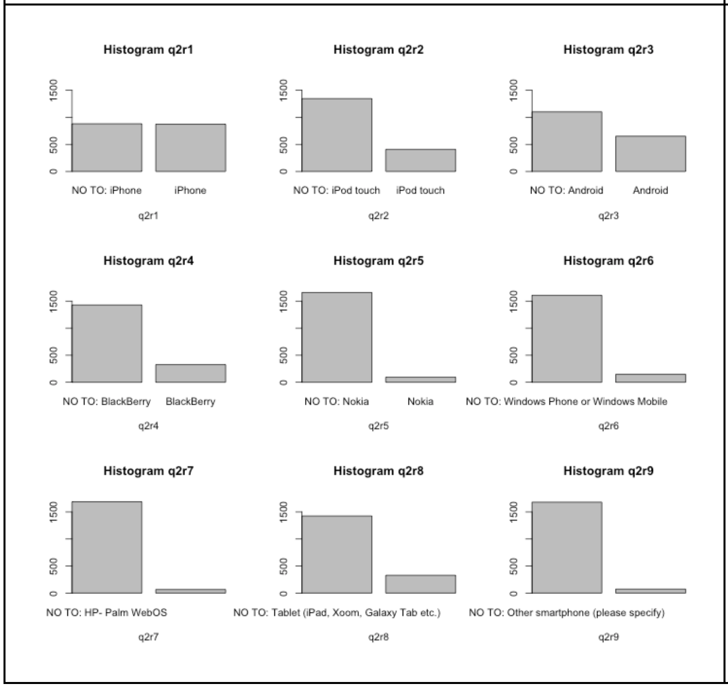
Of course, there are many other methodological considerations that have to be taken into account when using survey data in order to trust the validity and representativeness of the information collected. I think we all got a good taste of this in Predict 402, but this is a good reminder that it's important to raise questions about the collection methodology when inheriting a set of survey data when you weren't involved in the initial survey design!

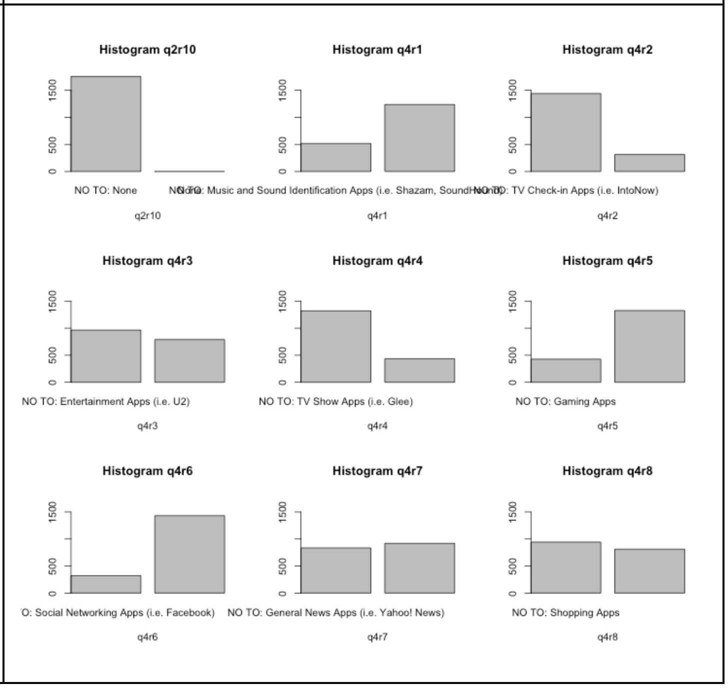
APPENDIX

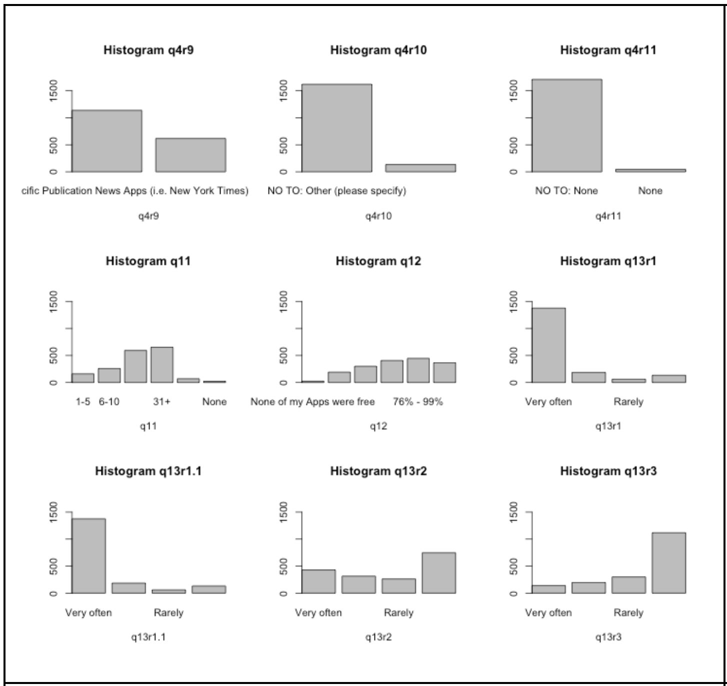
A. EXPLORATORY DATA ANALYSIS

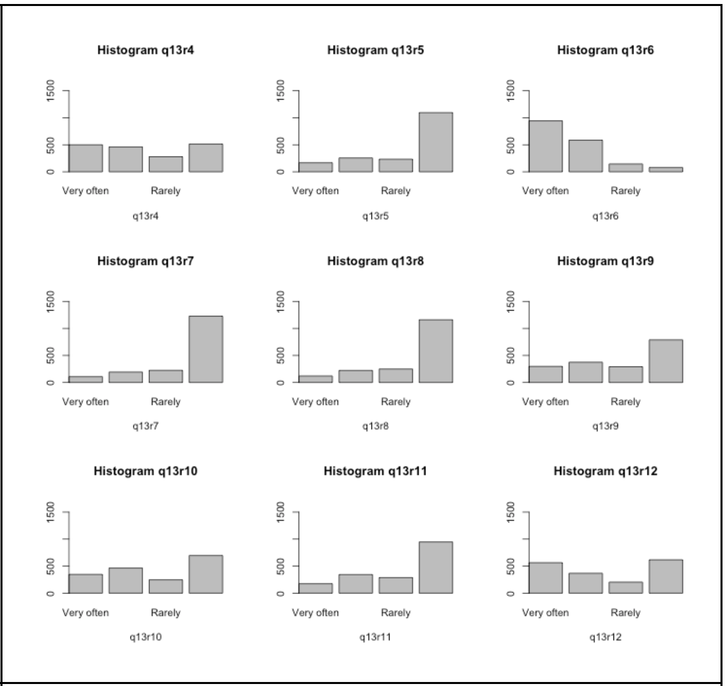


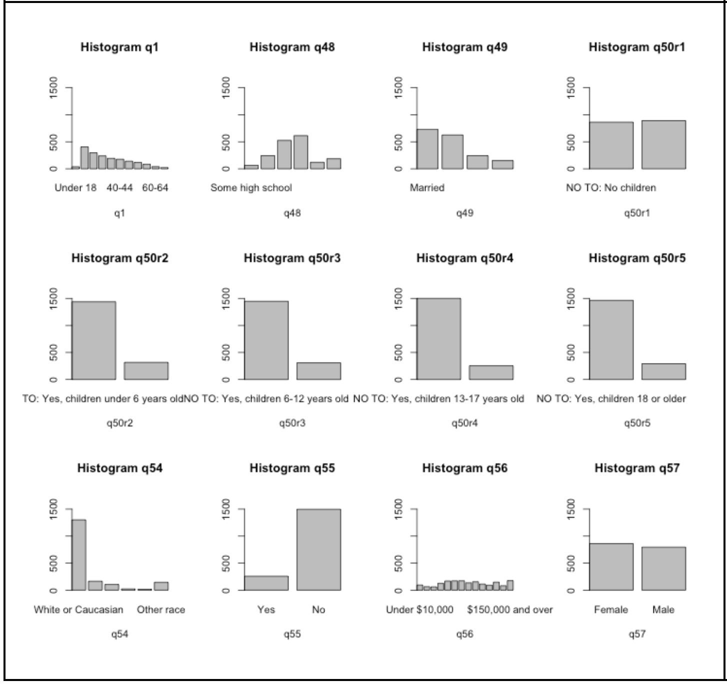


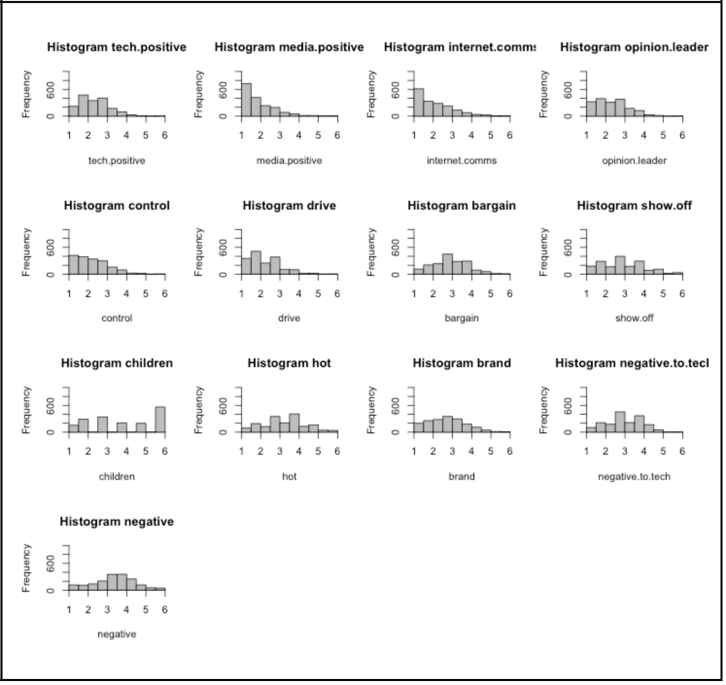


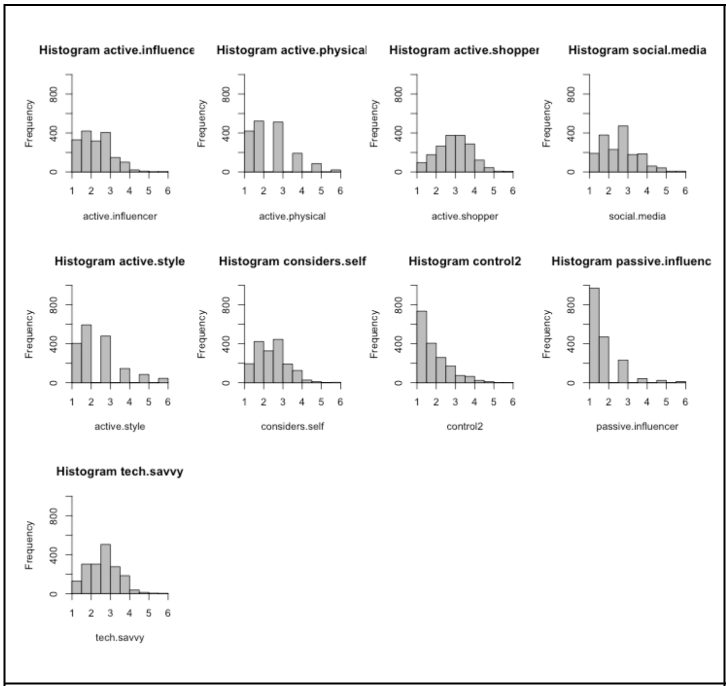


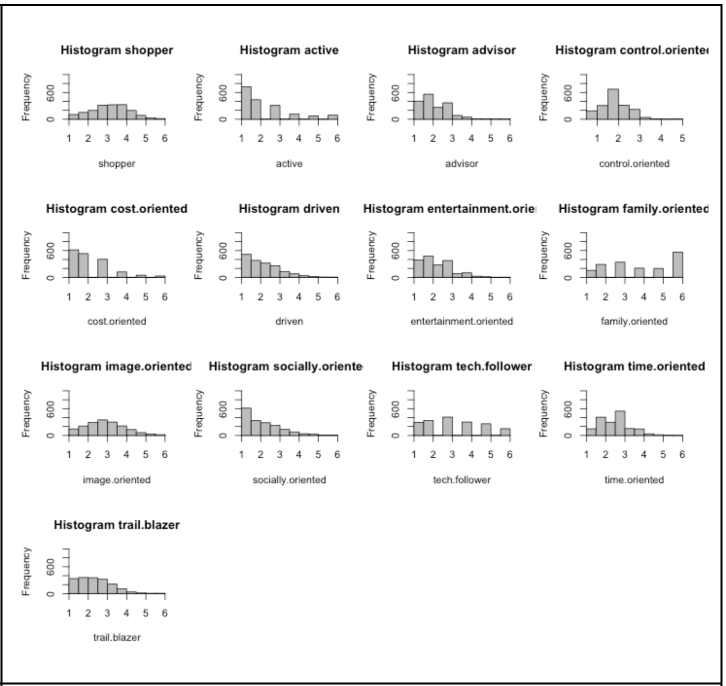

Histogram q4r9
cific Publication News Apps (i.e. New York Times)
q4r9
Histogram q4r10
NO TO: Other (please specify)
q4r10
Histogram q4r11
NO TO: None
None
q4r11
Histogram q11
1-5
6-10
31+
None
q11
Histogram q12
None of my Apps were free
76% - 99%
q12
Histogram q13r1
Very often
Rarely
q13r1
Histogram q13r1.1
Very often
Rarely
q13r1.1
Histogram q13r2
Very often
Rarely
q13r2
Histogram q13r3
Very often
Rarely
q13r3
0
500
1500

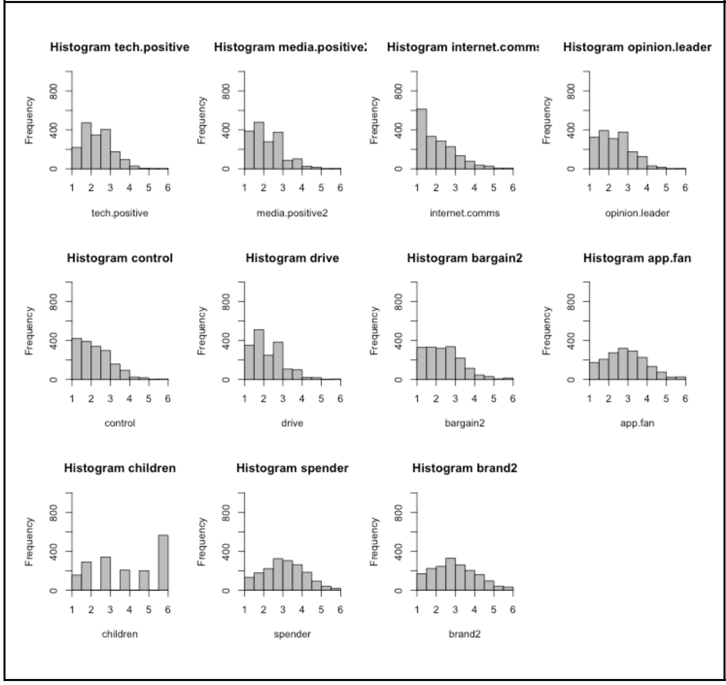

Histogram q13r4
q13r4
Histogram q13r5
q13r5
Histogram q13r6
q13r6
Histogram q13r7
q13r7
Histogram q13r8
q13r8
Histogram q13r9
q13r9
Histogram q13r10
q13r10
Histogram q13r11
q13r11
Histogram q13r12
q13r12
Very often
Rarely
0
500
1500


Histogram q1
Under 18
40-44
60-64
q1
Histogram q48
Some high school
q48
Histogram q49
Married
q49
Histogram q50r1
NO TO: No children
q50r1
Histogram q50r2
TO: Yes, children under 6 years old
q50r2
Histogram q50r3
NO TO: Yes, children 6-12 years old
q50r3
Histogram q50r4
NO TO: Yes, children 13-17 years old
q50r4
Histogram q50r5
NO TO: Yes, children 18 or older
q50r5
Histogram q54
White or Caucasian
Other race
q54
Histogram q55
Yes
No
q55
Histogram q56
Under $10,000
$150,000 and over
q56
Histogram q57
Female
Male
q57
0
500
1500

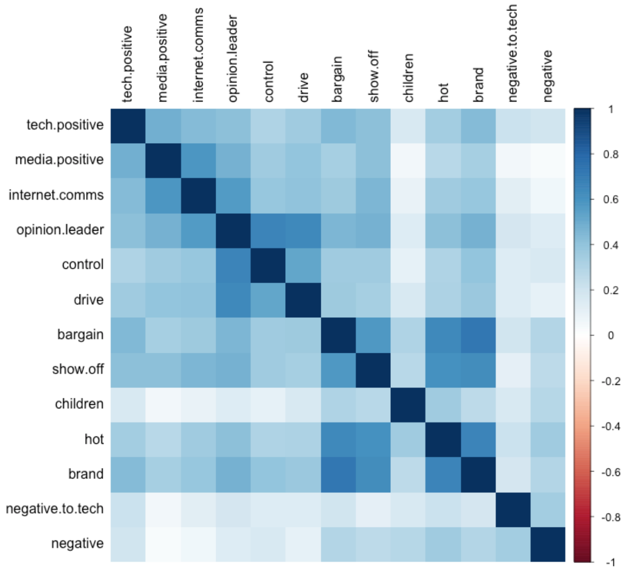

Histogram tech.positive
tech.positive
Histogram media.positive
media.positive
Histogram internet.comms
internet.comms
Histogram opinion.leader
opinion.leader
Histogram control
control
Histogram drive
drive
Histogram bargain
bargain
Histogram show.off
show.off
Histogram children
children
Histogram hot
hot
Histogram brand
brand
Histogram negative.to.tecl
negative.to.tech
Histogram negative
negative
Frequency
0
600
1
2
3
4
5
6

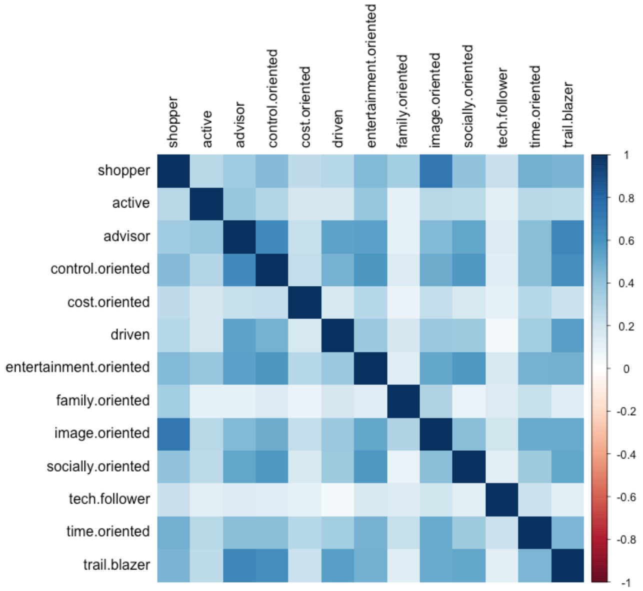

Histogram active.influence
Histogram active.physical
Histogram active.shopper
Histogram social.media
Histogram active.style
Histogram considers.self
Histogram control2
Histogram passive.influenc
Histogram tech.savvy
Frequency
0
400
800
1 2 3 4 5 6
active.influencer
active.physical
active.shopper
social.media
active.style
considers.self
control2
passive.influencer
tech.savvy

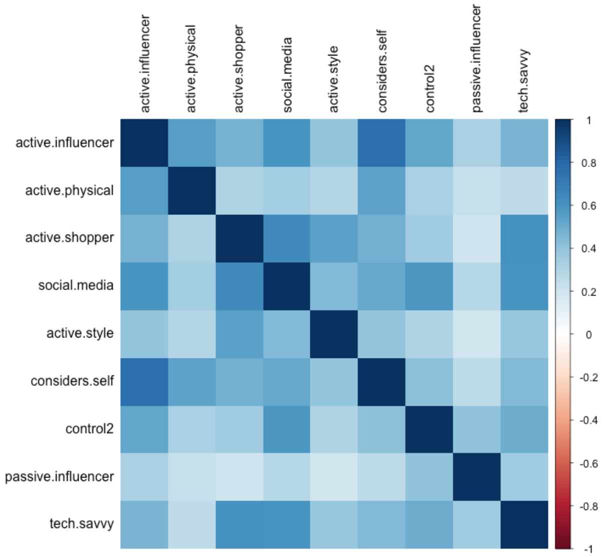

Histogram shopper
Histogram active
Histogram advisor
Histogram control.orientec
Histogram cost.oriented
Histogram driven
Histogram entertainment.orie
Histogram family.oriented
Histogram image.oriented
Histogram socially.oriente
Histogram tech.follower
Histogram time.oriented
Histogram trail.blazer
Frequency
0
600
1 2 3 4 5 6
1 2 3 4 5
shopper
active
advisor
control.oriented
cost.oriented
driven
entertainment.oriented
family.oriented
image.oriented
socially.oriented
tech.follower
time.oriented
trail.blazer


Histogram tech.positive
Histogram media.positive2
Histogram internet.comms
Histogram opinion.leader
Histogram control
Histogram drive
Histogram bargain2
Histogram app.fan
Histogram children
Histogram spender
Histogram brand2
Frequency
0
400
800
1 2 3 4 5 6
tech.positive
media.positive2
internet.comms
opinion.leader
control
drive
bargain2
app.fan
children
spender
brand2

B. VARIABLE REDUCTION



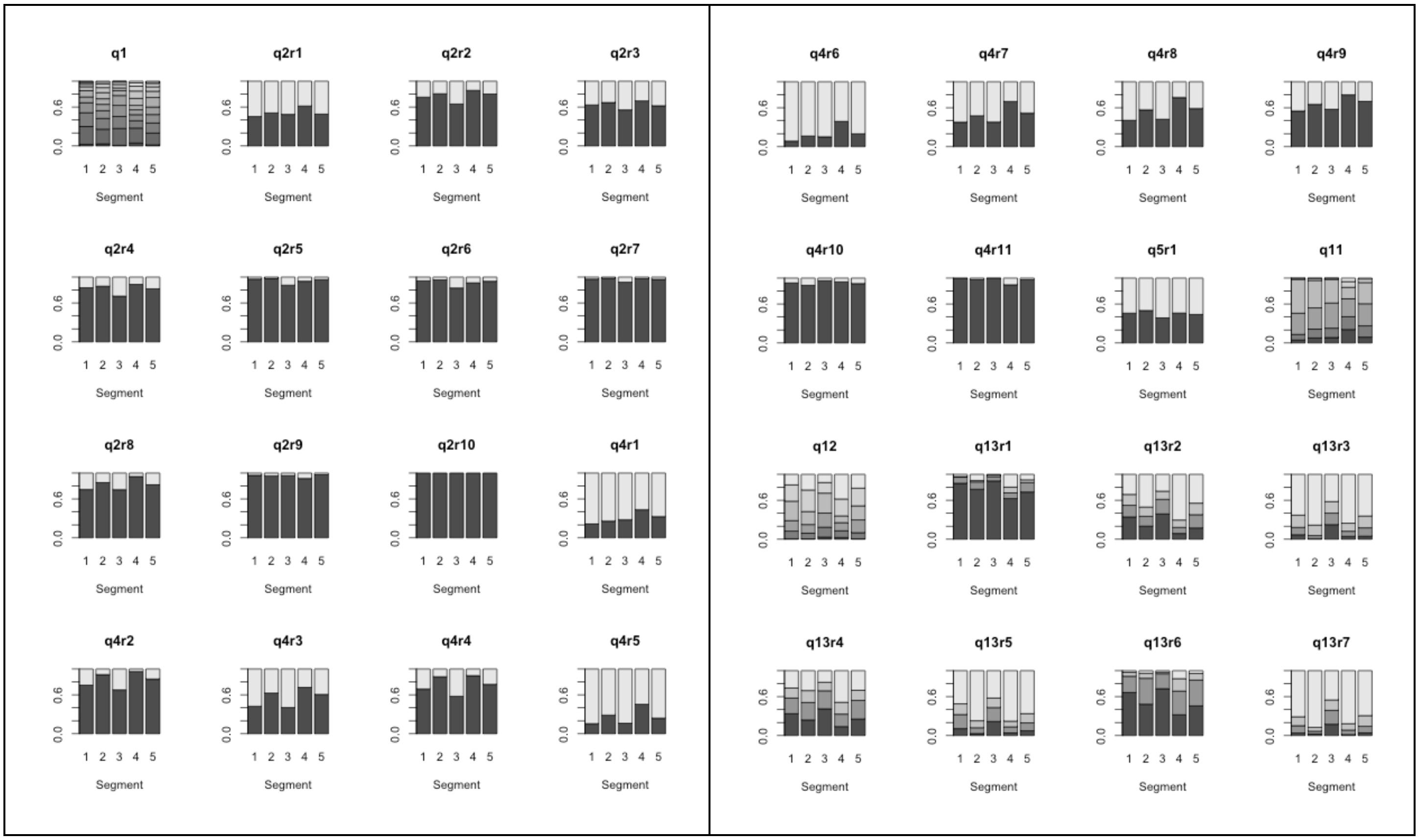


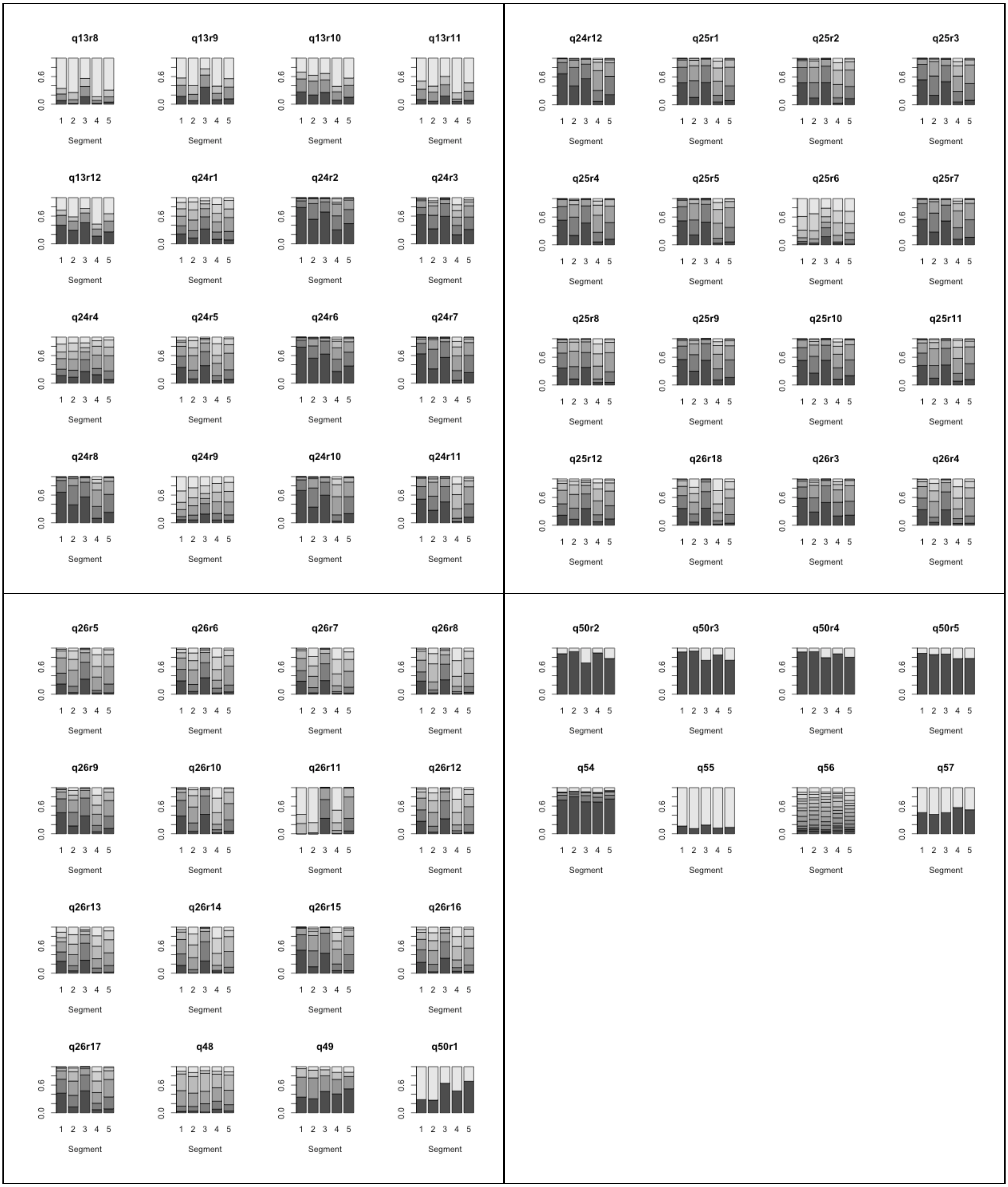


This schema was recommended by a classmate in the discussion forum, but was not used for modeling in the end since my basis variables performed better.

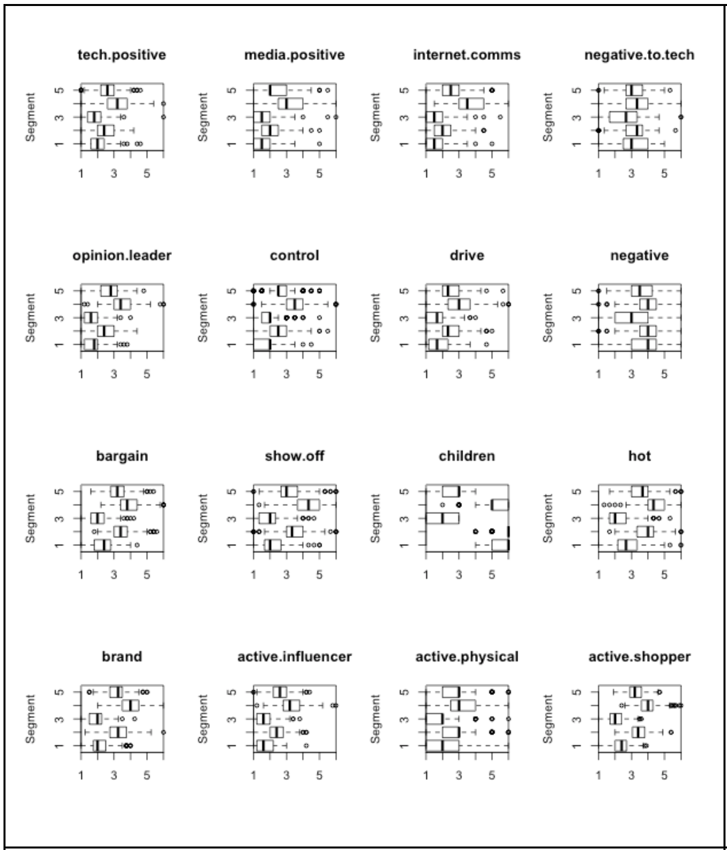
C. SEGMENTATION PROFILING - FREQUENCY PLOTS

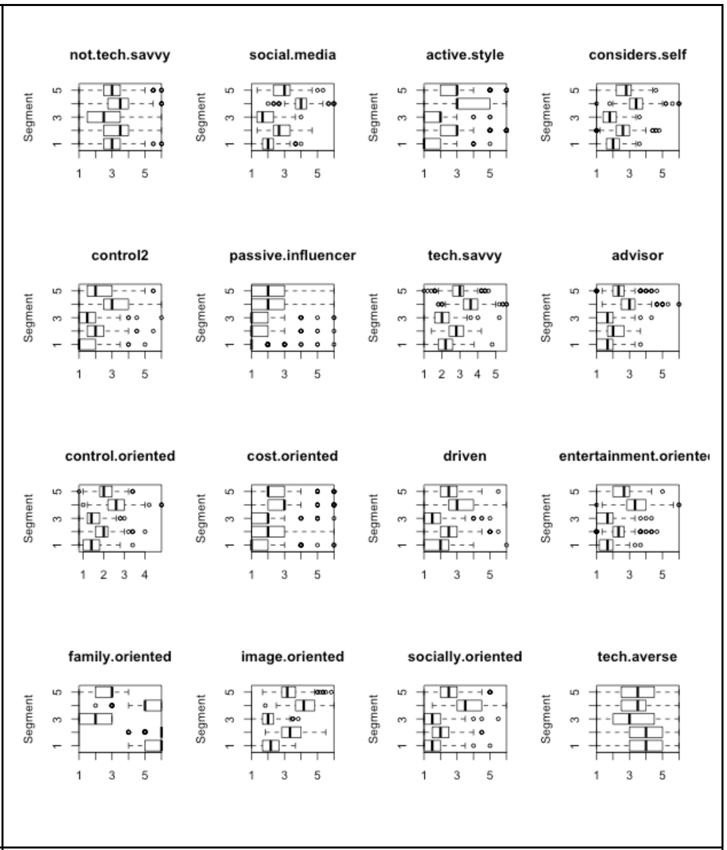
Frequency Bar Plots by Cluster (1-5) for Original Variables

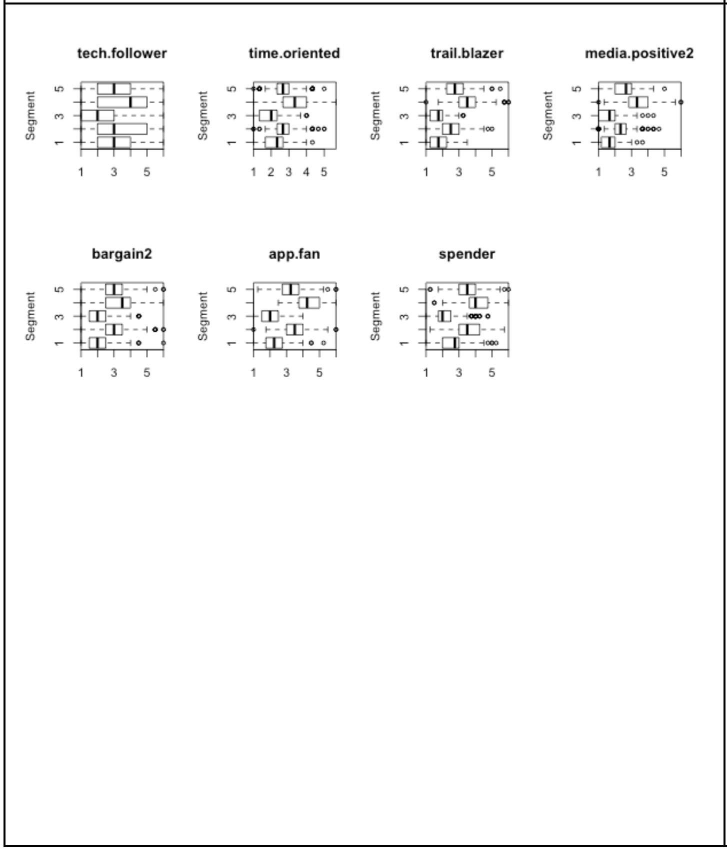




Box Plots by Cluster (1-5) for Derived Variables


tech.positive
media.positive
internet.comms
negative.to.tech
opinion.leader
control
drive
negative
bargain
show.off
children
hot
brand
active.influencer
active.physical
active.shopper
Segment
1
3
5
1
3
5


not.tech.savvy
social.media
active.style
considers.self
control2
passive.influencer
tech.savvy
advisor
control.oriented
cost.oriented
driven
entertainment.oriente
family.oriented
image.oriented
socially.oriented
tech.averse
Segment
1
3
5
1 2 3 4 5
1 2 3 4


tech.follower
time.oriented
trail.blazer
media.positive2
bargain2
app.fan
spender
Segment
1
3
5
1 2 3 4 5