Exercises 1

Probability Practice

Part A.

For this problem, I will define event A to be a clicker answering Yes, and I will define event B to be a clicker answering truthfully. The rule of total probability gives:

$$P(A) = P(A|B) * P(B) + P(A|B^c) * P(B^c)$$

We are looking to find $P(A|B)$, the probability that a truthful clicker answered yes, so we can rearrange the equation:

$$P(A|B) = \frac{P(A) - P(A|B^c) * P(B^c)}{P(B)}$$

We have that $P(A) = 0.65$. $P(A|B^c)$ is the probability that a random clicker answered yes, so we know that $P(A|B^c) = 0.5 * 0.65 = 0.15$. And then $P(B) = 0.7$. This gives us:

$$P(A|B) = \frac{0.65 - 0.15}{0.7} = \frac{0.5}{0.7} = 0.714$$

The fraction of people who are truthful clickers that answered yes is 0.714.

Part B.

In this problem, I will define event A to be that someone has the disease, and I will define event B to be that someone tests positive. We are looking for the probability of someone having the disease given that they tested positive, $P(A|B)$, and we can use Bayes' rule to calculate this.

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

We know that $P(A) = 0.000025$, and the *sensitivity* gives $P(B|A) = 0.993$. $P(B)$ can be broken down into the rule of total probability to say that:

$$P(B) = P(B|A) * P(A) + P(B|A^c) * P(A^c)$$

And this becomes

$$P(B) = 0.993 * 0.000025 + (1 - 0.9999) * (1 - 0.000025) = 0.0001248$$

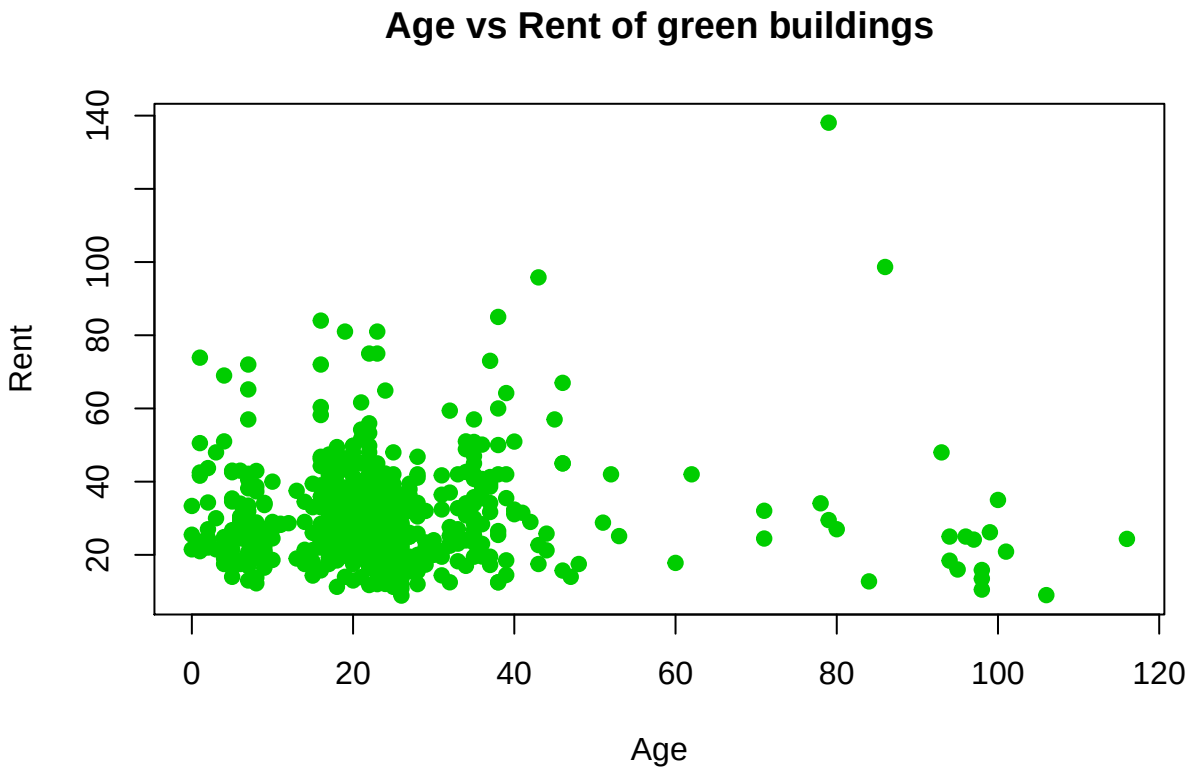
Then

$$P(A|B) = \frac{0.000025 * 0.993}{0.00001248} = 0.1989$$

The probability that someone who tests positive will have the disease is 0.1989. This seems like a very low reliability for a test, and it would be unwise to implement this test universally. There are going to be many people, over 80% actually, that think they have the disease when in fact they do not. This can cause not only emotional stress but financial burden, because those false positive testers will probably take steps to get treatment, consult other doctors, etc. and these things cost a lot of money. Ultimately this test is not good enough to implement in practice because it performs too poorly and will cause undue burden to the patients.

Exploratory analysis: green buildings

The analysis made by the stats guru is not comprehensive and fails to consider a few key points, and can therefore by improved in several ways. His analysis assumes that all green buildings are lumped together when calculating the rent per square foot per year. This will not work because all buildings will have some different properties that will affect this rent, such as the geographic location, age, number of stories, and building class. It would then be beneficial to look at buildings with similar properties to the one that the Austin real estate developer is looking to build. Unfortunately, we can't look at similarities in the geographical area of the building because we don't have the data, but we can look at some of the other factors. An important thing to note is that I decided to include all of the data and not remove any outliers, because as mentioned by the stats guru, the median value should be robust to outliers anyway.



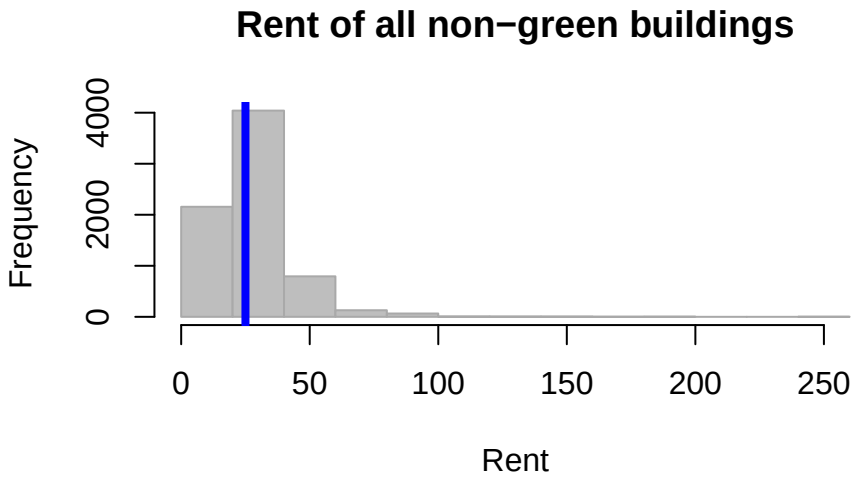
It can be seen that age doesn't really play a role in the rent price of green buildings, and that is confirmed when looking at the correlation between the two variables. Another factor to consider is the class of the building.

Here is the distribution of rent prices in class A green buildings overlayed with the distribution of rent in class B green buildings.

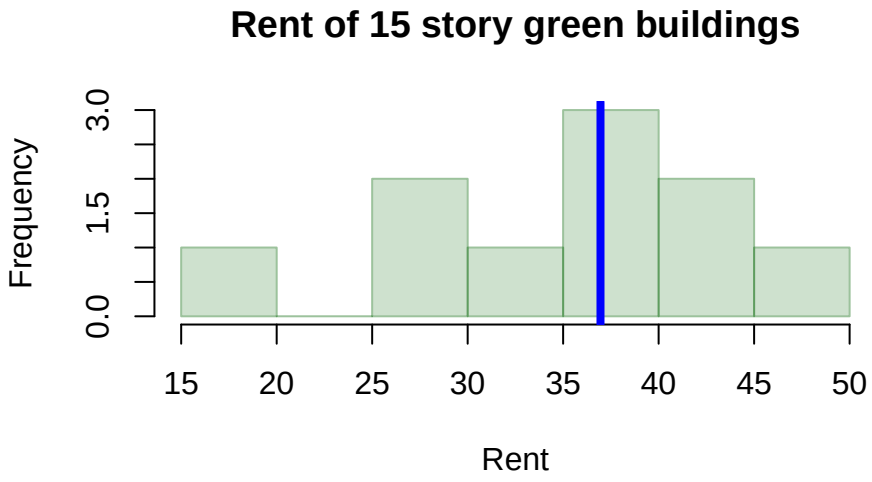
The blue line shows the median value of rent for class A, and the red line shows the median value of rent for class B. It can easily be seen that a class A building is likely going to bring in more rent, so the income

expected from the new building investment will depend on what class of building it will be.

Finally, an important factor to consider will be the number of stories of the new building. Because it has been specified to be 15 stories, I looked at all of the green buildings in the dataset that also had 15 stories, and compared their rent prices to that of all non-green buildings.



Here the median rent of all non-green buildings is shown to be 25, in dollars per square foot per year. Now looking at the green buildings with 15 stories:



Here the median value of rent is about 37 - much higher than that of non-green buildings. This could actually mean that the value of investing in a green building with 15 stories is much higher than expected, and that the developer could expect to recuperate costs of the green building premium much quicker than previously thought. It can't be directly calculated however, because of the other factors involved. For example, the fact that the building will be on East Cesar Chavez may mean it could bring in less rent than anticipated due to its location, but that can't necessarily be quantified here. My recommendation for the developer would be to look at buildings with similar features in that area of Austin to get an idea for how that will impact the rent, and then a more comprehensive assessment can be made as to the value of the investment.

Bootstrapping

In order to analyze the risk/return properties of each of the five ETFs, I first took a look at some of the summary data for these ETFs.

```
##    ClCl.SPYa            ClCl.TLTa            ClCl.LQDa
##  Min.   :-0.0984477   Min.   :-0.0504495   Min.   :-0.0911111
##  1st Qu.:-0.0038653   1st Qu.:-0.0055729   1st Qu.:-0.0018986
##  Median : 0.0006588   Median : 0.0006122   Median : 0.0004234
##  Mean   : 0.0004155   Mean   : 0.0002552   Mean   : 0.0002064
##  3rd Qu.: 0.0055835   3rd Qu.: 0.0058283   3rd Qu.: 0.0024613
##  Max.   : 0.1451977   Max.   : 0.0516616   Max.   : 0.0976772
##    ClCl.EEMa            ClCl.VNQa
##  Min.   :-0.1616620   Min.   :-0.1951372
##  1st Qu.:-0.0084284   1st Qu.:-0.0065609
##  Median : 0.0006993   Median : 0.0006952
##  Mean   : 0.0009638   Mean   : 0.0005033
##  3rd Qu.: 0.0088460   3rd Qu.: 0.0075746
##  Max.   : 1.8891250   Max.   : 0.1700654
```

I wanted to specifically look at the min and max statistics for each ETF, as that would give me an idea of how the ETF typically performs in terms of risk and return - with min being a measure of risk, and max being a measure of return.

We can see that the ETFs with the least amount of risk, or lowest absolute value of the minimum, are TLT and LQD - not surprisingly these are the bonds which would be expected to be less risky. SPY also has pretty low risk compared to the very high risk ETFs - EEM and VNQ. The riskiest ETF is the real estate, followed closely by emerging markets. In terms of the return, emerging markets does offer the highest possible return in terms of the max as well as the mean and median returns - so EEM is going to be the high risk/high return ETF. TLT and LQD offer pretty low returns in terms of the maximum, and based off of this and their risk measure, we can say these are the safe ETFs. SPY is somewhere in the middle, offering a pretty good return with not quite as high of risk.

For my analysis, I also simulated 4-week trading for each ETF individually to get the average final wealth and 5% value at risk if I had just taken one ETF in my portfolio. The result of this analysis gives similar results to looking at the summary data, finding TLT and LQD to have lower value at risk (VAR) and lower returns, while EEM and VNQ have much higher VAR and higher returns, with SPY somewhere in the middle. We can also look at a correlation matrix of these ETFs:

```
##             ClCl.SPYa  ClCl.TLTa  ClCl.LQDa  ClCl.EEMa   ClCl.VNQa
## ClCl.SPYa  1.0000000 -0.4383751 0.12463988  0.3900771  0.76802863
## ClCl.TLTa -0.4383751  1.0000000 0.42083612 -0.1613449 -0.25039738
## ClCl.LQDa  0.1246399  0.4208361 1.00000000  0.1024610  0.09029288
## ClCl.EEMa  0.3900771 -0.1613449 0.10246098  1.0000000  0.27962938
## ClCl.VNQa  0.7680286 -0.2503974 0.09029288  0.2796294  1.00000000
```

It's important to consider these correlations when choosing a portfolio for several reasons. To get a safe portfolio, you can diversify by having ETFs that are negatively correlated with each other, that way if one ETF does poorly, the other ETF will be doing well and can compensate. Another strategy would be to pick safe ETFs that are positively correlated with each other, that way you know you can rely on all of your ETFs. For an aggressive portfolio, you may want to pick ETFs that are highly positively correlated with each other to maximize your chance of high returns.

With all of these considerations, I will pick a "safe" portfolio consisting of SPY, TLT, and LQD. The TLT and LQD funds are very safe bets that minimize risk and have low return, and the SPY is medium as far as risk and return but it is negatively correlated with TLT so it should compensate if the bonds do poorly. I will go with a distribution of 30% in the SPY, 40% in the TLT, and 30% in the LQD fund.

For my "aggressive" portfolio, I will put 60% in the high risk/high return EEM fund, and 40% in the VNQ, which is also risky and is positively correlated with EEM.

Using an "even split" portfolio, the average return after using bootstrap resampling to estimate the 4-week trading period is:

```
## [1] 1013.788
```

And the 5% value at risk is:

```
##       5%
## 6200.862
```

Now looking at the "safe" portfolio, average return is:

```
## [1] 615.9337
```

And the 5% value at risk for the safe portfolio is:

```
##       5%
## 3058.656
```

Finally, for the "aggressive"" portfolio, the average return is:

```
## [1] 1673.782
```

And the 5% value at risk for the aggressive portfolio is:

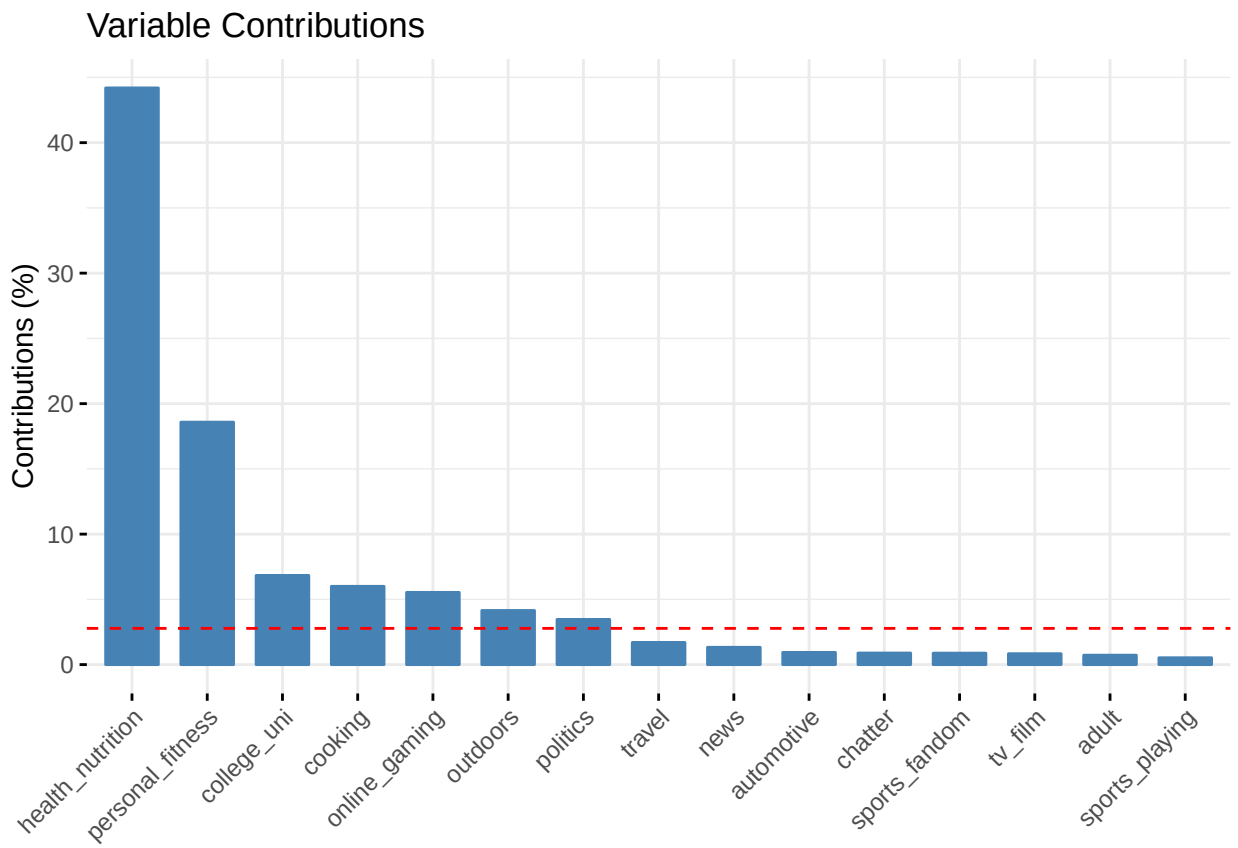
```
##       5%
## 12809.15
```

The results are pretty much as expected - the safe portfolio had the lowest average return and the lowest VAR, the aggressive portfolio had the highest average return and highest VAR, and the even split portfolio was in the middle. My recommendation for which portfolio to choose from will depend on the preference of the investor. An investor who is confident in the market and is looking to capitalize on gains and high returns should choose the aggressive portfolio - the average return is almost 3 times more than the safe portfolio and about 1.5 times higher than the even split. However the risk is about twice as much as the even split and 4 times greater than the safe portfolio, so I would advise this investor to proceed with caution.

Overall, the best option for a more conservative investor is going to be the safe portfolio, which has half the risk of the even split while still getting about 2/3 of the return that the even split makes. This portfolio would be a strong bet to ensure that your money keeps growing over time, and even when the markets do poorly, you won't be suffering too much loss.

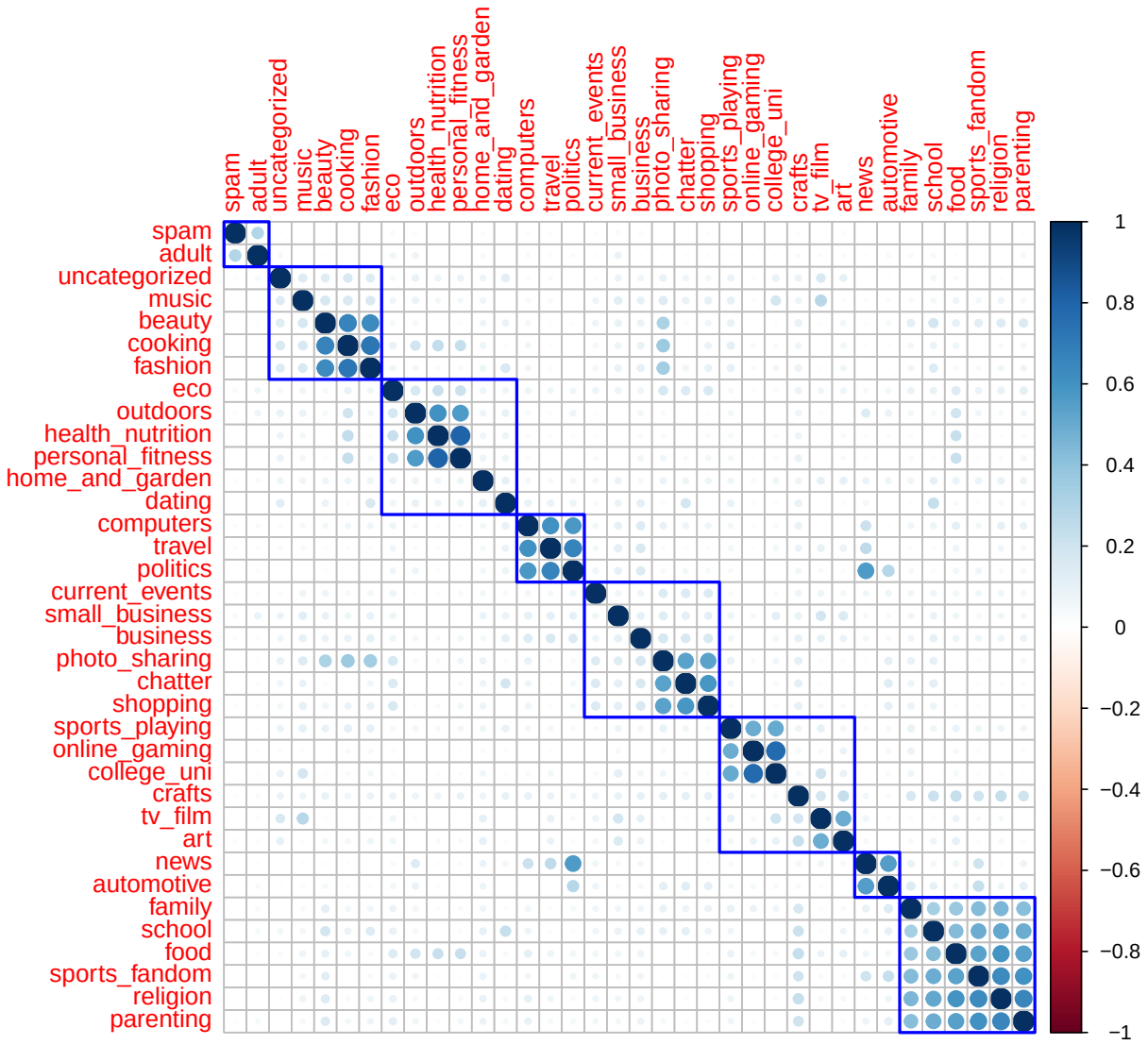
Market Segmentation

The goal of this report is to identify interesting market segments that appear to stand out among NutrientH2O's social-media audience. The first step was to explore the data and try to identify which interest categories stand out the most. This plot is a measure of importance of the top interests, or their individual contributions to the overall picture.



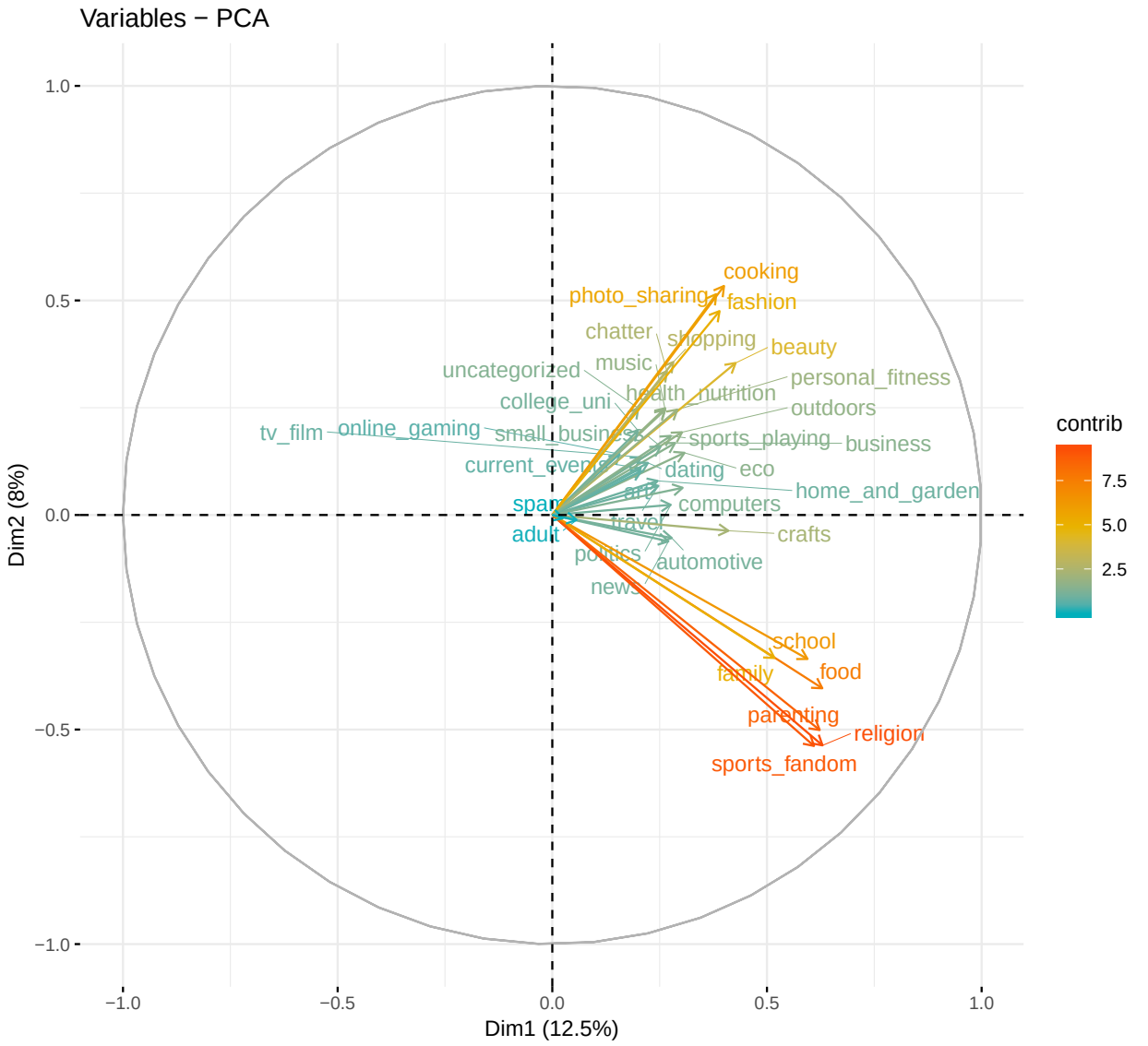
By this it appears that, not surprisingly, health and nutrition plays a large part in the conversations of NutrientH2O's Twitter followers, as well as fitness, cooking, and outdoor activities. And there appears to be a relatively strong influence by college students.

Next, examining the correlations between the interest categories - it's possible to group together some of the followers' interests into particular market segments through hierarchical clustering.



So there are several groups of interets that appear to be correlated together, and form several clusters, or market segments. The large group of interests correlated together in the bottom right is an unsurprising find - these are all things that a lot of people spend a majority of their time on - school, food, family, religion, sports - they're typically high priorities. Another interesting segment from this chart is the group with eco, outdoors, health and nutrition, fitness etc. that is also not surprising to see for this particular group of followers - those would all be people likely to consume the Nutrient H2O product.

One more way to identify market segments is through principal component analysis, and the results can be examined here:



The plot shows relative strengths and directions of the various interests - basically grouping them together as well as showing the prevalence of each interest. Again, seeing similar interests grouped together as from the correlations - a strong contribution from the group of family, school, sports, religion etc. and in this case there is a strong contribution from the cooking, fashion, beauty, and photo-sharing group. Also the health/fitness group can be seen together again, so there is strong commonality in the market segments across the different methods of analysis.

Overall, there can be four major market segments identified in the data:

1. People interested in health/nutrition, fitness, and the outdoors
2. People focused on high priority core values such as family and school
3. People interested in creative endeavors such as photography, beauty, and fashion
4. A group of "others" - comprising everything else such as news and tv, gaming, cars, business, and the adult and spam categories