Predictive Models: Exercise 1

Due on August 6, 2017 at 3:10pm

Professor James Scott

Matt Barrett **Timothy Lai** **Brett Scroggins** **Meyappan Subbaiah**

Problem 1

Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3.

After a trial period, you get the following survey results: 65% said Yes and 35% said No.

What fraction of people who are truthful clickers answered yes?

Solution:

Look for **P(Yes | Truthful)** = ?

Knowns:

1. Two Categories: Random Clicker (RC), and Truthful Clicker (TC)
2. P(Yes | RC) = P(NO | RC) = 0.5
3. P(RC) = 0.3 ; P(TC) = 0.7
4. P(Yes) = 0.65 ; P(No) = 0.35

Using the rule of total probability,

$$
\begin{aligned}
P(Y) &= P(Y, RC) + P(Y, TC) \\
P(Y) &= P(Y \mid RC) * P(RC) + P(Y \mid TC) * P(TC) \\
0.65 &= (0.5)(0.3) + P(Y \mid TC) * 0.7 \\
0.65 &= 0.15 + P(Y \mid TC) * 0.7 \\
0.5 &= P(Y \mid TC) * 0.7 \\
P(Y \mid TC) &= \frac{0.5}{0.7} \\
P(Y \mid TC) &= 0.714
\end{aligned}
$$

Problem 2

Imagine a medical test for a disease with the following two attributes:

1. The *sensitivity* is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.
2. The *specificity* is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.

In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).

Suppose someone tests positive. What is the probability that they have the disease? In light of this calculation, do you envision any problems in implementing a universal testing policy for the disease?

Solution:

Knowns:

1. Sensitivity is 0.993 ; P(Positive — Disease) = 0.993
2. Specificity is 0.9999 ; P(Negative — No Disease) = 0.9999
3. P(Disease) = 0.000025

Look for **P(Disease | Positive)** = ?

Using Bayes rule: $P(A \mid B) = \dfrac{P(A) * P(B \mid A)}{P(B)}$

First we find P(Positive) using the rule of total probability.

$$P(Positive) = P(Positive \mid Disease) * P(Diseases) + P(Positive \mid NoDisease) * P(NoDisease)$$
$$P(Pos) = (0.993)(0.000025) + (0.0001)(0.999975)$$
$$P(Pos) = 0.0001248225$$

Now we have all the terms to plug into Bayes rule.

$$P(Disease \mid Positive) = \frac{(0.000025)(0.993)}{(0.0001248225)}$$
$$P(Disease \mid Positive) = 0.19888 \approx 0.2$$

Explanation: Based upon the results found, the universal testing policy does not seem effective enough to warrant testing across the population. This is heavily due to the fact that 80% of positive results are false positives, and the very small fraction of the general population that are expected to actually be afflicted with the disease. However, if widespread testing was determined to be necessary despite poor cost-effectiveness, the results from this test would be reliable enough to meet the testing needs. This is mainly due to the very, very low likelihood of false negative testing. Additionally, although 4 of the 5 positive results would turn out to be negative, providing a second round of this test to verify would result in drastically reduced false results without significant risk of false negatives.

Problem 3

Exploratory Analysis: Green Buildings

Developer thoughts: I began by cleaning the data a little bit. In particular, I noticed that a handful of the buildings in the data set had very low occupancy rates (less than 10% of available space occupied). I decided to remove these buildings from consideration, on the theory that these buildings might have something weird going on with them, and could potentially distort the analysis. Once I scrubbed these low-occupancy buildings from the data set, I looked at the green buildings and non-green buildings separately. The median market rent in the non-green buildings was $25 per square foot per year, while the median market rent in the green buildings was $27.60 per square foot per year: about $2.60 more per square foot. (I used the median rather than the mean, because there were still some outliers in the data, and the median is a lot more robust to outliers.) Because our building would be 250,000 square feet, this would translate into an additional $250000 x 2.6 = $650000 of extra revenue per year if we build the green building.

Our expected baseline construction costs are $100 million, with a 5% expected premium for green certification. Thus we should expect to spend an extra $5 million on the green building. Based on the extra revenue we would make, we would recuperate these costs in $5000000/650000 = 7.7 years. Even if our occupancy rate were only 90%, we would still recuperate the costs in a little over 8 years. Thus from year 9 onwards, we would be making an extra $650,000 per year in profit. Since the building will be earning rents for 30 years or more, it seems like a good financial move to build the green building.

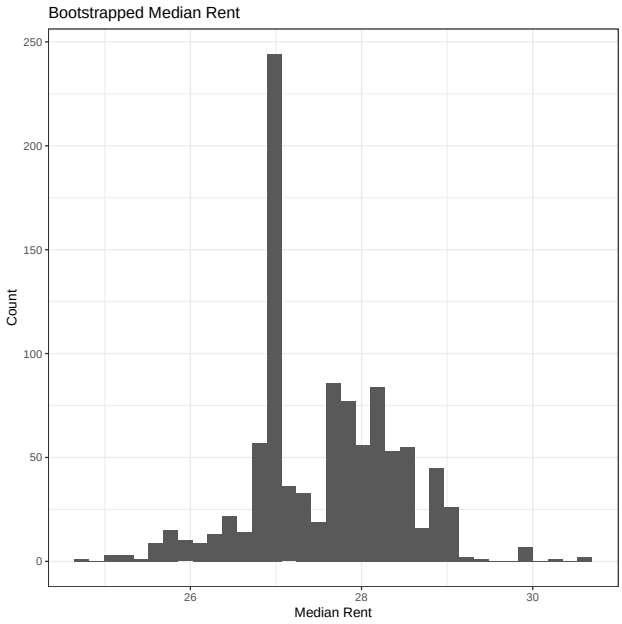
She has therefore asked you to revisit the report, so that she can get a second opinion.

Do you agree with the conclusions of her on-staff stats guru? If so, point to evidence supporting his case. If not, explain specifically where and why the analysis goes wrong, and how it can be improved. (For example, do you see the possibility of confounding variables for the relationship between rent and green status?)

Solution:

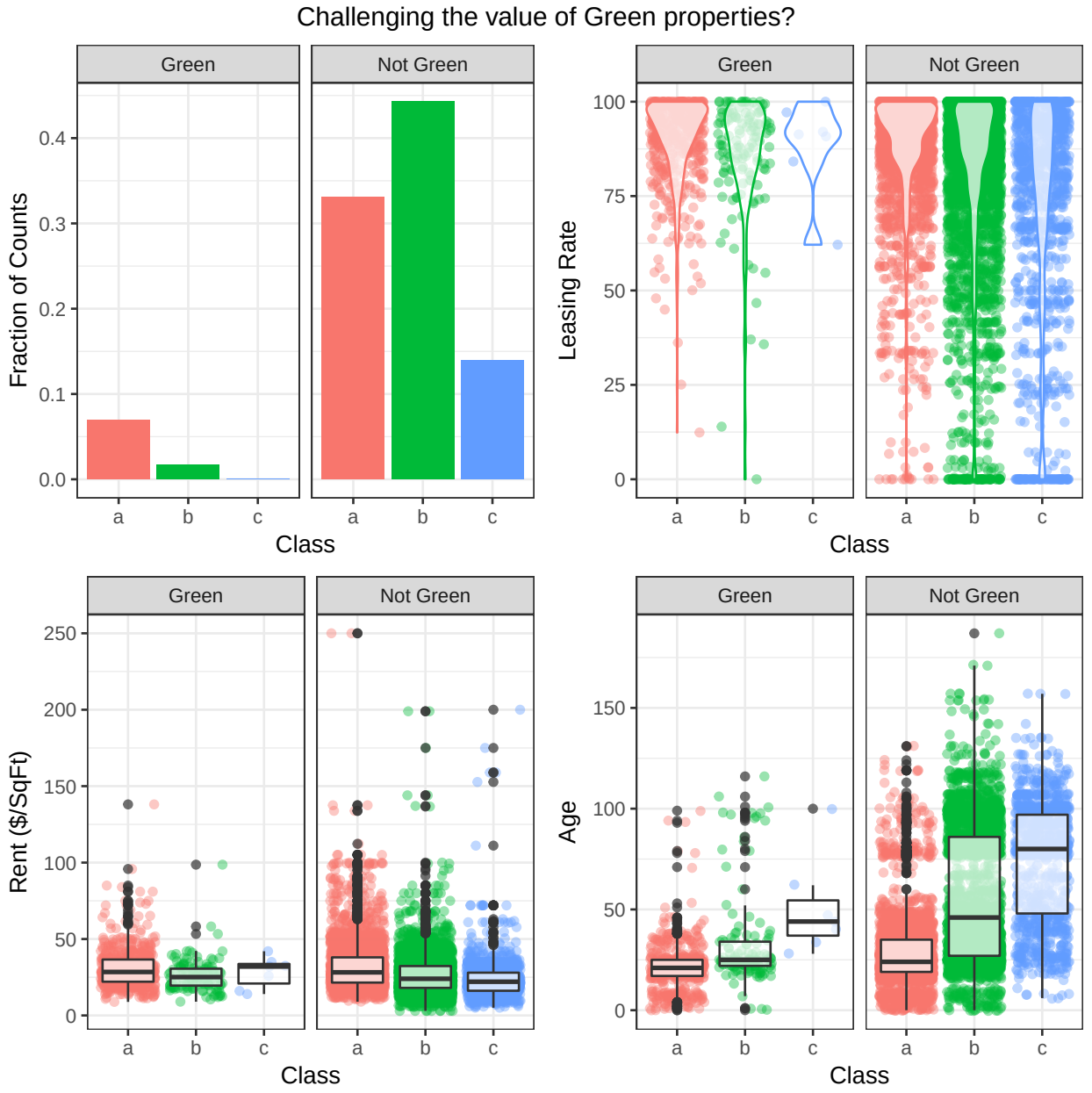
The Excel guru has determined that constructing green buildings will have a strong commercial upside. The guru suggests that there is a premium to go green, and the additional revenue received in rent collections will provide a quicker pay-off than building non-green buildings. However, the guru's analysis and recommendations were likely oversimplified. To verify the guru's findings further before committing to such large investments, further exploratory data analysis was completed.

Problem 3 continued on next page...



According to the Excel guru's initial assessment of the data, there is a $2.60 premium for green buildings. We decided to run a bootstrap on green buildings only in order to verify this statement. From the plot above, we see that there is a high variability between median rents of green buildings, as evidenced by the non-normal distribution. Therefore, the median is not a strong metric to use in this case, and was not a good metric for which to recommend whether or not green buildings are viable from a business perspective.

Problem 3 continued on next page...



Next, we sought to identify whether or not there is a significant premium in rent for "green" buildings. The four plots presented each provide some correlated insight to suggest that there is in fact no such premium. In the top left plot, the normalized fraction of counts of the number of each class of buildings is shown. It is evident that the majority of buildings in the data set are not green buildings (in fact, less than 10% of buildings are "green"). It is also clear that buildings that are green were skewed heavily towards class "A"; although these characteristics are not directly related. In looking at the bottom left plot of rent vs. class, we are also able to identify that there is not a significant increase in average rent between green and non-green buildings of the same class. Therefore, other factors have a stronger influence on defining the rent. In the top right plot depicting leasing rate vs. class, we see that the majority of green buildings have a 75% or greater leasing rate. However, a more important factor in determining leasing rate is clearly class. The mean leasing rate for both Classes "A" and "B" are similar, whether the buildings are green or non-green. Lastly, in the bottom right plot depicting age and class, we see that the majority of Class "A" buildings are significantly younger in comparison to Class "B" and "C". The so-called "higher end" buildings are clearly

younger, and it is therefore apparent that a combination of age and class would be more important factors towards determining leasing rate and rent as opposed to basing determination from the classification of the building as "green".

Table 1: Median Rent per Class

Class	Green Rating	Median
A	Green	$28.44
A	Not Green	$28.20
B	Green	$25.10
B	Not Green	$24.00
C	Green	$32.00
C	Not Green	$22.06

As evidenced above, being "green" ultimately is not the main predictor in whether or not there is a premium in rent. Rather, the class of the building is a much stronger predictor and should have been presented to the developer when determining the viability of constructing new buildings. To quantify the median rents between classes, the table above is presented. From the table, we can see that the difference between median rents for green vs. not green Class A buildings only shows a 24 cent difference. There is not much added commercial benefit to building a green Class A building vs. a "not green" Class A building. Since class is defined by high quality finishes, state of the art features, and strong marketability, constructing a "Class A" building without additional green features would already be a strong investment, given the earlier data on high leasing rate and eliminating the need for additional building cost to get to "green status".

According to the Excel guru, the certification for "green" would cost an estimated $5 million based on $100 million building costs. Since we have determined that there would be little to no benefit from "green" certification as opposed to class determination, it seems more fiscally responsible to invest in ensuring class "A" certification than in "green" certification.

Problem 4

Bootstrapping:

Consider the following five asset classes, together with the ticker symbol for an exchange-traded fund that represents each class:

1. US domestic equities (SPY: the S&P 500 stock index)
2. US Treasury bonds (TLT)
3. Investment-grade corporate bonds (LQD)
4. Emerging-market equities (EEM)
5. Real estate (VNQ)

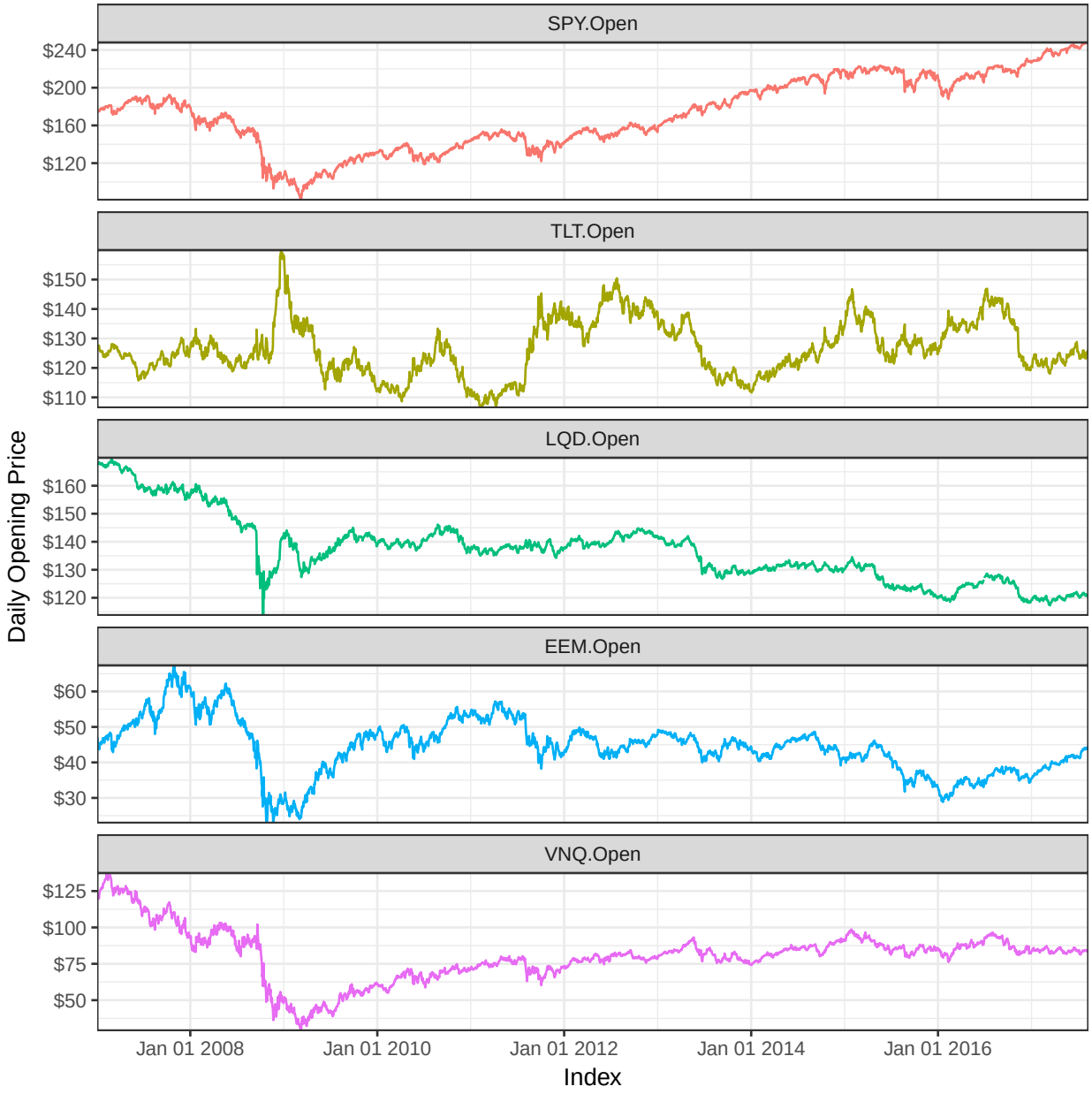
Suppose there is a notional $100,000 to invest in one of these portfolios. Write a brief report that:

1. marshals appropriate evidence to characterize the risk/return properties of the five major asset classes listed above.
2. outlines your choice of the "safe" and "aggressive" portfolios.
3. uses bootstrap re-sampling to estimate the 4-week (20 trading day) value at risk of each of your three portfolios at the 5% level.
4. compares the results for each portfolio in a way that would allow the reader to make an intelligent decision among the three options.

Solution:

For this problem, we will be looking at five stocks and creating a portfolio based on wealth distribution within these stocks. First, to determine the variability of the stock, the plot below was created to see how stock prices gained and lost from January 2007 to June 2017.

Problem 4 continued on next page...



As can be seen above by looking at the US domestic equities (SPY) stock, the stock market generally goes up over this time period. What is different though is the amount of volatility in each respective stock. As can be seen, emerging-market equities (EEM) and real estate (VNQ) are the more variable stocks in this portfolio, making them riskier investments. Conversely, when observing US Treasury bonds (TLT) and investment-grade corporate bonds (LQD), these seem to be much safer investments as their stocks are much quieter in terms of movement over time.

Problem 4 continued on next page...

Table 2: Stock Volatillity

Beta over Time	name
-0.33	TLT
0.04	LQD
1.32	EEM
1.34	VNQ

To characterize the risk/return profiles of each asset class, the beta coefficient was calculated for each ETF relative to the S&P 500 index. β is a measure of the volatility of a security in comparison to the market. For the case of the 5 ETF's presented, SPY was chosen to be the index, as it is directly correlated with S&P 500 ($\beta = 1$). If β is greater than 1, the security is expected to be less volatile and be a safer choice for those with a "conservative" mindset.

In calculating the coefficients for the four asset classes (again, SPY used as the index), we see that EEM and VNQ show the highest volatility. This is expected, as emerging markets and real estate generally have higher risk/reward profiles. On the other hand, bonds are generally thought of as safe choices for investors who are risk averse. In ranking the volatility form greatest to worst (and thus characterizing risky to conservative), we have classified VNQ and EEM as "risky", while TLT and LQD are "safe".

For our 20 day trial period estimating the 5% value-at-risk, we ran three separate portfolios. The associated weights are outlined below:

1. Even: 20% to each stock
2. Safe: 40% TLT and LQD, and 20% SPY
3. Risky: 40% EEM and VNQ, and 20% SPY

Table 3: Risk Assessment Portfolio

	SP	TLT	LQD	EM	RE	val
Even	0.20	0.20	0.20	0.20	0.20	-6127.95
Safe	0.20	0.40	0.40	0.00	0.00	-3066.33
Risky	0.20	0.00	0.00	0.40	0.40	-11928.34

What the table above shows is what one would expect - the risky portfolio has a much higher value-at-risk while the safer portfolio has a much lower value-at-risk. On the opposite end of the spectrum, the risky portfolio could potentially produce at a much higher rate when comparing with the safe portfolio if the market conditions are right. Lastly, as can be seen, the 20% splits portfolio tends to land near the middle for both of these factors.

With all of this in mind, the question is how should one invest in these five stocks? While the risky portfolio clearly has much more risk involved, in the long run, could be much more progressive than the other portfolios. This would be recommended for a more patient investor who could wait out the ups and downs of the market. Conversely, the conservative portfolio would be recommended for the risk averse investor. Finally ,the 20% splits (equal) fund creates a balance between these two options. While it does not have the same return on investment that the risky fund does, it produces a much higher return than that of the safe portfolio.

Problem 5

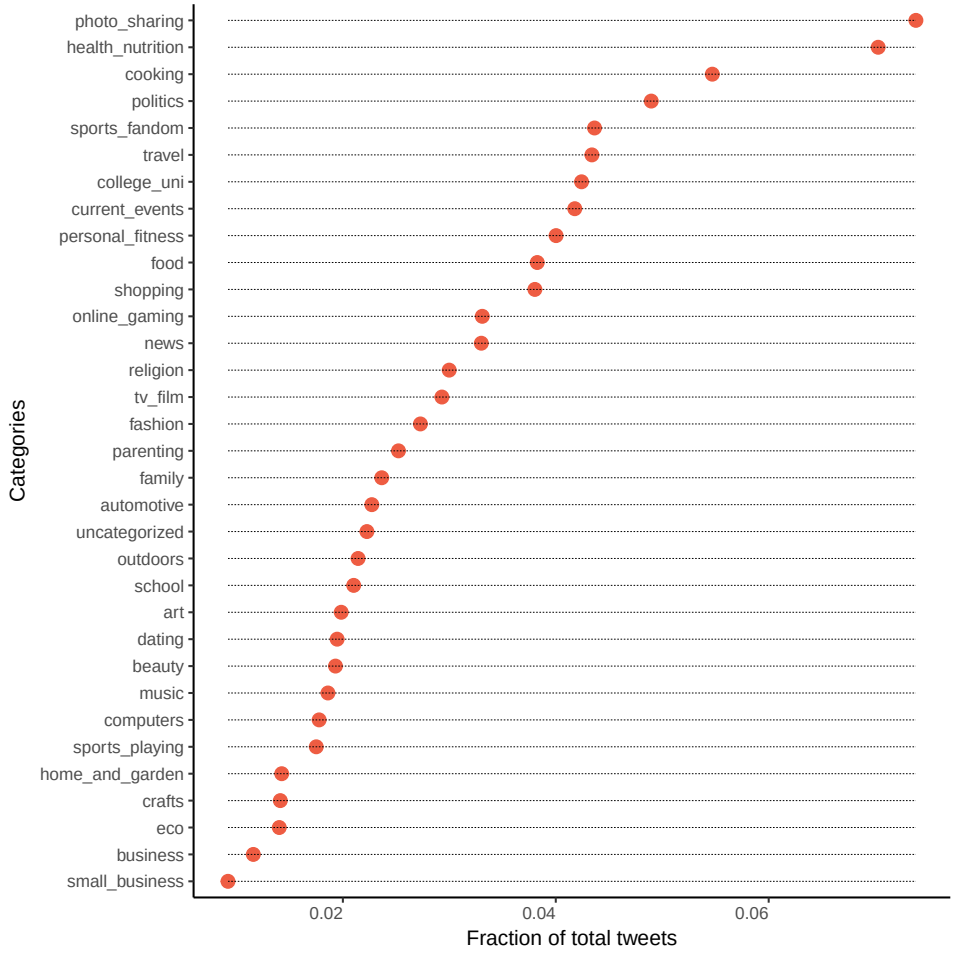
Market Segmentation:

Consider the data in social_marketing.csv. This was data collected in the course of a market-research study using followers of the Twitter account of a large consumer brand that shall remain nameless—let's call it "NutrientH20" just to have a label. The goal here was for NutrientH20 to understand its social-media audience a little bit better, so that it could hone its messaging a little more sharply.

Your task to is analyze this data as you see fit, and to prepare a (short!) report for NutrientH20 that identifies any interesting market segments that appear to stand out in their social-media audience. You have complete freedom in deciding how to pre-process the data and how to define "market segment." (Is it a group of correlated interests? A cluster? A latent factor? Etc.) Just use the data to come up with some interesting, well-supported insights about the audience.

Solution:

For this exercise, we will attempt to identify market segments that would be potential targets in marketing a new NutrientH2O "water" product. To do this, we first began by doing preliminary data cleaning for the social marketing data. In looking at the observation points, we noticed that there were several categories that could be removed for purposes of reducing dimensions - the "turk index", "spam", and "adult". Since there are many tweets in the data that represent unsolicited advertising and explicit content, we decided to remove these counts from our data set to focus on more relevant data for a potential marketing audience. Additionally, we noticed that most tweets fell under the "chatter" category, which for the purposes of this exercise also added no value and would have been classified as "noise". With the remaining 33 categories, we moved on to check variable importance so we could try to focus in on different clusters of groups that could each be uniquely targeted.



As can be seen in the plot above, it is clear that photo sharing and health nutrition are the most commonly tweeted about topics. While allocating resources to marketing towards these two categories would be effective, there are still significant other factors that need to be accounted for. After those top two factors, there is a wide range of additional factors from cooking (at approximately 5.5% of tweets) to art (at approximately 2% of tweets) from all followers. Seeing as a majority of tweets fell within these percentages, it was important to try and broaden the marketing to more categories. Clustering was done in order to group similar users and their tweets to better utilize interesting marketing segments for product promotion.

Table 4: Tweet Categories by Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
politics	religion	health_nutrition	cooking	college_uni	tv_film
news	parenting	personal_fitness	fashion	online_gaming	art
travel	sports_fandom	outdoors	beauty	sports_playing	shopping
computers	food	eco	photo_sharing	tv_film	current_events
automotive	school	food	music	music	uncategorized

Once a useful subset was made, the information was scaled and a seed was set to provide reproducibility. This data was then clustered using k folds of 6 total clusters. We then identified the centers of these clusters to determine the categories that fell into each grouping. Once each cluster was defined with a unique set of categories, we would be able to make inferences based on tweets that fell into similar categories.

Problem 5 continued on next page...

In examining each cluster in more detail, we wanted to be able to infer descriptions for each groups of users to potentially market the product. The first cluster included 686 users, whose main tweets focused on a combination of travel, news, politics, computers, and automotive. We concluded that this cluster may comprise of middle aged, white collar adults. The second cluster included 761 users, focused on sports, food, family, religion, school, and parenting. This group could be inferred to be conservative religion and sports first families. The third cluster included 889 users who mainly tweeted about health nutrition, personal fitness, and the outdoors. The next two categories for this cluster were eco and food, but neither of these had high enough magnitude to merit value despite being in the top 5 for this cluster. This would certainly be a group for NutritionH2O to target, as it likely consists of athletes or health conscious individuals. The fourth cluster included 568 users who were interested in photo sharing, cooking, beauty and fashion. The final category for this cluster's top five was music, but the magnitude again was very low for this category. This group likely consists of women and could potentially be a marketing target in terms of branding the product. The fifth cluster included 440 users who mostly tweeted about online gaming, college/university, and sports. The final two categories of this cluster's top five was with music and TV/film. These categories were low in frequency magnitude, and didn't appear to influence the cluster. This group likely comprised of individuals aged 18-25 and would likely be receptive to trying a new "healthy" water product. Finally, the last cluster included 4538 users who had minimal activity - these users were likely lurkers who we would need more information about in order to market the product. All five of the listed categories in this cluster were low in tweet magnitude, as such the listed categories are of little to no value.

To conclude, we have clustered Twitter users and identified different marketing audiences to target. The two obvious clusters include Cluster 3 - the "athletes", and Cluster 5 - the "millenials" who are also interested in sports. Ultimately the data analyses yields information for current followers, and additionally an audience to target moving forward to further grow the product.