STA 380, Part 2: Exercises 1

Turn these in by 5 PM on Friday, August 7th. Prepare your reports using RMarkdown so that they are fully reproducible, carefully integrating visual and quantitative evidence with prose. You should submit your work by sending me a link to a GitHub page where the final report has been stored – preferably in Markdown format but PDF is OK too, especially if you want to include mathematical expressions in the manner described here, since GitHub doesn't do math very well. Also include a link to the raw .Rmd file that can be used to reproduce your report from scratch.

You can either e-mail me the link to send a message through Canvas, but please use the subject line "STA 380 Homework 1: Lastname, Firstname" so that I can sort my inbox easily. (Obviously use your own first and last names in the subject.)

Note: I want your report to be fully reproducible. Of course, it would seem that, by its very nature, one thing that cannot be reproduced exactly is a Monte Carlo simulation. But in fact you *can* reproduce such a simulation, if you specify a "seed" to the underlying random number generator. Thus these two sets of 10 normal random numbers are different (try copying and pasting to an R console):

```
rnorm(10)
rnorm(10)
```

But these two are the same, because in each case we reset the random-number seed to be the same thing:

```
my_favorite_seed = 1234567
set.seed(my_favorite_seed)
rnorm(10)
set.seed(my_favorite_seed)
rnorm(10)
```

You can use this fact to your advantage to create fully reproducible Monte Carlo simulations in RMarkdown, by setting the seed at the very beginning of the file.

Exploratory analysis

Consider the data in georgia2000.csv, which contains Georgia's county-level voting data from the 2000 presidential election. You might recall that the 2000 election was among the most controversial in history, and turned on an esoteric set of issues surrounding voting machines, vote counts, and the Equal Protection Clause of the Constitution.

This file contains the following information for all 159 counties in Georgia: - votes: number of votes recorded
- ballots: number of ballots cast
- equip: voting equipment (lever, optical, paper, punch card)
- poor: coded 1 if more than 25% of the residents in a county live below 1.5 times the federal poverty line; coded 0 otherwise.
- perAA: percent of people in the county who are African-American
- urban: indicator of whether county is predominantly urban (1)
- atlanta: indicator of whether the county is in Atlanta (1)
- gore: number of votes for Gore
- bush: number of votes for Bush

Your goal is to investigate the issue of vote undercount, or the difference between the number of ballots cast and the number of legal votes recorded. There can be many different reasons for undercount. Voters may

have chosen not to vote for any presidential candidate; they may have voted for more than one candidate, in which case their votes were disqualified; they may have misunderstood the instructions on the ballot; or the equipment may have simply failed to register their choices.

Your goal is to make any plots, tables, or numerical summaries that you believe illuminate two issues: (1) whether voting certain kinds of voting equipment lead to higher rates of undercount, and (2) if so, whether we should worry that this effect has a disparate impact on poor and minority communities. Note: while you should feel free to use any of the tools from the "supervised learning" half of the course, this is intended mainly as a warm-up exercise in visual and numerical story-telling. Keep it concise.

Bootstrapping

Consider the following five asset classes, together with the ticker symbol for an exchange-traded fund that represents each class: - US domestic equities (SPY: the S&P 500 stock index) - US Treasury bonds (TLT)
- Investment-grade corporate bonds (LQD)
- Emerging-market equities (EEM)
- Real estate (VNQ)

If you're unfamiliar with exchange-traded funds, you can read a bit about them here.

Download five years or so of daily data on these ETFs, using the functions in the `fFimport` package. Explore the data and come to an understanding of the risk/return properties of these assets. Then consider three portfolios: - the even split: 20% of your assets in each of the five ETFs above.
- something that seems safer than the even split, comprising investments in at least three classes. You choose the allocation, and you can certainly invest in more than three assets if you want. (You can even choose different ETFs if you want.)
- something more aggressive (again, you choose the allocation) comprising investments in at least two classes/assets. By more aggressive, I mean a portfolio that looks like it has a chance at higher returns, but also involves more risk of loss.

Suppose there is a notional $100,000 to invest in one of these portfolios. Write a brief report that:
- marshals appropriate evidence to characterize the risk/return properties of the five major asset classes listed above.
- outlines your choice of the "safe" and "aggressive" portfolios.
- uses bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of your three portfolios at the 5% level.
- compares the results for each portfolio in a way that would allow the reader to make an intelligent decision among the three options.

You should assume that your portfolio is rebalanced each day at zero transaction cost. That is, if you're aiming for 50% SPY and 50% TLT, you always redistribute your wealth at the end of each day so that the 50/50 split is retained, regardless of that day's appreciation/depreciation.

Clustering and PCA

The data in wine.csv contains information on 11 chemical properties of 6500 different bottles of *vinho verde* wine from northern Portugal. In addition, two other variables about each wine are recorded: - whether the wine is red or white
- the quality of the wine, as judged on a 1-10 scale by a panel of certified wine snobs.

Run both PCA and a clustering algorithm of your choice on the 11 chemical properties (or suitable transformations thereof) and summarize your results. Which dimensionality reduction technique makes more sense to you for this data? Convince yourself (and me) that your chosen method is easily capable of distinguishing the reds from the whites, using only the "unsupervised" information contained in the data on chemical properties. Does this technique also seem capable of sorting the higher from the lower quality wines?

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.

A bit of background on the data collection: the advertising firm who runs NutrientH20's online-advertising campaigns took a sample of the brand's Twitter followers. They collected every Twitter post ("tweet") by each of those followers over a seven-day period in June 2014. Every post was examined by a human annotator contracted through Amazon's Mechanical Turk service. Each tweet was categorized based on its content using a pre-specified scheme of 36 different categories, each representing a broad area of interest (e.g. politics, sports, family, etc.) Annotators were allowed to classify a post as belonging to more than one category. For example, a hypothetical post such as "I'm really excited to see grandpa go wreck shop in his geriatic soccer league this Sunday!" might be categorized as both "family" and "sports." You get the picture.

Each row of social_marketing.csv represents one user, labeled by a random (anonymous, unique) 9-digit alphanumeric code. Each column represents an interest, which are labeled along the top of the data file. The entries are the number of posts by a given user that fell into the given category. Two interests of note here are "spam" (i.e. unsolicited advertising) and "adult" (posts that are pornographic, salacious, or explicitly sexual). There are a lot of spam and pornography "bots" on Twitter; while these have been filtered out of the data set to some extent, there will certainly be some that slip through. There's also an "uncategorized" label. Annotators were told to use this sparingly, but it's there to capture posts that don't fit at all into any of the listed interest categories. (A lot of annotators may used the "chatter" category for this as well.) Keep in mind as you examine the data that you cannot expect perfect annotations of all posts. Some annotators might have simply been asleep at the wheel some, or even all, of the time! Thus there is some inevitable error and noisiness in the annotation process.

Your task to is analyze this data as you see fit, and to prepare a 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.