# STA 380, Part 2: Exercises 1

Turn these in by the start of class on Monday, August 8th. 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 the following two sets of random numbers 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.

## Probability practice

### Part A.

Here's a question a friend of mine was asked when he interviewed at Google.

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?

### Part B.

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

## Exploratory analysis: green buildings

### The case

Over the past decade, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious buildings. There are both ethical and economic forces at work here. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. In this context, the decision to invest in eco-friendly buildings could pay off in at least four ways.

1. Every building has the obvious list of recurring costs: water, climate control, lighting, waste disposal, and so forth. Almost by definition, these costs are lower in green buildings.

2. Green buildings are often associated with better indoor environments—the kind that are full of sunlight, natural materials, and various other humane touches. Such environments, in turn, might result in higher employee productivity and lower absenteeism, and might therefore be more coveted by potential tenants. The financial impact of this factor, however, is rather hard to quantify ex ante; you cannot simply ask an engineer in the same way that you could ask a question such as, "How much are these solar panels likely to save on the power bill?"

3. Green buildings make for good PR. They send a signal about social responsibility and ecological awareness, and might therefore command a premium from potential tenants who want their customers to associate them with these values. It is widely believed that a good corporate image may enable a firm to charge premium prices, to hire better talent, and to attract socially conscious investors.

4. Finally, sustainable buildings might have longer economically valuable lives. For one thing, they are expected to last longer, in a direct physical sense. (One of the core concepts of the green-building movement is "life-cycle analysis," which accounts for the high front-end environmental impact of acquiring materials and constructing a new building in the first place.) Moreover, green buildings may also be less susceptible to market risk—in particular, the risk that energy prices will spike, driving away tenants into the arms of bolder, greener investors.

Of course, much of this is mere conjecture. At the end of the day, tenants may or may not be willing to pay a premium for rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market.

The file greenbuildings.csv contains data on 7,894 commercial rental properties from across the United States. Of these, 685 properties have been awarded either LEED or EnergyStar certification as a green building. You can easily find out more about these rating systems on the web, e.g. at www.usgbc.org. The basic idea is that a commercial property can receive a green certification if its energy efficiency, carbon footprint, site selection, and building materials meet certain environmental benchmarks, as certified by outside engineers.

A group of real estate economists constructed the data in the following way. Of the 1,360 green-certified buildings listed as of December 2007 on the LEED or EnergyStar websites, current information about building characteristics and monthly rents were available for 685 of them. In order to provide a control population, each of these 685 buildings was matched to a cluster of nearby commercial buildings in the CoStar database. Each small cluster contains one green-certified building, and all non-rated buildings within a quarter-mile radius of the certified building. On average, each of the 685 clusters contains roughly 12 buildings, for a total of 7,894 data points.

The columns of the data set are coded as follows:

- CS.PropertyID: the building's unique identifier in the CoStar database.

- cluster: an identifier for the building cluster, with each cluster containing one green-certified building and at least one other non-green-certified building within a quarter-mile radius of the cluster center.

- size: the total square footage of available rental space in the building.

- empl.gr: the year-on-year growth rate in employment in the building's geographic region.

- Rent: the rent charged to tenants in the building, in dollars per square foot per calendar year.

- leasing.rate: a measure of occupancy; the fraction of the building's available space currently under lease.

- stories: the height of the building in stories.

- age: the age of the building in years.

- renovated: whether the building has undergone substantial renovations during its lifetime.

- class.a, class.b: indicators for two classes of building quality (the third is Class C). These are relative classifications within a specific market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.

- green.rating: an indicator for whether the building is either LEED- or EnergyStar-certified.

- LEED, Energystar: indicators for the two specific kinds of green certifications.

- net: an indicator as to whether the rent is quoted on a "net contract" basis. Tenants with net-rental contracts pay their own utility costs, which are otherwise included in the quoted rental price.

- amenities: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.

- cd.total.07: number of cooling degree days in the building's region in 2007. A degree day is a measure of demand for energy; higher values mean greater demand. Cooling degree days are measured relative to a baseline outdoor temperature, below which a building needs no cooling.

- hd.total07: number of heating degree days in the building's region in 2007. Heating degree days are also measured relative to a baseline outdoor temperature, above which a building needs no heating.

- total.dd.07: the total number of degree days (either heating or cooling) in the building's region in 2007.

- Precipitation: annual precipitation in inches in the building's geographic region.
- Gas.Costs: a measure of how much natural gas costs in the building's geographic region.

- Electricity.Costs: a measure of how much electricity costs in the building's geographic region.

- cluster.rent: a measure of average rent per square-foot per calendar year in the building's local market.

**The assignment**

An Austin real-estate developer is interested in the possible economic impact of "going green" in her latest project: a new 15-story mixed-use building on East Cesar Chavez, just across I-35 from downtown. Will investing in a green building be worth it, from an economic perspective? The baseline construction costs are $100 million, with a 5% expected premium for green certification.

The developer has had someone on her staff, who's been described to her as a "total Excel guru from his undergrad statistics course," run some numbers on this data set and make a preliminary recommendation. Here's how this person described his process.

> 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.

The developer listened to this recommendation, understood the analysis, and still felt unconvinced. 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?)

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 an exercise in visual and numerical story-telling. Tell your story primarily in plots, and don't feel like you have to run a dozen different regression models. 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 several years of daily data on these ETFs, using the functions in the `fFimport` package. Go back far enough historically so that you get both good runs and bad runs of stock-market performance. Now 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.

## 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.