STA 380, Part 2: Exercises

Due: by end of the working day (5:00 PM US Central time) on Monday, August 16.

Prepare your report on the problems below using RMarkdown so that they are fully reproducible, carefully integrating visual and numerical evidence with prose. You may work solo, or in groups of 4 or fewer people. You can self-organize groups via Canvas.

Note: the option to submit as a group is intended to give you an incentive to get to know some of your classmates. The idea is for y'all to work together on *all* the problems and to learn from each other, not to divide up the individual problems.

Submit via Canvas under the "Assignments" tab. You can submit in one of two ways:
1. A link to a GitHub repo where the final report has been knitted and stored *in Markdown (.md) or PDF format*. (Knitting to .md format is actually best because it displays nicely in a browser. But PDF is acceptable.) Make sure your repo is publicly accessible.
2. A PDF file uploaded via Canvas.

Either way, your knitted submission file *must also include a link at the top of the document* to your GitHub repo where the raw .Rmd file can be found. If we cannot find the .Rmd file, you will not receive full credit.

Notes: - Do not knit to .html, which won't render properly on GitHub.
- Do not include raw R code in your knitted document. *That's what the .Rmd file is for.*
- Do not send six different sets of links, one for each problem. We want a single document.
- Do not directly e-mail the instructor directly with your reports. We will ignore any e-mailed submissions.
- If you need to include mathematical expressions in your report, you can use LaTeX, which I encourage you to learn anyway. Alternatively, you can just handwrite the math, snap a photo, and include the image in the final report. This is a simple, low-overhead option.
- We 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. That's OK — you can try figuring out how to set a seed for your simulation so that it is fully reproducible, or you can just accept that it will be a little bit different next time the script is compiled.
- 10 points will be deducted for each day (or partial day) that your submission is late. One minute late = one day late!

Grading criteria:

- Did you make an honest, concerted attempt at each problem?
- Did you attempt to address all parts of the question?
- Did you include enough detail on what you actually did so that a well-informed reader could understand your analysis in detail? (You won't receive full credit if it's not clear what steps you actually took in your analysis.)
- Did you include properly annotated figures/tables where appropriate?
- Did you write up your solution professionally, with an actual narrative flow (good), or did you just copy and paste a bunch of R code without much in the way of explanation (bad)?
- Did you use sensible procedures to answer a given question?
- Did you make any significant technical mistakes?

Visual story telling part 1: 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 goal

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. Do you see the possibility of confounding variables for the relationship between rent and green status? If so, provide evidence for confounding, and see if you can also make a picture that visually shows how we might "adjust" for such a confounder. *Tell your story mainly in pictures, with appropriate introductory and supporting text.*

Note: this is intended mainly as an exercise in visual and numerical story-telling. While you can run a regression model if you want, that's not the end goal here. Telling a story is. Keep it concise.

Visual story telling part 2: flights at ABIA

Consider the data in ABIA.csv, which contains information on every commercial flight in 2008 that either departed from or landed at Austin-Bergstrom Interational Airport. The variable codebook is as follows:

- Year all 2008
- Month 1-12
- DayofMonth 1-31
- DayOfWeek 1 (Monday) - 7 (Sunday)
- DepTime actual departure time (local, hhmm)
- CRSDepTime scheduled departure time (local, hhmm)
- ArrTime actual arrival time (local, hhmm)
- CRSArrTime scheduled arrival time (local, hhmm)
- UniqueCarrier unique carrier code

- FlightNum flight number
- TailNum plane tail number
- ActualElapsedTime in minutes
- CRSElapsedTime in minutes
- AirTime in minutes
- ArrDelay arrival delay, in minutes
- DepDelay departure delay, in minutes
- Origin origin IATA airport code
- Dest destination IATA airport code
- Distance in miles
- TaxiIn taxi in time, in minutes
- TaxiOut taxi out time in minutes
- Cancelled was the flight cancelled?
- CancellationCode reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
- Diverted 1 = yes, 0 = no
- CarrierDelay in minutes
- WeatherDelay in minutes
- NASDelay in minutes
- SecurityDelay in minutes

- LateAircraftDelay in minutes

Your task is to create a figure, or set of related figures, that tell an interesting story about flights into and out of Austin. Provide a clear annotation/caption for each figure, but the figure should be more or less stand-alone, in that you shouldn't need many, many paragraphs to convey its meaning. Rather, the figure together with a concise caption should speak for itself as far as possible.

You have broad freedom to look at any variables you'd like here – try to find that sweet spot where you're showing genuinely interesting relationships among more than just two variables, but where the resulting figure or set of figures doesn't become overwhelming/confusing. (Faceting/panel plots might be especially useful here.) If you want to try your hand at mapping, you can find coordinates for the airport codes here: https://github.com/datasets/airport-codes. Combine this with a mapping package like ggmap or usmap, and you should have lots of possibilities!

Portfolio modeling

Background

In this problem, you will construct three different portfolios of exchange-traded funds, or ETFs, and use bootstrap resampling to analyze the short-term tail risk of your portfolios. If you're unfamiliar with exchange-traded funds, you can read a bit about them here.

The goal

Suppose you have $100,000 in capital. Your task is to:
- Construct three different possibilities for an ETF-based portfolio, each involving an allocation of your $100,000 in capital to somewhere between 3 and 10 different ETFs. You can find a big database of ETFs here.
- Download the last five years of daily data on your chosen ETFs, using the functions in the `quantmod` package, as we used in class. Note: make sure to choose ETFs for which at least five years of data are available. There are tons of ETFs and some are quite new!
- Use bootstrap resampling to estimate the 4-week (20 trading day) value at risk of each of your three

portfolios at the 5% level.
- Write a report summarizing your portfolios and your VaR findings.

You should assume that your portfolios are rebalanced each day at zero transaction cost. For example, if you're allocating your wealth evenly among 5 ETFs, you always redistribute your wealth at the end of each day so that the equal five-way split is retained, regardless of that day's appreciation/depreciation.

Notes: - Make sure the portfolios are different from each other! (Maybe one seems safe, another aggressive, another very diverse, etc...) You're not being graded on what specific portfolios you choose... just provide some context for your choices.
- If you're unfamiliar with value at risk (VaR), you can refer to any basic explanation of the idea, e.g. here, here, or here.

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 concise 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, and be clear about what you did.

Author attribution

Revisit the Reuters C50 corpus that we explored in class. Your task is to build the best model you can, using any combination of tools you see fit, for predicting the author of an article on the basis of that article's textual content. Describe clearly what models you are using, how you constructed features, and so forth. Yes, this is a supervised learning task, but it potentially draws on a lot of what you know about unsupervised learning, since constructing features for a document might involve dimensionality reduction.

In the C50train directory, you have 50 articles from each of 50 different authors (one author per directory). Use this training data (and this data alone) to build the model. Then apply your model to predict the authorship of the articles in the C50test directory, which is about the same size as the training set. Describe your data pre-processing and analysis pipeline in detail.

Note: you will need to figure out a way to deal with words in the test set that you never saw in the training set. This is a nontrivial aspect of the modeling exercise. You might, for example, consider adding a pseudo-word to the training set vocabulary, corresponding to "word not seen before," and add a pseudo-count to it so it doesn't look like these out-of-vocabulary words have zero probability on the testing set. Or you might simply ignore those new words, at a possible cost in performance.

This question will be graded according to two criteria:
1. the clarity of your description. We will be asking ourselves: could your analysis be reproduced by a competent data scientist based on what you've said? (That's good.) Or would that person have to wade into the code in order to understand what, precisely, you've done? (That's bad.)
2. the test-set performance of your best model, versus the best model that your instructors can build using tools we have learned in class.

Association rule mining

Use the data on grocery purchases in groceries.txt and find some interesting association rules for these shopping baskets. Pick your own thresholds for lift and confidence; just be clear what these thresholds are and how you picked them. Do your discovered item sets make sense? Present your discoveries in an interesting and concise way.

Notes: - Like with the first problem: this is an exercise in visual and numerical story-telling. Do be clear in your description of what you've done, but keep the focus on the data, the figures, and the insights your analysis has drawn from the data.
- The data file is a list of baskets: one row per basket, with multiple items per row separated by commas. You'll have to cobble together a few utilities for processing this into the format expected by the "arules" package. This is not intrinsically all that hard, but it is the kind of data-cleaning and pre-processing wrinkle you'll encounter frequently on real problems, where your software package expects data in one format and the data comes in a different format. Figuring out how to bridge that gap is part of the assignment, and so we won't be giving tips on this front.