MSDS 6306: Introduction to Data Science (401/402/403)

Case Study 02 – Final Exam

Due: Week of December 3rd, 2017 (One Hour Before Your Live Session)

Submit the link to the GitHub repository via the space provided for in the Case Study 02 page in 2DS. No late submissions will be accepted!!

Description

You have finally finished data collection; it involved various measures of procrastination and the qualities these folks have. You plan on pitching a proposal to a company of your choosing. The details of the company, the objectives of the repo, and some of the questions are up to you; however, you should make sure that you answer at minimum the questions posed in the **Tasks** section. Though you get some leeway as data scientists, there are some baseline questions that the company wants to know about the data they funded. This is the resulting data set from the study, tabulated by Qualtrics. It is not entirely well-cleaned and will need some manipulation to make it useful.

Procrastination.csv

- Age: The participant's age in years.
- Gender: The gender the participant identifies as (Male or Female)
- Kids: Binary, whether they have kids or not.
- Edu: Education level
- Work Status: What kind of job are they working?
- Annual Income: All converted to dollars.
- Current Occupation: A write-in for occupation.
- How long have you held this position?: Years: Number of years in this job.
- How long have you held this position?: Months: Number of months in this job.
- Community: Size of community
- Country of Residence: The country where the person holds citizenship.
- Marital Status: Single, Married, Divorced, Separated, etc.
- Number of sons/Number of daughters: integer number of children.
- All variables starting DP – the Decisional Procrastination Scale (Mann, 1982)
- All variables starting AIP – Adult Inventory of Procrastination (McCown & Johnson, 1989)
- All variables starting GP – the General Procrastination scale (Lay, 1986)
- All variables starting SWLS – the Satisfaction with Life Scale (Diener et al., 1985)
- Do you consider yourself a procrastinator?: a binary response
- Do others consider you a procrastinator?: a binary response

Note: The above is not a full codebook. As Master's students, you may have to do a little domain research as to what the DP, AIP, GP, and SWLS are with your SMU library access to journal articles. The original CSV gives the individual questions, so make your codebook detailed.

Deliverable

A GitHub repository and peer-grading email.

a. Introduction to the project. The introduction should not reference a project, persay. No part of this should be informal. It should be appealing-looking as well.

b. The introduction needs to be written as if you are presenting the work to someone who has given you the data to analyze and wants to understand the result. Pretend it's a presentation for a client. This may take some imagination of whom your client might be. **If it sounds like a student presentation, that is not acceptable**.

c. As you do your 'sales pitch', briefly explain what your code is doing (both procedurally and for the customer). The explanations should appear as a sentence or two before or after each code chunk. Even though you will not be hiding the code chunks (so that I can see the code), you need to pretend that the client cannot see them.

d. R code for answers concerning the tasks below. Make it in RMarkdown file format and always include echo=TRUE and include=TRUE for charts. **Keep the .md file so I can readily see everything on GitHub! Your .md file should mirror the .RMD file.**

e. Give clear, explicit answers to the questions. Just the code to answer the questions is not enough, even if the code is correct and gives the correct answer. You must state the answer in a complete sentence outside the code chunk.

f. Once you're finished, be sure to frame a conclusion to the project. The presentation does not stop when you're done with the questions. Find a way to wrap it up: summarize your findings from this exercise. Again: **the file must be readable in GitHub. In other words, dont forget to keep the md file!!**

g. You should expand your repository with at least this RMarkdown file, the input file(s), and a README.md that describes the purpose of the project and codebook. The repo can be structured how you like, but it should make sense and be easily navigated. If things are not clear from the root directory, you will lose points.

h. You're working on this with one other person. I expect that you both will do equal work. Both of you will need to push, add, commit, and pull the GitHub. If I do not see equal effort (and remember, GitHub tracks commits!), I reserve the right to penalize either or both students. Part of your grade will also be contingent on your partner's evaluations of your effort. This is a collaborative project, so take it seriously and plan ahead with your teammate.

Tasks

Tip: I recommend in the code block commenting which number the code answers. It will help me find your answers more readily. Something as simple as # 3B helps me.

1. Formulate your Repo (20%)

- **a** The client wants this to be reproducible and know exactly what you did. There needs to be an informative Readme, complete with several sections, as referenced in Live Session. Give contact information, sessionInfo, and the objective of the repo at least.

- **b** Procrastination.csv is a large data set, and it will need its own Codebook, formatted in an approachable way. Make sure you describe peculiarities of the data by variable and what needs transforming. However, do not make it too long either.

- **c** Create a file structure that is accessible and transparent. Document it in the root directory, ideally in the Readme.

2. Clean your Raw Data (10%)

- **a** Read the csv into R and take a look at the data set. Output how many rows and columns the data.frame is.

- **b** The column names are either too much or not enough. Change the column names so that they do not have spaces, underscores, slashes, and the like. All column names should be under 12 characters. Make sure you're updating your codebook with information on the tidied data set as well.

- **c** Some columns are, due to Qualtrics, malfunctioning. Prime examples are the following columns: "How long have you held this position?: Years", Country of residence, Number of sons, and Current Occupation.

    - **i** Some have impossible data values. Detail what you are doing to fix these columns in the raw data and why. It's a judgment call for each, but explain why. For example, most people have not been doing anything for over 100 years. For the "Years" columns, round to the nearest integer.

    - **ii** Somehow, "Number of sons" was labeled with Male (1) and Female (2). Change these incorrect labels back to integers.

    - **iii** There are no "0" country of residences. Treat this as missing.

    - **iv** Current Occupation has no "please specify" or "0." Treat them as missing. Some jobs are quite similar. Use judgment calls to make overwrite them into the same category. It does not have to be 100% accurate, but right now "ESL Teacher" would not be counted as "teacher" if there were unique counts.

d Make sure your columns are the proper data types (i.e., numeric, character, etc.). If they are incorrect, convert them.

e Each variable that starts with either DP, AIP, GP, or SWLS is an individual item on a scale. For example, DP 1 through DP 5 are five different questions on the Decision Procrastination Scale. I've reverse-scored them for you already, but you should create a new column for each of them with their mean. To clarify, you'll need a DPMean column, an AIPMean column, a GPMean column, and a SWLSMean column. This represents the individual's average decisional procrastination (DP), procrastination behavior (AIP), generalized procrastination (GP), and life satisfaction (SWLS).

3. Scrape the Human Development Index tables online (20%) (https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index#Complete_list_of_countries). **Read what it is**, because you'll have to explain it to your clients. *BIG TIP:* This one's going to be weird looking—it's real scraping. Make sure you're looking between the webpage and R while you do this to find your footing. Unit 9's HW might help.

a You will notice there are several sections, but you should only worry about the Complete List of Countries section of this Wikipedia entry. There are 8 tables in this section, but you should pull these eight, clean them as to be usable, and then find a way to bind them into one singular table. You only need *Country* and *2016 Estimates for 2015* (you can call this HDI) columns for the final table.

b Create a new column for this final scraped table which categories the Countries like the original page (*Very high human development, High human development, Medium human development, Low human development*). After these categories, output a csv file of this table to your repository.

c Merge this data frame to the *Country of Residence* column of Procrastination.csv so that your data now has an HDI column and HDI categories (Very high human development, etc.).

4. Preliminary Analysis (10%)

a Remove all observations where the participant is under age 18. No further analysis of underage individuals is permitted by your client. Remove any other age outliers as you see fit, but be sure to tell what you're doing and why.

b Please provide (in pretty-fied table format or similar), descriptive statistics on Age, Income, HDI, and the four mean columns (DP, etc.). Create a simple histogram for *two* of these seven variables. Comment on the shape of the distribution in your markdown.

c Give the frequencies (in table format or similar) for Gender, Work Status, and Occupation. They can be separate tables, if that's your choice.

d Give the counts (again, pretty table) of how many participants per country in descending order.

e There are two variables in the set: whether the person considers themselves a procrastinator (yes/no) and whether others consider them a procrastinator (yes/no). How many people matched their perceptions to others' (so, yes/yes and no/no)? To clarify: how many people said they felt they were procrastinators and *also* said others thought they were procrastinators? Likewise, how many said they were not procrastinators and others *also* did not think they were procrastinators?

5. Deeper Analysis and Visualization (20%)

a *Note:* You should make all of these appealing looking. Remember to include things like a clean, informative title, axis labels that are in plain English, and readable axis values that do not overlap.

b Create a barchart in ggplot or similar which displays the top 15 nations in average procrastination scores, using one measure of the following: DP, AIP, or GP. The bars should be in descending order, with the number 1 most procrastinating nation at the top and 15th most procrastinating at the bottom. Omit all other nations. Color the bars by HDI category (see 3B). Use any color palette of your choice other than the default.

c Create another barchart identical in features to 5B, but use another one of the three variables: DP, AIP, or GP. How many nations show up both in 5B's plot and 5C's? Which, if any?

d Is there a relationship between Age and Income? Create a scatterplot and make an assessment of whether there is a relationship. Color each point based on the Gender of the participant. You're welcome to use lm() or similar functions to back up your claims.

e What about Life Satisfaction and HDI? Create another scatterplot. Is there a discernible relationship there? What about if you used the HDI *category* instead and made a barplot?

6. Outputting to CSV format – Make sure there are no index numbers (10%)

a The client would like the finalized HDI table (3A and 3B)

b The client would like the Tidied version of the original input to be output in the repository, including the merged HDI data (3C).

c The client would like a dataset (or two) that shows the Top 15 nations (in 5B and 5C), as well as their HDI scores.

d All output should be in plain English or translated in the Codebook.

7. Peer Grading (10%)

a Please wait to do this after you have submitted the project, not during. It is still due by the same deadline, but unless the situation is absolutely egregious, give your teammate a chance to improve. If it is that bad, reach out.

b Send a review of your performance **and** your teammate's performance to my email (tptibbett@gmail.com). Honest (but not rude) appraisal is fair game. I will not show them their feedback, but keep in mind that unkindness reflects on my assessment of **you**, not them. Give them a score from 0 to 10. Most people should fall between a 6 to 10 (D to A) range. 5 or lower (F) is reserved for particularly difficult teammates. Grade your own effort on this metric, too.

c Failure to turn in a review will result in a zero for this section and will not affect your partner's grade.

d This metric should not be a measure of your partner's coding ability, but the amount of effort they put in. If they are well-meaning but make mistakes, that is not cause to dock them points. It should be reserved for no-shows, careless work, or laziness. Keep in mind that this is a two-week project, so pace yourself. If a person cannot immediately finish the Case Study by Day 2, you might need to be patient with them.

e Make yourself available and accountable. By signing up for this course, you committed and will *need to make time* to work on the project. This is the hardest assignment yet. We are all busy adults; you must manage your time well and meet halfway on your partner's schedule.

f **Finally:** The grades, in the end, are up to me. Make your case with a convincing argument in a review. If I disagree with your assessment, I may grant the person more (or fewer!) points. Your peer grade is not necessarily reflective of your peer's choice.

Remember, no late assignments will be accepted, as this is a final exam. It's been a pleasure teaching you this semester! Congratulations on making it through your first semester in the Masters of Data Science program. Good luck!