CODE ▾

AMOD 5250 - Final Project

Course Instructor: Jamie Mitchell

Due : Dec 15, 2021

For each day (to a max of 5) early you submit your report, you will receive a bonus corresponding to 1% of your mark.

Project outline

The Process

The final project in this course is an exploratory data analysis project to be completed and then presented as a final report. I would recommend doing your initial exploratory analysis within a basic R Script file (just executing blocks of code as you try things). Once you've found all the interesting things you need, you can use the code you wrote to put together an R Notebook to generate the actual report.

Leave all the code visible in your notebook, but your report should read correctly if the *hide all code* option is chosen in the viewable HTML.

For appearance sake, your report should suppress all message, errors and warnings.

Please keep in mind that your final result should be a report, not just a Markdown file that outputs code and graphs.

Also keep in mind that you should be generating a reproducible report. Any data file with the same structure should produce a correct report. In other words, be sure to use *inline code* rather then hard-coding values in discussions.

The Data

For your final project, and it's associated paper, you will be working with the data from the 2019 American Community Survey (ACS) 5-year Public Use Microdata Samples (PUMS),which is a sample of the actual responses collected by the American Community Survey between 2014-2018, and split into **population** and **household** characteristics.

You can choose to work with either the population or household data. You are not required to use both (unless you want to, in which case you would join them based on the `SERIALNO` variable). (In previous years, as a general rule, students who've chosen to use the population data (or both) have tended to produce better reports because there isn't as much variation in the housing data as you'd expect)

Each subject zip contains 4 files (a,b,c,d), that represent a complete set of records. You will need to combine them to complete your dataset (each can be loaded into a separate dataframe, then combined). Because they aren't separated in any logical manner, you must use all 4 files. If you find the dataset is to big for your computer, please find a meaningful way to pare the data down, rather then just using one.

The links to the data files can be found on Blackboard.

All the relevant information about the data files can be found here: https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.2019.html (https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.2019.html)

However, I've also include the ones your most likely to need on Blackboard. To start, take a look at the data summary & definition files, so that you can decide which dataset (or both) you want to use before downloading.

Within the data file of your choice you are free to examine whatever relationships interest you.

You are not expected to use all of the provided variables in your analysis. It suffices to choose a total of ~10-20 variables and to perform a thorough analysis using just those variables.

Once you filter down your dataset, be sure to remove the full dataframes from your workspace to free up memory! You can export your paired down data to it's own file you can use while working (but be sure to leave the code in your markdown to load and filter the original files, since that's what I would need to do to re-run your report)

Things of note about the ACS

- To protect privacy, several of the variables are top and/or bottom coded (the list is on Blackboard). If you choose to use any of these variables it is probably worth considering in your discussion how the top/bottom coding may have affected your results.

- Because the data is collected over 5 year, to be completely accurate, dollar values would be adjusted for inflation. The variables `ADJHSG` (housing) and `ADJINC` (income) provide the relevant adjustment factor, so you should do the appropriate mutation.

- Because the PUMS are microdata, each record represents a sample of the population. Both files contain a weighting value to account for the fact that individuals are not sampled with equal probably (people who have a greater chance of being sampled have a lower weight to reflect this). These are essentially a frequency count for each row.

  - WGTP: PUMS household weights
  - PWGTP: PUMS person weights
- Properly using the weights in the ACS files will affect almost all analysis you do.
  - To get counts within groups, you would `sum(PWGTP)` rather then `count()`.
  - When graphing **ggplot2** had a weight `aes()` parameter which allows you to apply the weight to your graph.
  - When doing statistical calculations you need to use functions that account for weights.
  - `lm()` has a weight parameter.
  - The **hmisc** package contains several weighted statistical estimates (wtd.mean, wtd.var, etc)
  - The **weights** package contains functions for weighted comparisons (wtd.chi.sq, wtd.cor, wtd.t.test, etc).
  - **Note:** When using some functions in **weights** and **hmisc** you need to indicate the type of weight by setting a parameter (*mean1* and *normwt* respectively) to **FALSE** since the weight values in ACS are essentially frequency counts. (this is not the default behavior in either library so you must actively set this parameter when using functions that have it)
- That all being said, although incorporating the weight (and inflation adjustment) is essential for getting correct and meaningful results, it does add complexity. Since the inaccurate results you would receive from ignoring these factors doesn't essentially matter in demonstrating your ability to do quality analysis (any more then if we were to do this with completely made up data), you can choose to ignore the weights, which will only affect your complexity mark, assuming you include the impact of your choice in the relevant sections.

Project requirements

Note: Sections 1, 2 ,3 are based on both the quality of your analysis/write up, but also contribute to the code marks based on and the underlying R-code you used to do it.

Your end-product for the project will be an R Markdown report of approx 2000-3000 words that contains at least the following sections:

1. Data summary (10 marks)

You should begin by describing the data you have available. Other then a brief description of what the dataset is, only cover the data you actually used. You will want to display tabular summaries of means and proportions where appropriate.

Your score for this section will be based on the following criteria:

- Your description of the dataset itself
- The knowledge you convey regarding the overall make-up of the data you chose to use
- Meaningful pre-processing: variable names, factor variables, and factor level names
- Insightful graphical and tabular summaries of the data
- Proper labeling of figure axes and table columns
- Discussion of the graphical and tabular summaries.

Note: Figures and tables that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output. Make sure that your discussion could easily be understood by a first year student trying to learn more about your subject matter.

2. Methodology (5 marks)

In this section you should provide an overview of the approach you took to exploring and analyzing the data. This is where you tell the story of how you got to your main findings. It's too tedious to carefully format plots and tables for every approach you tried, so you can also use this section as a place to explain the various types of analyses that you tried, and where appropriate put unused code in an appendix at the end that can be referred. (it should be readable and documented, but doesn't have to be pretty). Sometimes you put the most work into something that doesn't end up contributing to your final results, and including it this way means it will be considered as part of your mark.

You should address at least the following questions, when relevant:

- How did you deal with missing values? What impact does your approach have on the interpretation or generalizability of the resulting analysis?
- How did you deal with outliers? What impact does your approach have on the interpretation or generalizability of the resulting analysis?
- How did you deal with weights & income adjustment values? What impact does your approach have on the interpretation or generalizability of the resulting analysis?
- Did you produce any tables or plots that you thought would reveal interesting trends but didn't?
- What relationships did you investigate that don't appear in your findings section?
- What's the analysis that you finally settled on? What relationships do you investigate in the final analysis?

Note: Figures and tables that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output. Make sure that your discussion could easily be understood by a first year student trying to learn more about income inequality between men and women.

3. Findings (15 marks)

In this section you give a careful presentation of your main findings. You should provide, where appropriate:

- Tabular summaries (with carefully labeled column headers)
- Graphical summaries (with carefully labeled axes, titles, and legends)
- Regression output + interpretation of output + interpretation of coefficients **and/or**
- Assessments of statistical significance (output of tests, models, and corresponding p-values)

As part of your analysis you are expected to do some statistical testing (regression model, t-test, chqisq test,etc). Note: When running regressions, you should discuss whether the standard diagnostic plots indicate issues with the model (trends in residuals, variance issues, outliers, etc.). You will not receive full credit for your regression unless you clearly display and discuss the diagnostic plots. For other statistical states you must also intelligently describe/evaluate the outcome.

Note: Figures and tables that are presented without accompanying description/discussion will receive at most half credit. To earn full credit, you must describe what each table/figure is showing and discuss any key takeaways. In other words, it is not sufficient to simply display R output. You must also provide thoughtful discussion of the output. Make sure that your discussion could easily be understood by a first year student trying to learn more about your subject of investigation.

4. Discussion (5 marks)

In this section you should summarize your main conclusions. You should also discuss potential limitations of your analysis and findings. Are there potential confounders that you didn't control for? Are the models you fit believable?

You should also address the following question: How much confidence do you have in your analysis? Do you believe your conclusions? Are you confident enough in your analysis and findings to present them to policy makers? (You will not be deducted points for saying that you are unsure of your analysis. This is just something I want you to reflect upon.)

5. Overall Code Quality & Complexity (15 marks - 5 and 10 respectivly)

Although this won't appear as an actual section in your report, it will be based on the code behind the scenes that generates the report (and anything in the appendix). For quality you will be will be marked on things like:

- good, consistent coding style
- appropriate use of variables
- appropriate use of functions
- good commenting
- good choice of variable names
- appropriate use of inline code chunks

You complexity mark will be based on the difficulty and variety of the analysis you do. (i.e. A report full of bar charts will do poorly here). If you try things that don't product meaningful results, they can still be included in your complexity mark as long as you include your work in your appendix.

Package requirements

You can assume that all of the packages covered in class are installed. Your markdown file should include **installation commands** for any other packages, and must **load** all packages necessary (including those used in class).

Collaboration policy

While you may work in small groups to discuss on appropriate topics of investigation, statistical methodology and graphical/tabular summaries, each of you will be required to produce and submit their own code and final report (or one report for you and a partner). You may not copy someone else's code or write-up. You may not enable someone to copy your work by sharing your code or write-up. No two exploratory analysis should cover the exact same set of approaches or consider all the same variables.

All student are expected to comply with the Trent policy on academic integrity. This policy can be found online at www.trentu.ca/vpacademic/academic-integrity/graduate-academic- integrity-policy.

Any submitted project that is deemed by the instructor to be in violation of the collaboration policy will receive a score of 0.

You must list your collaborators at the top of your project submission.

Kaggle Warning

The ACS was used as a Kaggle dataset in 2013, 2014 and 2015. Although looking through the kernels created with this set may give you some ideas, incorporating any part of a Kaggle kernel directly will be considered academic dishonesty. Only repeating things done in a Kaggle Kernal, even if it's all your own code, will result in a poor mark, since you're only illustrating your ability to replicate someone else's analysis. It's also worth noting that some Kaggle kernels used the weight and dollar adjustments, some do not. Evaluate what you see carefully.

Final Submission

You may assume that the person grading your report will have the same files provided to you in their working directory (with the original names). Any other files that you need in order for your Rmd file to knit should be submitted along with your report (although there shouldn't really be any). You will be responsible for submitting:

- The Rmd file that generates your analysis
- The resulting report in HTML form

A note about the file sizes

These files are **big**. To simplify, I would suggest doing your initial data load and elimination of any columns/rows you aren't going to use in an R script (Rather then RMarkdown which has extra overhead). Then export your filtered dataframe to a new csv file, which you can use for the rest of your project. The code you use to load, filter, and export should be included in a *visible* but *non-executing* code block in your appendix.

As long as you provide this code in your appendix, you don't need to provide your new csv file with your submission.

Also because of the size, your going to probably want to consider using **caching** with your code blocks, but please make sure to pay attention to **dependencies**.

To save yourself significant headache, be sure to use a R Notebook, rather then a straight R Markdown file. This will allow you to "compile" your report in stages if you don't have a fast computer (and make changes to straight text without re-running all your analysis!)

Do NOT leave your report to the last minute. Working with large files like this can be time consuming!