DSC 520 Final Project Template

Name: Matt Burns

Date: 2/27/19

Section 1

Explain what your interests are in the data sets identified.

The analysis pertains to state tax dollars per capita. I work with a handful of people that live in Wisconsin and they believe that their taxes are much lower. My hypothesis is that the Illinois taxes are in line with other states except for property taxes.

The analysis will be broken into two parts.

The first part looks at the amount of taxes adjusted for population . I pulled two datasets to accomplish this. The first data set is the number quantity of taxes pulled by state and state population.

The second looks at outcomes from taxes. I pulled infrastructure data, education data and life expectancy,

I will only look at the 50 states. I.e. Puerto Rico and D.C. are beyond the scope of this project.

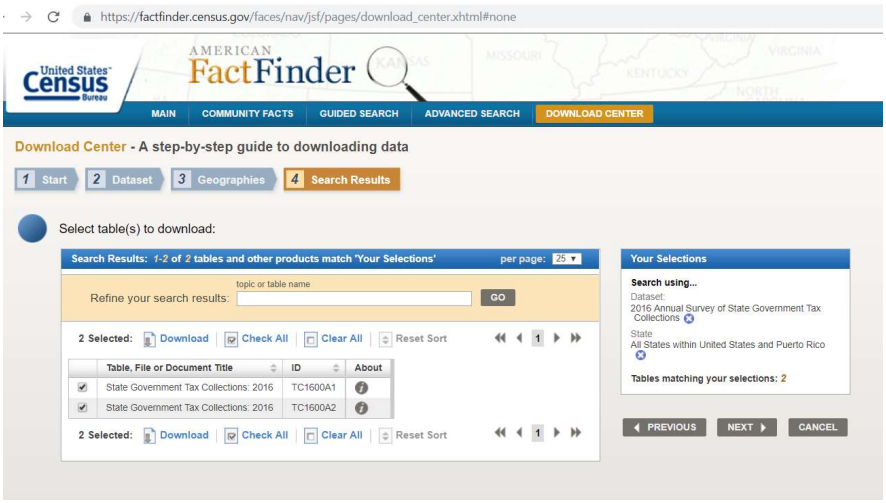
Identify the Packages that are needed for your project.

I will use the following packages for the analysis:

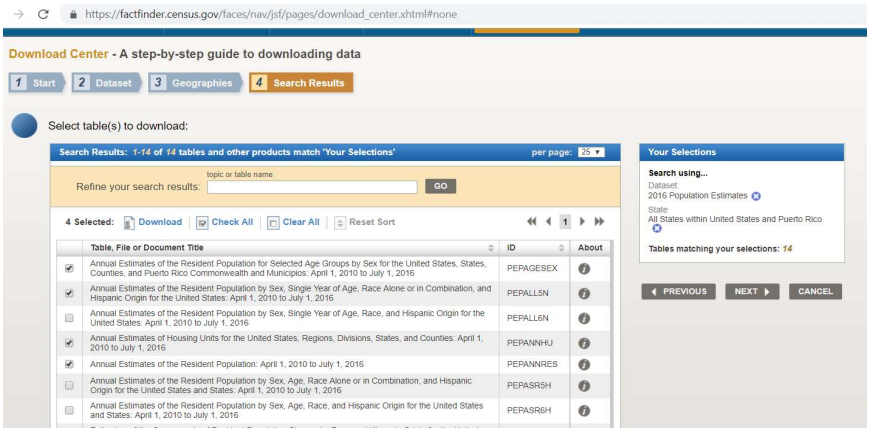
- ggplot2
- readxl

Original source where the data was obtained is cited and, if possible, hyperlinked.

Taxes



Population



Industry employment

SOURCE: 2016 County Business Patterns. For information on confidentiality protection, sampling error, and nonsampling error, see http://www.census.gov/programs-surveys/susb/technical-documentation/methodology.html.

Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).

The tax data was collected in 2016 and reflects the tax dollars collected by each state and tax type. There are 31 types of taxes per state and 51 states (it classified DC as a state). The data is pretty clean, but it does have "X" for some types of taxes for some states. I will classify the "X" as a 0.

The population data was collected in 2010 and then estimated for the subsequent years. It reflects the number of residents that each state split by gender and also includes the median age. It has 51 states (DC and Puerto Rico). I data looks clean and complete, so I will use total people (not male or female) for 2016. All other data will be stripped from the file.

Section 2

- Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?

Every state has unique tax policies. Ideally, they are designed to accommodate nuances in the states' population, economy and goals. Cynically, the taxes could be designed to benefit the politicians and their donors. How different are the sources and amounts of taxes and do they correlate with different outcomes?

- Provide a concise explanation of how you plan to address this problem statement.

I will need to work a two-step process. First, I will have to detail the revenue sources and amounts for each state. The second will be to look at potential outcomes / benefits from the taxes. I.e. assuming all revenues are spent, what is being obtained from payments.

- Discuss how your proposed approach will address (fully or partially) this problem.

My approach can identify situations where tax rates are too high and outcomes are not being improved. It could also identify the opposite situation where not enough investments is being made.

A difficult analysis will be determining the optimal type of taxes. When I analyze the data, I'll see in anything pops, but it may not be straightforward.

- List at least 6 research questions you aim to answer.

Total Tax Rates

- What are the are they states with the highest and lowest total tax rates?
- Is there a correlation between tax rates and educational attainment?
- Is there a correlation between tax rates and ASCE's percentage of roads in bad condition?
    - Is weather an explanatory variable
    - Is population an explanatory variable
- Is there a correlation between tax rates and life expectancy?

Types of Taxes

- Which types of taxes do the high tax states use?
- Which types of taxes do the economically developed state use?

- Explain how your analysis may help the consumer of your research findings.

This research can help with tax policies. Which types of taxes are the most effective and what is the ideal amount?

- What types of plots and tables will help you to illustrate the findings to your research questions?

I will make scatter plots with average revenue per resident on the x axis and various outcomes on the y axis. This would be:

  - Education as a function of taxes
  - Infrastructure as a function of taxes
  - Infant mortality as a function of taxes

- What do you not know how to do right now that you need to learn to answer your research questions?

Section 3

- Data importing and cleaning steps are explained in the text and in the DataCamp exercises (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.

I understand that tidy data usually preferred, but it is easier for me to think through a data frame that has 50 rows (one for each state) and multiple columns.

All of the is organized by state alphabetical order and only 50 state are included.

If N/As are identified, they are removed.

Some taxes are marked with an X. They will be changed to 0.

- With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```
Classes 'tbl_df', 'tbl' and 'data.frame':       50 obs. of  33 variables:
 $ Geography                                                          : chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
 $ Total Taxes                                                        : num  1.00e+07 8.97e+05 1.37e+07 9.43e+06 1.55e+08 ...
 $ Property Taxes                                                     : num  352378 111736 943008 1097908 2513157 ...
 $ Sales and Gross Receipts Taxes                                     : num  5106102 260846 8315805 4590072 51602180 ...
 $ General Sales and Gross Receipts Taxes                             : num  2596223 0 6300443 3314363 39189007 ...
 $ Selective Sales and Gross Receipts Taxes                           : num  2509879 260846 2015362 1275709 12413173 ...
 $ Alcoholic Beverages Sales Tax                                      : num  210535 42430 72281 55164 368699 ...
 $ Amusements Sales Tax                                               : num  6 10306 0 56031 0 ...
 $ Insurance Premiums Sales Tax                                       : num  317657 64400 545124 192020 2561932 ...
 $ Motor Fuels Sales Tax                                              : num  526763 48773 898234 479879 5000539 ...
 $ Pari-mutuels Sales Tax                                             : num  1401 0 160 2616 14537 ...
 $ Public Utilities Sales Tax                                         : num  695626 4027 22337 0 714623 ...
 $ Tobacco Products Sales Tax                                         : num  180301 67918 317331 230527 840034 ...
 $ Other Selective Sales and Gross Receipts Taxes                     : num  577590 22992 159895 259472 2912809 ...
 $ License Taxes                                                      : num  507479 120529 482362 396891 10275132 ...
 $ Alcoholic Beverages License                                        : num  4224 1919 7416 4624 57406 ...
 $ Amusements License                                                 : num  0 0 0 473 16767 ...
 $ Corporations in General License                                    : num  162117 0 18342 26703 75066 ...
 $ Hunting and Fishing License                                        : num  22931 29500 35059 26579 104698 ...
 $ Motor Vehicle License                                              : num  213550 38000 228970 163023 3996089 ...
 $ Motor Vehicle Operators License                                    : num  33964 0 31373 21825 296160 ...
 $ Public Utilities License                                           : num  14443 838 0 8351 674660 ...
 $ Occupation and Business License, NEC                               : num  56250 46957 159454 143422 5027281 ...
 $ Other License Taxes                                                : num  0 3315 1748 1891 27005 ...
 $ Income Taxes                                                       : num  3966640 67457 3906722 3231617 90655530 ...
 $ Individual Income Taxes                                            : num  3492904 0 3336174 2781458 80753345 ...
 $ Corporations Net Income Taxes                                      : num  473736 67457 570548 450159 9902185 ...
 $ Other Taxes                                                        : num  90352 336801 32524 114345 145715 ...
 $ Death and Gift Taxes                                               : num  0 0 0 3 330 ...
 $ Documentarty and Stock Transfer Taxes                              : num  43730 0 17328 38844 0 ...
 $ Severance Taxes                                                    : num  46622 336801 15196 48340 68500 ...
 $ Taxes, NEC                                                         : num  0 0 0 27158 76885 ...
 $ Population Estimate (as of July 1) - 2016 - Both Sexes; 18 years and over: num  3766477 554567 5299579 2283195 30157154 ...
```

- What do you not know how to do right now that you need to learn to import and cleanup your dataset?

There are blanks for property taxes. I believe Colorado has property tax. Why is this blank? Plus, I need to more work on my results data.

Section 4

Discuss how you plan to uncover new information in the data that is not self-evident.

How each state uses the various types of taxes for funding is not obvious. One would think as one type of taxes rise, others would fall. I will run a regression of tax revenue per capita for each type of tax against each other. Assuming a zero-sum game, there should be many taxes that crowd each other out and have a negative correlation. I need to think through if I want to include total taxes in the regression.

What are different ways you could look at this data to answer the questions you want to answer?

I plan on using tables ranking the variables, plots illustrating the relationship, and linear regression quantifying the relationship. Examples will include:

- Ranking total state tax revenue
- Graphing outcomes vs. revenues
- Quantifying the relationship between tax revenue and road quality

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information, learn and use an R package not covered so far in the course?

The most obvious way I will need to adjust the data is to create per-capata variables. This will adjust for the variance in state size.

For now, I only concern I have regarding joining data frames is linear regressions. Since I'm only dealing with only 50 rows and I've ensured they're in exactly the same order in each dataframe, this won't be a big deal.

The aggregate function might be helpful in understating all of the taxes types if I need to split the data into subsets and compute summary stats.

How could you summarize your data to answer key questions?

I will be as succinct as possible. I'll show the appropriate table, graph or regression and then explain what the conclusions are.

What types of plots and tables will help you to illustrate the findings to your questions?

Tables will be helpful to rank order the states by many metrics. On the revenue side, I can rank the states by various tax levels. On the outcome side, I can rank by outcomes including health, education and infrastructure.

What do you not know how to do right now that you need to learn to answer your questions?

There isn't anything obvious that I need to learn except maybe machine learning, but I'm not sure 50 states constitute "big data".

Do you plan on incorporating any machine learning techniques?

No, I do not. However, I am eager and nervous to learn more about those techniques.

Section 5 (write a coherent narrative that tells a story with the data as you complete this section.)

Summarize the problem statement you addressed.

I want to determine two things. First, which states are paying the highest taxes. I accomplished this by running a very simple regression that controlled for population size. Out of curiosity, I also checked only state income taxes. Second, I wanted to determine if the taxes improved healthcare, infrastructure and education.

Summarize how you addressed this problem statement (the data used and the methodology employed).

On the revenue side, I first utilized plots of tax revenue as a function of population size. Second, I took the residuals and adjusted them for population, ranked them and plotted them. This process was repeated for state income taxes as well.

On the social outcomes side, I made three very simple regressions to quantify the relationship between state taxes and education, health and infrastructure.

4. Summarize the interesting insights that your analysis provided.

- The most shocking thing was that Illinois was right on the linear model and the residuals were very low.
- California and New York had the largest variance above the linear model. ← True for total and income taxes
- Texas and Florida were well below the model. ← True for total and income taxes
- Taxes have a small correlation with roads (infrastructure proxy) and bachelor degrees (education proxy).
- Taxes have a strong correlation with infant mortality (heath)

5. Summarize the implications to the consumer (target audience) of your analysis.

I'd like to say to my colleagues that moved to Wisconsin that they are not avoiding taxes by leaving the state. They may be paying lower taxes, but that is most likely that they moved from the suburbs with expensive schools, expensive park districts and expensive libraries to exurbs with a revenue generating outlet mall, a Jelly Belly distribution center and a renaissance faire.

6. Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

I'd love to spend more time on the regressions. It would be interesting to try to dial those in with more data and more attempts with data that could control for age, racial mix, population density et al.