**School Data Project**

Name: Lara Clasen

This set of data comes from the State of Minnesota database and tells us employment information of graduate-level students following their programs.

What is the target audience for this research?

* Residents of Minnesota looking to pursue graduate programs

Identify the Packages that are needed for your project.

* I will need readr to read in my CSV files, ggplot2 to visually represent the data, Knitr and pastecs to create a nice RMarkdown representation of the analysis, shiny to further improve the interactivity of the analysis.

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

* https://mn.gov/deed/data/data-tools/graduate-employment-outcomes/

Source data is thoroughly explained.

* Methodology: https://mn.gov/deed/data/data-tools/graduate-employment-outcomes/method-geo.jsp

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

* The problem statement that I would like to address is whether in the 3rd year post-graduate program, students are working in the industry related to their degree. I am also interested in the 24-month post-graduate median hourly wage for the same graduates. The 3rd year industry data is relevant to understand whether the graduate degree is likely to take an individual into that industry long-term, and compare that likelihood to other areas of study. The 24-month data is relevant to short-term satisfaction, for those who might be more interested in how different area graduate degrees can boost their pay more immediately. There should be a lot of interest into these questions as graduate school costs increase and, when our time is so valuable, we need to know if this investment will pay off.

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

* I will need to clean the data in order to remove missing values as there appear at first glance to be many. I would like to remove variables that are not relevant to my questions. I will need to look at averages across industry and, more than likely, do some combining of data to account for fields that are very similar to one another. The data that I am looking at initially is only for the year 2016 and so I could remove the Year variable entirely, although it is possible that I will expand my analysis to include other years as well. I may need to exclude results with only a few graduates, or combine them with a like field, so that those low numbers don’t unintentionally skew my overall results.

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

* My approach will be to clean and narrow down my data set as much as possible without introducing bias or accidentally excluding relational information for the variables that I *do* want to analyze. The limitations of this analysis will come into play when I am limited by the definitions of fields in the study and the way that the study was conducted. I only have so much control over the data as it was reported. For example, because programs with less than 10 graduates were suppressed in the data, I cannot conclude anything helpful for those areas of study.

List at least 6 research questions you aim to answer.

1. Did students obtaining a degree in the Computer Sciences in 2016 earn a higher hourly wage than students obtaining a degree in Education?
2. Were most students who had earned a graduate degree in the Computer Sciences still working in the field 3 years after graduation?
3. How much more per hour did Computer Science graduates earn when compared to Statistics graduates 24 months after graduation?
4. Was there a higher number of Computer Science graduate program students as compared to Liberal Arts program graduates?
5. What is the average hourly rate 24 months after graduation for the sciences versus the arts?
6. When comparing Mathematics graduates and Statistics graduates, what is the difference in hourly wages 24 months after graduation?

Explain how your analysis may help the consumer of your research findings (recall you target audience from Section 1).

* My analysis and research questions will aid Minnesota residents looking to pursue a graduate degree by shedding light on the pros and cons of doing so. By looking at hourly wages 24 months after graduation, an individual in this audience would better know if this degree would be valuable for them short-term. If we take that information and combine it with the knowledge of whether they can expect to remain in the field after 3 years, this is a pretty good view into whether this is something they will not regret undertaking.

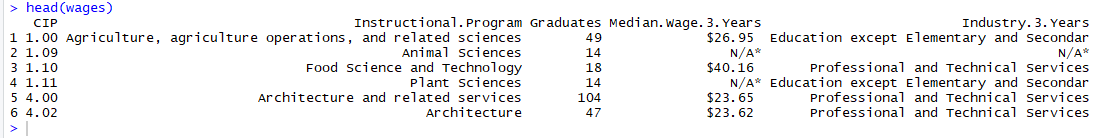
What types of plots and tables will help you to illustrate the ﬁndings to your research questions?

* I will use scatterplots to compare many of the variables that I’ve described in my research questions. I will also likely use some histograms when attempting to combine the various fields into similar ones. As our book mentions, scatterplots show the most basic step in judging relationships between variables and they’re a great place to start. From there, I will likely create a table with a chunk of my data to show trends, and a more complex ggplot to show several variables on one chart.

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

* I need to learn more about combining like variables. I have used the merge() function this week but unfortunately it created too large of a process for my computer to handle. I look forward to attempting this function with my project data set as I believe it will be a more manageable task.

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.



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

I still need to learn some more about manipulating my data frames into easily readable data. I also need to become more comfortable working with missing data and deciding at what point to ignore all observations that contains some missing data.

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

* I will run tests and analysis between a few sets of variables to determine if perhaps there are relationships that would not be readily assumed just from looking at the data set or at what the different variables are.

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

* You can look at this data visually, using plots, charts and graphs, or look at the statistical description of the data using the statdesc function, or by running tests between variables based on the questions asked.

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? Explain.

* Yes, I plan to slice the data to narrow it down to my relevant variables because this data set contains many variables that either won’t be relevant to my questions or that have a large number of missing values and would not provide valuable information anyway. I will likely also create new variables as I hone in on a range of certain variables or seek to ignore outliers.

How could you summarize your data to answer key questions?

* I could use the summary function which will provide us with a jumping off point with the data set and give some direction as to where to go with running tests and looking for relationships.

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

* I anticipate using histograms to show initial and basic information, but also the ggplot2 package to display more than two variables. I will use alpha to better visualize clumps of data as well as filling with color based on a variable to show multiple levels of the information.

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

* I feel fairly confident in my analysis ability at this point, but I am still hoping to learn some more about cleaning the data set, in particular how to remove missing data without skewing the set too much.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

* I’m certain there will be machine-learning incorporated into my project. I plan to utilize the techniques that we have learned thus far in the course, and hope to look back in the textbooks to refresh my memory on all of the mechanisms we have at our fingertips.

Overall, write a coherent narrative that tells a story with the data as you complete this section.

* My data set was obtained from the State of Minnesota’s Graduate Employment Outcomes tool. This tool offers insight into the data behind institutional education programs and their associated job outcomes. This includes growth trends, wages, full-time versus part-time employment information, and prospects. The variables that I used in my analysis involved wages and associated number of years post-graduation.

Summarize the problem statement you addressed.

* The problem statement that I was interested in learning more about was whether a graduate’s first job played an important role in that of their career jobs down the road a few years.

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

* To analyze this question, I looked at the wages earned after two years out of school and in the work place in comparison to those same graduates’ wages after 4 years in the workplace.

Summarize the interesting insights that your analysis provided.

* My analysis results generally surprised me, with evidence that your first position wages may not play much of a role in your future wages. At least they do not play a significant role in the increase or differentiation of those future wages.

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

* The implications could be large here. As someone who is currently enrolled in a graduate program (noting that not all participants in the data set are in the same such program) it is of particular interest to me, and likely would be to my classmates as well. Specifically these results are encouraging because coming out of an educational program you may not possess the experience that many higher-level employees seek; We see here that by taking a financially lesser position you are not implicitly hindering your chances of gaining a more valuable role just a couple of years later.

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

* Limitations of analysis for this data set arise via a combination of two things: my novice R abilities, and the difficulties of pulling data from this particular public tool. If the data set were arranged in an ideal manner I feel that my analytical success could have been improved; On the other hand, if my skills in R were more advanced I likely could have better defined my data set. If I could return to the beginning of the project I would probably have chosen a more straightforward data set for my first R project.