Sophie Perez

Final Project 115

https://data.world/data-society/student-alcohol-consumption

For this project, my motivating question was to find out if there were any correlations between alcohol, and a student's performance. This was brought forth after attending WSU and observing my peer's day to day life. Though there was limited data on the subject, I was able to find studies from around the world. As there are different age limits to drinking varying by country, I then needed to clean my data to a certain age range. Therefore I created a subset that only showed me students from the age range of 18 to 22 - to get it a close to our situation as possible.

The source of my information came from World.Data and was based around 2 colleges in Portugal. It wasn't until recently did I realize how limited the data was on this subject -especially when it came to wanting to make an analysis for ourselves and not just reading on the subject through a .org or .gov.

I started out cleaning the data to match the age range of the average college students that go to Washington State University - therefore, the circumstances could be as close to ours as possible. I then had to make the data more concise, considering there were way too many students and variables to make any displays within RStudio. This included me taking out variables. Though, I found once I had set the age range to >= 18 and <= 22, that took away almost mre than half my set and it became a lot more easier to work with. As for variables, I removed things like family size, who their main guardian was, how they had to travel to school, what classes they were in; and really all other relevant data that didn't give me information on their school performances, patterns, and consumption.

As for techniques learned throughout this course, I was able to filter, create visualizations, and clean the data. These were all things very new to me, as this was my first course that I have ever had to code.

I ran into a lot of problems while working with this set. The main one being there were hardly any numeric variables. These categorical variables were then able to be changed to either 0s or 1s in order to be read in RStudio. I then was able to sweep the data by figuring out out to filter all of the variables, like mentioned earlier, this included me taking out unnecessary information. The main thing I wanted to focus on was their drinking patterns and how it could impact their studies. This included focusing on things like; absences, drinking during the week or weekend, and how they performed in classes. This process was a lot of trial and error, as i found that the majority of factors could still be used as correlations and I needed to still find a way that they could still be incorporated.

What was surprising to me, it I wasn't able to find many correlations between outside factors of school and their drinking trends. In the future, I'm interested in finding more studies that pertain to the "why?" portion of a questions, and not just the "what?". For example, how would the students rate their happiness? What were their reasons for choosing the school? I know for WSU, many students come here because their parents and their grandparents also used to walk these halls. This university isn't just a learning space, it's a tradition for many.

Overall, I learnt from my data that each variable and person within your set has their own motivating factors, and people can quickly become outliers - these are the people that i would want to focus on in the future. Through this process I have learned how to quickly find correlations between people and their practices. Though this class was a struggle for me, it has been one that I can also look back on and deem it as one of the most valuable.