Austin Hudgins
CSE 578 Data Visualization
Progress Report

In the field of academia, enrollment is one of the highest priorities. This creates a great opportunity to showcase data analysis through marketing profiles. With these marketing profiles, companies like XYZ Corporation can use data visualization to find key factors that can accurately predict vital information, in this case, income allowing UVW College's marketing team to send targeted advertising and bolster their main priority, enrollment. Without this visualization it would be difficult if not impossible to see connections and correlations in the huge swath of data like the census collects, causing UVW College's marketing team not to be nearly as effective.

So far, I believe I am about 50% of the way done with the marketing profiles. I was able to clean and transform the data, get 3 out of the 5 user stories created, 2 multivariate and 2 univariant graphs, and, used 6 different attributes out of the possible 14 given in the census dataset. The user stories seem to be the easiest portion so far. Since the prompt is to make marketing profiles to help UVW College's enrollment, I have been considering what specifics each department in UVW College would need to know to help them with the goal of enrollment. An example of this would be the Financial Aid department, which would normally offer scholarships or grants to lower-income students or students who excel in class, since we have no academic records, I focused on the former. What type of factors, like native Country or Gender, could cause a student to have lower income, this would allow the Financial Aid department the ability to narrow their search down drastically and offer students that fit the criteria scholarships or grants to enroll in UVW College because this could be a key deciding factor on whether they enroll in UVW College or go somewhere else. As for the graphs I have been trying to use different ones so it's not receptive, but they also show accurate data. For the multivariate graphs, I have made a line graph that compares average capital gain per age for incomes over 50k and at or under 50k, as well as a mosaic plot that compares the native country, gender, and income group like the example I mentioned above. Regarding the univariant graphs, I have not fully settled on which ones I am going to use but I created a bar graph, pie chart, and donut graphs for the most common occupations in the income groups. Focusing on different types of graphs has allowed me to use most of the attributes I need and don't see any future problems getting to a total of 8 different attributes.

The problems I have encountered so far started when I tried to manipulate the data. Throughout this course, we have focused on querying data from SQL databases and then using that data to visually graph different information. However we were given a large CSV for the project, not a SQL database, I started by attempting to make a function to convert it to a database but that was taking too long so I decided to just read the CSV into a data frame using pandas, it has more than enough tools to sort, filter, or query the data for what the project is asking me to do. Then after that when I was creating my line graph mentioned previously, I noticed huge spikes in capital gain at specific ages, this was caused by a lot of data for capital gains being '99999'. It might be that a lot of people gained exactly that much money, but I think it's more likely to be the max for the field or an error in the data. To deal with this I made sure to remove outliers from the data that was 2 or more standard deviations away from the mean. The last major problem I have encountered so far was the amount of nominal or categorical data. This isn't necessarily a problem with the data set, I just prefer working with quantitative data. This causes me to double check I am using the right

type of data for the given graph, so I don't accidentally show any relationships between data that isn't truly there.

To finish the marketing profiles, I need to create at least 1 more multivariate graph, settle on which univariant graphs I want to use, write 2 more user stories, use at least 2 more attributes that haven't been used yet, and put all the information into a formatted paper. For the next multivariate graph, I am leaning towards seeing the relationship between education, work class, and income. This would also complete the use of 8 different attributes. The univariant graph seems to be a lot more important than I originally thought because you are only focusing on the one variable outside of income, you must be able to make it show insightful information. I think one of the strongest connections to income would be education level, while I have not looked at that data specifically, I think a donut or pie chart of this information would do well to show this connection if it exists.

Going forward I only see 2 major challenges, the first being the formatted paper and the second is deciding what graph to make my third multivariate graph. The formatted paper seems to be the most daunting task of them all, for this, I am most likely going to use ACM format. The first issue is more of my lack of experience converting data to a research paper. To combat this, I plan on searching and reading some other researcher's papers so I can see what they do well or what I would do differently from them. ASU also has a great writing center, so going to talk with them would also be a great solution. When it comes to making my third multivariate graph, I think I will have to revisit module 3, multivariate analysis, so I can make sure the chosen graph is appropriate for the data.

This is all to say I am on track and think I will be able to deliver very good marketing profiles to UVW College's marketing team. I have solved the issues I have encountered so far and have been able to complete about 50% so far. While I do have more to go, I am confident I can that completed and have a plan for future problems that could come up.