# Data Visualization Paper

Aryan Kumar
kumar.aryan@gmail.com

## I. Goals and a Business Objective

I have been tasked by a company to develop effective marketing profiles on people that the company can then sell to their customers. For these profiles I was tasked to work with UVW college and analyze the best way for them to bolster enrollment. The college has picked out salary as a main demographic to base their marketing on. To build these profiles the data that was provided came straight of the United State Census Bureau and with a focus on $50,000 a middle point for salary. I believe this number was chosen has been around the median income. So with that as a sort of expected value for salary they suggest using many other variables like age, gender, education status, marital status, occupation and just about any other variable that could be correlative to determine which variables make the most impact on people making either more or less than that amount.

With all of that contextual information provided there can seem like there are many different paths to provide an analysis but fortunately the college gave a very direct development goal and that goal was to make an application that can figure out all the factors necessary to properly predict a person's income so that they can have a list of factors to use in their own creation of a model. They also want the product to output income based on a set of inputed parameters so that way the marketing can be alternated to specifically suit each unique individual.

## II. Assumptions

For my analysis I assumed that all the salary values that were provided in the census data were standardized to USD even if the countries that each people were from wasn't the same.

I also assumed that the data provided was accurate and didn't have any elements with errors or any data points that could be misleading or vague.

Additionally, I assumed that the data was relevant to the current time period or as close as possible. An example of data that wouldn't be relevant would be data gathered during either a big economic swell or a big economic downturn as both of these data points would be outliers in the grand scheme of economic trends in the country.

Finally I assumed that some parts of the data would be incomplete as the census is data gathered from surveys sent out to people and people won't always fill out every data point. Because of this and the fact that I wouldn't expect the census department to handle unknowns values and taint the dataset by adding data that wasn't explicitly sent in by citizens, I built on that assumption to assume that this would be a task that I would have to handle through the creation of a model to predict unknown values, some way of finding the most common value and swapping that out for the unknown values or simply dropping those values when doing my analysis and graph creation.

## III. User Stories

To focus my analysis, I came up with five user stories. Two of them were univariate analysis meaning I just compared the salary with one other variables while the other three were multivariate analysis where I compared multiple variables to salary.

The first user story was that as a member of the UVW marketing team I want to analyze the distribution of hours worked and how they differ for individuals making less or more than 50k to help UVW college demonstrate if higher education can lead to people working less hours each week or if higher income is directly collected to higher working hours.

The second user story was that as a member of the marketing team at UNW college I want to look at the relationship between occupation and salary count to best help UVW college see which occupations would be the most financially efficient to target through advertisements.

The third user story was that as a member of the UNW college marketing team I want to to a multivariate analysis of the relationship between work class, age and income. Through this UCW college will be provided with a age range to target and a specific work class to target along with the the occupation from the previous user story.

The fourth user story was that the marketing director wants to observe the relationship that both higher education and capital gain had on the amount of income that people made. With this UCW college can advertise the potential benefits of higher education on a higher base salary along with increased capital gains or angle their advertising differently if the data shows a different conclusion.

The fifth and final user story was that as a member of the marketing team for UNW college I want to analyze how family structure and education not just education number impact a person's income. For family structure I will use the relationship feature and this will help UCW college understand the best types of

family structure to target to demonstrate the benefits of getting higher education as they will have a point of comparison for the impact of different education level among each type of family structure.

## IV. VISUALIZATIONS

The first variable I decided to compare against salary was hours per week. It seems obvious that the variables would be positively correlated meaning that as the number of hours that someone works goes up then the amount of money that they make should go up as well. My analysis pretty easily confirmed that hypothesis and to demonstrate that I made a box plot. Observations from the plot included:

- The average person making over 50k worked more hours than the average person making less than 50k

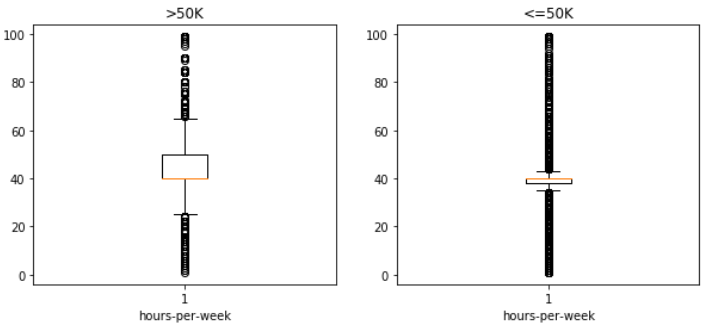
- For people making less than 50k there seems to be more people at the highest amount of hours worked.

- The mean for above 50k was about 46 hours and the mean for below 50k was about 39 hours.

- To make more money people above 50k worked on average 7 extra hours each week.

- The standard deviations also showed that there's slightly more variation for those making less than 50k compare to those making more.

- Those that made less than 50k had a standard deviation of about 12 and those that made more than 50k had a standard deviation of about 11. So it isn't too significant but it is something that the next user story can build on for a deeper analysis of how working conditions compare for these two income groups.




For the next user story I generated two pie charts for people making more or less than 50k a year that broke down the exact percentages of people that worked each occupation. Observations from the pie charts included:

- For those making more than 50k they were more concentrated in two occupations

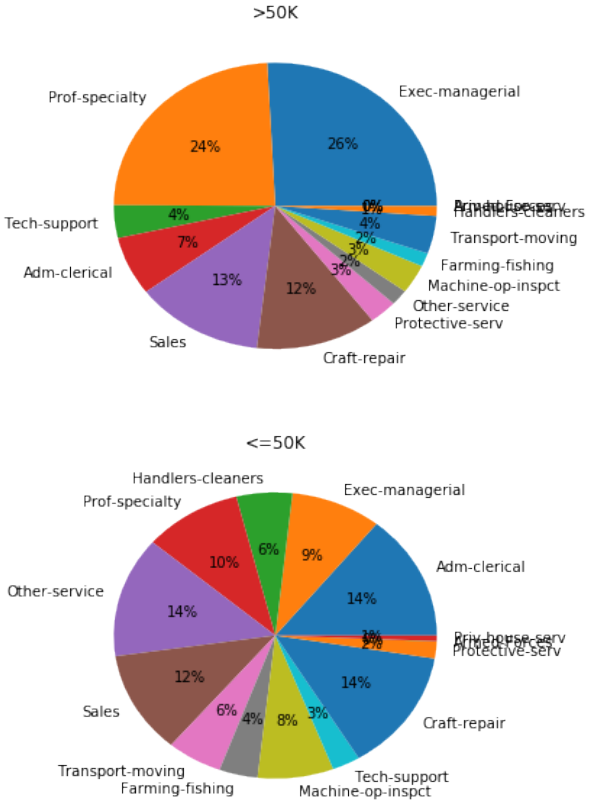
- For people making less than 50k their occupations were more evenly distributed across many different occupations.

- This builds on the idea from the last plot that those making less than 50k have more variation whereas those making more than 50k seem to have more established and stable working conditions.

- For those making more than 50k the two most common jobs were executive management related positions at 26% and professional specialization at 24%.

- For those making less than 50k the most common job was a three way tie between craft repair, other service and administration clerical all at 14% then the next most common was sales at 12%.

- Just through a quick look at the most common jobs for the two income groups already shows the various different career paths that exist for those making less than 50k.



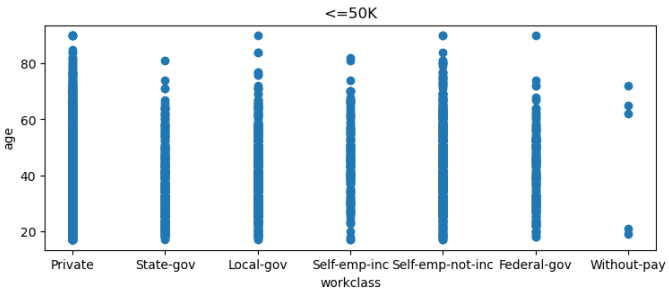

To visualize the third user story I made a scatterplot to show the age vs work class relationship. Observations from this scatterplot included:

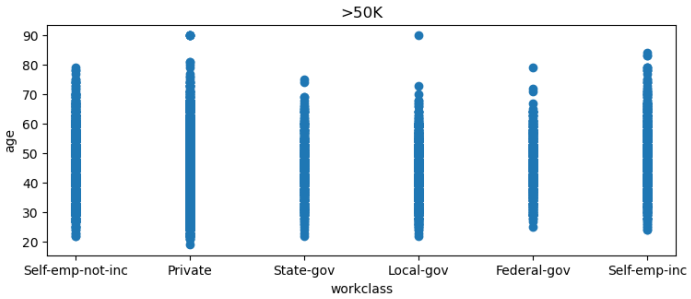
- Those that made over 50k tended to work at older ages compared to those who made less than 50k.

- As shown in the two scatter plots, people making over 50k were the only one with points at the age 90 but for the less than 50k graph the age axis didn't get to 90.

- Those making more than 50k had no one in the without pay section but those that made less than 50k had a couple points in that work class.

- The work classes that made for most of the income gains in those making more than 50k versus those making less were private and local government sector of work class.




For the fourth user story the visualization that I thought would best represent what I was trying to prove was another scatter plot but this time placing both the above and below 50k on the same graph but changing the colors to really highlight the difference in capital gain and the difference in the amount of people that existed for higher education number.

From that scatter plot graph the observations that can be made are:

- With capital gain for those that made less than 50k they all had capital gain below 10k

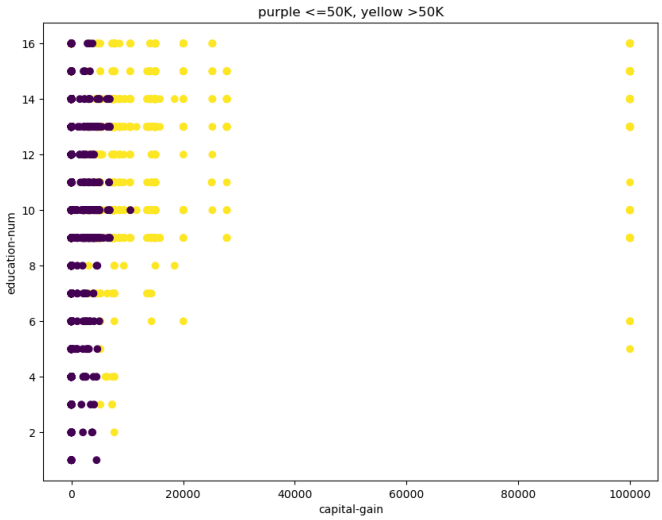
- For those that made more than 50k across all education ranges they had higher capital gain values

- Many have capital gain values as high as 100k but

- Even without those outliers, people making more than 50k consistently had capital gain values ranging around 10-20k instead of being below 10k.

- People that made more than 50k also consistently had higher amounts of education with a lot more having a education number going from 9-16.

- People who made less than 50k were more concentrated around 5-10




- Some people making less than 50k even had education numbers lower than 2

- For those making more than 50k no one had education numbers that low.

For the fifth user story I went with a mosaic plot to analyze how relationship status and education impact income. The analysis for education here is different from education num as it lets UCW college directly advertise the exact types of degrees that they offer instead of vaguely pointing to higher education. Additionally since relationship status included things like husband or wife, this analysis also ended up adding gender as an additional variable which was unintended but lead to this being a more complex analysis than the user story initially suggested.
Observations from this mosaic plot include:

- A bachelor's degree and being a husband is the combination that leads to people making over 50k the most

- For those that made more than 50k across all education ranges they had higher capital gain values

- Only being a high school graduate is the most common education level for people making less than 50k.

- Not being in a family is the most common type of relationship for those making less than 50k

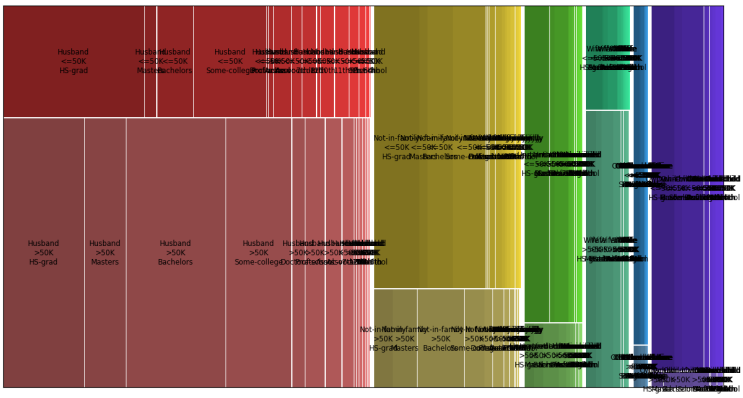
- Comparatively very few people who were not in a family ended up making more than 50k.

- Wife also seemed to be on the lower side for both those making less and more than 50k which suggests that there's a gender element to income in addition to the other two factors for this user story.

- The education level most common for wife was high school graduate.

- For husbands making more than 50k the most common education level was split between bachelors and some college.

- For husbands making less than 50k the most common education level was between high school graduate and some college.



Through these five user stories I ended up analyzing eight attributes directly along with a ninth one indirectly out of the fourteen that existed in the dataset. As a recap those eight attributes are hours per week, occupation, work class, age, capital gain, education num, relationship and education itself. There was also the indirect analysis of gender as a result of the relationship feature in the fifth user story.

## V. QUESTIONS

The first question I had was a general one about the which of the various functions that were at my disposal would be best to use as I wasn't as experienced with many of these libraries. This also led to questions about if I was fully understanding of if there were any unknown functions that could make my application better. This initial development question got alleviated through reading the documentation, revising previous assignments for this class to refresh my memory of how I had used some of the functions in the past and by merging these two approaches I believe that everything I have done so far is one of the best approaches possible.

I then had questions about how exactly I could take a raw text file that was provided in the assignment and convert it to csv. I could've just manually inputted the data but that was tedious and as someone tasked to build an application it seems like I can use the development skills needed for that to also automatically convert a txt file that I pass in and then output a csv file with the category headers that match up with the listed variables in the census data using python libraries as a initial test of how equipped I can be to even complete building the application. After that each task related to actually analyzing the data that I completed had its own set of issues to grapple with. But this was all also one big learning experience for how I could better parse the data for a future project and that was when I started the graph creation process and data analysis process with code and learned that actually raw txt files are perfectly file to use as well so this entire process could be scrapped to make the process more efficient for the customer as they can just give a txt file directly and excess code doesn't have to be used.

For the first task of filling in the unknown data with predicted values, I had a question of what the best option would be between the variety of existing methods. This question stemmed from the extra research I did into the documentation of the python libraries that I planned to use as there were so many models available across so many python libraries to handle unknown values. Ultimately, I tried a series of models and used mean squared error to calculate the loss produced by each of them and they had higher loss than what I got using linear regression. So that based on that linear regression looked like the clear best model for the job. However, this method only worked for columns that were numerical so even though I had coded this part in the progress report, I was unable to use this for most of the unknown data and had to resort to just dropping those values and choosing user stories that didn't rely on those columns as much.

A final question I had different from the initial ones I had was the best way to represent categorical data numerically. This was one I was unable to answer in the development of this application but will be something I discuss more in the future plans section.

## VI. FUTURE PLANS

For future improvements to the model things like linear regression for each variable separately could be done to narrow down the best variables to use for create univariate and multivariate graphs from. That was part of my original plan but I ran out of time to do the appropriate deep data analysis to identify those variables so I had to settle for picking out what logically seemed like the best variables instead of the factually best ones.

The mosaic plot graph made for the fifth user story could also be spaced out better to make it more readable. This would require taking more time to research the various spacing options that exist for each type of graph and properly investing better UI options for every style of graph that could be made so that the customer can request whichever style they want and get a good looking and readable graphs.

As mentioned in the question section there was a struggle in properly making linear regression predictions for columns that were categorical so I ended up having to drop those values and chose columns that didn't rely as much on those unknown values. For future data analysis it would be good to properly integrate that linear regression model to work with categorical data instead of just the numerical ones and that way that extra column data could be unlocked for proper analysis instead of being ignored.

Another future plan for improving on this application relates to the categorical data conversion to numerical data question. With the given dataset there was a fixed range of the values that each category would be and that range of values was given to us beforehand so the numerical conversions could be hard coded. However, for scalability of the application it would be good to have a plan to create some dynamic way to convert those values. This way any dataset with different categorical data can be inserted and progress can be made to a proper software system that can work with any type of data to produce similar relationship graphs to the ones I made in the visualization section automatically.