Problem statement:

- Create a marketing profile using US census bureau data, with a focus on the $50k salary benchmark.
- Evaluate how factors such as age, gender, education level, marital status, job, etc. for individuals can be used to determine if they earn above and below $50k.
- Group fields that are most relevant in being able to predict salary.

Describe the progress you have made so far, including the background work you completed.

I have created a list of both technical and business assumptions that will be included in the final report when it is handed to the "team" who will oversee building the application. For instance, I assume that the data is correct, and that the audience is capable of interpreting visualizations such as pixel based displays.

I have created 6 user stories that I am prioritizing for the project deliverable as well. For example one of the stories is: *Customer wants to understand the correlation between salary and hours-per-week.*

In these user stories I am using 9 different attributes: *education_num, marital status*, *work class*, *hours-per-week*, *occupation*, *sex*, *race*, *age* and *salary.*

I have created visualizations for all of the stories. Additionally, I have drawn conclusions on these stories that are supported by the visualizations. For example *Customer wants to understand the correlation between salary and education.* I first looked at each value count of education values to get a sense of the distribution. I then separated the total count for each education value into 2 separate columns so I could see greater than and less than $50k. From there I knew that a bar chart would be good where each education value has 2 lines (for greater than and less than). What I ultimately used to represent this was a count plot from the seaborn library. What became apparent, and what I was able to conclude was that education is a reliable predictor of salary, as the percentage greater than $50k was notably higher after education_num = 13.

Summarize the specific tasks you have completed to date.

- Defined the purpose querying (to find factors influencing wage).
- Created user stories to drive the querying and visualization creation.
- Identified the data types in the dataset by reading adult.names. This included a full list of column descriptions and datatypes: for example, capital-gain was labeled as continuous data, and education values including bachelors and preschool were revealed to be categorical.
- Loaded the data for analysis with python. Excluded values that were unknown which the authors of the data had marked with “?”.
- Used data exploration techniques to identify patterns. Created box plots and histograms for the numerical data to visualize distribution of over and under $50k for these fields. Did the same with pie charts for the categorical data (country, sex, race, etc.)
- Drawn conclusions on these visualizations. For example, I have seen that all the attributes that I have chosen in my user stories can in some way be shown to have some correlation with salary.

Discuss issues you have encountered thus far and your plan for solving them.

ISSUE: all the fields in my user stories appear to be an indicator of salary, so I need to enhance my visualizations to be able to tell a more descriptive story. RESOLUTION: incorporate mosaic plots, pixel based displays, parallel coordinate plots and word clouds.

ISSUE: At first I was going to use field ‘education’ instead of ‘education num’, but realized that the education num would be better because it can be more intuitively sorted. This allowed me to be able to better summarize the results, in being able to illustrate that education above 13 appears to be an indicator of higher salary. RESOLUTION: use field ‘education-num’ instead of ‘education’

ISSUE: I wanted to use a histplot to visualize the relationship of hours per week on salary. After I grouped by hours per week and got the counts of salary greater than and less than for each value of hour per week, I normalized the value counts to get the counts in percentage form. I then plotted and realized the hist plot was grouping the hours per week into groups of 10, so each bar went as high as the sum of the percents. I was expecting to get each hour per week as its own bar on the plot right away. RESOLUTION: To get each hour per week as its own bar I converted the data type of the hours per week field to a string. This revealed a right skewed data set.

ISSUE: Some data points seem strange – for example about 70% of people working 76 hours per week are above $50k, but 0% of people working 77 hours per week are above $50k. Looking into the records that have hours per week = 77, I confirm there are 6 records where this is the case. RESOLUTION: update my assumptions to include the validity of the data.

ISSUE: There were some minor errors in finding my way around the data that have already been resolved. For example, renamed the field “class” that was described in the adult.names file as that is a reserved word in python. RESOLUTION: Rename “class” field to “salary” (could have worked around by also changing df.class to df[“class”])

Summarize the tasks you have yet to complete and how you intend to approach them. Be specific.

TASK: Complete remaining items part of the final report. STEPS TO COMPLETE: As I progress in enhancing my existing visualizations and in creating new visualizations, keep a list of questions that arose during the project progression, the solutions implemented,

TASK: Enhance visual elements. STEPS TO COMPLETE: pick the right visualization type, label axes and add titles, remove unnecessary elements, highlight important patterns with markers, make sure there is appropriate scaling, and add text annotations of the findings.

TASK: Pick the visualization that best represents each attribute STEPS TO COMPLETE: For each attribute in each user story, take inventory of the data type (categorical, numerical ordinal, etc), the distribution of the data in that field (i.e. are there outliers?), and in the case of multivariate analysis the relationship between the different variables (i.e. multiple numerical values?).