CSE 578 - Final Project Report

Youssef Serag, *Student, Arizona State University*

I. Goals and objectives

THIS project aims to aid the UVW college in determining what characteristics help determine if individuals earn an annual salary of one of two, more than $50,000 or less than or equal to $50,000. The end goal of this analysis is to aid the college in developing marketing plans to approach individuals with persuading enrollment plans towards their degrees. In this project, we, an analyst working at XYZ company, have been hired to come up with marketing profiles for the ideal candidate to help with the goal that UVW college has in mind. The main goals of this project are as follows. The user is tasked to come up with five separate user stories to come up with a marketing profile. Furthermore, these five user stories must be compromised of 8 different variables, three of which must be multi-variate analysis and two of which are univariate. Once the user, in this case the analyst, is able to come up with five unique user stories, they are tasked to come up with at least five story-telling visualizations to support their claim. They must use best practices mentioned in the course to allow their visualizations to be perceived in a better manner by their audience. Finally, once the initial analysis is done and the main requirements are fulfilled, the user is tasked to come up with tasks that can be done to ensure better results that are to be given to the team during the handoff of the project.

II. Assumptions

A. Technical Assumptions

In reference to technical assumptions related to the project, one can assume that the data provided by the United States Census Bureau is of high quality, meaning that it is a comprehensive and high-quality dataset representing the entirety of individuals around the world. Due to the nature of the project and the mentioning that there are missing data points, it is safe to assume that the dataset will require data preprocessing to ensure any screen real-estate is not consumed by incomplete samples. Finally, it is safe to assume that the features within the dataset have some sort of correlation to the income feature. Even though creating a machine learning model is not required for this project, it can be leveraged in the tasks that could be done in the future to see whether the features used for the analysis are among the most positively correlated to the income feature or not.

B. Business Assumptions

When it comes to the business assumptions for this project, it is assumed that the targeted features from the dataset have some sort of correlation to the income feature, which would mean that once the marketing profiles features have been completed, we expect to have a higher enrollment rate. Similar to the assumptions made in the previous section, the user is assuming that the dataset was chosen with the features being high candidates of indicating income of individuals. The statement made in the previous sentence is different than the initial one, as it highlights the importance and power of the features, while the initial statement focuses on the impact of the results that will be generated by the results.

III. User stories

A. User Story Overviews

For the five user stories, we chose the following charts: a grouped bar chart, a heat-map plot, a parallel-coordinates plot, a box-whisker plot, and a histogram. The reason behind the selection of these charts is due to the variety of data types within the dataset. Furthermore, it is important to mention that these charts were not the initial selection, but after further analysis and trials, these final five selections were the ones that returned the best results. When it comes to the variables used for the different visualizations, we used Age, Education level, Hours-per-week, Marital status, Sex, Relationship status, Capital loss, and Capital Gains. The Age, Hours-per-week, Capital loss, and Capital gain variables are all the numerical type, while the Education level, Marital status, Sex, and Relationship status variables are of the categorical data type. The reason for mentioning this is that these different variable types require us to represent the attributes and metrics using a variety of different charts, which at times, requires to transform the data using a Label Encoder or a similar process. The details of such processes will be talked about in the sections below.

B. User Story 1: As an analyst for the XYZ Company, I want to see if the zero and non-zero occurrences of capital gains & losses have any impact on one's income.

This visualization is one of the three multi-variate visualizations. On the Y-axis, we are using the occurrences for capital gains and losses, and on the X-axis are the two categories of income, less than $50,000 and more than or equal to $50,000. The reason why we were interested in these two variables were due to them being the only variables that have an opposite, as all other variables/features are unrelated to each other.

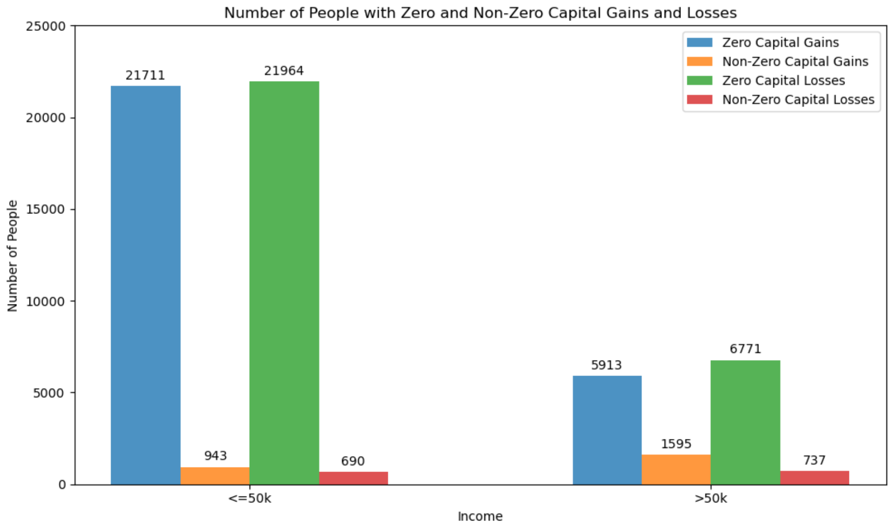

Figure 1: Grouped Bar Chart (Capital Gains & Losses VS. Income)

The figure above is the visualization generated from the data. As seen, the categories of “non-zero Capital Gains” and “Non-Zero Capital Losses” are almost identical, and the difference between the categories of “Zero Capital Gains” and “Zero Capital Losses” are five times more and three times more respectively between the “less than or equal to $50,000” and “more than $50,000” categories. The reason why this chart is valid is due to us not using the actual values of either gains or losses, but we used the occurrences of either zero or non-zero values occurring in each feature. To further improve this visualization, best practices mentioned in the lectures were added. For example, a grouped bar chart was used over a stacked bar chart due to the ambiguity it would lead between the different categories of losses and gains. Furthermore, as mentioned in the lectures, the issue with a basic bar chart is that it is difficult to know the exact value of each bar when placed next to each other. To tackle this issue, I’ve placed the values for each bar on top of their respective bar. To create the visualization, shown above, the following steps were taken:

1. We first split the entire dataset into two categories, based on each entry’s income class.
2. Create a custom dictionary to store the results of our chart. We create 4 keys in our custom dictionary, which are:
   a. Zero Capital Gains
   b. Zero Capital Losses
   c. Non-Zero Capital Gains
   d. Non-Zero Capital Losses

   Each category key in the dictionary has an array of two elements, representing the two different income classes.
3. We then iterate over the respective Dataframes to sum up the occurrences.
4. When it comes to plotting, we used the custom data structure’s data to feed it into the matplotlib functions to create the bar plot.
5. Moving onto the non-graphical components, we then added the y-axis, x-axis, and title labels to the figure, as well as the legends, which included the color code of each category of the bars plotted.
6. Finally, as mentioned in the paragraph above, to enhance the understanding of the data presented in the grouped bar chart, we included the value of each bar on top of each bar. Furthermore, similar to “R’s Pretty” algorithm, there is a matplotlib function named “tight_layout”, which shrinks the plot labels to ensure that is within the values of the data.

Looking at the data, we can conclude that the majority in either income bracket reported that they do have zero capital gains. Even though we cannot come to a direct conclusion why that is, a possible reason might point to the majority of the population not contributing to investments. Additionally, it is shown that the trend between the four bars is consistent in either income bracket. The main difference in numbers is due to the number of samples being in the “less than or equal $50,000” & “more than $50,000” bracket, where each one has 22,654 and 7,508 samples respectively. When it comes to inferring which income bracket a sample is more likely to be in, there seems to be a higher likelihood of belonging to the “more than $50,000” bracket if the sample has non-zero capital gains.

C. User Story 2: The UVW College is curious if the Relationship and Marriage Status of an individual affect which income bracket, they belong to.

This visualization is the second multi-variate visualization, where the Y-axis represents the categorical values of the relationship feature, and the X-axis being the categorical values of the martial status feature. For this specific user story, we have two separate visualizations, one for each income bracket. Furthermore, the reason we were interested in these variables was to observe that there is/are a certain value(s) in either category where most of the samples present themselves in either income bracket.

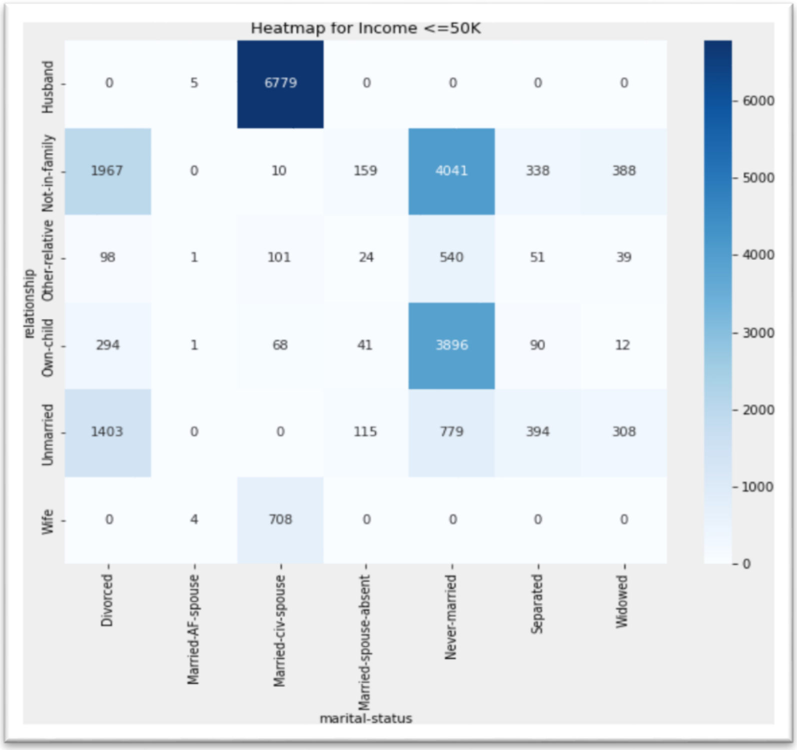

Figure 2: Heat Map (Relationship & Marital Status VS Income) for “less than or equal to $50,000” income bracket

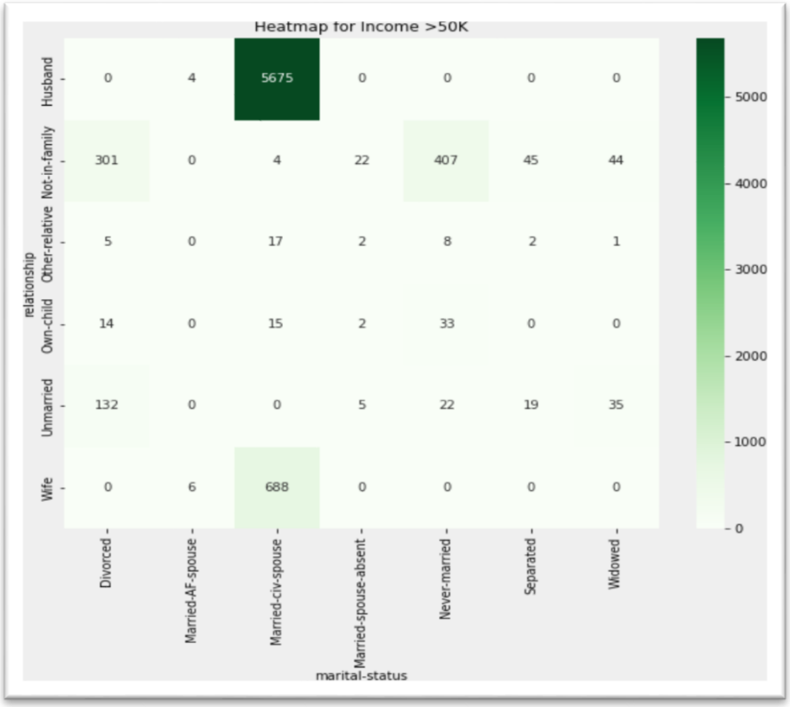

Figure 3: Heat Map (Relationship & Marital Status VS Income) for “more than $50,000” income bracket

This visualization is the second multi-variate visualization, where the Y-axis represents the categorical values of the relationship feature, and the X-axis being the categorical values of the martial status feature. For this specific user story, we have two separate visualizations, one for each income bracket. When it comes to the selection, the reason we were interested in these variables was to observe that there is/are a certain value(s) in either category where most of the samples present themselves in either income bracket. Furthermore, the availability of two different but related variables would allow us to see home in on a specific user or set of users than where would it not be possible if we looked at each variable on its own. Due to the nature of a heat map, there was not much that we could do further improve the visual effects of the chart; however, the legend added on the right of the plots helps the user better understand how many occurrences are in each category. Additionally, similar to the case of the grouped bar chart, including the numbers of the occurrences in each cell, we are able to realize the prevalence of some of the categorical values. To achieve the visualizations mentioned above, we took the following steps:

1. We first copied the original Dataframe over to a new variable.

2. Label encode the income feature to zero or one.
3. We call the "pivot_table" function on each Dataframe copy to be able to aggregate the data to prepare it for the heatmap function.
    a. We index the Dataframes on the relationship column.
    b. The columns for the calculations would be the "Marital Status" feature.
    c. We indicate that we are counting the values of the "Income" feature and include that if there are no values for any categorical value, it would be defaulted to zero.
4. Use the Matplotlib library functions to plot the heatmap and add the x-axis, y-axis, title, and legend labels.

From the two visualizations mentioned above, we can see that there are a couple combinations among the categorical values that stood out. For example, in the "<=50k" income bracket, we can see that 6,779, 29.9 %, of the samples in that income bracket, are of the class "Husband" & "Married-civ-spouse." On the other hand of that income bracket, we can see that 50% of the data are in the following categories: 35% is shared between the "Not-in-family" and "Never-married" classes & the "Own-child" and "Never-married" classes and ~15% shared between the "Not-in-family" and "Divorced" classes & the "Unmarried" and "Divorced" classes. If we move onto the ">50k" income bracket, we can plainly see that 5,675 samples in that income bracket are of the class "Husband" & "Married-civ-spouse." These number of samples represent 75.5% of the samples in the income bracket. With that being said, we can infer that most of the marketing resources for the ">50k" income bracket can be placed for one class mentioned above, and for the "<=50k" bracket, the resources can be placed in the other 4 classes.

D. User Story 3: The admission office at UVW College is interested in finding if there is a correlation between hours per week worked and Sex related to income.

For the final multivariate visualization, we chose the Y-axis to be the number of occurrences of individuals in a specific category, and in the X-axis, we had two variables, two bars for the different sexes and along the x-axis were the number of hours an individual worked. Due to the visualization being multivariate, we had two different plots for the differentiating of the income brackets. The interest in these variables and type of plot is to discover if there was a peak in the number of hours as well as a difference between the number of the genders in income brackets.

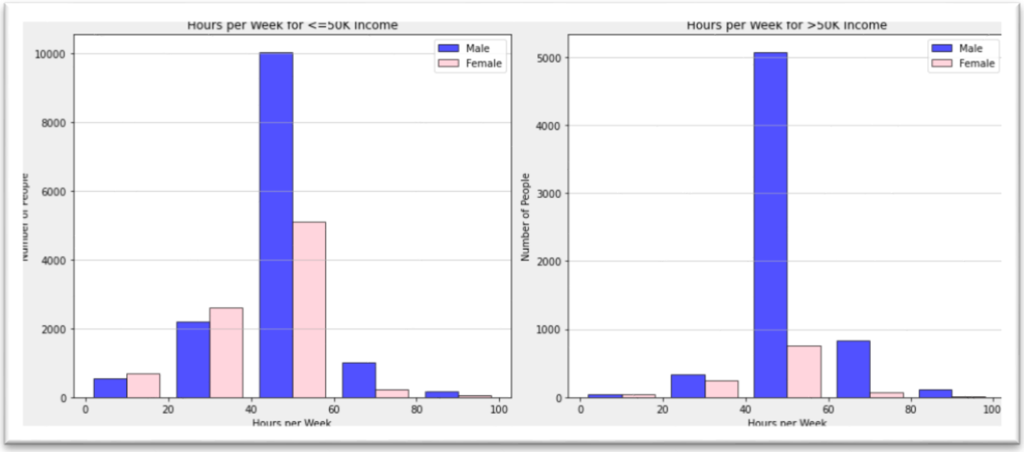

Figure 4: Histogram displaying HPW & Sex VS Income

To enhance the visualization, we created five bins for the data to be displayed in, as we wanted to limit the number of different hours being displayed on the x-axis, which allowed the data points to be more cohesive. To achieve this visualization, we used the following steps:

1. We defined the five bins, which are increments of twenty beginning from zero all the way up to one hundred.
2. Next, we split the dataset based on the different income brackets.
3. We then called the histogram function twice, once for each income bracket.
4. Finally, we added the X-axis, Y-axis, and title labels, as well as used the "tight_layout" function which was used in the grouped bar chart.

From the visualization, we can see the majority of the data points lie within the forty to sixty hours-per-week mark. Furthermore, on the histogram on the right, we can see that the majority of the data points, specifically 75.3% of data points in the higher income bracket, belong to the male gender. Additionally, we can see that the number of data points dramatically decreases after the forty to sixty range. So we can conclude, that a good indicator that an individual is in the higher income bracket if they are a male and work between forty to sixty hours-per-week.

E. User Story 4: The Dean of the UVW College is interested in seeing the correlation between income bracket and education levels of individuals.

This visualization is the first univariate visualization. On the Y-axis we have the number of occurrences of entries in each of the unique values of the "Education" feature, and on the X-axis, we have the different education levels. The reason we chose the parallel coordinates plot is due to the nature of the data. This type of chart allows us to see any spike in the data at any level. Furthermore, due it to being a form of a line chart, we can plot both income brackets on the same plot unlike the heat map plot in the previous user story. The interest of this plot arises from finding where most entries are located within the different levels of education available within the dataset.

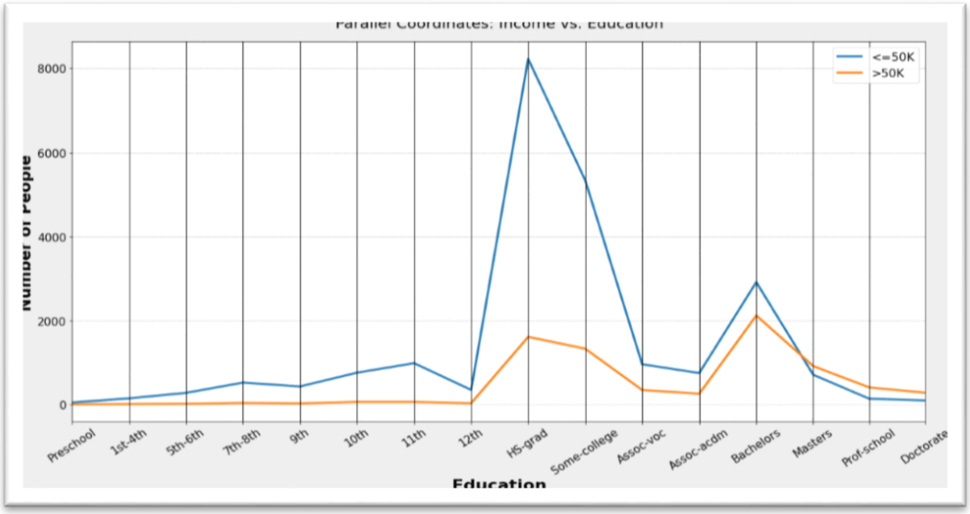

Figure 5: Parallel Coordinate Plot of Education VS Income

As seen in the figure above, there are two peaks in the data when it comes to either income bracket. The most distinct peak is the one at the Highschool graduate degree with around 8,000 entries. Even though this chart may not be able to steer us too much into the right direction, it allows us to see where most individuals in the dataset are in terms of education. To further enhance this visualization, we organized the x-axis categorical values in increasing orders of education, which enhances the flow of the chart. The steps taken to achieve this visualization are as follows:

1. We first defined an ordered list of education levels that are present in the dataset.
2. We then split the dataset based on the two different income brackets.
3. Next, we counted the occurrences of each education level in the respective income brackets.
4. Most importantly, we iterated over the education levels to ensure that all education levels are in both lists, and if not, we place a default value of 0 for that respective education level.
5. We then sorted the resulting lists based on the education level defined in the first step.
6. Transformed the data into Dataframes.
7. Finally, used library functions to plot the parallel coordinates plot and add the x-axis, y-axis, title, and legend labels.

From the results mentioned in the plot, we can conclude that most individuals that earn "<=50k" are high-school graduates, which compromises 36.3% of all the entries in that income bracket. On the other hand, for the ">50k" bracket, you had 1,617 entries, which compromises 21.5% of the entries in the income bracket. Most of the samples in the ">50k" income bracket lies within the "Bachelors" education level, which comes out to be 2,126 entries or 28.3%. Furthermore, in the "Masters" education level and the levels after, the ">50k" income bracket has more entries than the "<=50k" income bracket. This allows us to conclude that if we are to have samples in those levels, they are most likely to be of the higher income bracket. Even though this user story did not steer us in a certain direction, it has given us more details on what values in the "Education" feature would affect the income classification.

F. User Story 5: As an analyst for the XYZ Company, we are interested in finding out if there is a correlation between the age and income bracket of an individual.

This visualization is the third and final univariate visualization. On the Y-axis, you have the numerical variable of age, and on the X-axis, you have the two types of income brackets. The reason behind the selection of the box-and-whisker plot is that it works well with numerical variables, and it allows us to see where the majority of the data points lie, as well as how the values are skewed between the income brackets.

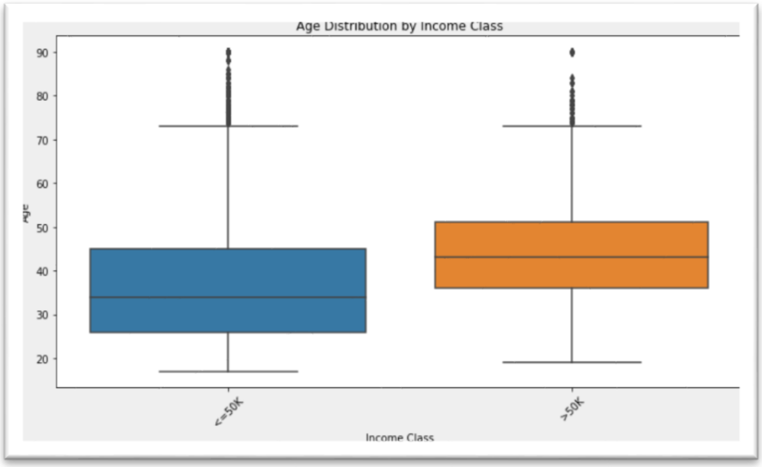

Figure 6: Box-and-Whisker Plot of Age VS Income

The steps taken to obtain this visualization are as follows:

1. Split the dataset based on the two different income brackets.
2. Use the Matplotlib function titled “boxplot.”
3. Add the x-axis, y-axis, and the title label charts.

From the results mentioned in the plot, we can see that the most samples in the lower income bracket lie between 25-45, while the higher income bracket is more concentrated between 35-50. Furthermore, we can see that the 4th quartile and outliers are the same in both income brackets. We can conclude that entries above the age 45 are more likely to be in the higher income bracket.

Questions arose and how did you answer them

IV. QUESTIONS WHILE WORKING

A. Questions that arose

During the initial phases of the project, we were interested in exploring the “Native-Country” feature to implement a heat map to see what we can discover. To our surprise, the results achieved were repetitive between the two income brackets, as we were nearly identical plots.

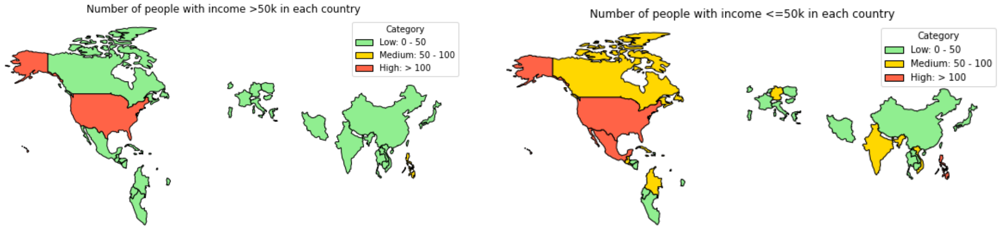

Figure 7 and 8: Heat Maps plotting Country VS Income

B. Solutions Implemented

After further investigation, we discovered that this was due to the majority of the data points being in the same categorical value. In the case of the “Native-Country” feature, the most prominent value was the “United States” value, which makes up around 91% of the dataset. This led us to take preventive measures during the selection process of the eight variables for the user stories, as we did not want to result in misguiding visualizations where the feature is being dominated by one value.

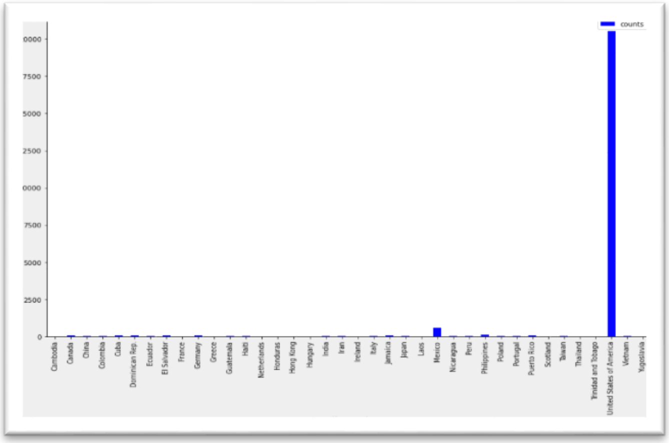

Figure 9: Bar chart displaying that 91% of the data points are located within the USA

V. FUTURE WORK

When it comes to the future work for this project, we can leverage the power of machine learning models to see which features are the best correlated to the income bracket feature. As mentioned by the professor that we are not being judged or graded on the implementation of any machine learning techniques used in this project, it would, however, be beneficial for future work to leverage models to see what would result in the best accuracy, as we have already implemented the data preprocessing steps needed. Another technique that can be used would be to establish a correlation matrix that would identify the most recurring values of features in either income bracket. From there, we could identify the features and the values for the respective time bracket and inform the UVW College how to plan accordingly.