Systems Documentation Report

Roles and Responsibilities

The team consists of 3 graduate students: Quan Le, Francis Kim, and Nathan Kelly. We are all computer science majors with an interest in data visualization. Every member has a responsibility to the team to submit assignments on time in a correct fashion. Each member has a responsibility to communicate with other members. Our members have to document their visualizations and work done on this project.

Team Goals and Objectives

The main goal of this project is to study the census data that was given, and try to find the major determining factors for income. We do this by creating data visualizations to make the patterns and trends easy to see for us and others. Our objective is to predict which factors are the biggest contributors to income, and then we try to back up our predictions through the visualizations we create. One we create our visualizations, we will analyze them for trends and patterns. Our next objective is to try to rank the factors by greatest contribution to determine the largest contributors. A sub objective is to create the visualizations that are easy to read for other people to show our work.

Assumptions

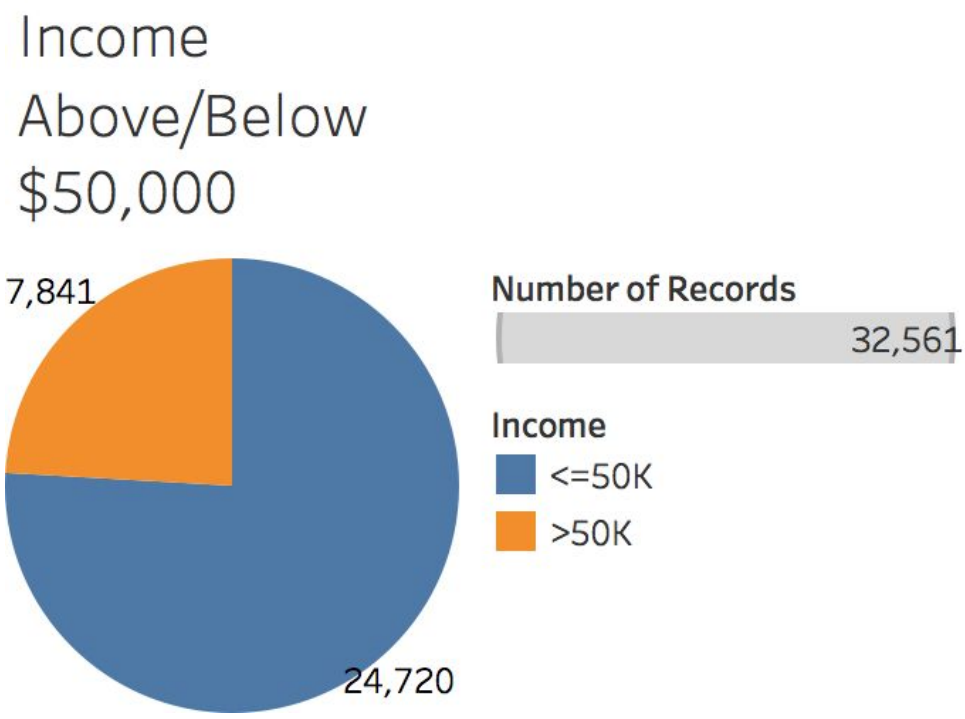
Some assumptions that we have to make is that the data set we are given was correctly recorded and that there are minimal mistakes in the data. Another assumption is that

the data set we are working with is the only data we will look at. We assume that we only have two tools to use to create visualizations, and the tools are Tableau and Python libraries. We are working under the assumption that the project is being presented to others, so the data must be easy to read and identify.

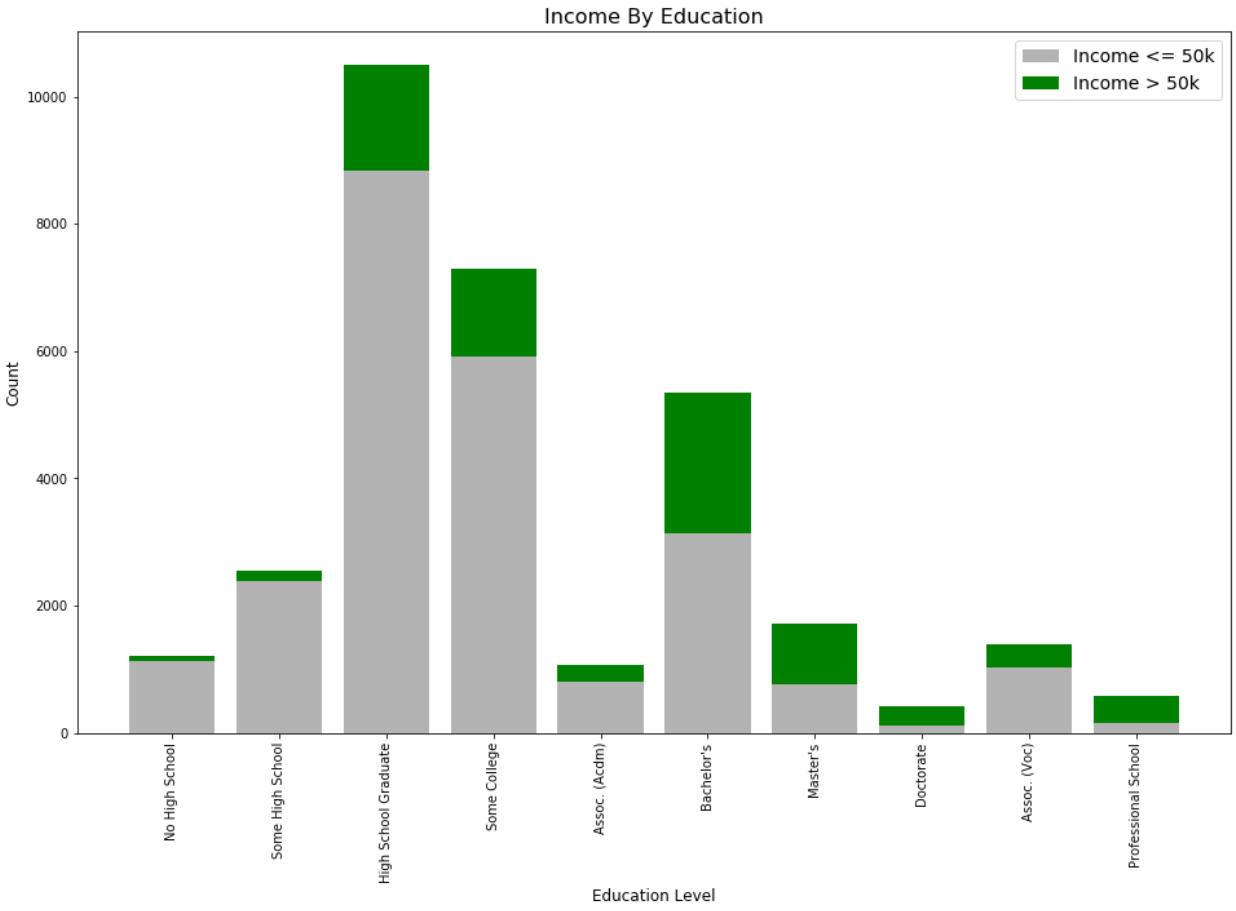
User Stories

As a user I want to be able to see the data in a clear and concise manner. The data should be easily readable and should not be confusing to distinguish data items. This means that there should be proper titles, axes labels, and legends when necessary. Colors mays be used to help distinguish data sets, and make everything more visually pleasing and legible.

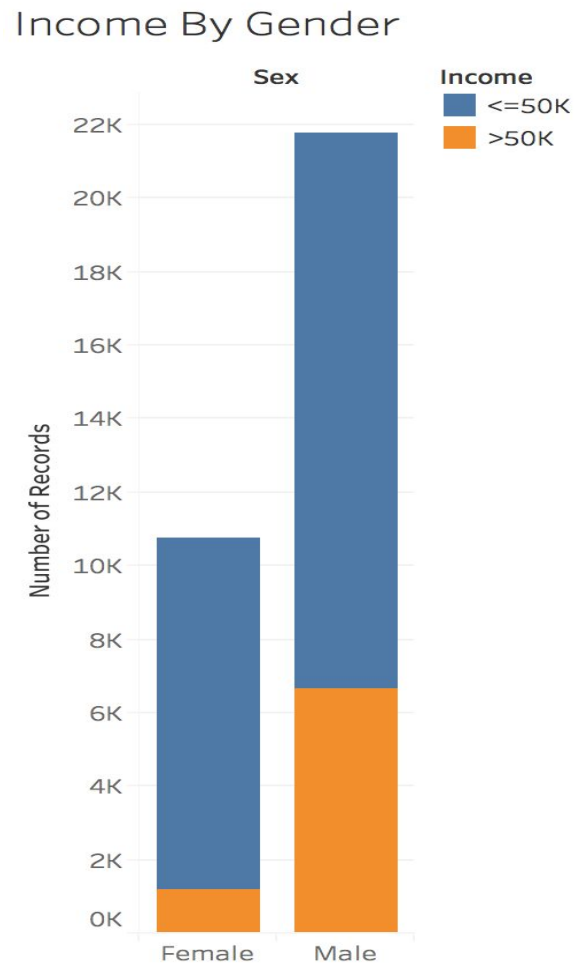
Visualizations



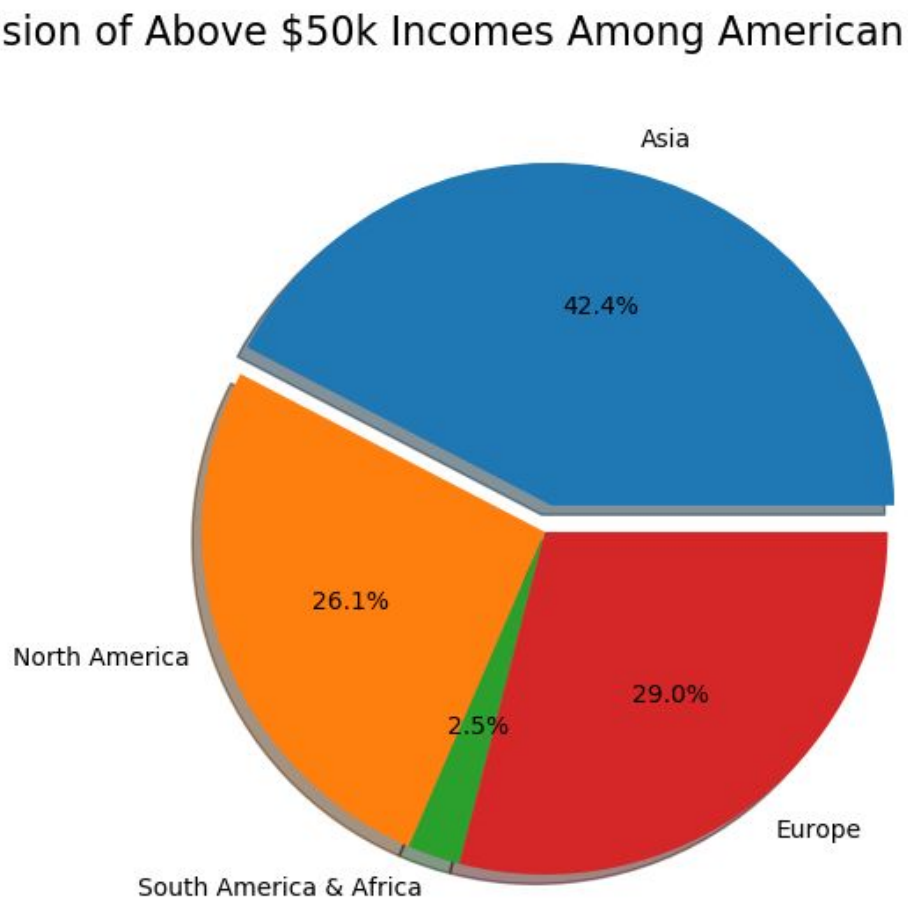
We created this visualization because it is a good baseline to see how many people are above and below $50,000 income. The pie graph shows us that approximately 25% of the sampled data is above $50,000 in income.



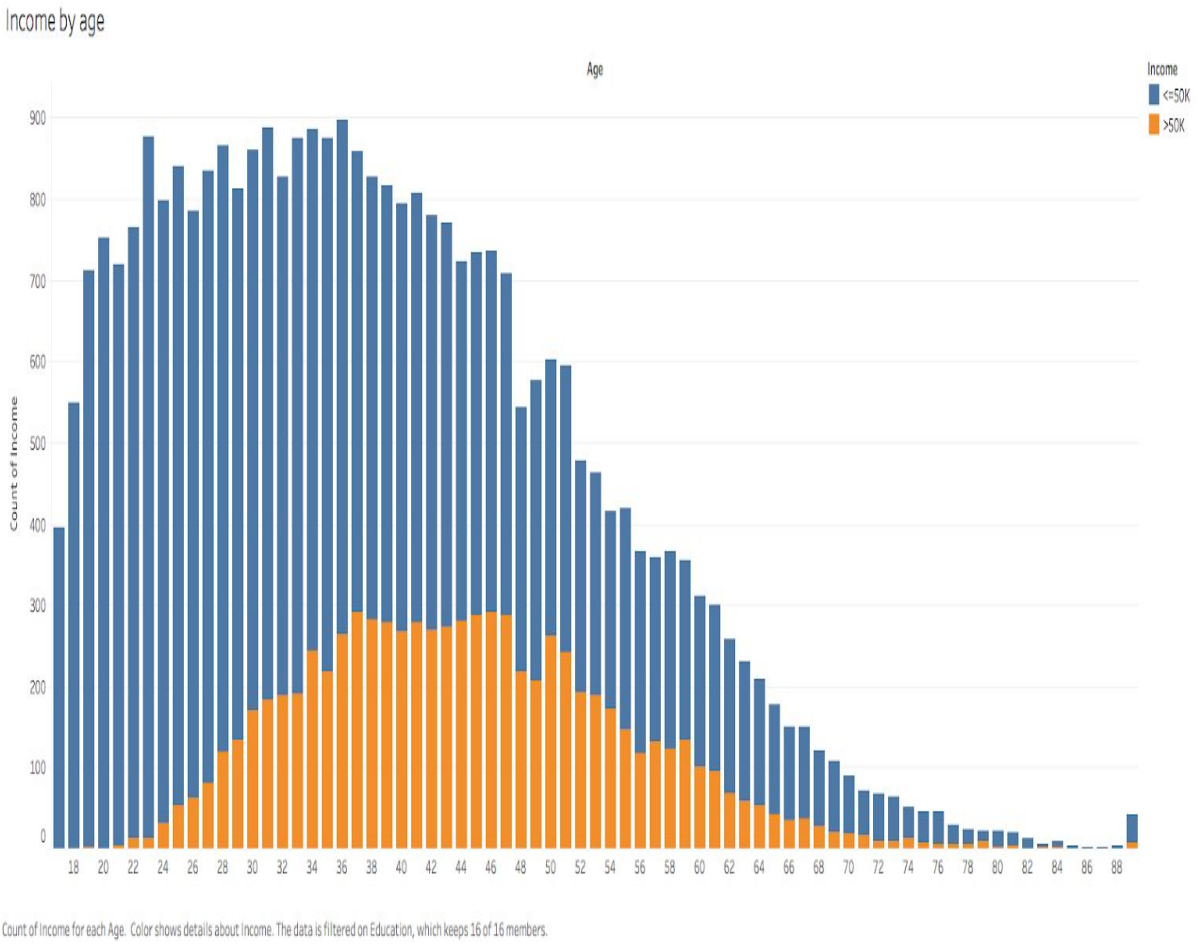
We chose to create this visualization because we predicted that education would be a major contributing factor to the income. We can see that the groups of people with higher level education have a larger proportion of people that make more than $50,000. We can clearly see that the groups of people with masters, doctorate, and professional level education hold a majority of jobs that make more than $50,000. All other groups in the data set have a majority that make less than $50,000. The trend appears that the higher level of education a person receives, the more likely the are to make more than $50,000.



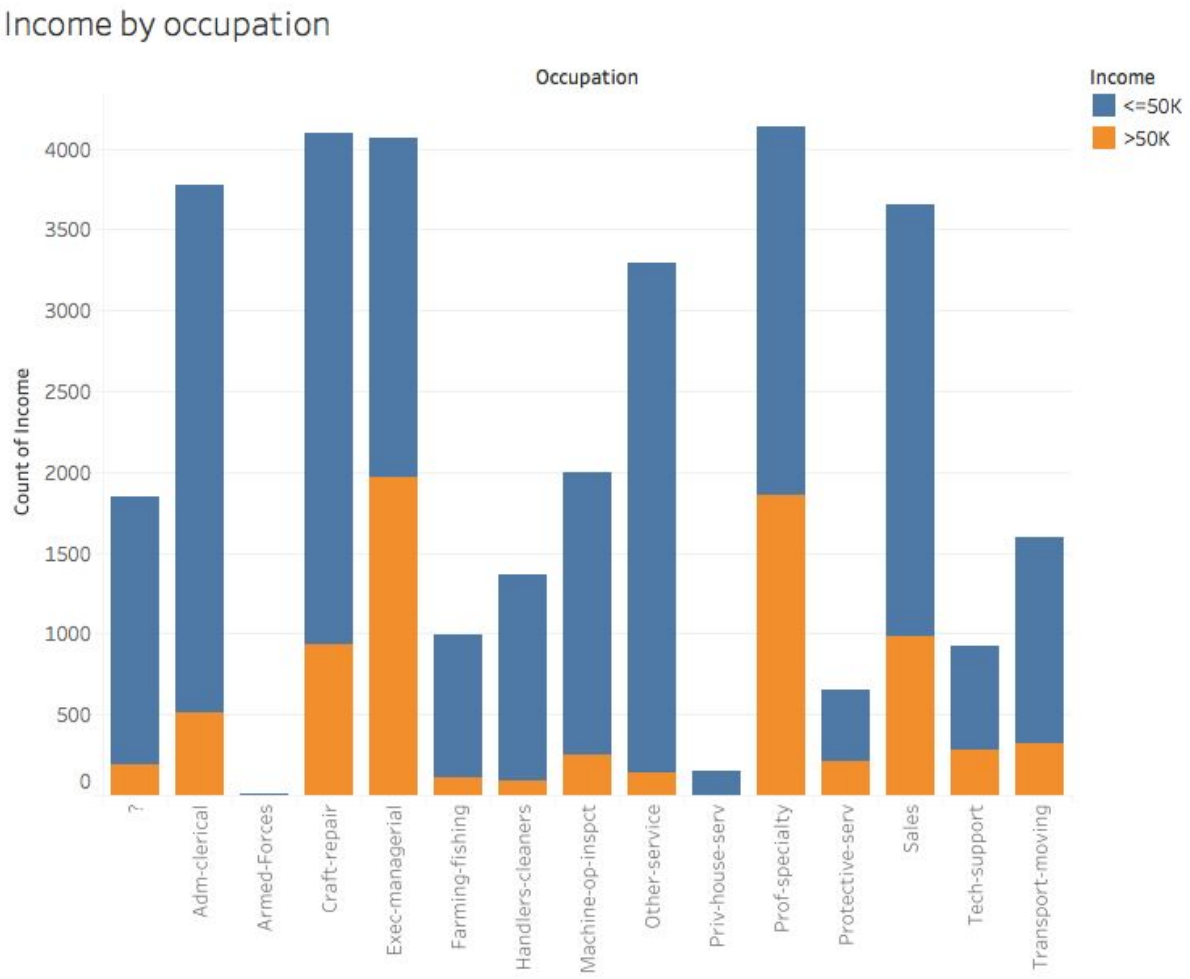
An interesting contributor that we wanted to look at was whether or not gender would affect income. The data makes it difficult to see a trend because the the men are in the majority of income less than $50,000 and income greater than $50,000. It is also difficult to determine because there are more samples with males in the data set than there are women in the data set, so it is harder to determine whether the sample size is large enough to say that there is a clear trend.



We wanted to see which immigrant groups or continent of origin would have the highest income among the data set. It turned out the Asia had the highest percentage of high income earners, with Europe and and people from North America excluding Americans almost closely tied.



We wanted to see the income range depends on people's ages. Age from 30 to 50 have a lots of people who earned over 50k, and age from 18 to 23 have relatively small number of people who earned over 50k. In addition, even if there are small amount of people who work after age 60, there are high percentage rate of people who earned over 50k due to their experiences.



Count of Income for each Occupation. Color shows details about Income. The data is filtered on Education, which keeps 16 of 16 members.

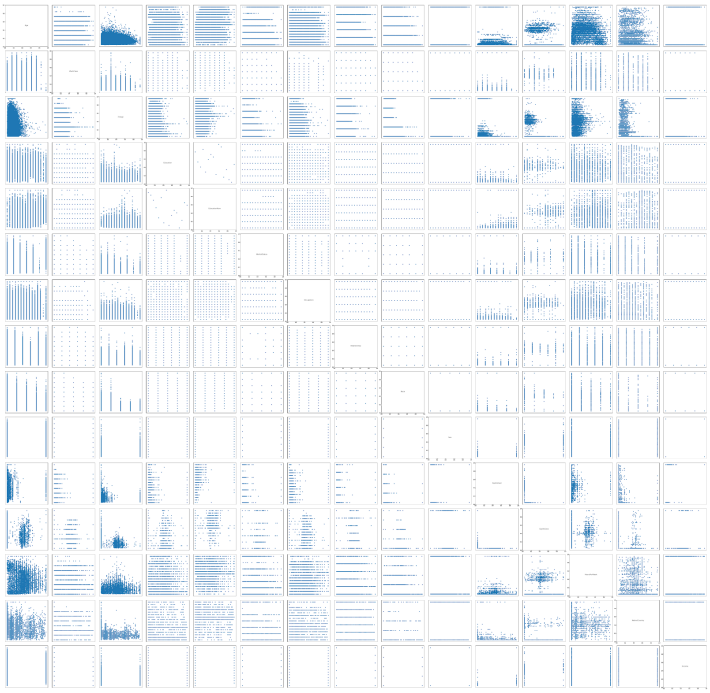
We wanted to see the range of income corresponding to different occupations. There are a lot of people who work in craft-repair, exec-manager, and prof-speciality, and also those occupations have relatively many people who earned over 50k income. Sales and Adm-clerical is also having high percentage of people who earned over 50k income.

Questions.

The first question we had was what factors did we think were the biggest contributors to income. We discussed among ourselves and tried to pick our top 5 data columns that we thought would affect the income the most. Next we had the question of how we would determine whether or not our predictions were correct. We did this by creating graphs and data visualizations to see whether or not there were any trends in the data that would give us a good idea whether or not the factors were big or small contributors.

Not Doing

As there was quite a lot of fields surveyed in this census, it was not possible to deeply evaluate each and every one. A scatterplot matrix was used to see if we could quickly hone in on the features that mattered most without having to invest too much time into dead ends.



However, since we were dealing with a lot of nominal and ordinal data, and our primary focus was two discrete values (<=50k & >50k) rather than continuous, this low-investment exploration did not help much. For each field there are many ways to evaluate as well, such as rolling up from country to continent in our evaluation of immigrant incomes. Given more time it would be beneficial to explore more approaches like this.

Conclusion

We found that the largest contributor to income was education level. The higher the level education a person had, the more likely it was that they would make more than $50,000. It was hard to find whether or not gender had anything to do with income determination since the data had more males than females. An interesting thing we

found in the data was that the country of origin showed that people from Asia were more likely to have a higher income, with Europe following behind in second.