Project Milestone 4: System Documentation Report for

Scatterplotter (Team 38)

Team members:
James Rodriguez
Priyanka Sugawe
Quency Hadisubroto
Rakesh Krishnan Thaliyil Veetil
Sreenivas Nagaraja
Yingbing Wang

Roles and responsibilities

Product Owner: UVW College (is looking to bolster the enrollment process).

The target goal for each team member is to come up with visualizations for the attributes in the dataset in order to figure out which of them are the best predictors for income level. To start, our group held meetings in which we discussed how to proceed with the project requirements. We decided that the best place to start is to first analyze which attributes would intuitively be the best predictors for income level. We choose to visualize: sex, race, marital status, education, education number, occupation, age, capital gain, hours per week. Then we divided the task of visualizing these attributes amongst 6 of us.

Owner	Title/Subject	Graph
Priyanka Sugawe	Sex, Race, Marital Status	Pie Chart
Quency Hadisubroto	Education, Occupation, Sex	Bar Chart
Sreenivas Nagaraja	Education num, Age	Box Plot
James Rodriguez	Age, Sex, Education	Mosaic Plot
Rakesh Krishnan Thaliyil Veetil	Age, Capital gain, Hours per week	Scatter Plot
Yingbing Wang	Education-num, hours per week, Capital gain	Parallel Coordinate Plot

We then worked individually on getting our part of visualizations done. We decided that each team member would be in charge of one type of visualization method. Then we assign multiple attributes to be visualized by each team member. As you can see, some attributes are visualized using more than one graph. We decided to do this in order to see which graph can capture the data distribution the best.

In addition to creating visualizations, each of our team members are responsible for the following:
James Rodriguez: Explained project requirements. Analyze attributes and come up with which graphs are best for each one. Proposed the attributes for Mosaic plot.

Priyanka Sugawe: Created shared drive documents. Schedule team meetings. Recommended the attributes for Pie chart. Wrote Progress Report.
Quency Hadisubroto: Came up with the team name. Explained project requirements. Analyze attributes and come up with which graphs are best for each one. Proposed the attributes for the Bar chart. Wrote System Documentation Report. Worked on Executive Summary Report.
Rakesh Krishnan Thaliyil Veetil: Schedule team meetings. Analyze attributes and come up with which graphs are best for each one. Wrote System Documentation Report.
Sreenivas Nagaraja: Proposed the features which can be utilized in a box plot. Analyze graphs and decide which graphs are best for attributes chosen. Decide which attributes are best for determining income level. Worked on Executive Summary Report.
Yingbing Wang: Explained project requirements and expected results. Presented features that can be used in Parallel Coordinate plot. Calculates mean, median and standard deviation value for education number, and hours of work.

Team goals and a business objective.

Our team's most important goal for this project is to come up with clear, meaningful and useful graphs for each attribute in the dataset. The business objective is to figure out which of the attributes in the dataset could be the best predictor for income level amongst the population. In this project, income level is divided into two categories, <=50K and >50K. The business, UVW college, plans to use this information to bolster their enrollment. In figuring out which attributes are best predictors for salary, we are building an application that will predict the income of an individual based on a few input parameters. Our goal is to be able to perform tailored marketing when reaching out to these individuals. Let's say a person has a full time job and is a single father, our marketing efforts can be tailored towards flexible scheduling and online options.

Assumptions

- **Completeness and Comprehensiveness:** Incomplete data gives inaccurate results in the data visualization and prediction analysis. Our team assumed that the dataset is accurate and there are no gaps in data collection which might possibly lead to a partial view of the overall picture to be displayed. It is important to understand the complete set of requirements that constitute a comprehensive set of data to determine whether or not the requirements are being fulfilled.
- **Dataset is accurate and precise:** Our team assumed that the dataset given to us was accurate and precise . We assumed that the data does not contain erroneous information like outliers, corrupted data in any of the columns or rows. Without understanding how the data will be consumed, ensuring accuracy and precision could be off-target or more costly than necessary.
- **Time and Relevance:** Our team assumed that the data was collected at the correct data intervals or at the right moment in time. Data collected at irregular intervals could give inaccurate information and decisions.
- **Feature Selection:** We assume that the features which contribute the most towards class prediction can give us more interpretable patterns. So, the majority of the visualizations and analysis is done on the selected features which are unbiased.
- **Legitimacy and Validity:** We presume that the dataset is legitimate and validated. For example, in the given dataset, items such as gender are typically limited to a set of options and open answers are not permitted. Any answers other than these would not be considered valid or legitimate based on the dataset's requirement.

User Stories

User Story #1: As a member of the UVW marketing team, I want to know if the **age of an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #2: As a member of the UVW marketing team, I want to know if the **sex of an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #3: As a member of the UVW marketing team, I want to know if the **education achieved by an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #4: As a member of the UVW marketing team, I want to know if the **marital status of an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #5: As a member of the UVW marketing team, I want to know if the **race of an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #6: As a member of the UVW marketing team, I want to know if the **capital gain of an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #7: As a member of the UVW marketing team, I want to know if the **hours worked per week by an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

User Story #8: As a member of the UVW marketing team, I want to know if the **occupation of an individual** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

Visualizations

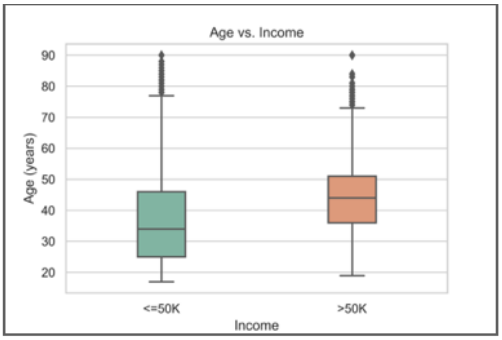

Fig.1

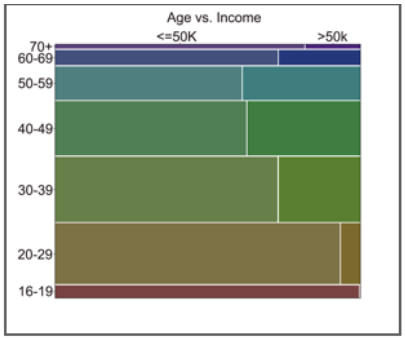

Fig.2

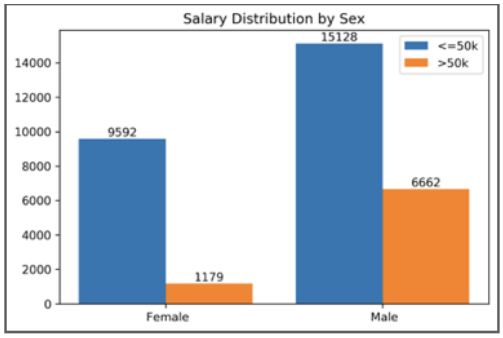

Fig.3

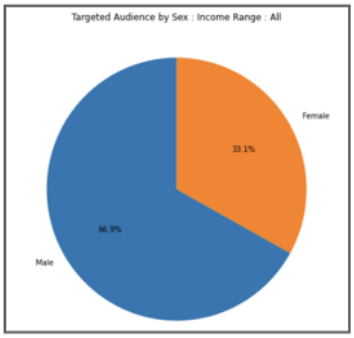

Fig.4

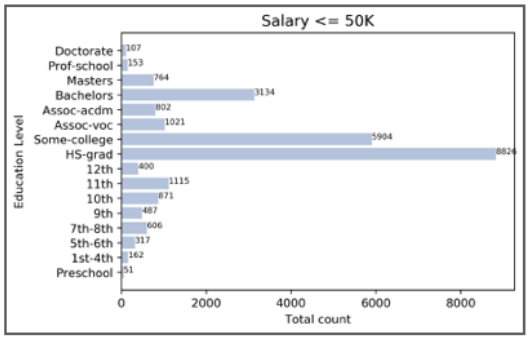

Fig.5

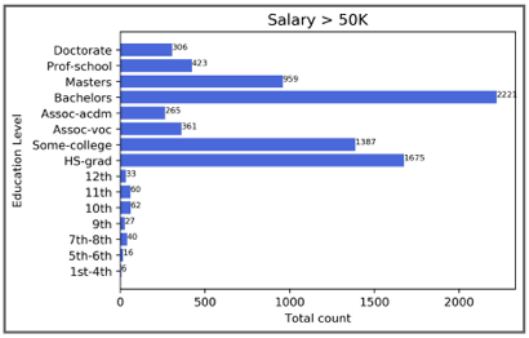

Fig.6

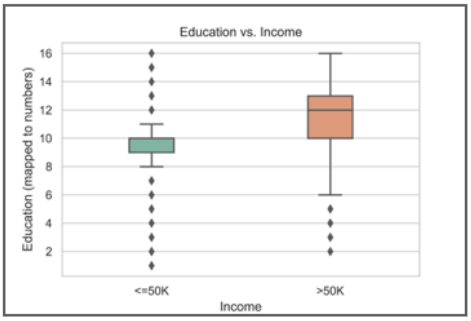

Fig.7

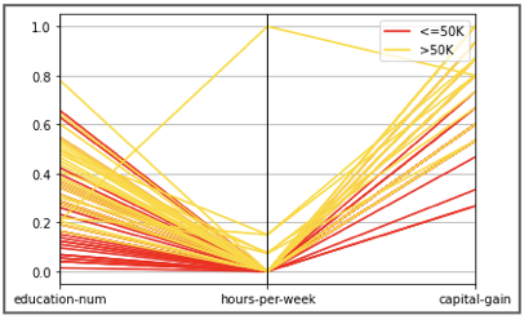

Fig.8

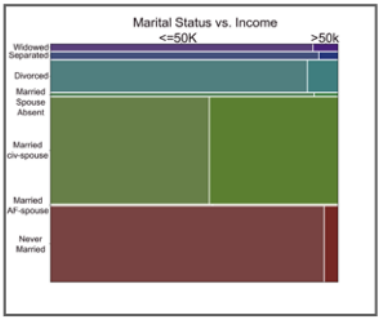

Fig.9

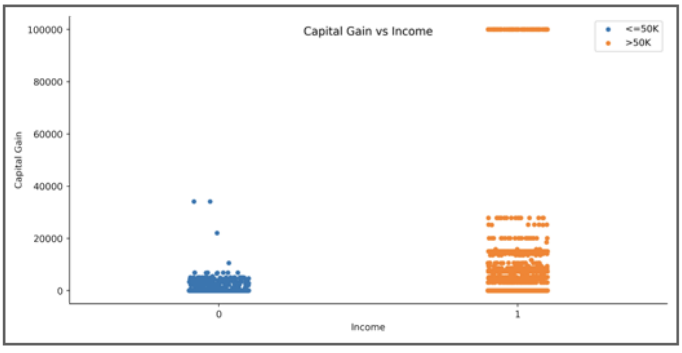

Fig.10

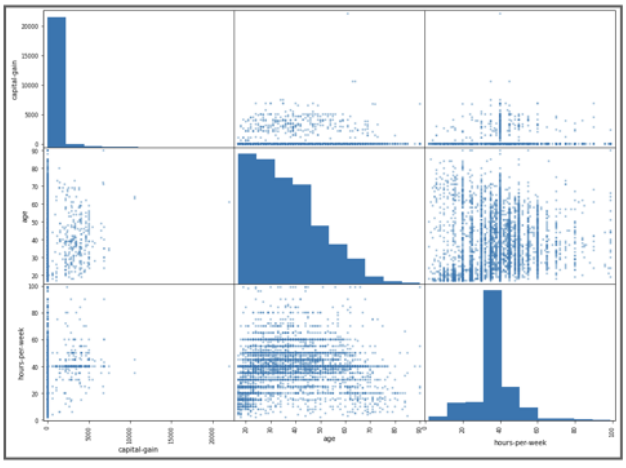

Fig.11

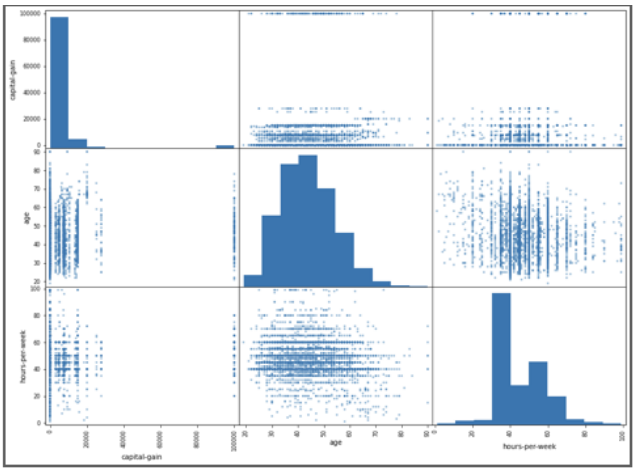

Fig.12

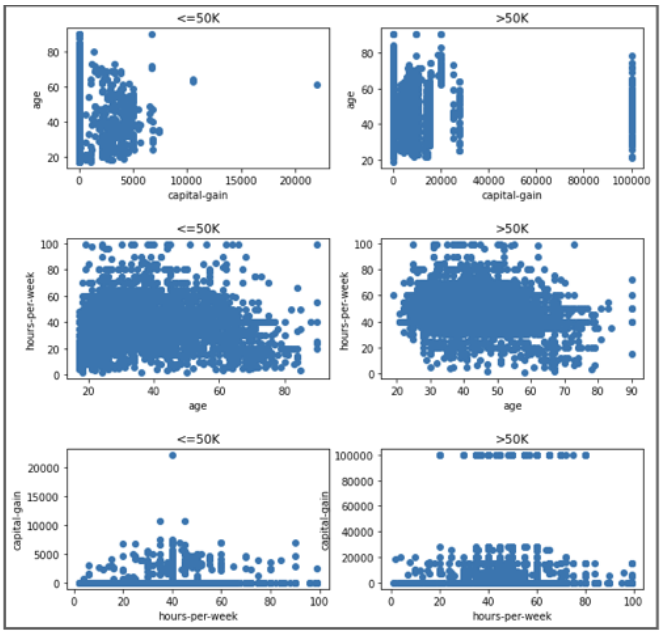

Fig.13

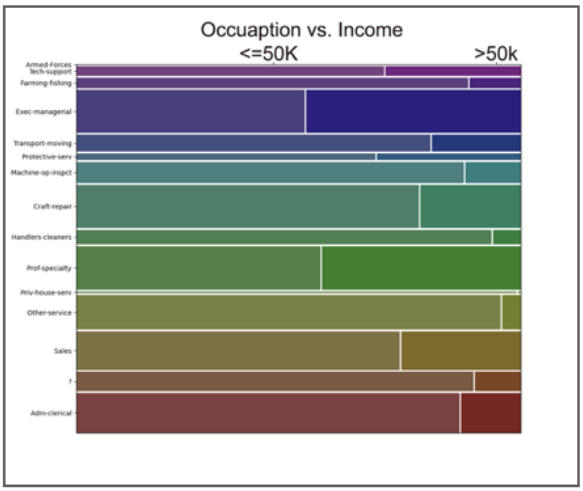

Fig.14

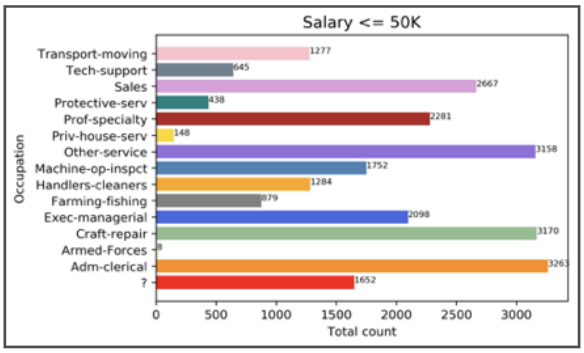

Fig.15

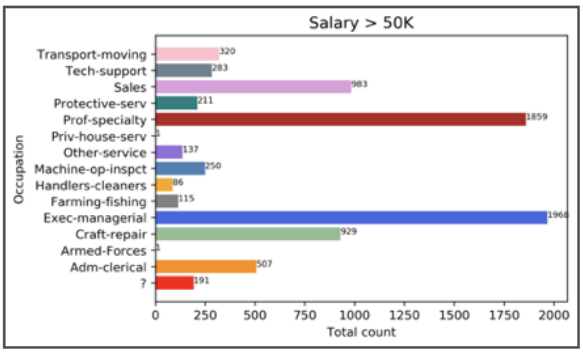

Fig.16

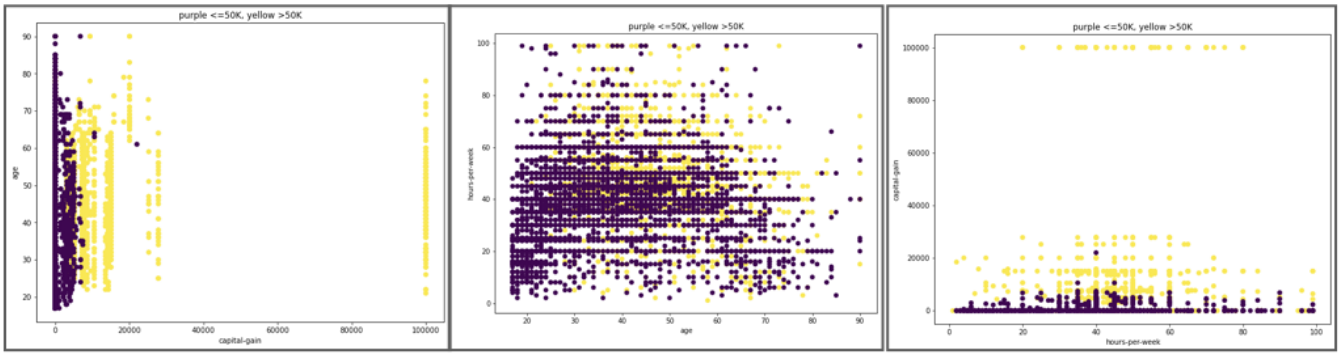

Fig.17

Fig. 1 **Age:** Compares age ranges in terms of their incomes. (E.G. People at an elder age, or under their thirties tend to have less people with <-50k income compared to other age ranges)

Fig. 2 **Age:** Overall data indicates that the population earning less than 50K is younger with a median age of 34 years vs. 44 years for those earning above 50K. In both the populations, there are also outliers as indicated and population earning less than 50K tend to work for lot more years compared to the other group

Fig. 3 **Sex:** The resulting bar chart indicates that there are more people with wages <=50K in both sex cateogries. However the ratio for Females is 8:1 in favor of a salary of <=50k, whereas for Males, the ratio is around 2.5:1. Furthermore, although the ratio between males and females is 2:1, the number of males with a salary >50k is about 6 times the number of females making that salary. From this we can conclude that people making >50k would very likely be males.

Fig. 4 **Sex:** As shown in the Pie chart, we can see that 66.9% target audience is male and 33.1% audience is female. If we compare the audience gender with the income range greater than 50K, there is 85% male population. 38.8% female audience has income range less than or equal to 50K.

Fig. 5&6 **Education:** The resulting bar chart indicates that many people with salary less than 50k have an education level of hs-grad, followed by some-college and bachelors. Whereas for people with salary greater than 50k, many have an education level of bachelor degree followed by hs-grad, some-college and Masters. We can use education level as an indicator for whether a person makes less than or more than 50k a year in this way: If the person has a bachelor degree, they're more likely to have a salary of greater than 50k. If a person has a HS diploma, they're more likely to make less than 50k. We can also conclude that most people with education level of Masters, Prof-school and doctorate, make salary greater than 50k.

Fig. 7 **Education:** The box showed the median education number is 12 for income > 50K, 9 for <50k group

Fig. 8 **Education:** From the parallel coordinate plot, we can see that the yellow lines and the red lines can be distinguished using the combination of these three features.The result shows the higher education num, the hours-per-week go lower, and the capital-gain get higher. The income >50K group has higher education and capital gain but lower capital gain than the income <=50K group.

Fig. 9 **Marital Status:** Seeing if having a specific marital status leads to a majority of <=50k or >50k

Fig. 10 **Capital gain:** The capital gain plots show that the capital gain is less than 10k if the salary is <=50K and the capital gain is more 1000 to 5000 when the age is less than 70. If age is above 70 the capital gain is close to 0. The capital gain is 100k when the salary is more above > 50k and till age 70. Most of the capital gains are in the range of 0 to 20k till age 65-70.

Fig. 11&12 **Capital gain & hours per week:** Scatter plot matrix of hours per week, capital gain and age divided using salary <=50k (fig. 11) and >50k (fig. 12). Some inferences can be made from this graph. Age: people in the age range of 30-55 get a salary greater than 50k and as age increases the salary reduces and hours per week reduces. Capital gain: If capital gain is high, it is more likely that the person earns higher than 50k. Hours per week: Most people who work 40 hours per week have a salary greater than 50k.

Fig. 13 **Capital gain, hours per week & age:** The above plot shows the ages below 60-55 do work hours per week more than 40 hours for salary less than 50 k. Ages above 70 the hours per week gradually decreases and at

age 80 the hours per week is less than 20.For the next plot if the salary is greater than 50K the age range from 20 to 70 the hours per week is more than 60-70 hours per week. For age above 70 the hours per week are much lesser. The above plot shows the scatter plot for ages vs hours per week with purple dots as salary less than 50 k and yellow dots with salary greater than 50k. The hours per week is more for people with a salary greater than 50 k when compared to the salary with less than 50k. As the age increases the hours per week reduces.

Fig. 14 **Occupation:** Displays what occupations have a majority of individuals over 50k income.

Fig. 15&16 **Occupation:** From the resulting barchart, we can conclude that people with occupations of prof-specialty and exec-managerial are most likely to make a salary greater than 50k a year.

Fig. 17 **Capital gain and Age:** There is separation between the two classes and few outliers. Individuals with higher capital gain earn more than $50k income salary.

Questions

In the first stages of working on this project, we only assigned each team member with one type of visualizations (pie chart, bar chart, scatter plot, box plot, mosaic plot, parallel coordinate plot). We assumed that each team member would come up with visualizations of the 14 attributes in the dataset. Then we realized that not all attributes can be represented with every graph type we have. Then we asked the question, which attributes are best represented with which graphs? The team had confusion on which dataset to use since there were 2 datasets i.e, adult.data and adult.test dataset. We decided to use the adult.data dataset.

Not doing

In the scope of this project, we decided to visualize only 9 out of the 14 attributes in order to answer the 8 user story questions asked. Education has two attributes, education number and type of education. We chose not to look at income correlations with these following attributes: relationship, work class, capital loss, fnlwgt, and native-country at this stage. We also chose not to use machine learning methods to predict income based on input parameters at this stage. For future developments, putting effort into using machine learning methods would be ideal.