Team 10 – Special Data Force Milestone 4 – Systems Documentation Report

Rasoul Barri, Victoria Kirkman, Sai Soujanya Parasa, Ryan Slaney, Thien Tran

Abstract—This Systems Documentation Report outlines the progress made by Team 10, 'Special Data Force', on the CSE 578 Data Visualization group project. It describes the roles and responsibilities of all personnel involved in the project, provides the teams goals, objectives and assumptions, communicates user stories and their associated visualizations, reviews questions that the team had during the development process, and concludes with a list of tasks that were not completed within the time constraints of this project.

Keywords— Data Visualization, user story, pie chart, bar chart, histogram, mosaic plot, scatter plot, box and whisker plot, stacked bar chart, choropleth map, roll-up, categorical variables, continuous variables, US census data, US income classification.

I. ROLES AND RESPONSIBILITIES

A. Roles

For this project, the team nominally worked as a "team of data analysts" [1] for XYZ Corporation, a company that "uses data to develop marketing profiles on people" in order to sell them "to numerous companies for marketing purposes" [1]. XYZ Corporation was hired by UVW College, a local educational institution "looking to bolster enrollment" using "salary as a key demographic to determine criteria for marketing its degree programs" [1]. Three main categories of roles were identified for this project: Product Owner, Team Members, and Stakeholders.

- *PO (Product Owner) / End user* was UVW College. This role in reality was filled by the course instructor and graduate student assistants. PO is responsible for approving the project conclusion as well as defining its acceptance criteria.
- *BA (Business Analyst, Thien Tran)* is responsible for discovering the business objectives and their technical requirements. BA makes sure that the project scope can meet the deadline set by the client, UVW university.
- *PM (Product Manager, Ryan Slanley)* is responsible for overseeing the documentation of the project. PM ensures the product specifications meet the technical requirements.
- *SM (Scrum Master, Victoria Kirkman)* helps to drive meetings forward. SM focuses on the value-added pieces of the backlog and refines the understanding of individual tasks.
- *DE (Data Engineer, Sai Soujanya)* processes data: format, structure and consolidate the data into a tabular for analysis.
- *QA (Quality assurance specialist, Rasoul Barri)* performs quality tests on the dataset.

B. Responsibilities

Legend	Responsible	Accountable	Consulted	Informed

Activity/ Role	BA	PM	SM	DE	QA	PO
Team goals and business objective	R	I	A	I	I	
Report revising, consolidating and proofing	A	R	A	I	I	
Visualizations	R	R	R	R	R	
Decide on plans of attack	I	A	R	I	I	
Meeting	R	R	R	R	R	
Progress report	R	I	I	I	I	C
Coordinate plots, mosaic plots, star plots	I	I	I	I	R	
Focus on relevant attributes	A	A	A	A	R	
Backlogs	R	A	A	I	I	
Box and whisker plots, scatter plots, pie charts	A	A	R	A	A	
Histograms	A	A	A	R	A	
Executive report	R	R	R	R	R	I
System documentation report	R	R	R	R	R	I

II. TEAM GOALS AND BUSINESS OBJECTIVE

A. Team Goals

The team's goals were driven by the Project Description [1] and included:

- Analyze the dataset to determine the most important attributes affecting the enrollment.
- Process the provided US Census Bureau data.
- Collect user stories from the stakeholders.
- Conduct initial exploratory analysis to find the demographic attributes with the strongest correlation to salary.
- Create visualizations showing correlation between demographic data and salary in response to user stories.
- Pass the correlation to the development team that "will be in charge of building and maintaining the application that the marketing department will use" [1].
- Deliver the Intermediate Project Report, Executive Summary, System Documentation Report.

B. Business Objective

The business objectives were tied to the profits of both stakeholders, in this case both UVW College and XYZ Corporation.

UVW College's objective was to increase enrollment, which will increase revenue. The college's plan to bolster enrollment was to focus its efforts using "salary as a key demographic

to determine criteria for marketing its degree programs" [1]. XYZ Corporation's objective was to develop and sell successful marketing profile(s) to UVW College to assist them in their goal. Presumably, the contract between UVW College and XYZ Corporation was written such that XYZ Corporation would profit in a manner proportional to the success of its marketing profiles.

III. ASSUMPTIONS

The team made a series of technical and business assumptions to aid in development. During the duration of the project, these assumptions were refined as clarification was received from the stakeholders.

- The analysis is based on the given dataset adult.data and adult.test. Other data won't be used. The adult.names file contains all the metadata.
- The CSE 578 Instructor and Graduate Student Assistants hold the role of Product Owner and Stakeholder, representing both XYZ Corporation management and UVW College marketing team.
- Corporation XYZ does not need to engage UVW College in every phase of the project.

IV. USER STORIES

The Atlassian project management literature describes a "user story" as "the smallest unit of work in an agile framework" [2]. It is an "informal, general explanation of a software feature written from the perspective of the end user or customer" [2]. In this project, the user stories were framed from the perspective of the UVW marketing team and were as follows:

1. As a member of the UVW marketing team, I want to know if the Education 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.
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.
3. 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.
4. As a member of the UVW marketing team, I want to know if the Native Region 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.
5. 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.
6. 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.
7. As a member of the UVW marketing team, I want to know if the Workclass 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.
8. As a member of the UVW marketing team, I want to know if the Capital Gains 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 .

V. VISUALIZATIONS

This section contains the visuaizations corresponding to the attributes listed in the user stories above. Each visualization is paired with background information, a conclusion that notes whether each attribute is a reliable indicator for predicting the income of an individual, and a recommendation as to whether it should be included in the team's prediction tool. Each visualization corresponds with a user story above, e.g. the first visualization included below corresponds to the first user story in the list in Section IV of this document.

1) Education

The bar chart can be inferred by looking at the median lines that the individuals with a high education number are more likely to earn a salary greater than 50K. This feature could be one of the most striking ones to determine an individual's income for developing the application *(Figure 1)*.

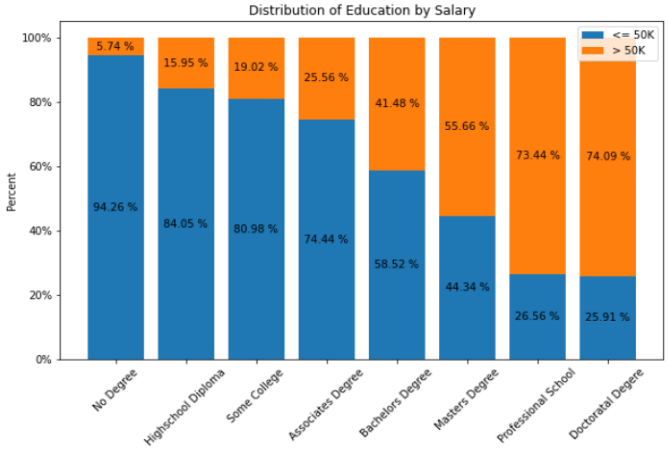

Fig. 1. Stacked Bar Chart of Education with respect to Salary

2) Sex

This pie chart provides some meaningful insight with a quick look at the visualization. When looking at each gender, it can be seen that among all females, a very high proportion (of females) have a salary of <=50K, more so than the proportion of men who make <=50K. This data indicates a correlation between sex and salary, and this wage gap is a known, shameful phenomenon that continues to be an issue in modern society. Because of this correlation, it is recommended that the application team include sex in their prediction tool *(Figure 2)*.

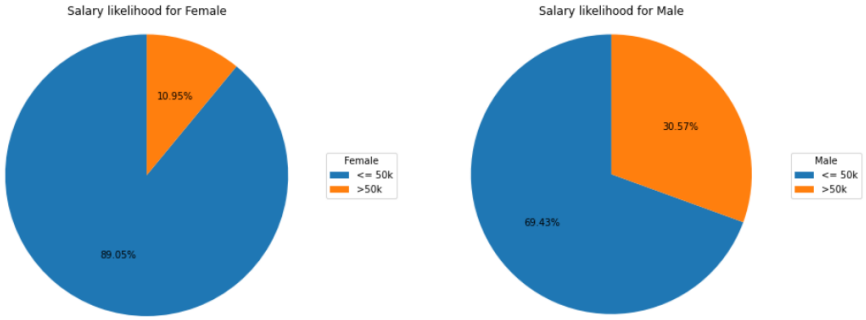

Fig. 2. Pie Chart of Sex with respect to Salary

3) Marital Status

This plot reveals that a high proportion of individuals whose marital status is 'Married' earns a salary >50K when compared to the remaining categories of marital status. Also, 95.4% of the individuals who never married fall under the class <=50K. This feature could be one of the most determining ones to develop the application *(Figure 3)*.

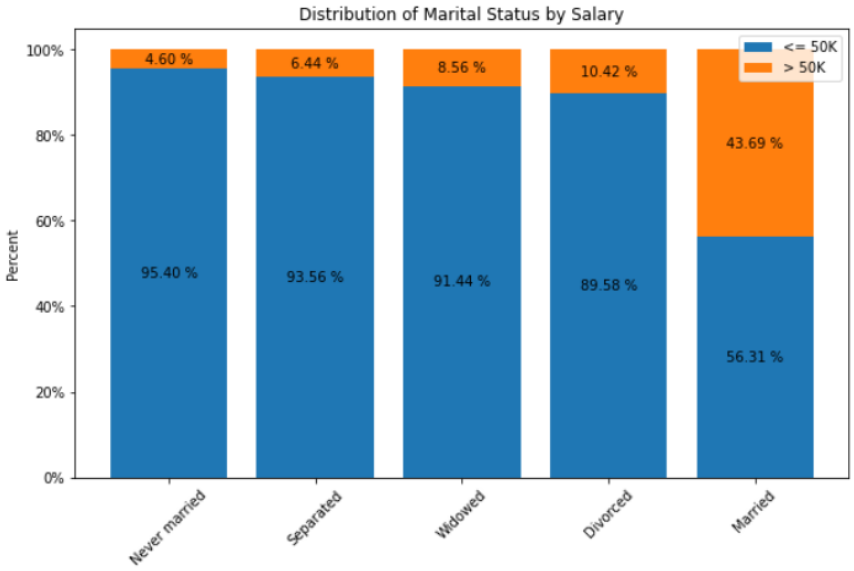

Fig. 3. Stacked Bar Chart of Marital Status with respect to Salary

4) Native Region

The choropleth map shows the existence of a strong spatial autocorrelation and partial autocorrelations of the spatial statistic variable Native Region. The countries with high percentages of wealthy individuals having been living in the US tend to be from these regions. This attribute is recommended to be included in the application *(Figure 4)*.

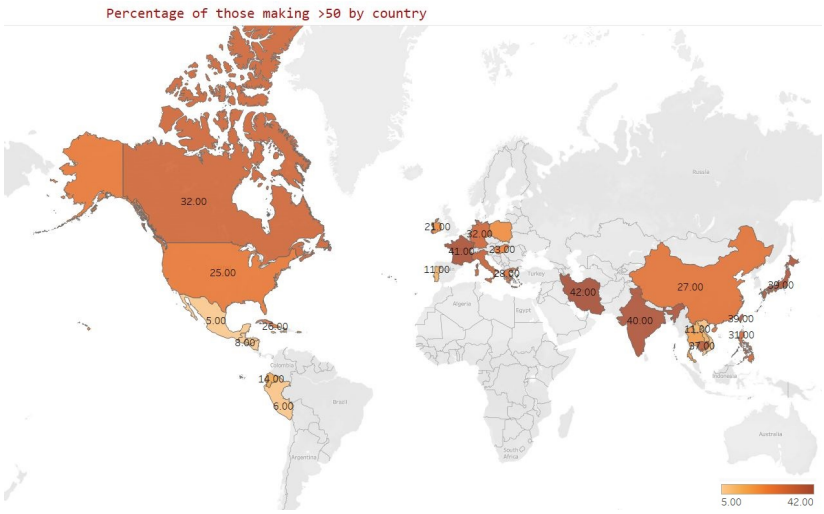

Fig. 4. Choropleth Map of Native Country with respect to Salary

5) Age

The inference made from this visualization of the box-and-whisker plot plotted against Age and Salary is; the median line for the category salary >50K is higher that is between the ages 45 and 50 when compared to the one in salary <=50K that is between the ages 35 and 40. Thus, older individuals are more likely to earn a salary >50K. To develop an application to identify the factors that determine an individual's income, 'Age' could serve as one of the top 8 features *(Figure 5)*.

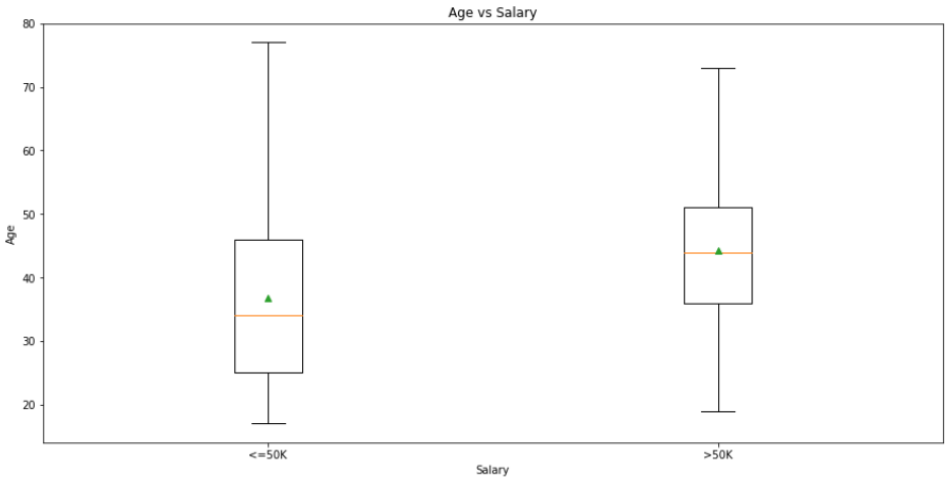

Fig. 5. Box and Whisker Plot of Age with respect to Salary

6) Occupation

From the plot, the individuals who have occupations 'Exec-managerial' and 'Prof-speciality' make an earning more than 50K. It is recommended that this feature be used to develop the application by binarizing the categories of interest *(Figure 6)*.

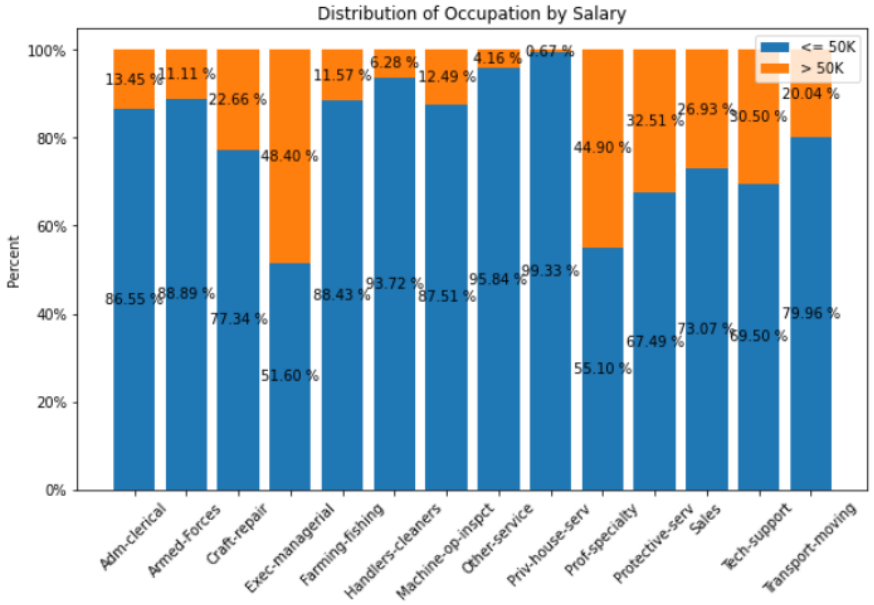

Fig. 6. Stacked Bar Chart of Occupation vs Salary

7) Workclass

Of all the individuals in 'Self-emp-inc.', 55.73% of them earn a salary greater than 50K. This category of the workclass has the highest percentage of individuals that fall under the class salary >50K. Those who never-worked and without pay have a strong correlation (100%) with making less than 50K. A few of these categories can be binarized for use *(Figure 7)*.

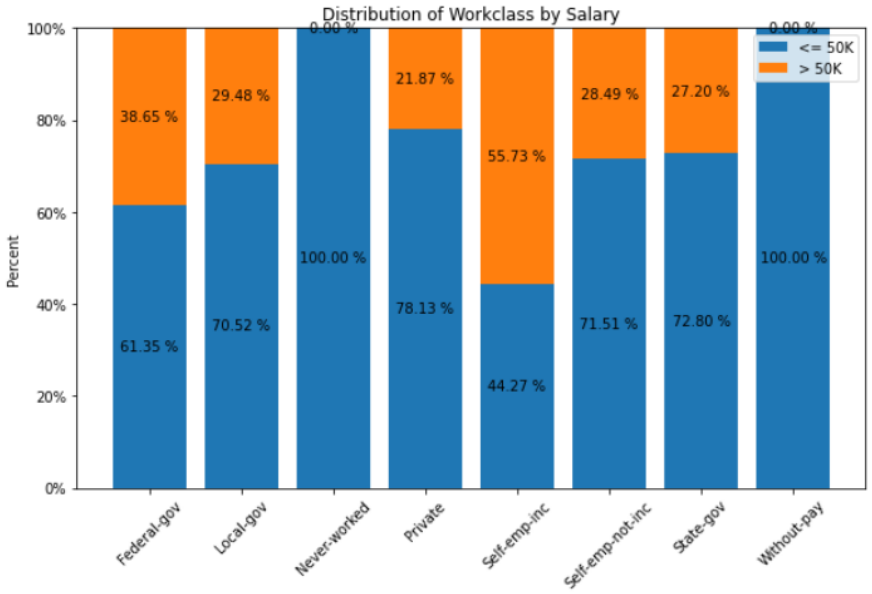

Fig. 7. Stacked Bar Chart of Workclass with respect to Salary

8) Capital Gain

This scatter plot in Fig. 8 shows the bivariate analysis of capital gain and salary. Scatter plots were chosen to represent the attribute 'capital gain' because this attribute was composed of continuous, numerical data. The majority of those <50K have capital gain amount less than $7,500. Although the attribute might be statistically significant, it could represent the same underlying property of the label variable. The formula of an individual taxable income is:

taxable income = income +/- capital gains/losses +/- interest

Therefore, it is not recommended to include capital gains in the prediction tool *(Figure 8)*.

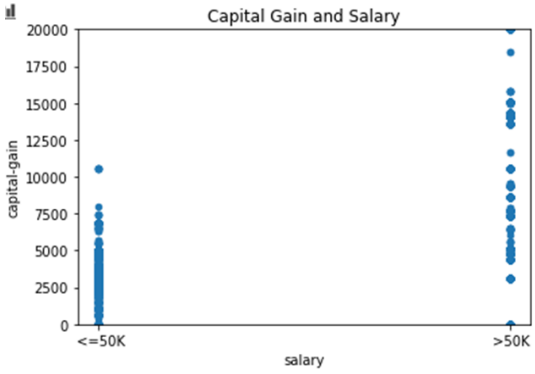

Fig 8. Scatter Plot of Capital Gain vs Salary

VI. QUESTIONS

During this project, many questions arose. The following is a summary of the questions the team had, and the solutions that were implemented:

- Question: What format should be used for the reports? Answer: There was no specified format. The team used a modified version of the IEEE template.
- Question: Does the goal include predicting future income (do we need to consider future income)? Answer During the Week 4 live event and in the Discussion Forum, the instructor indicated that this was not a goal for the project.
- Question: Is the >= $50,000 / < $50,000 the class label? Answer: Yes – this was confirmed in Class Forum Week 2. However, the use of machine learning and class labelling was determined to be outside of the scope of the project by the Product Owner and communicated in the Course's Week 4 Live Event.

VII. NOT DOING

This section lists the tasks that were not completed by the team by the final submission of the report, but need to be completed in the future in order to build a successful application. They are ordered by descending priority, i.e. the first item in the list is the most important and should be worked first, in subsequent builds of the marketing profile.

- The following attributes were not explored due to time constraints, but should be explored at a future date in order to build a more robust predictive model the correlations between: race salary, relationship education-num, hours worked per week and salary.
- We exclude variable fnlwgt because it is an id variable unsuitable for analysis.
- Creating a roll up and drill down on the Occupation attribute.
- Combinations of attributes were not considered in this iteration of the project. Exploring each attribute in isolation was a good starting point, however a more robust application would benefit from visualizations in which attributes are joined to show their combined impact on salary.
- The following visualizations were not included in the data exploration; however, their inclusion in future builds may enhance understanding of the data.
- Star charts, mosaic plots, and parallel coordinate plots – these multivariate analysis plots were not used in the initial analysis of the data but will be extremely useful to provide future insights for the application.

REFERENCES

[1] Project Description Document:
https://www.coursera.org/learn/cse578/supplement/B4Lrb/course-project-introduction

[2] https://www.atlassian.com/agile/project-management/user-stories

Appendix

CODE

(Included in the project zip file)