CSE 578: DATA VISUALISATION COURSE PROJECT FINAL REPORT

Aditya Kumar Verma
MS Computer Science
Arizona State University
Tempe, United States of America
akverma6@asu.edu|1225121638

***Abstract*—In this study, I used the 1994 US Census dataset to analyze the relationship between various demographic attributes and income. The goal was to assist UVW College in targeting their marketing efforts towards individuals earning more than $50K annually. Eight key attributes—age, education, marital status, relationship, occupation, work class, race, and sex—were scrutinized through a series of visualizations. Despite challenges along the way, strategic solutions ensured the project's success. The resulting insights provide a valuable resource for UVW College, empowering them to craft more targeted and effective marketing strategies.**

I. Goals

The principal goal of this project was to carry out an exhaustive data exploration and analysis that divulges the convoluted relationship between selected demographic attributes and an individual's income level. The ultimate intent is to illuminate the critical factors that shape whether an individual's income exceeds or remains below an annual income threshold of $50,000.

II. Business Objective

The intelligence gleaned from this data analysis has profound strategic significance for UVW College. The objective of deriving these insights is to bolster the efficacy of the institution's marketing campaigns, shape their admissions strategy, hone the design of their academic programs, and facilitate more effective diversity initiatives. A detailed understanding of the drivers behind income levels will empower UVW College to tailor their strategies and initiatives more effectively, thereby amplifying their impact.

III. Assumptions

During the course of the project, I worked under certain assumptions:

- The dataset chosen for this analysis accurately reflects the larger population that UVW College is targeting
- The cut-off income level of $50,000 is a fitting benchmark to differentiate between 'low' and 'high' income classifications for the purposes of UVW College's objectives
- The data provided is dependable and has been recorded and presented without any inherent bias or skew
- The demographic attributes selected for analysis - namely age, education, marital status, relationship, occupation, work class, race, and gender - exert a tangible influence on a person's income level. The reasons for the attributes selected are given below:

1. *Age:* It's generally observed that income levels tend to rise in the early to mid-career stages and decrease in later stages.
2. *Education:* The level of education has been strongly correlated with earning potential.
3. *Marital Status & Relationship:* Studies suggest that marital status and the type of relationship play significant roles in economic prosperity.
4. *Occupation & Work Class:* The industry in which one works and the nature of their occupation contribute significantly to income disparities.
5. *Race & Gender:* These socio-demographic factors have historically influenced earning potential, making them important considerations in our analysis.

IV. User Stories

- User Story #1: As a marketing strategist, I want a granular understanding of the income distribution across different age and education groups. This data will help tailor effective marketing strategies, targeting degree programs to specific income groups
- User Story #2: As an admissions officer, I want to analyze the income distribution across various marital statuses and relationships. Understanding how marital status and the nature of the relationship (like spouse, unmarried, own-child, etc.) relate to income can help us create tailored campaigns that

resonate with specific relationship demographics

- User Story #3: As a program planner, I need to examine how occupation and work class correlate with income. Such an understanding will help me design academic programs and courses that align with the job markets and potential earnings, making our programs more desirable
- User Story #4: As a diversity committee member, I wish to observe the intersection of race and gender in relation to income. Knowledge about income disparities among different racial and gender groups would inform our initiatives, driving our goal to increase diversity and inclusion in our academic programs
- User Story #5: As an academic planning team member, it is crucial to understand how the level of education influences income. This insight will guide our curriculum design process, ensuring alignment with job market trends and potential earnings

V. Visualizations

The journey of arriving at the most effective visualization for each scenario was meticulous and involved several layers of experimentation and decision-making.

A. *User Story #1: Age vs. Income (Marketing Strategist)*

My initial approach was to represent the data using histograms and box plots to understand the distribution of income across different ages. However, I quickly realized the need for a more nuanced visualization that encapsulated not just the median income across ages, but also the distribution of income within each age group. This led me to adopt the violin plot, a sophisticated visualization tool that elegantly captures both these dimensions. While a box plot provides quartile information, and a histogram displays density estimation, a violin plot exhibits both, giving a more detailed distribution of data. Additionally, it is more effective for comparing distributions between different categories, displaying the density of data points around different values of age, which box plots don't provide. This visualization serves as a lucid representation of income distribution across ages and offers a nuanced understanding that can help target demographics with higher income potential more effectively. It thus clearly ties back to the business objective of enabling targeted marketing strategies.

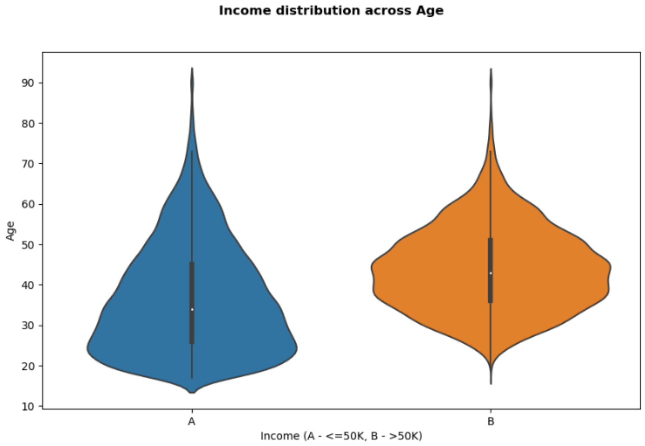

Figure 1 Violin Plot depicting Income distribution across Age

As depicted in Figure 1, the violin plot shows not just the spread of the income but also the probability density at different values. This allows the marketing strategist to have a fuller picture of income trends across age groups. Upon examining the plots, a pattern emerges that suggests the majority of individuals earning less than $50k fall within the 20-30 age bracket. Conversely, those with earnings exceeding $50k are predominantly clustered within the 35-50 years range. Another insightful takeaway from these visual representations is the apparent correlation between age progression and the probability of earning more than $50k. This likelihood typically is higher up until individuals reach their retirement age.

B. *User Story #2: Marital Status & Relationship vs. Income (Admissions Officer)*

With the introduction of marital status and relationship variables, the complexity of the data increased significantly. I experimented with a variety of visualizations, including grouped bar plots, before settling on the Parallel Categories Diagrams (Parcats). Parallel Categories Diagrams prove to be superior over grouped bar plots for visualizing Marital Status and Relationship vs. Income because they can effectively illustrate interactions among multiple categories. A grouped bar plot tends to fall short when there are numerous categories to be compared. In contrast, Parallel Categories Diagrams can handle a larger number of categories with interrelationships and provide a more comprehensive visualization. The Parcats visualization excels at depicting the interconnectedness of marital status, relationship type, and income levels. The comprehensive insights offered by this visualization will help the Admissions Officer in strategizing and implementing targeted outreach efforts that resonate with specific demographics. This multivariate plot allows the admissions officer to see not only the individual categories but also how these categories interact with each other.

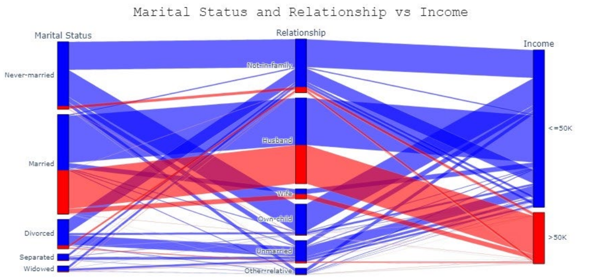

Figure 2 Parallel Categories Diagrams of Marital Status and Relationship vs Income

Evaluating the Parallel Categories Diagrams plot, shown in Figure 2, leads us to discern that the majority of individuals with incomes exceeding $50k are married. Those individuals with a marital status other than 'married' generally exhibit earnings below the $50k threshold. This visualization offers invaluable insights to the Admissions Officer, assisting them in tailoring outreach efforts while considering potential students' relationship status. Furthermore, the plot implies a near-even distribution of married wives' incomes, with about half earning more than $50k and the other half falling below this threshold. The clearer and bigger representation of the figures are added to the end of the report in the Supplementary Materials section.

C. User Story #3: Occupation & Work Class vs. Income (Program Planner)

For the third user story, I grappled with representing multi-level categorical data that included occupation, work class, and income. To demonstrate the correlation between occupation, work class, and income, I used a heatmap. Heatmaps prove to be particularly proficient in the visual portrayal of Occupation, Workclass, and Income correlation, courtesy of their capacity to denote data density via varying color intensity. This visualization technique facilitates the effective absorption of the relationship information embedded in the categories. Traditional methods like bar plots or pie charts may find themselves inadequate to handle the high-dimensional complexity of the data; however, heatmaps seamlessly encapsulate this information, offering a compact yet visually appealing representation. The heatmap provides a clear visual summary of complex relationships, showing how different combinations of occupation and work class relate to different income levels. This aids the program planner by showing which combinations align with higher incomes. These insights will be invaluable in developing academic programs that prepare students for high-earning professions and work environments.

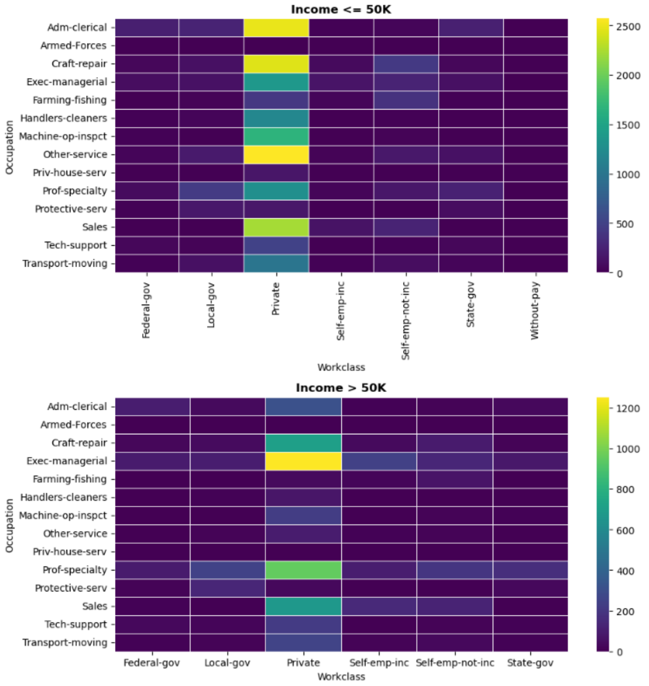

Figure 3 Heatmap depicting Occupation & Work Class vs. Income

The heatmap presented in Figure 3 delineates the distribution of individuals across various occupations and work classes, divided into two income groups: earning less than $50k (depicted in the first subplot) and earning more than $50k (represented in the second subplot). The heatmap reveals the following crucial insights:

1. Except for the 'Private' work class, other work classes do not exhibit a significant correlation with the income. This insight suggests that the Program Planner might want to prioritize the 'Private' job sector when devising academic programs
2. Among the occupations in the private job market, the 'Exec-managerial' role emerges as the most lucrative. This observation could motivate the Program Planner to consider the scheduling of management-focused events and webinars within the college's academic offerings as an experimental initiative

D. User Story #4: Race & Gender vs. Income (Diversity Committee Member)

I experimented with point plots and FacetGrid plots to understand income disparities across race and gender intersections. However, I settled on a stacked bar chart, as it better illustrates the income distribution within each race, split by gender. A Stacked Bar Chart shines when visualizing categorical data like Race and Sex against Income due to its capacity to depict parts of a whole effectively. Unlike the FacetGrid plot and Point plot, a Stacked Bar Chart showcases the distribution of income across different categories of race and sex in a single, unified visualization. It makes the composition of the data explicit and allows us to compare individual segments as well as whole categories with ease. This chart allows the Diversity Committee to understand the discrepancies and patterns within and across racial and gender groups. This in-depth understanding of income distribution across different racial and gender groups will enable the Diversity Committee to identify and bridge income gaps, thus contributing to the institution's diversity and inclusivity objectives.

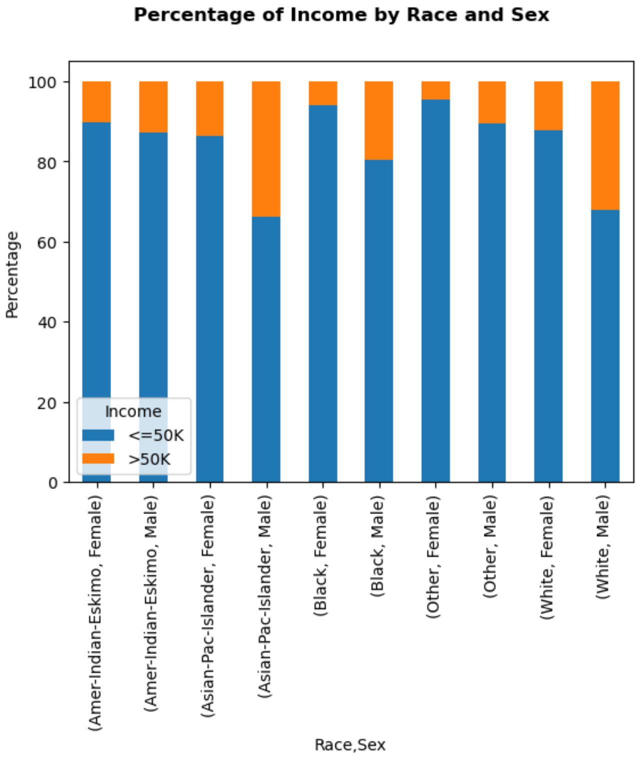

Figure 4 Stacked Bar Chart depicting Percentage of Income by Race and Sex

Figure 4 offers a granular view of income distribution across different races and genders. It becomes evident that irrespective of race or gender, a majority of individuals earn less than $50k. Notably, nearly 30% of Asian-Pac-Islander and White Males earn over $50k. This observation suggests that the diversity committee might not need to prioritize these demographics in their initiatives.

Instead, their focus could be redirected towards the female population, particularly those of the Black and 'Other' race categories. In every community, females earning more than $50k do not comprise more than 20% of the income bracket. Therefore, the diversity committee could strategize and implement initiatives that empower the female population, especially those from the Black community. These efforts would enhance the institute's commitment towards fostering diversity and inclusivity.

E. *User Story #5: Education vs. Income (Academic Planning Team Member)*

I initially tried pie charts and treemaps but ultimately chose donut charts to illustrate the distribution of income across different education levels. Donut charts are effective in comparing parts of a whole, where the entire donut represents the total population, and the slices represent people with different education levels. Donut charts provide an uncluttered and polished method to contrast categorical proportions, such as Education versus Income. They offer a more streamlined visualization than pie charts, mainly due to their distinct, empty central area. This void enhances their utility, serving as a space for additional data or labels, increasing informativeness without adding clutter. In comparison, treemaps effectively visualize hierarchical data but may lack intuitiveness, especially when it involves comparing proportions. Interpreting the rectangle sizes in treemaps can potentially complicate visualization, particularly with closely-matched category proportions. Conversely, donut charts represent proportions as distinct circular arcs, simplifying segment distinction and comparison. Therefore, their clarity and immediate comprehension make donut charts a potent tool for exhibiting income distribution across varied educational strata in our study. The donut chart provided a concise yet comprehensive overview of the data, aiding the academic planning team. These provide an easy understanding of income distribution across education levels, thereby aiding the Academic Planning Team in tailoring curriculum design.

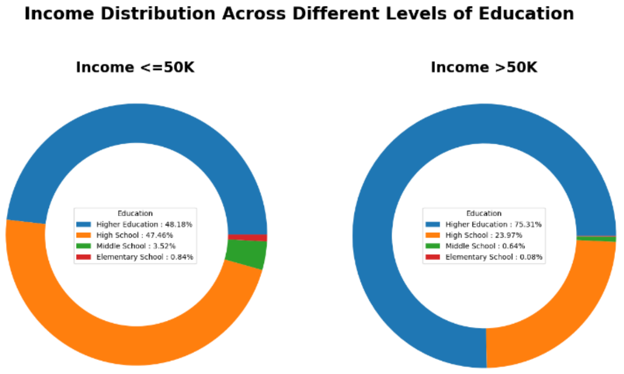

Figure 5 Donut Charts displaying the Income Distribution across different levels of education

Analyzing the donut charts reveals an intriguing pattern — approximately 75% of high-income earners (> $50k) have pursued higher education. This valuable insight informs the Academic Planning team of the significance of higher education in securing high-paying jobs. As such, they might consider introducing accelerated 4+1 programs, thereby encouraging students to pursue advanced degrees.

A secondary observation to be drawn from these charts is the relatively low proportion (about 24%) of employees educated up to high school level. Given this, it becomes clear that to stay relevant in today's competitive job market and to maintain bargaining power, attaining higher education is not just a boon but nearly a necessity. In effect, considering higher education and graduate programs significantly increases the likelihood of securing a more lucrative income.

VI. Questions

Throughout the project, several questions and challenges arose that required meticulous thought and innovative solutions:

A. How do I best represent the distribution of income across various demographic variables?

To answer this question, I first delved into understanding each demographic variable individually and then studied how they related to income. For variables such as age, which is continuous, violin plots served as the best representation, highlighting both the distribution and median income for each age group. For categorical variables such as education, I used donut charts to show the proportion of individuals within each education level that fell under the 'high' or 'low' income category.

B. Which visualization tools can handle multiple variables and their interactions most effectively?

The complexity of the question increased as we started dealing with multiple variables and their interactions. I leveraged the capabilities of more advanced visualization tools, like the Parallel Categories Diagrams (Parcats) and FacetGrid. The Parcats visualization was particularly effective in depicting the interconnectedness of marital status, relationship type, and income levels. FacetGrid enabled us to create a multi-plot grid that depicted income distribution within each racial group, further broken down by gender.

C. How do I ensure the chosen visualizations are easily interpretable by the stakeholders?

As I worked with increasingly sophisticated visualization techniques, the readability of the plots became a crucial concern. To ensure that the visualizations were clear and interpretable, I consistently referred back to the user stories, ensuring that each visualization effectively answered the user's query. This approach allowed me to choose visualizations that, despite their complexity, were tailored to the stakeholder's perspective and thus remained easily understandable.

D. How to ensure the visualizations are aesthetically pleasing and engaging?

The aesthetic aspect of the visualizations was another important factor that could affect the engagement and interest of the stakeholders. I used a variety of techniques to enhance the appeal of the charts, such as adding bold titles and varying colors for pie and donut charts. This not only improved their visual appeal but also made them more readable.

E. How did I identify and handle the missing values in the dataset?

The dataset contained '?' marks to denote missing values. Identifying missing values is crucial as they can affect the accuracy of our analysis. In this case, I removed the records with '?' values in their columns. This approach provided an accurate analysis of the data.

F. How did I unify redundant categories in 'Education' and 'Marital-Status' columns?

The 'Education' and 'Marital-Status' columns contained similar categories that could be unified to streamline the data. For instance, categories like ' 11th', ' 10th', ' 9th', ' HS-grad' in 'Education' were combined into 'High-School', and ' Married-civ-spouse', ' Married-spouse-absent', ' Married-AF-spouse' were grouped into 'Married'. This helped reduce the dimensionality of the data and simplified the subsequent analysis.

G. How did I handle the diverse range of values in the 'Education' column?

The 'Education' column had diverse categories, which I decided to bin into broader categories to simplify the analysis. For instance, I grouped ' Preschool', ' 1st-4th' as ' Elementary School', ' 5th-6th', ' 7th-8th' as ' Middle School', ' 12th', 'High-School' as ' High School', ' Assoc-acdm', ' Assoc-voc', ' Bachelors', ' Doctorate', ' Masters', ' Prof-school', ' Some-college' as 'Higher Education'. This way, the values were more manageable, and their relation to income became clearer.

The journey through these challenges provided valuable learning experiences and led to the development of a comprehensive and insightful visualization project that addresses UVW College's requirements effectively.

VII. Not Doing

While the scope of this project was considerably large and extensive analysis was performed, due to time and resource constraints, several areas were left untouched that might be worth exploring in future iterations of the project. These include:

1. *Granular Demographic Delineation:* Our analysis was confined to the consideration of certain demographic factors like age, education, marital status, relationship, race, and sex. A more meticulous dissection of these demographics, involving segmentation into more refined categories, could yield nuanced insights.

2. *Longitudinal Examination:* The absence of a temporal facet within our dataset precluded the exploration of temporal trends in income distribution. Incorporating datasets with temporal data could provide a dynamic

perspective on income evolution, thereby amplifying the depth of our analysis.

3. *Textual Analytics:* A significant portion of our dataset comprised textual features. Implementing Natural Language Processing methodologies on these text features could unlock additional layers of insights, an opportunity that remained unexplored in the present undertaking.
4. *User Feedback Integration:* Instituting a mechanism for collecting user feedback could exponentially enhance the utility of the application. Gaining insights into the marketing team's requirements and preferences could enable the tailoring of the application to optimally align with their needs.

These enhancements and additions can significantly expand the potential use-cases and effectiveness of the application. Although they were not implemented in the current iteration of the project, they should be considered high-priority tasks for future development.

VIII. CONCLUSION

Through rigorous exploration and analysis, I have illuminated the intricate relationship between various demographic factors and income levels. The generated visualizations, tailored for specific stakeholders at UVW College, provide crucial strategic insights. By unraveling the complex influences that shape income, the college can hone their strategies and initiatives, thereby achieving their objectives more effectively. The insights gained have profound implications for marketing strategies, academic program planning, admissions, and diversity initiatives, helping UVW College create a more vibrant, diverse, and successful student body.

REFERENCES

[1] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," in Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007. [Online]. Available: https://matplotlib.org/

[2] J Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010). [Online]. Available: https://pandas.pydata.org/

[3] Numpy and Scipy Documentation, [Online]. Available: https://docs.scipy.org/doc/

[4] M. Waskom et al., "Seaborn: v0.5.0 (November 2014)", [Online]. Available: https://seaborn.pydata.org/

[5] Plotly Technologies Inc. "Collaborative data science." Montréal, QC, 2015. [Online]. Available: https://plotly.com/python

SUPPLEMENTARY MATERIAL: VISUALIZATION FIGURES

User Story #1: Age vs. Income (Violin Plot)

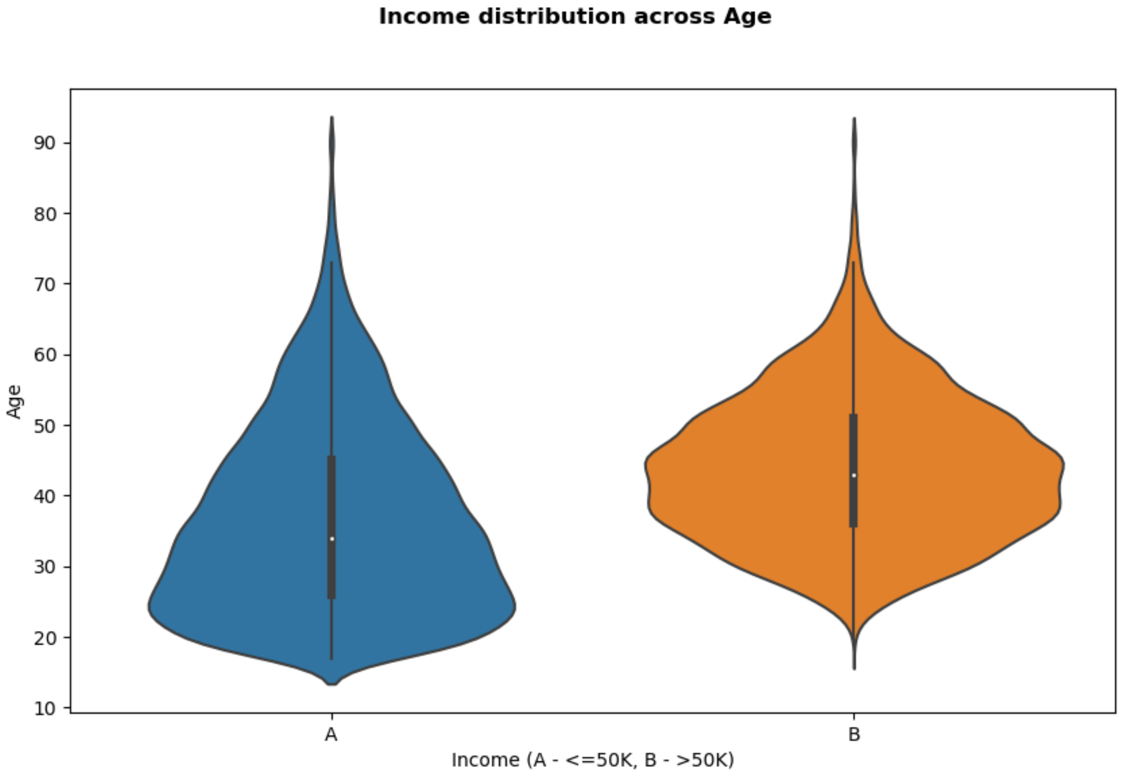

Figure 6 Enlarged version of Figure 1 Violin Plot

User Story #2: Marital Status & Relationship vs. Income (Parallel Categories Diagrams)

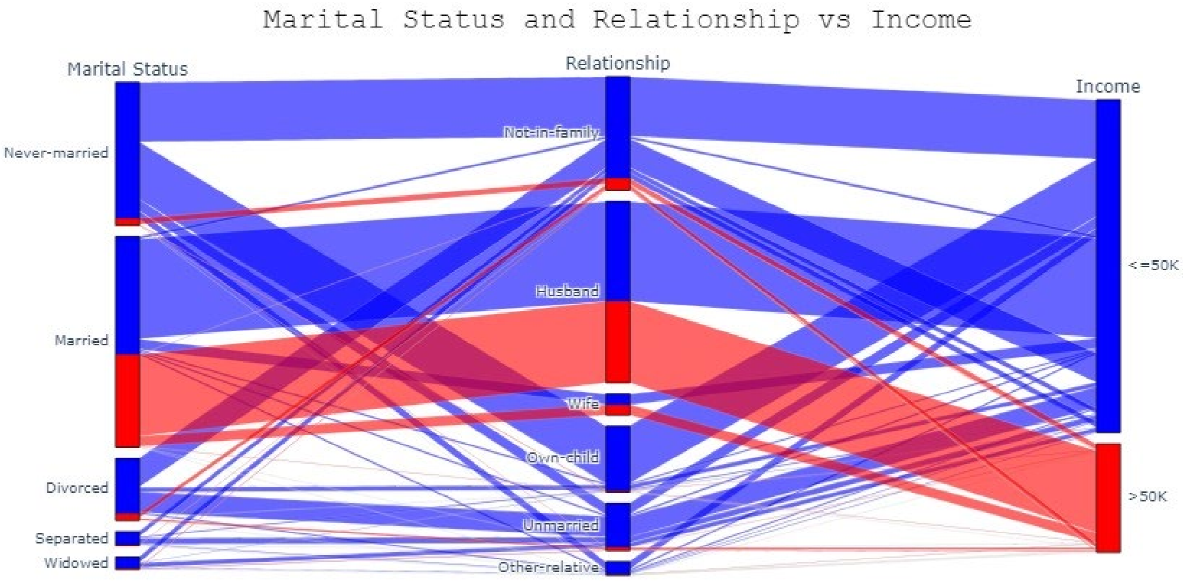

Figure 7 Enlarged Version of Figure 2 Parallel Categories Diagrams plot

User Story #3: Occupation & Work Class vs. Income (Heatmap)

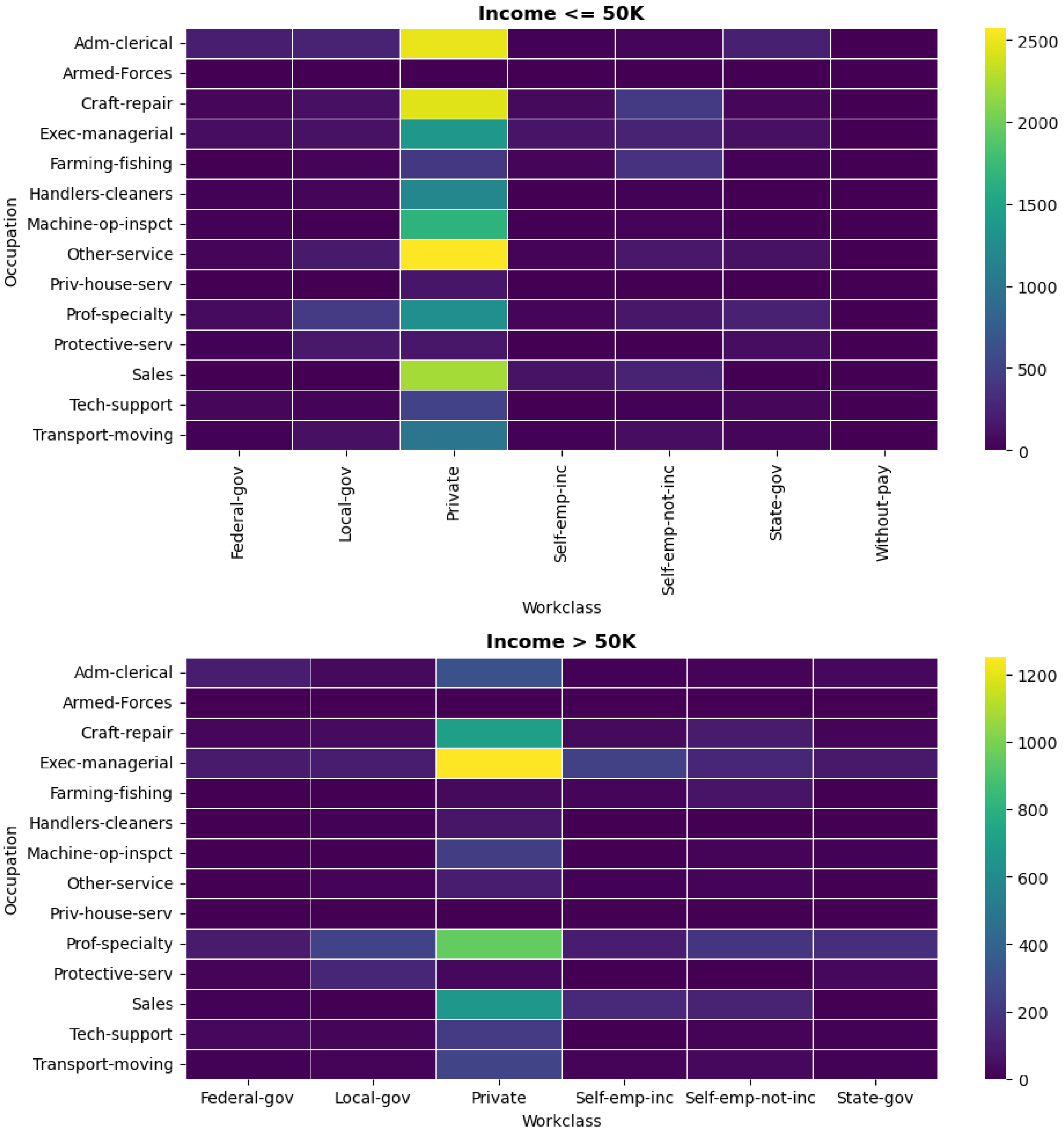

Figure 8 Enlarged Version of Figure 3 Heatmap

User Story #4: Race & Gender vs. Income (Stacked Bar Chart)

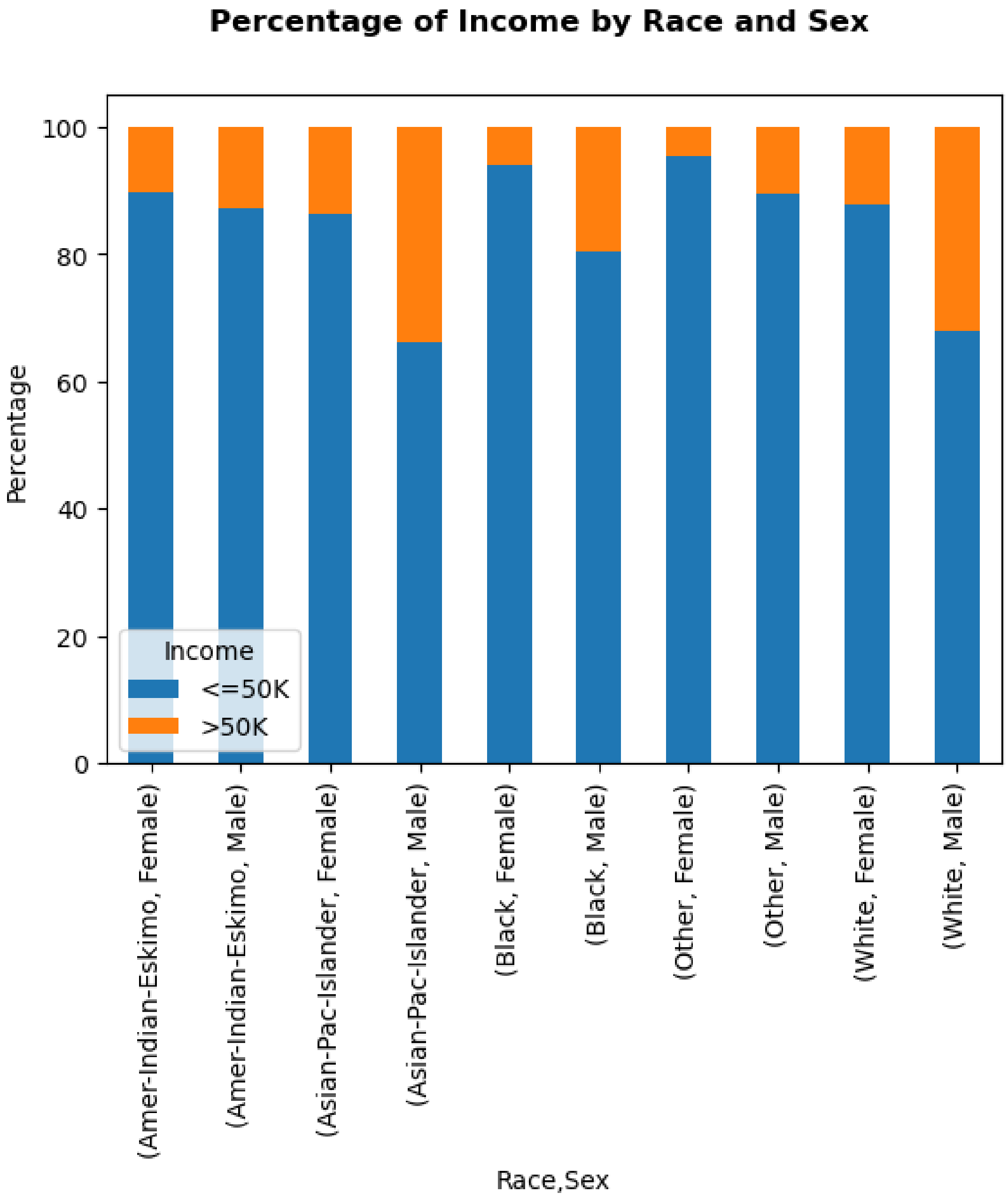

Figure 9 Enlarged Version of Figure 4 Stacked Bar Chart

User Story #5: Education vs. Income (Donut Charts)

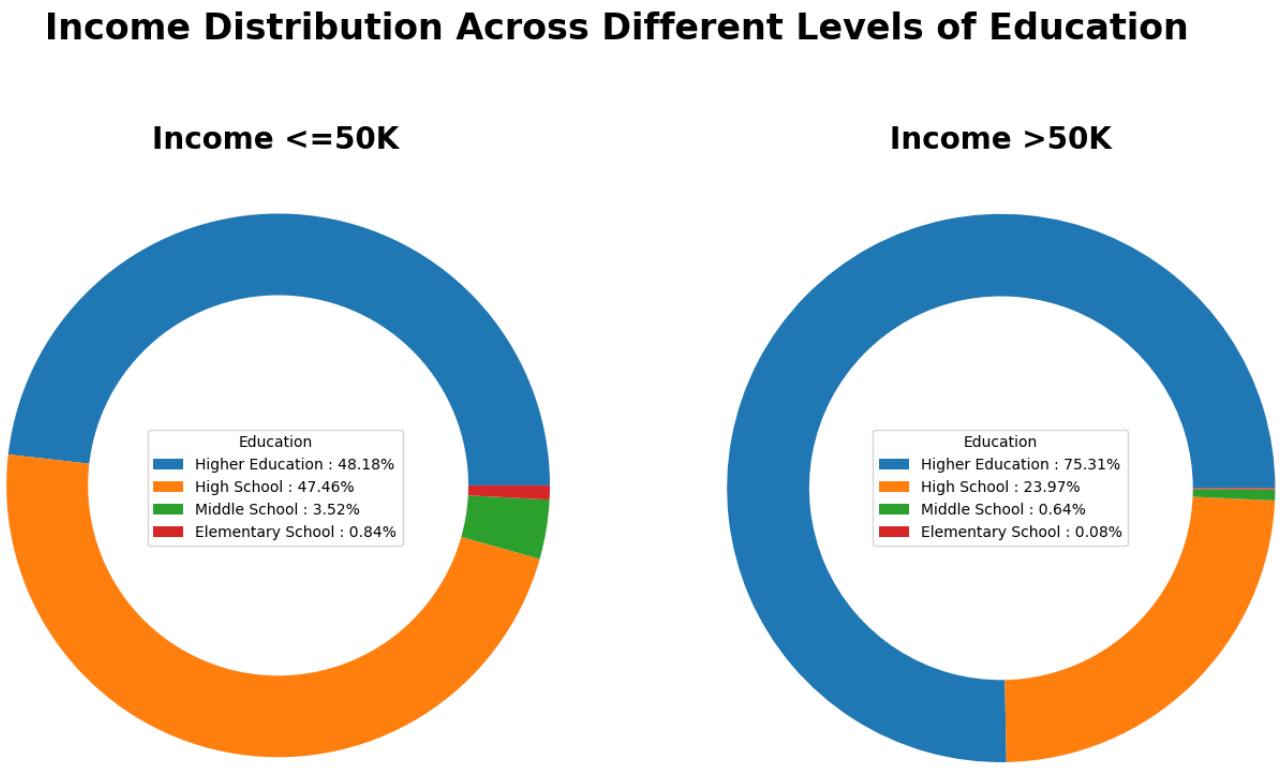

Figure 10 Enlarged Version of Figure 5 Donut Charts