CSE 578: Data Visualization

Course Project Progress Report

Goals and Business Objective

Goals

- Define the most important findings from the analysis conducted.

- Provide actionable insights to support the customer’s marketing efforts effectively.

Business Objective

- Enable the development team to build and maintain an application based on the findings and insights derived from this report.

---

Assumptions

- The data provided is accurate, clean, and relevant to the questions posed.

- Stakeholders understand the basic principles of the visualizations used.

- The target audience for the visualizations is the marketing team, who will rely on the insights for decision-making.

- All visualization tools and techniques align with the course’s design principles.

- Implementation of future features by the development team will depend on this report’s insights.

---

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.

User Story #2

- As a marketing director, I want to understand the combined impact of Age, Hours Per Week, and Capital Gain on income.

User Story #3

- As a data analyst, I want to see income variations across different education levels and work classes.

User Story #4

- As a member of the UVW team, I want to identify which factors most strongly correlate with high-income individuals.

User Story #5

- As a manager, I want visualizations that clarify which factors predict low-income categories effectively.

---

Visualizations

Visualization 1

- User Story: User Story #1

- Type: Univariate Visualization

- Description: Bar chart comparing age groups against income categories.

- Design Process:

- Step 1: Data grouping by age brackets.

- Step 2: Income labels segmented for visualization.

- Step 3: Bar chart formatted with color-coded income categories.

- Conclusion: Older individuals tend to fall into higher-income brackets more frequently than younger individuals.

Visualization 2

- User Story: User Story #2

- Type: Multivariate Visualization

- Description: Scatter plot matrix of Age, Hours Per Week, and Capital Gain vs. income.

- Design Process:

- Step 1: Filter dataset for the selected variables.

- Step 2: Create scatter plots to identify trends.

- Step 3: Add regression lines for each pair of variables.

- Conclusion: Higher values in Hours Per Week and Capital Gain positively correlate with higher-income categories.

Visualization 3

- User Story: User Story #3

- Type: Multivariate Visualization

- Description: Stacked bar chart of education levels by work class vs. income.

- Design Process:

- Step 1: Aggregate data by education level and work class.

- Step 2: Color-code income categories for stacked bars.

- Step 3: Label axes and include legends for clarity.

- Conclusion: Individuals with higher education levels and white-collar work classes are more likely to be in the high-income category.

Visualization 4

- User Story: User Story #4

- Type: Multivariate Visualization

- Description: Heatmap showing correlation coefficients of all attributes with income.

- Design Process:

- Step 1: Calculate correlation matrix.

- Step 2: Plot heatmap with a clear color gradient.

- Step 3: Highlight significant correlations for emphasis.

- Conclusion: Capital Gain and Education are among the most significant predictors of income.

Visualization 5

- User Story: User Story #5

- Type: Multivariate Visualization

- Description: Box plot comparing occupation types and income categories.

- Design Process:

- Step 1: Group data by occupation type.

- Step 2: Plot income categories for each occupation.

- Step 3: Include outliers and median lines for clarity.

- Conclusion: Certain occupation types, such as executive roles, are strongly associated with higher incomes.

---

Questions

Questions During Project Progression

1. Which attributes should be prioritized for analysis?

2. How do multivariate relationships vary across different income brackets?

3. What level of data cleaning is required for accurate visualization?

4. Which visualization tools best align with the project goals?

5. How can we ensure all insights are actionable?

Solutions

1. Attributes were prioritized based on the problem statement and relevance to income.

2. Pairwise comparisons and correlation matrices were used for multivariate analysis.

3. Automated scripts ensured data cleaning for missing or inconsistent entries.

4. Tools like Python’s Matplotlib and Seaborn were chosen for flexibility and design.

5. Actionable insights were emphasized through targeted visualizations.

---

Not Doing

- Developing a prediction model for income classification.

- Evaluating the feasibility of recommendations.

- Building end-user applications.

- Exploring attributes outside the provided dataset.

---

Python Code

Submit the Python code as a separate PDF with appropriate documentation and comments for clarity.