Individual Portfolio Report

For CSE 578: Data Visualization

Hannah O. Ajoge
hajoge@asu.edu and ohuajo@gmail.com

I. Introduction

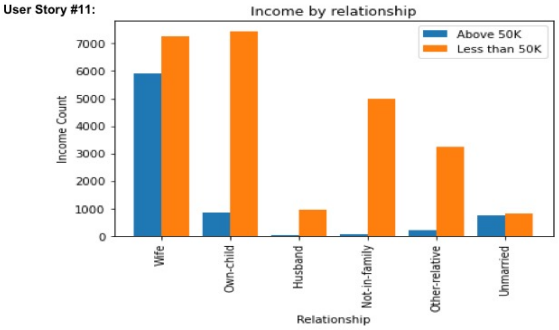
Data visualization is critical for exploring and communicating findings from diverse data types. Thus researchers have to organize and describe the design space of data visualizations strategically to inform design choices [1]. This strategic construction and analysis of visualization design space entail in-depth analysis of literature visualizations[1].

Thus, in this project, we used analysis of literature and visualization to identify six top attributes that are suitable for predicting income. These attributes are aimed at being a useful strategy for marketing and increasing enrollment for a hypothetical UVW College. In this team project, the stakeholders were thus the UVW college executives, to whom the project outcome was to be presented. The executives are the stakeholders because the executives have legitimate claims on the college, are affected, and can be affected by the college's objectives [2]. In an Agile team project, there are usually different roles like product owner or scrum master. However, our team adopted the Kanban software development strategy, this is more flexible than Scrum and thus no need for such roles [3]. It was then a joint responsibility by the team members to maximize efficiency and reduce the time taken to complete tasks. It was thus my responsibility and privilege to oversee the project activities as an Agile team facilitator.

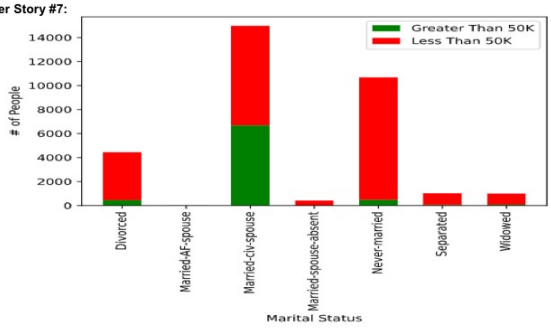
II. Description Of Solution

Our team communicated the importance of our selected attributes in a visually appealing way to the executive team. This was done by first numbered our variables in alphabetical order. Then we visualization the 13 user stories, viz: Age, Capital Gain-Loss, Total Education, Education-num, FNLWGT, Hours-per-week, Marital Status, Native Country, Occupation, Race, Relationship, Sex, and Working Class as user stories 1 to 13, respectively. Finally, we utilized the correlation coefficients to identify our six top user stories (attributes) in decreasing priority to be Marital Status, Relationship, Education, Age, Hours-per-week, and Sex. Below are our top six visualizations and the correlation matrix used for deciding the selection. Our code is available on https://github.com/ohuajo/Falcons_Income_Estimator.

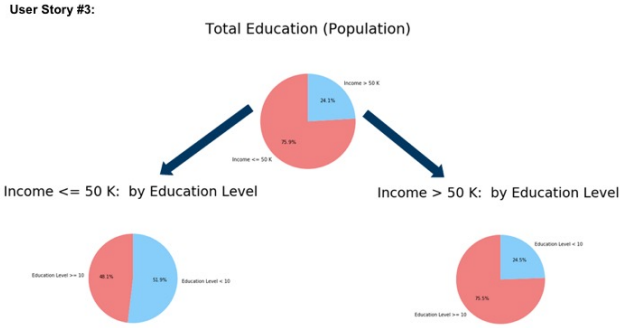
User Story #11:



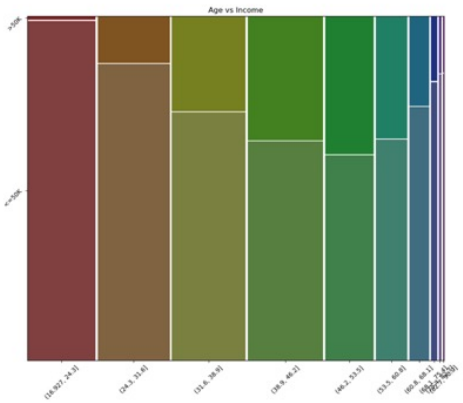
User Story #7:



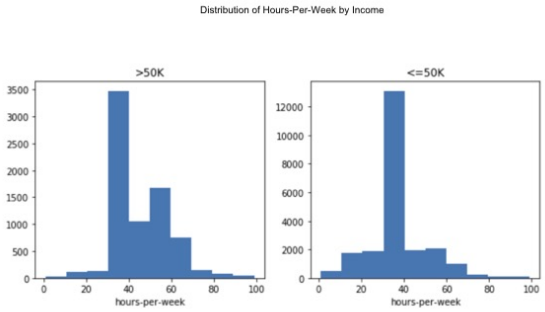
User Story #3:



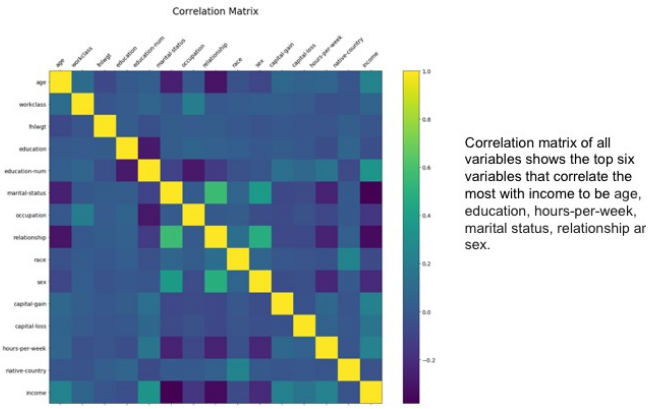
User Story #1:



User Story #6:



User Story #12:



Correlation matrix of all variables shows the top six variables that correlate the most with income to be age, education, hours-per-week, marital status, relationship ar sex.

III. Results

The visualizations above showed how the six top chosen attributes relate to the income of a group; and how the attributes will steer UVW's marketing strategies. Below are the interesting findings.

User Story 7 (visualization by Intzar Singh)

Marital status was selected as the top attribute because it had the strongest correlation with income. This attributed was visualized using a stacked bar chart. The chart shows what proportion of the people in each marital status category make more or less than the $50,000 threshold. The only group that shows a significant portion making more than $50,000 is the group of people who are married. This would make a compelling choice for further evaluation of how significant the difference in someone's marital status could be on their income level.

User Story 11 (visualization by Gad Asare)

Relationship had the second-strongest correlation with income. This attribute was visualized with grouped bar charts. From the chart, it could be concluded that wife and own-child were the most people who earned less than $50,000. However, wife also was the only group with a significant number of people making above $50,000.

User Story 3 (visualization by Pierre LeBlanc)

Total education was regarded as the third top attribute. This was visualized with pie charts. A pie chart was used to show that 75% of individuals earn less than $50,000. A pie charts of the two different income levels showed that having higher education (> level 10) increases your chance of earning above $50,000.

User Story 1(visualization by Michael Salzarulo)

Age ranked fourth among the top user stories. A mosaic plot showed that after age of the 46 the width of the bars rapidly deteriorates. The older the people the more likely to earn $50,000. Further analysis of age is warranted, especially with a third parameter to identify a deeper relationship to individual income.

User Story 6 (visualization by Jiteng Xu)

Hours-per-week ranked as the fifth most important variable. Histograms showed that the people who worked over 40 hours were more likely to earn above $50,000.

User Story 11 (visualization by Hannah Ajoge)

Sex ranked as the sixth top variable. Here I used a sunburst chart to show that males dominate both income groups. Males were more likely to earn above $50,000 (p= 0.0 by chi-square). Although this was the least correlated among the top six, it will make a good marketing strategy to market to both males and females.

As a result of these visualizations, we concluded that:

- Our visualizations have identified six top variables that strongly correlate with income.
- We showed that being married, being a wife, possessing a college degree, being younger than 46, working more hours, and being a male is associated with earning more than $50,000.

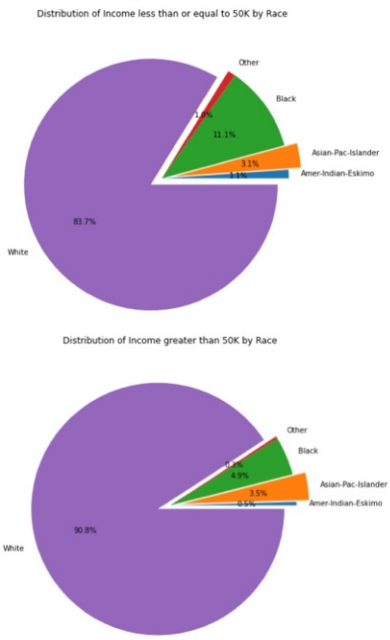
- Further analyzes of these top attributes with more complex analysis like k-means with L2-normalization, principal component analysis or an autoencoder will be necessary before implementation as part of UVW next marketing strategy.

IV. My Contribution

I attended all meetings; and I actively took part in brainstorming, strategizing and discussing the projects. I started the team slack channel, hosted the zoom meetings, co-chaired meetings, and functioned as an Agile team facilitator. I suggested our team name of Falcons because of my vision of speedy delivery. As an Agile team facilitator, my vision as a coach was to ensure that my team is high performing. From the agile coach's point of view, high-performing teams exhibit two qualities, delivery and continuous improvement [4]. Thus my focus was to deliver tasks in speedily and with continuous improvement. I was able to achieve this through enabling leadership, by enhancing the quality of interactions amongst team members.

In addition to my team-building activities, I initiated and drafted a significant portion of the system documentation and executive report. Out of the 13 visualizations produced by my team, I specifically assessed race (pie charts), sex (sunburst plot), and native country (choropleth maps) as factors that can be used to predict income. Among the three attributes I assessed, sex made it to be one of the top six attributes identified by the team. In addition to Sex, which I have already explained above, below are the other two visualizations I implemented and the interesting findings.

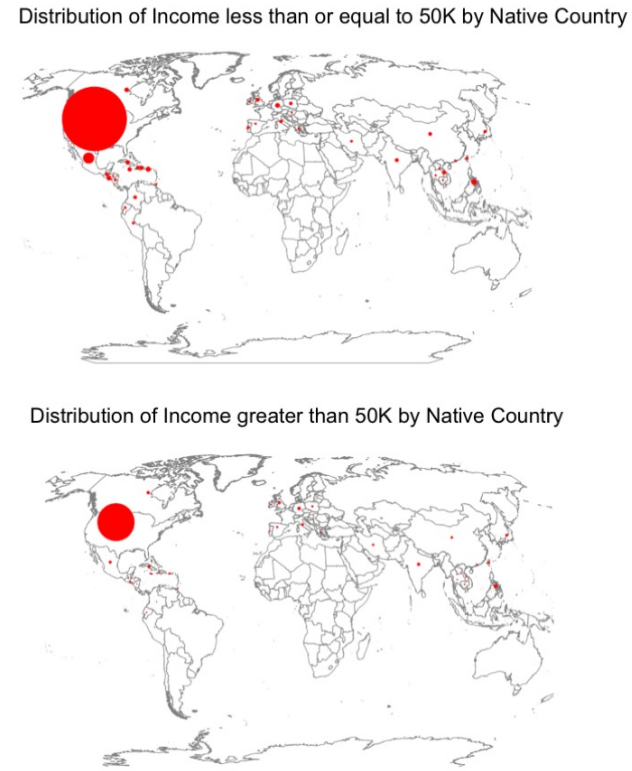
User Story #10:



User Story 10 (visualization by Hannah Ajoge)

Race was the next significant attribute among the three that I visualized. I used pie charts with exploded slices to show that the percentage of Whites and Asian-Pac-Islanders increases significantly from earning less than or equal to $50,000 to earning more than $50,000. This was found to be significant (p=2.30e-70 by chi-square). However, this was not a top priority because the significance could not be supported by the strength of the relationship with income.

User Story #8:



User Story 8 (visualization by Hannah Ajoge)

Here I used choropleth maps to show that the native country of the people who participated in the survey was predominantly the United States. The same pattern of native countries could be seen in the two groups of incomes. Most of the people came from North America, South America, Europa, and Asia. Participants did not come from Africa nor Australia. Though the choropleth map was aesthetically significant, the lack of a strong relationship with the income disqualified the attribute of being among the top six attributes considered to be important to the business objective.

As a result of my three visualizations, I concluded that :

- My visualizations have contributed to the identification of one of the six top variables that strongly correlate with income.

- I showed that being a male, being a white person, and being native to the United States is associated with earning more than $50,000.
- Further analyzes of these three variables and the five other top attributes with more complex analysis like k-means with L2-normalization, principal component analysis or an autoencoder will be necessary before implementation as part of UVW next marketing strategy.

V. LESSONS LEARNED

The new expertise I obtained through this project include the following:

1. I learned visualization methods like choropleth maps, exploded pie charts, treemaps, sunburst charts, and others I wasn't aware of previously.
2. I have also learned how to use Jupyter notebook to execute these visualizations. I also realized that the Python skills I have acquired in the course were also useful in another course I was taking at the same time.
3. Being an Agile facilitator in a team project setting. Although, managing people challenges is more of an art than a science; it is still a critical part of software development in a team setting [5]. The sources of team problems could be the organization, the project, the team, or an individual. An example of such problems includes decision-making [5]. Thus it is very important that a competent Agile coach/facilitator or Scrum master ensures effective team decision-making. In my role as a facilitator, I came across situations in which we had to make timely decisions. I usually restore to voting on the decision. To me, my most important experiential learning came from my leadership role. I made a deliberate effort to enhance my leadership skill and also to seek feedback from teammates individually (one on one) and as a group. I have also suggested connecting through LinkedIn and have been able to connect with two teammates. I intend to continue to collaborate with them and to consciously try to get feedback that can enhance my leadership skill.

My most important learning, as I have noted above, comes from my role as an Agile facilitator. I was excited to have worked with a group of five other teammates who were responsive, responsible, and cooperative. These great gentle people who made our teamwork fruitful in alphabetical order of first names are Gad Asare, Intzar Singh, Jiteng Xu, Michael Salzarulo, and Pierre LeBlanc

REFERENCES

[1] A. Crisan, J. L. Gardy and T. Munzner. "A Systematic Method for Surveying Data Visualizations and a Resulting Genomic Epidemiology Visualization Typology: GEViT." Bioinformatics 35.10 (2019): 1668–1676.

[2] M. H. Yip, and T. Juhola. "Stakeholder Involvement in Software System Development – Insights into the Influence of Product-Service Ratio." Technology in society 43 (2015): 105–114.

[3] E. L.-C. Law and M. K. Lárusdóttir. "Whose Experience Do We Care About? Analysis of the Fitness of Scrum and Kanban to User Experience." International journal of human-computer interaction 31.9 (2015): 584–602.

[4] G. Bäcklander. "Doing Complexity Leadership Theory: How Agile Coaches at Spotify Practise Enabling Leadership." Creativity and innovation management 28.1 (2019): 42–60.

[5] K. Conboy, S. Coyle, X. Wang and M. Pikkarainen People over Process: Key Challenges in Agile Development. IEEE software. [Online] 28.4 (2011): 48–57. Minna