Project Milestone 4 - Systems Documentation Report

Team 4 (The Falcons)

Gad Asare, Hannah Ajoge, Intzar Singh, Jiteng Xu, Michael Salzarulo and Pierre LeBlanc

Roles and responsibilities

The goal of this project is to increase enrollment for the hypothetical UVW College. The responsibility of our team is to identify factors (attributes) that could be used to predict income and present those factors to UVW executives using appropriate visualizations. Stakeholders can be defined as those who have legitimate claims on a company, can affect or are affected by the company's objectives [1]. In the context of this project, we can thus say that the stakeholders are the UVW executives. Since UVW executives are hypothetical figures, we can view the course/project staff as the stakeholders. This is because the staff can affect the project objectives and are responsible for objectives and rubrics of this project. Our team adopted the Kanban strategy of software development which is quite flexible in comparison to Scrum [2]. Because we adopted Kanban, there was no requirement for roles like product owner or scrum master. The strategy was thus to focus on maximizing efficiency by reducing the time taken to complete a project or user story. As it is typical of Kanban Agile practice, it was the responsibility of the entire Falcon team to collaborate on and deliver tasks.

Our first meeting on the 18th March 2021 resulted in adopting a team name (Falcons) and the following tentative checkpoints: (1) Every Saturday at 11 am we would have an "office hour" over zoom to discuss directions and progress. The office hour was an avenue to get things done and also an opportunity for team building and friendship. (2) We agreed to look at the projects individually and come up with directions by Tuesday (23rd of March 2021). We also agreed to get back via this slack channel as we came up with ideas. (3) Tentatively we agree to looking at hybrid ways of executing the tasks once we have direction. The hybrid way will be to have some people who will be sub-teams work together real time (synchronously) and others work individually on tasks asynchronously. Finally we adopted the asynchronous method. Tasks were not imposed but based on choice and capability.

Subsequent meetings were spent understanding the requirements and executing tasks. All team members attended meetings and actively participated in brainstorming, discussions and

strategizing. All team members participated in the visualization exercise by coding and presenting a minimum of two attributes. All codes were in Python 3, please see appendix for links to code. All members participated in writing/editing the system documentation and executive report. Below are the other roles peculiar to individual team members:

Gad Asare: Assessed occupation and relationship as factors that can be used to predict income. Gad presented a grouped horizontal bar chart and a regular grouped bar chart respectively to show that the factors could be used to predict income. Additionally Gad initiated a google doc and google slide for writing the system documentation and executive reports respectively.

Hannah Ajoge: Assessed race, sex and native country as factors that can be used to predict income. Hannah presented pie charts, sunburst plot and choropleth maps respectively to show that the factors could be used to predict income. Additionally, Hannah initiated the team slack channel, hosted the zoom meetings, co-chaired meeting and functioned as Agile team facilitator.

Intzar Singh: Assessed education-num and marital status as factors that can be used to predict income. Intzar presented box plot and stacked bar charts respectively to show that the factors could be used to predict income. Additionally Intzar spurred the team to action by ensuring that all team members were made aware of being part of team 4. He also co-chaired meetings.

Jiteng Xu: Assessed hours-per-week and fnlwgt as factors that can be used to predict income. Jiteng presented a box and whisker plot and histograms respectively to show that the factors could be used to predict income.

Michael Salzarulo: Assessed age and workclass as factors that can be used to predict income. Michael presented histograms and mosaics respectively to show that the factors could be used to predict income. Additionally Michael initiated and managed a github repository for the team coding collaboration and provided a chart on the distribution of income, grace our cover slide in the executive report.

Pierre LeBlanc: Assessed capital-gain, capital-loss and education as factors that can be used to predict income. Pierre presented scatter plots, box plots, and pie charts respectively to show that the factors could be used to predict income. Additionally Pierre provided a correlation plot

that aided in identifying the top 6 factors. This correlation plot is shown on the cover page and within the “Backup” slides in the executive report. He also co-chaired meetings.

Team goals and a business objective

The most important goals in our project were to make sure that we could communicate the importance of our selected variables in a manner that would be visually appealing to the executive team. With the amount of data that we had in tabular form, it would be hard to show the key insights without making use of a variety of the visualizations that we have shown. For example, it would be hard to see the distribution of education level between the two income groups without the use of a box plot. Another example is the choropleth map that highlights how the data differs between regions, the analysis of such data without this map would be impractical and inefficient.

Furthermore, we also wanted to emphasize which variables were the most relevant for further exploration. With so many different data columns, it would be easy to get lost trying to make sense of the relationship they have to the income of a group. However, by zoning in on a few key variables, we wanted to show which topics would be the most relevant for our study. In a world with limited resources, it is essential to make sure that we focus on the most important attributes of our data first and then expand as needed. Our visualizations show clearly how the variables chosen relate to the income of a group. The outcome of our project has fulfilled the given business objective of identifying the six top attributes that will steer UVW’s marketing strategies.

Assumptions

Assumption 1: All visuals were done within Python 3, assumption is that everyone has access to freely download Python.

Assumption 2: Data provided by “UVW College” is accurate and is the source of truth for this report. We did not check for erroneous patterns in the data.

Assumption 3: We are assuming the $50,000 income target is face value and did not research the area to check if possibly there might be a better income target to achieve for “UVW College”.

User Stories

The aim of this project is to identify attributes that will increase enrollment in UVW College (measured by change in income status). First, we numbered our variables in alphabetical order.

Then, we utilized the correlation coefficients (and fair representation of efforts) to identify our six top variables that will predict income level. Below are the top six user stories we prioritized, in decreasing order of importance. They are can be identified on our user stories slides which is attached in appendix.

User story 7: Marital Status

User story 11: Relationship

User story 3: Education

User story 1: Age

User story 6: Hours-per-week

User story 12: Sex

Visualizations

Our visualizations are available in the appendix. Below are each visualization, their significance and importance to the business objectives.

Visualization for User Story 1 (Age):
Based on the correlation matrix featured on the bottom left of the title slide; age was identified as one of the top 6 most positively correlated parameters to an individual's income. Further analysis is illustrated in the mosaic plot. Here we can see that income distribution is skewed to the left with respect to individuals who make greater than 50k per year while the converse is true for individuals who make less than or equal to 50k per year. We can observe the aforementioned positive correlation as age increases a greater number of individuals have an income over the 50k threshold. The last take away from this plot is that the number of individuals older than 46 rapidly decreases with respect to the rest of the data set; however there is a greater proportion of these individuals who have an income over the 50k threshold.

Visualization for User Story 2 (Capital Gain-Loss):
The box plots show the data spread of each data column within the capital-gain, loss fields. It is difficult to see any correlation here other than the capital gain for people making > 50K is substantially higher compared to people making < 50K. This suggests the capital-gain is generally greater for those people who make over 50K. Hence it would make more sense to market to that population that has higher capital-gains on average and within the Income > 50K group. The reason behind choosing the higher capital-gain group of greater than 50K is because this group generally appears to have more cash on hand as they have higher capital-gains. There is a similar pattern for capital-loss between the two different groups of income. The population who makes above 50K has a slightly higher capital-loss compared to those who make less than 50K.

Visualization for User Story 3 (Total Education):
The pie charts demonstrate that for people making over 50K, the majority have a higher education level (> Level 10), which shows partial or full college degrees. This is clearly shown with the pie chart on the right where ~3 out of 4 in the population who have higher education are in the group who make above 50K. This suggests the higher the education the higher the income level. Also, those who have a higher education (college degrees) have a higher probability to be within the greater than 50K group. This analysis is drastically different when looking at the pie chart on the left showing the Income <=50K group. Within the left pie chart the breakdown is almost evenly split. This suggests for the group making less than 50K education has less of an influence on income level. There are slightly more people within the lower education level compared to higher education level within the left pie chart. However, that difference does not add value to this analysis.

Visualization for User Story 4 (Education-num):
The box plot shows that the median education level of people who make more than $50,000 per year is much higher than the median for the group of people who make less than $50,000 per year. However what is more notable is that the first quartile value of the higher income group is higher than the median of the lower income group. I feel that this would be a great variable to explore further since there seems to be a clear distinction in the education levels of the two income groups.

Visualization for User Story 5 (FNLWGT):
For both classes of data fnlwgt, their distribution is similar as it is evident from the box and whisker plot. This feature may not help in distinguishing between the two classes of data.

Visualization for User Story 6 (Hours-per-week):
The histogram shows the most people working more than 40 hours a week belong to the ">50K" category. Also, in the >50k group it has less people who work less than 30 hours a week. This shows that people who spend more time working are likely to earn over $50,000. This could be good variable for future use in the analysis.

Visualization for User Story 7 (Marital Status):
This stacked bar chart shows what proportion of the people in each marital status category make more or less than the $50,000 threshold. As we can see from this chart, the only group that shows a significant portion making more than $50,000 is the group of people who are Married. I would also consider dropping the Married-AF-spouse column as there is not enough data to be represented on the same scale as the others. However, from the general data we can see that being, and staying married is a significant indicator of income. This would make a compelling choice for further evaluation as to how significant the difference of someone's marital status could be on their income level.

Visualization for User Story 8 (Native Country):
The choropleth maps shows 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. That is, 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 strong relationship with the income disqualified the attribute of being of importance to the business objective.

Visualization for User Story 9 (Occupation):
This grouped bar chart shows that among all occupations, Prof-specialty and Executive managers make up a chunk of the Above 50k category whilst other occupations such as Adm-clerial, Craft repair and other services contributed to majority of individuals with salary less than 50K

Visualization for User Story 10 (Race):
Pie Charts with exploded slices shows the percentage of Whites and Asian-Pac-Islanders increases significantly from earning less than or equal to 50K to earning more than 50K. This was found to be significant (p=2.30e-70 by chi square). However this was not a top priority because the significance coil not be supported by the strength of relationship with income.

Visualization for User Story 11 (Relationship):
This grouped bar chart shows income levels by relationship. From the chart, it can be concluded that wife and own-child make up the majority of the individuals making less than $50,000 with the 'Wife" being the only variable with a significant number of individuals making above $50,000. This attribute was selected as the second top attribute based on the strength of correlation [negative] with income.

Visualization for User Story 12 (Sex):
The sunburst plot shows that males are significantly more likely to earn more than 50K (p=0.0 by chi square). Males dominate both income brackets. While males significantly make up 85% (6662/7841) of people that earn greater than 50K, males only make up 61.2% (15128/24720) of people who make less than or equal to 50K. This analysis is the least correlated among the top six attributes. However it will make a good marketing strategy to both sex categories. To females, a motivation for crossing to a higher earning. To males, ambition to maintain the higher earning bracket.

Visualization for User Story 13 (Working Class):
This mosaic shows working class to income. At a glance we see the one category that is more likely to make >50k as self-emp-inc. Below the white break represents the distribution of the class that makes less than or equal to 50k, above the break represents the distribution of class that makes greats than 50k. Although there is a bias for self-emp-inc there is not much variation in the other working classes, we can see this by looking at the breaks where on average they nearly make a straight horizontal line. This means that there is not much value added when trying to predict income, therefore the working class parameter is not a good candidate for further analysis at this time.

Questions

Which Variables were the most relevant for our data?

This was one of the first questions that our team had. To properly answer this, we decided to make use of a correlation matrix to see which variables were the most strongly related to the income variable. This correlation matrix then helped guide our decision making process for the visualizations and the report presentation.

Not doing

Based on our top six visualizations we have identified key parameters that contribute to predicting the income of an individual. In the future we plan to use the parameters to develop a multivariate interactive user image. We have identified Tableau as a tool that can be used for rapid analysis where we can generate and publish interactive visuals. Although there is an overhead expense for using this tool the benefit of time saved will be greater than the cost of using the tool.

In addition to the creation of more telling visuals we also plan to continue the data exploration phase with some unsupervised learning methods. Namely using exploratory clustering such as k-means with an L2-Normalization to identify non-pictured relationships between variables. We will also use more complex algorithms in an attempt to segment the data. Complex algorithms such as support vector machines can help to identify thresholds between multivariate parameters and lead to a primitive prediction model.

Once sufficient data exploration is fulfilled the next step will be to identify a supervised learning classification algorithm that produces confident predictions of an individual's income. The supervised method will be iteratively derived from existing deep learning architectures starting with a single layer perceptron then expanding layers and types of layers. From there we will have some trade offs to measure such as computational complexity, runtime, predictive performance, and volatility. Each of these parameters will be weight and balanced in order to determine the best path forward.

The final phase will be to employ feature reduction algorithms in order to reduce input data, runtime, and computational complexity with the classification model in order to produce an efficient and reliable prediction model. Once again an iterative approach will be taken based on the top three performing models. We can reduce input dimensions with a Boltzman machine or reduce hidden layers with principal component analysis or an auto encoder.

Upon deployment of the software we intend to maintain performance by generating a feedback loop. Using feedback we can continuously monitor the software, search for failures, analyze anomalies, and ultimately improve performance. This strategy will also allow the team to adjust the model based on new and emerging trends such as the work from home culture as well as diverging trends like education becoming less correlated with income.

References

1. Yip, Man Hang, and Tomi Juhola. “Stakeholder Involvement in Software System Development – Insights into the Influence of Product-Service Ratio.” *Technology in society* 43 (2015): 105–114. Web.
2. Law, Effie Lai-Chong, and Marta Kristín 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. Web.

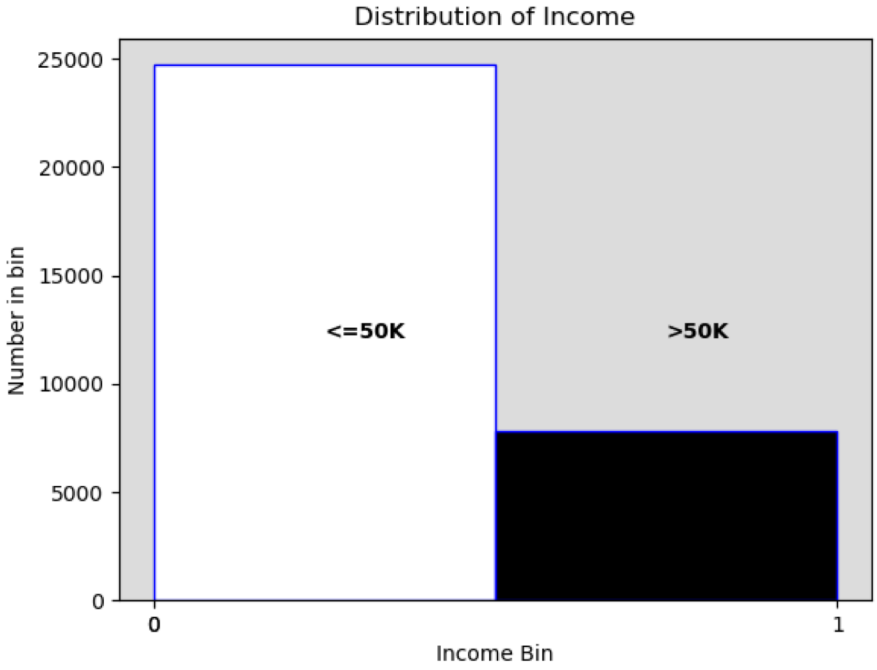
Appendix

Our code is attached with this system documentation. It is also available on GitHub and accessible with the following link:

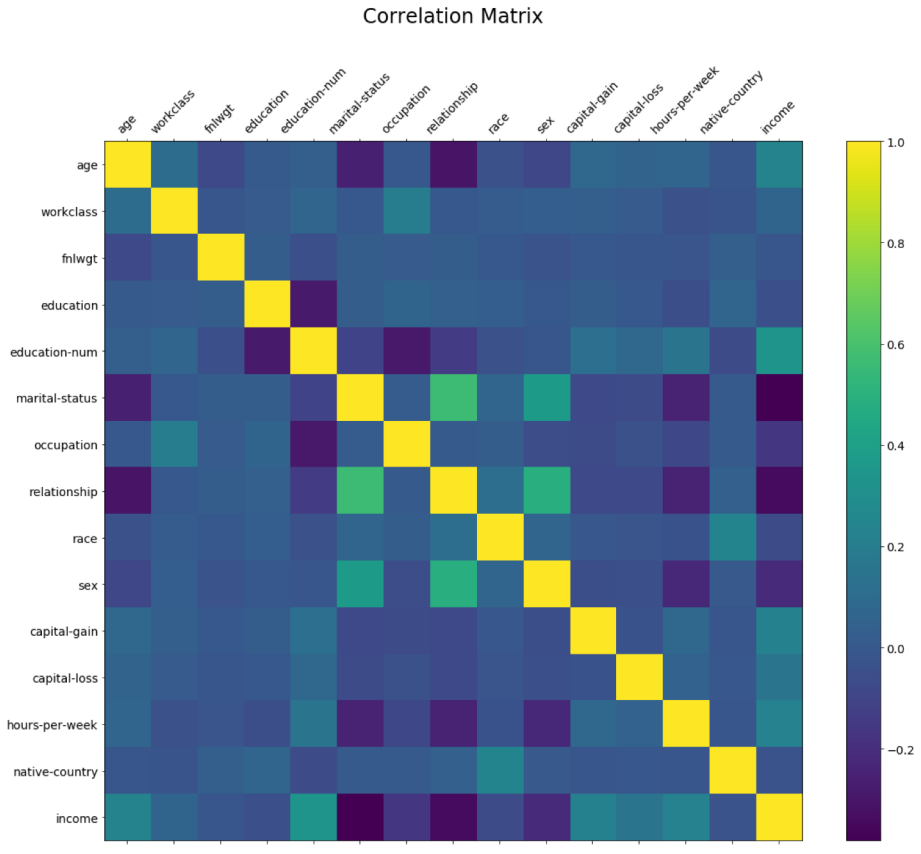
https://github.com/MichaelENGs/Falcons_Income_Estimator/tree/master/src

Below is a description of our Python Scripts in the repository.

	File Name:	Description:
1	Asare.py	Visualizations by Gad Asare
2	CSE578Project_JitengXu_Visuals.py	Visualizations by Jiteng Xu
3	HannahAjoge_Visualization.py	Visualizations by Hannah Ajoge
4	asuProject_capitalgainloss_visuals.py	Visualizations by Pierre LeBlanc (Pie Charts)
4	asuProject_capitalgainloss_visuals_scatter.py	Visualizations by Pierre LeBlanc (Scatter Plot)
4	asuProject_correlation_plot.py	Visualizations by Pierre LeBlanc (Correlation Plot)
5	salzarulo_visualize_data.py	Visualizations by Michael Salzarulo
6	visualizations_intzar.py	Visualizations by Intzar Singh

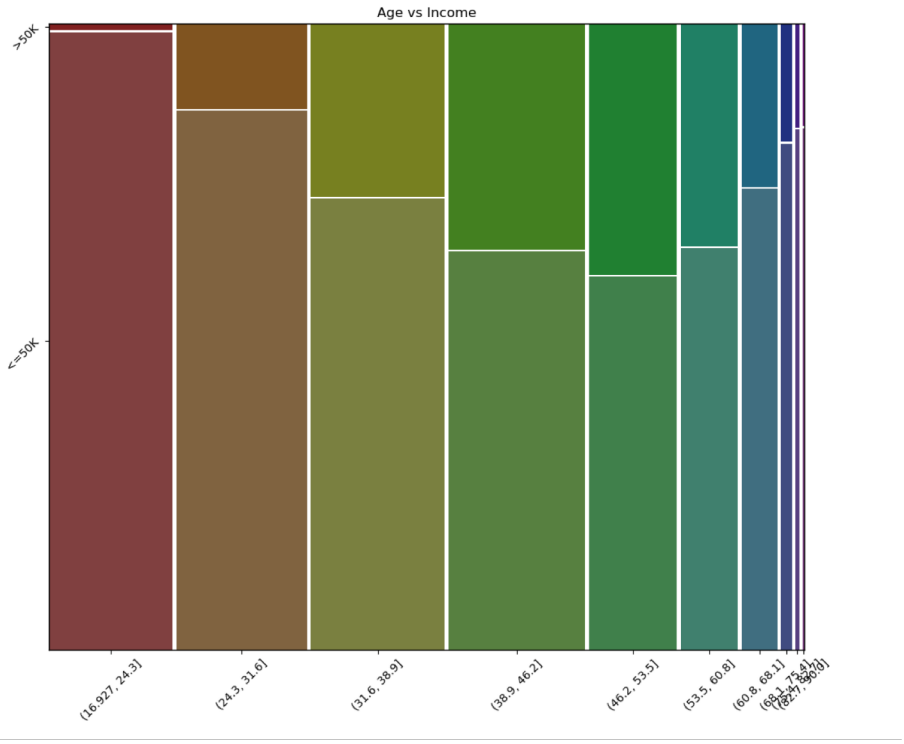


This histogram shows the distribution of Income for all samples in the data set. It's a clear indication that the majority of people make less than or equal to 50k.

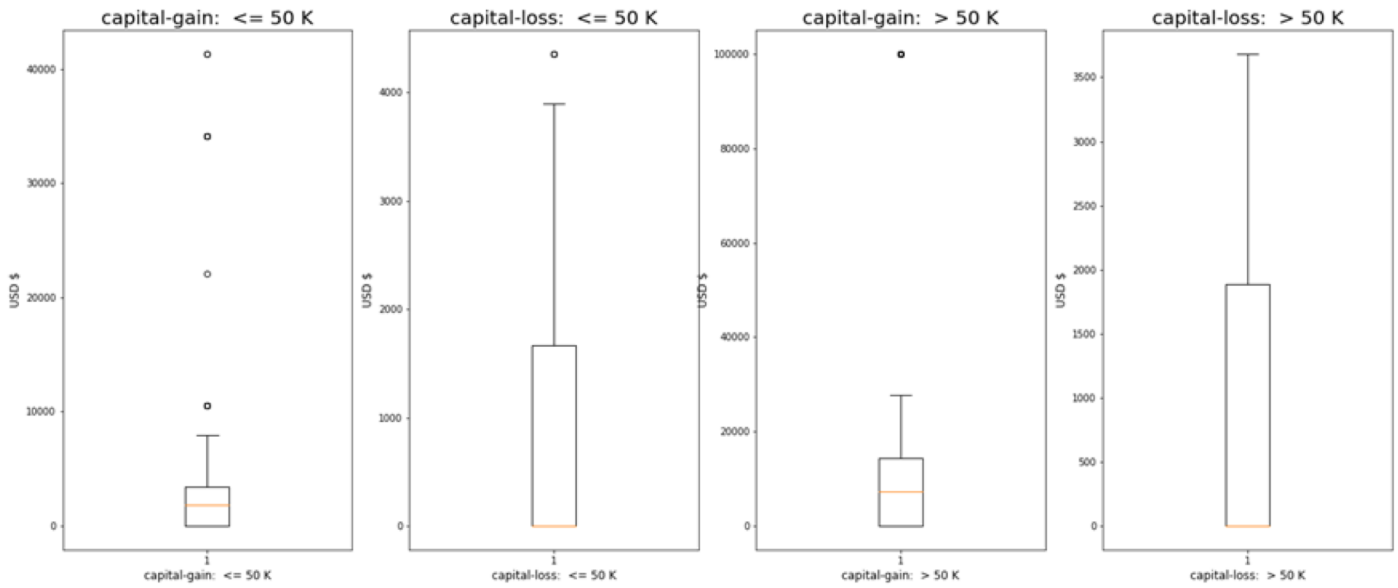


Correlation matrix of all variables shows the top six variables that correlate the most with income to be age, education, hours-per-week, marital status, relationship and sex.

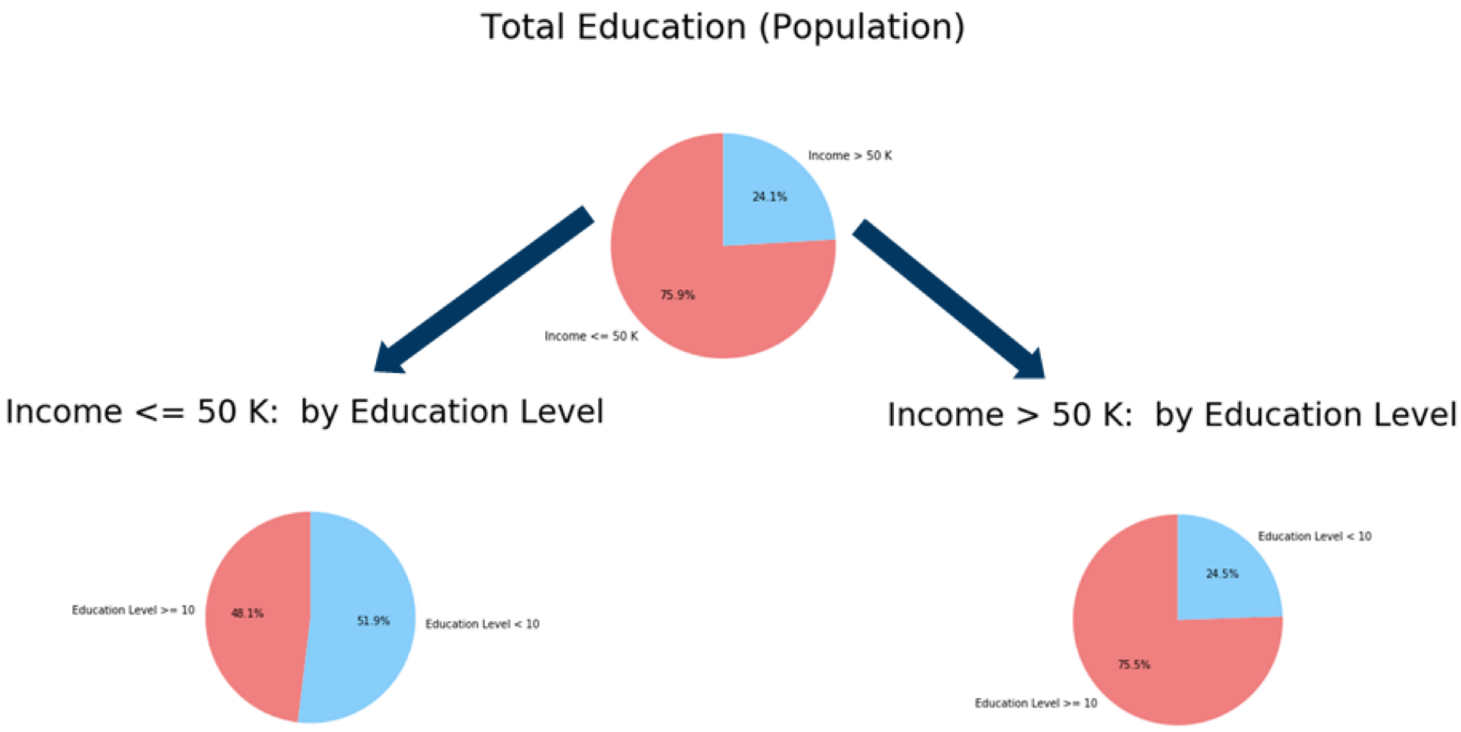
User Story #1:



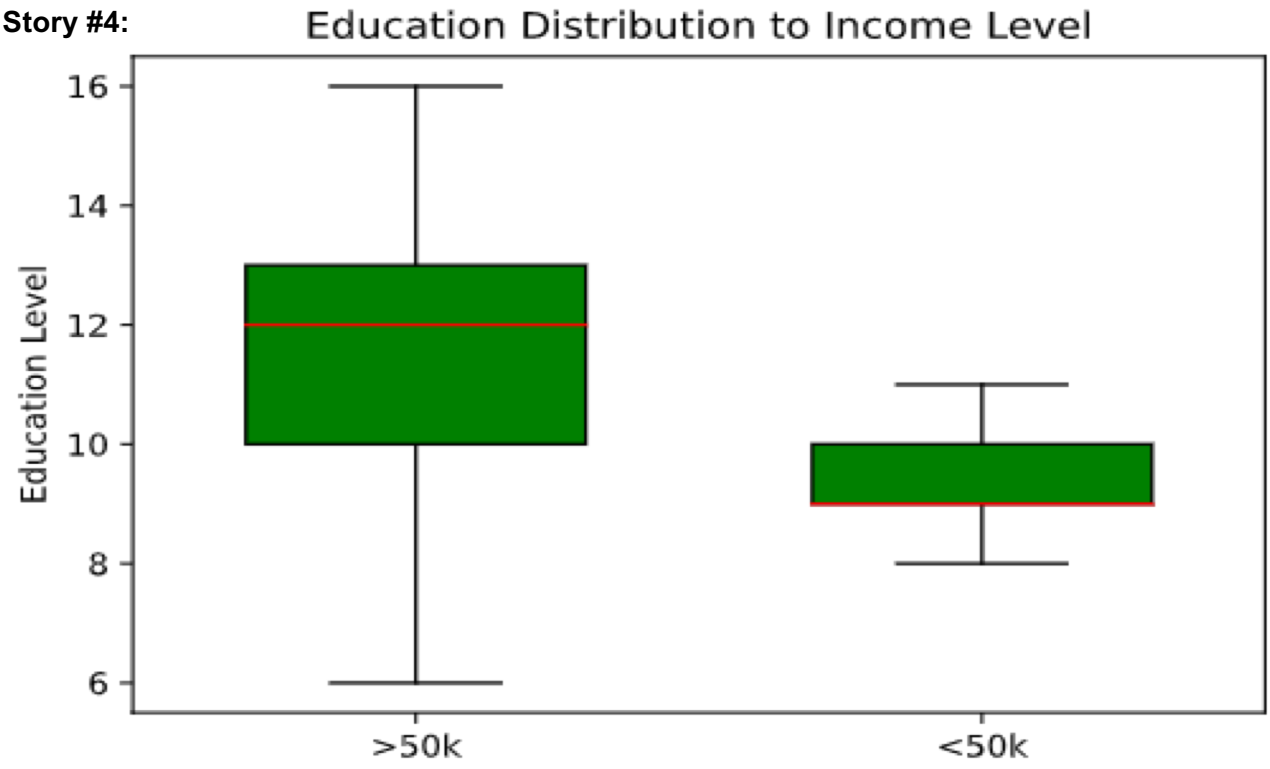
User Story #2:



User Story #3:

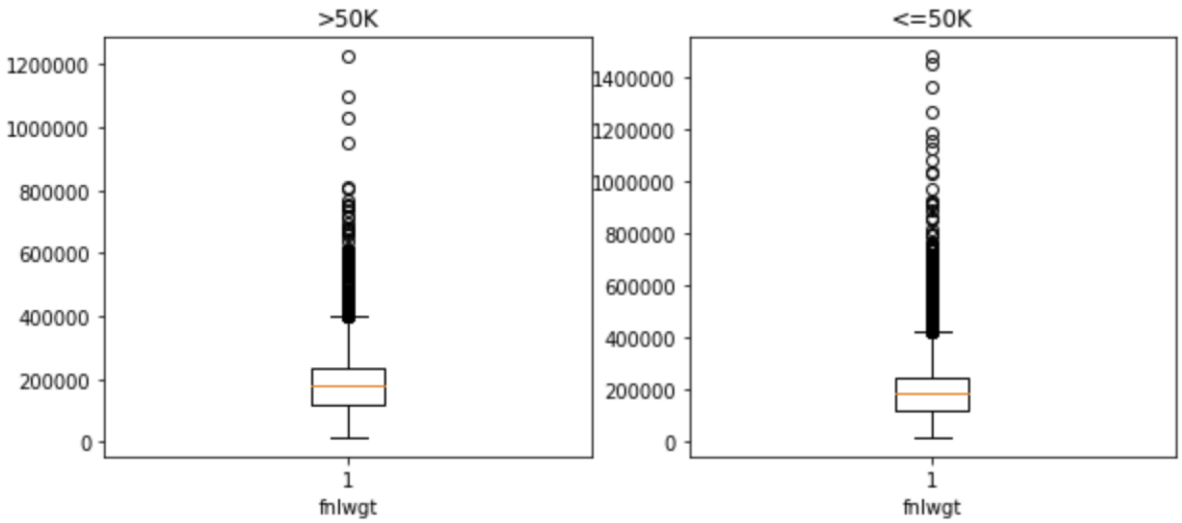


User Story #4:



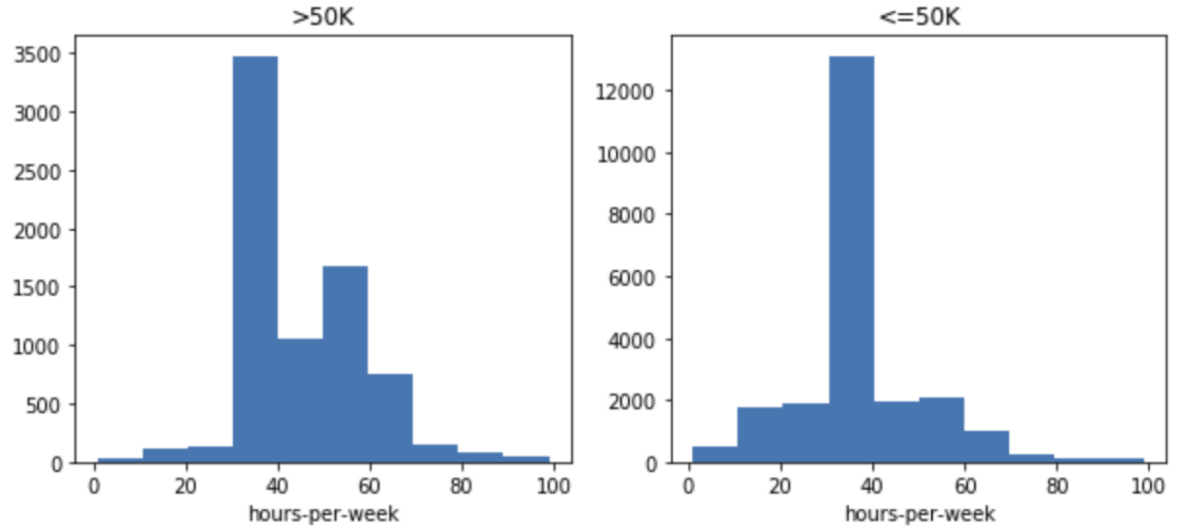
User Story #5:

Distribution of FNLWGT by Income

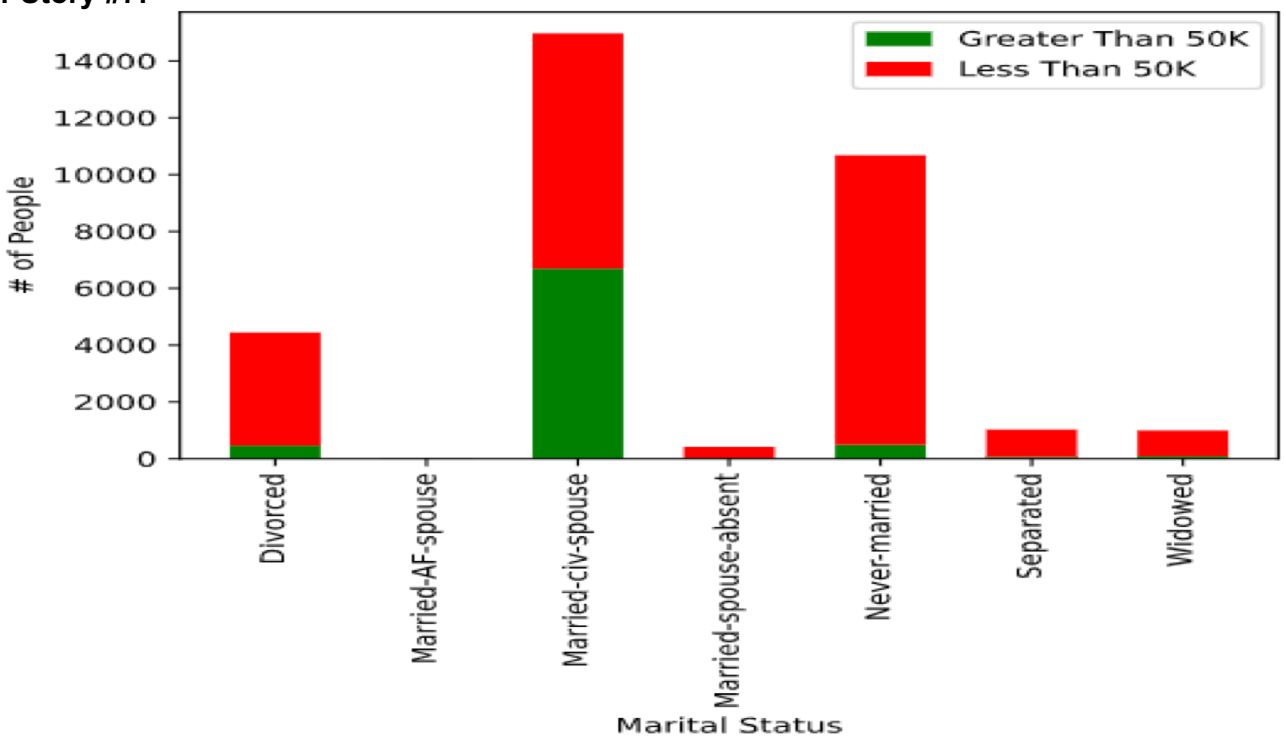


User Story #6:

Distribution of Hours-Per-Week by Income

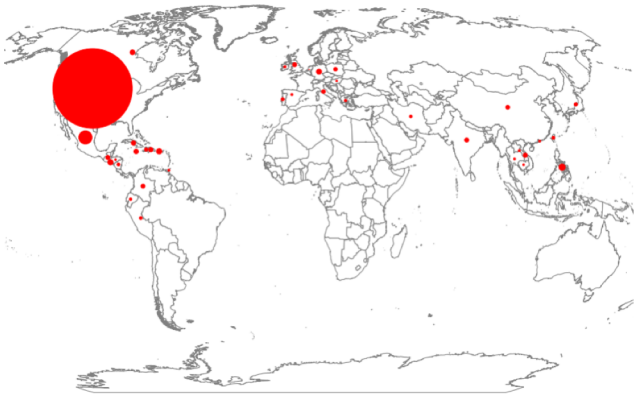


User Story #7:



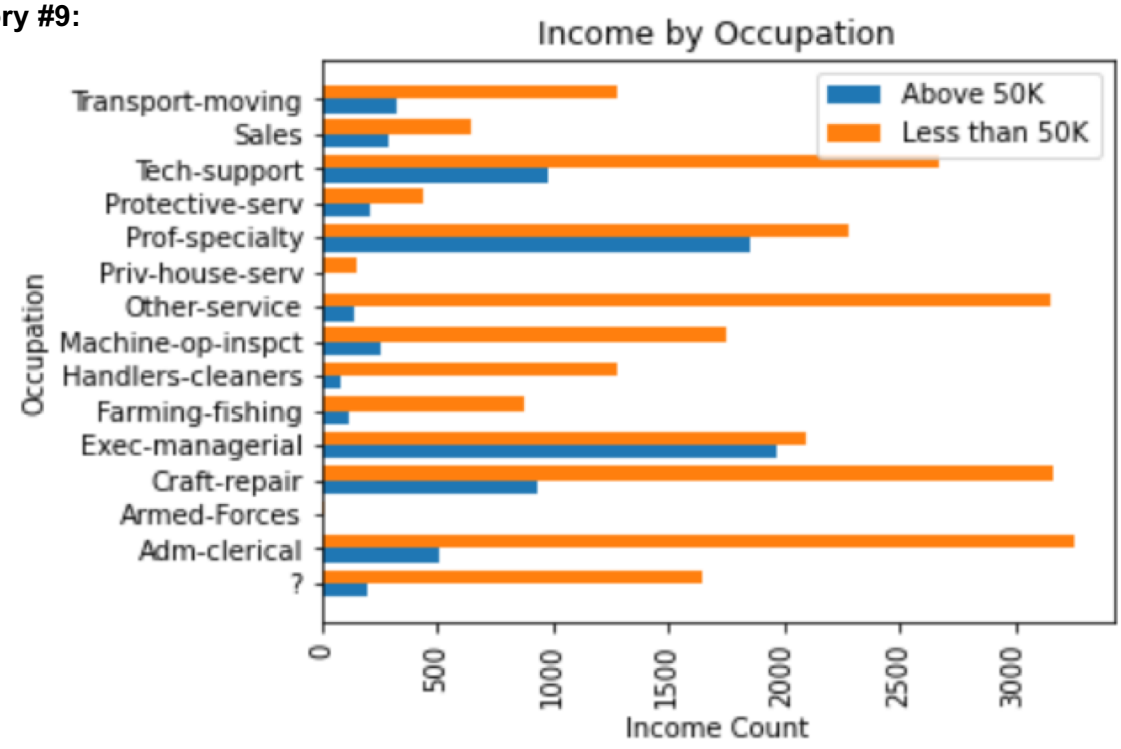
User Story #8:

Distribution of Income less than or equal to 50K by Native Country

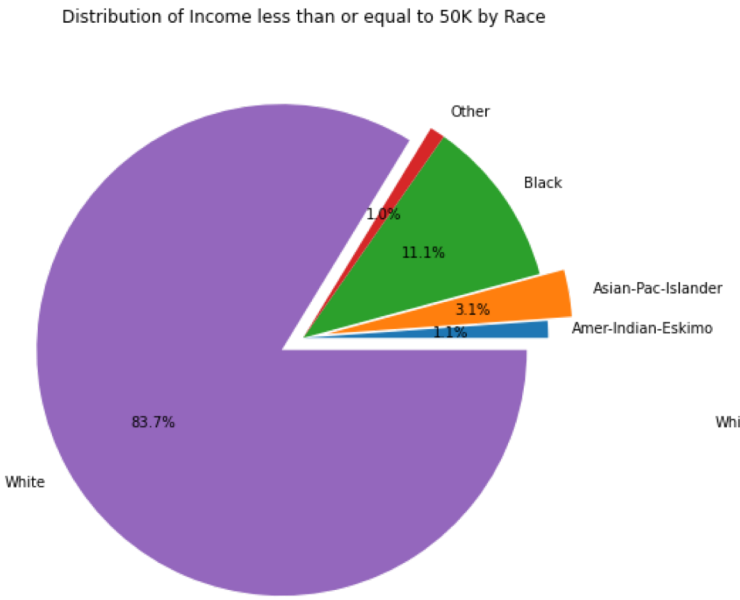


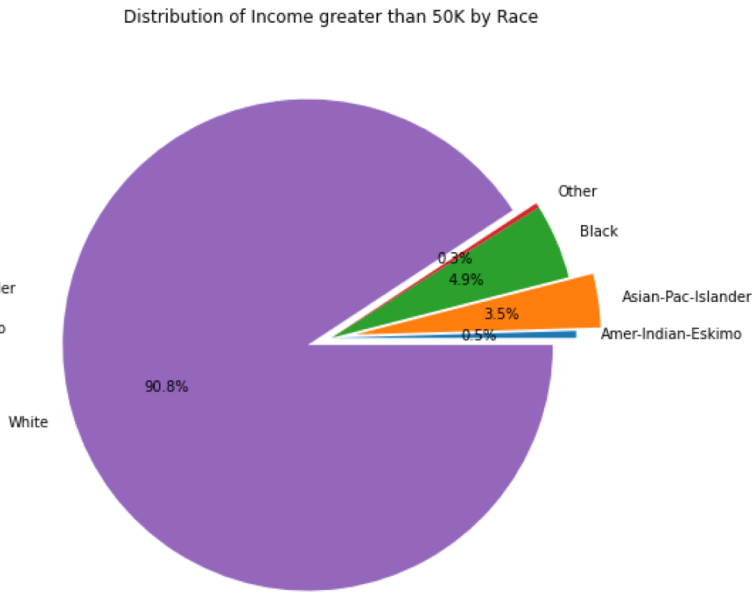
Distribution of Income greater than 50K by Native Country

User Story #9:

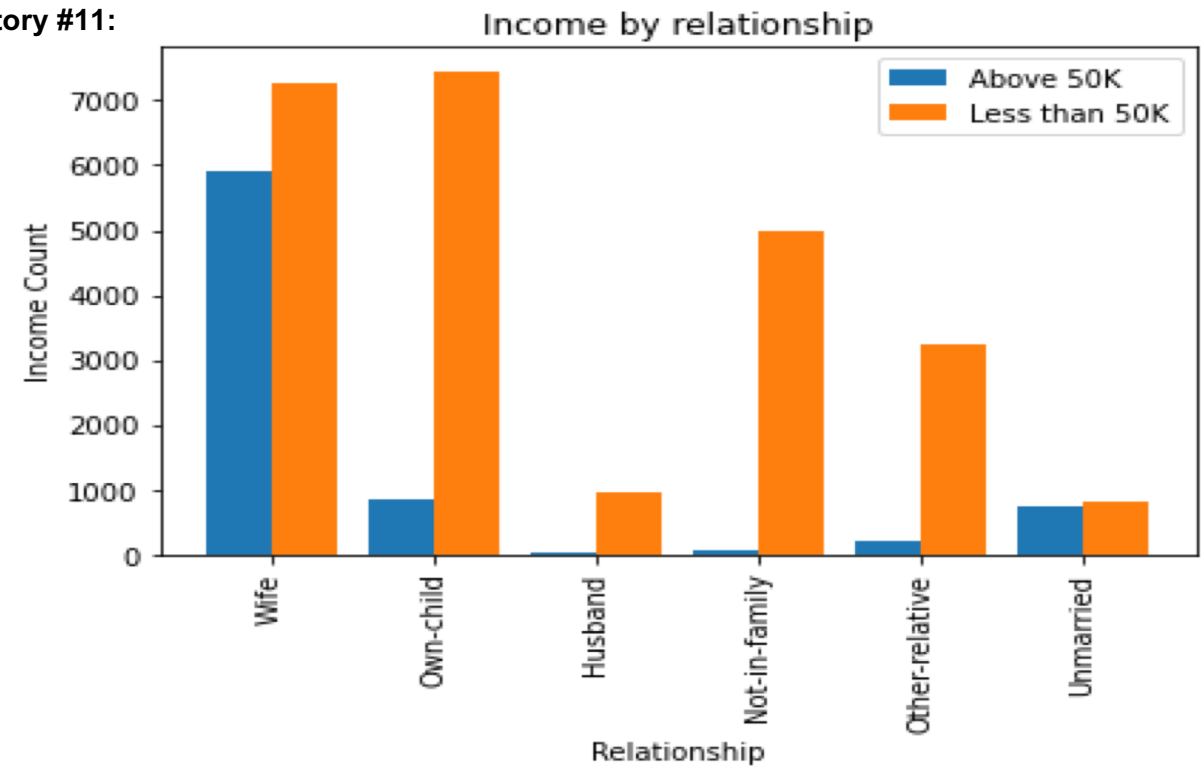


User Story #10:

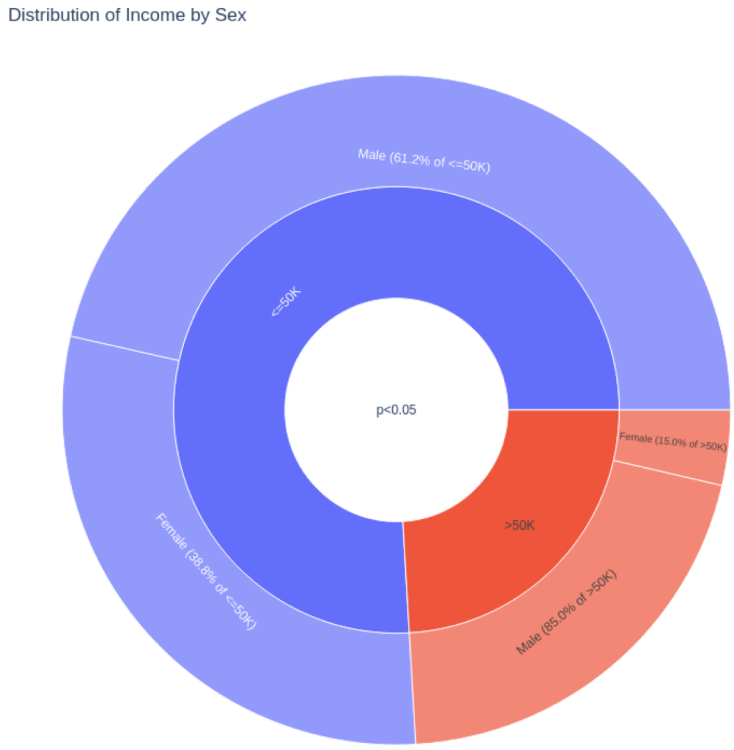




User Story #11:



User Story #12:



User Story #13: