CSE 578: Data Visualization Spring 2021
SYSTEM DOCUMENTATION REPORT

Submitted by: *Group 4*
Shivani Priya (1219492950) Atin Singhal (1217358454)
Varoon Parthasarathy (1216905976) Avinash Khatwani (1216635004)
Uttam Bhat (1217142043) Anil Kumar (1219401495)

Roles And Responsibilities

- Team Members : Shivani, Atin, Varoon, Avinash, Uttam, Anil
- Stakeholders : UVW College
- Product Owner : XYZ Corporation

As a team of data analysts at XYZ Corporation, the responsibility of the team is to develop a marketing profiles for UVW College. The college will use this information to increase their enrollment. Among the 15 features made available in the dataset by the United States Census Bureau, income is considered as the key to determine the criteria for marketing. $50,000 income is set as the key value for income. It is important to determine the relevant features which would play an important role to determine if the income of an individual is above or below $50,000. Thus, the team is responsible for determining such features. These features will be grouped together to develop a proposed model by the marketing team. Eventually, it will be used by the marketing team to predict the income of individuals so that the marketing could be done only for people who satisfy the criteria as relevant customers.
Thus, if the data shows that the majority of people with age 20-24 years having masters degree and occupation as working have salary above $50,000 then this information can be used by the college to market their degree program to people with similar features. Thus, the role of the analysts team is to give a list of such features which will be helpful.

For the project, the work of the team was done in different stages. Firstly, the data was cleaned. Any irrelevant characters or symbols like '?' were removed from the analysis dataset. White spaces were trimmed from the data for better analysis. Secondly, data exploration was done. Number of NULL characters were checked and the type of data to be analysed was noted. Next, data correlation was found to determine the list features highly correlated with the income. This was done as part of initial analysis. Lastly, data analysis and visualization was done for a set of univariate variables and multivariate variables to determine the list of features which will be relevant to predict the salary of a person.

Role of the team was:

1. **Shivani Priya** - Responsible for analysing the dataset and finding the correlation between features using various plots like scatter matrix and heatmap. I was also responsible for analysing attribute "Education" and concluding if the feature is relevant to determine the income of an individual.
2. **Atin Singhal** - Responsible for creating the github repository, helped enhance the heatmap for correlation matrix analysis. Also responsible for analysing the attributes "education-num" & "relationship" and determining if the features are relevant enough to help predict the income.
3. **Varoon Parthasarathy** - Responsible for analyzing the attribute Hours-per-week in the given data. Worked on the multivariate analysis of 'sex' v/s 'occupation' v/s 'income' in the given data.
4. **Avinash Khatwani** - Responsible for analysing attribute Age in the given data. Worked on the multivariate analysis of attributes 'capital-gain' v/s 'age' v/s 'income' and 'age' v/s 'Education-num' v/s 'income'.
5. **Uttam Bhat** - Responsible for analysis the sex and capital-gain attribute in the given data. Worked on the univariate analysis of said attributes to figure out how they might affect the prediction.
6. **Anil Kumar** - Responsible for analysing attributes Marital-status in the given data. Worked on the multivariate analysis of attributes 'capital-gain' v/s 'income' v/s 'hours-per-week'.

Team Goals and Business Objective

The primary objective of this project is to analyze and draw insights on the dataset to help create marketing profiles based on the income of people. These insights determine the important factors and also finds patterns in the data through various visualization techniques.
The analysis provided by us will then be used by the UVW College for marketing purposes. As a result they will be able narrow down the marketing to a particular set of individuals depending on their income.
Overall, the analysis will help the UVW University to attract more students and also reduce the cost of the marketing campaigns.

Assumptions

- **Accurate dataset** : We assume the dataset provided is accurate and precise. It is assumed the data is not misleading and does not contain erroneous information. Ensuring that the data is accurate can be expensive.
- **Relevance and Timeliness** : It is assumed that the data is collected at a relevant time. Data collected at an irrelevant point in time misrepresents a situation and causes inaccuracies in the analysis.
- **Valid data** : The dataset provided must be validated and legitimate. The data attributes must be limited to a set of pre decided options. No attribute must have value outside of these decided options. No open answers are allowed.
- **Attribute selection** : We have assumed the attributes that have the highest correlation and contribute most towards class prediction give us the best patterns. All the visualisations and analysis is done on these attributes and they are unbiased.
- **Complete data** : as is the case with any sort of large dataset, it is assumed that some parts of the dataset are incomplete. Using these rows for prediction can alter the results in a way that is undesired. We have decided it's best to remove these rows from the dataset for better and accurate prediction results.

Data Exploration:

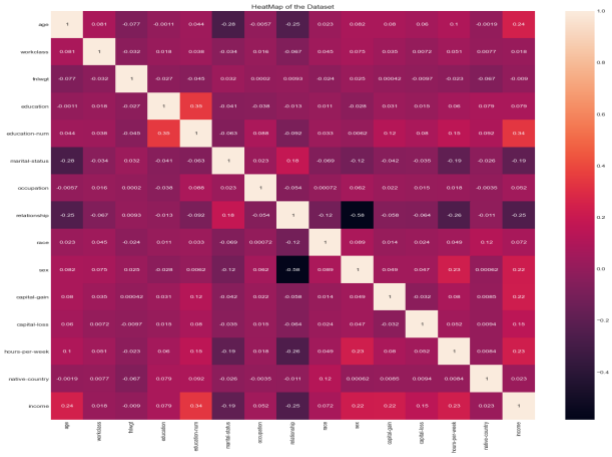
- **Missing values treatment**: The data consists of various corrupt data elements which would be creating hindrance in the data visualization. Data is processed to adjust missing values and other inconsistent values (‘?’) to make it compatible for the requirement.
- **NULL Value Detection**: Check if any NULL values are present in the dataset.
- **Variable transformation**: Convert categorical features into numeric values to determine the correlation with the income.
- **Data Correlation**: Correlation between all the features available in the dataset with respect to income was done to select the features for further univariate and multivariate analysis.

User Stories:

- Analysing the role of features using **univariate analysis** to understand the relevance of features to use them for creating the marketing profiles. Following features were used for the univariate analysis :
    - education-num
    - relationship
    - age

    - hours-per-week
    - capital-gain
    - sex
    - marital-status
    - education
- Analysing the role of features using **multivariate analysis** to understand the relevance of features to use them for creating the marketing profiles. Following features were used for the univariate analysis :
    - Capital-gain vs Hours-per-week vs Income
    - Sex vs Occupation vs Income
    - Capital-gain vs Education-num vs Income

Visualizations

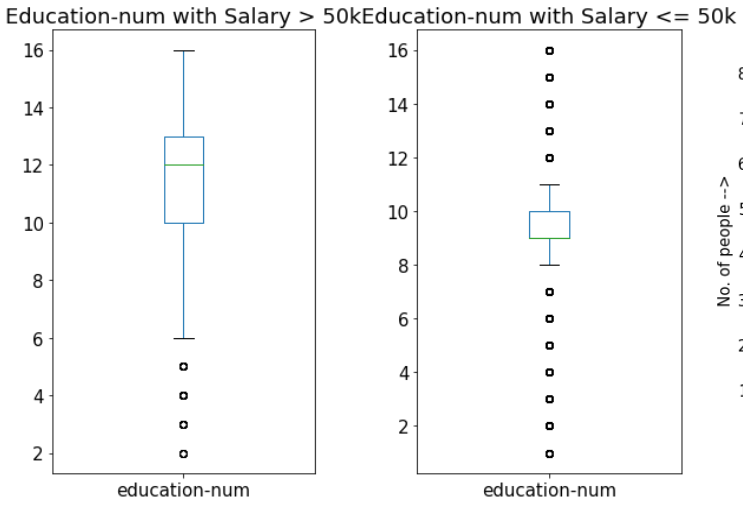


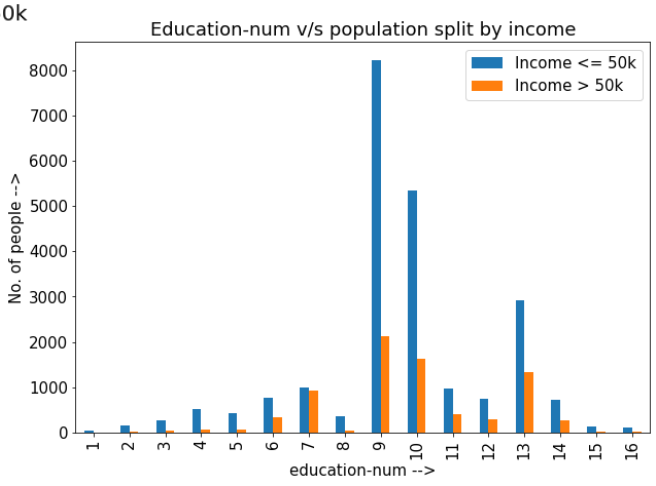
Correlation between features:

Correlation between features was determined using heatmap. Using the output of the heatmap features for univariate and multivariate analysis were chosen for analysis.

Education-num:

This feature has the highest correlation with income at 0.33 so we decided to explore and see if this will be relevant enough. Box plots and bar graphs were used for this purpose.



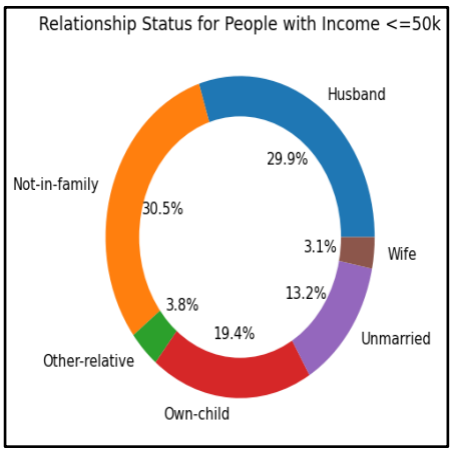


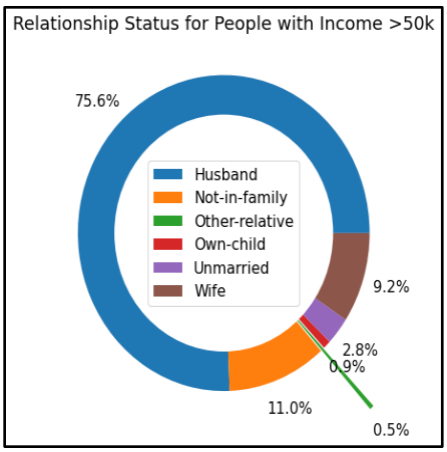
Inferences:

- People earning <=50k will be in the range of 8-11 years of education (with the median being 9), whereas people earning >50k will be in the range of 6-16 years of education (with the median being 12).
- People who have >10 years of education are more likely to earn >50k.
- Taking the above results into consideration, 'education-num' should be a good feature to include while training the model.

Relationship:

Relationship has the second highest correlation with income at -0.25. Even though the value of correlation is negative, it should not impact the end goal- selecting features to train a model for prediction. Donut plots were for the purpose of visualization.



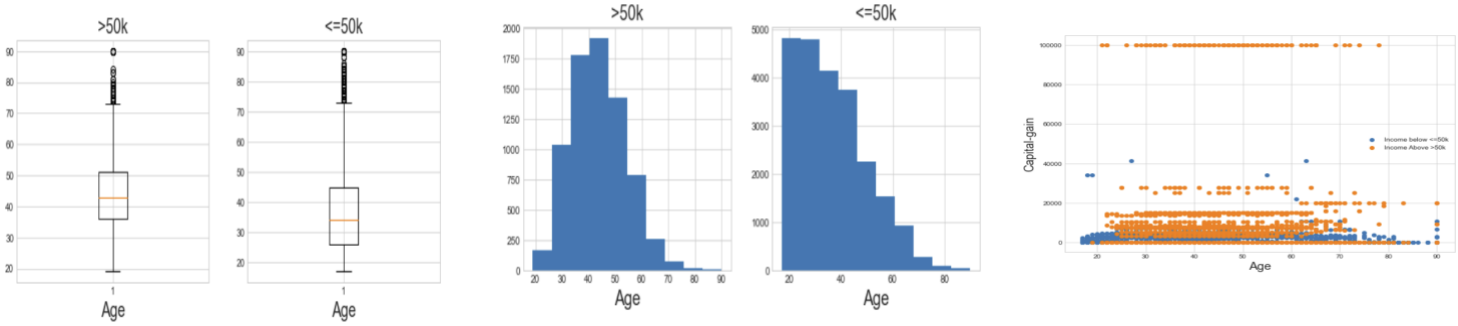


Inferences:

- Most of the people earning <= 50k are either husbands or not in a family.
- More than 3/4th of the people making > 50k are husbands.
- 'Husband' and 'Wife' are the only categories which have a higher probability amongst the people making >50k. The other categories (not-in-family, other-relative, unmarried and own-child) have a higher probability of being amongst the people making <50k.

- Since 'relationship' has a medium correlation with income and we can see patterns in the visualization, it should be a good feature to select for training the ML model.

Age:



Age has a high correlation of 0.24 with income. To further analyze the Age feature further, we used Box Plot and Bar Graph as shown below:

Inferences:

- Quartile 1 and Quartile 3 of Box Plot for Age Above 50 K has values 36 and 51. While, Box Plot for Age with Income below 50K has value 26 and 45 for Quartile 1 and Quartile 3. Similarly, for Income Below 50 K the median

lies for a lesser value compared to median for Age value of Individuals with Income Above 50K. This shows that individuals with a higher age are more likely to have an income above 50k when compared to the other class.

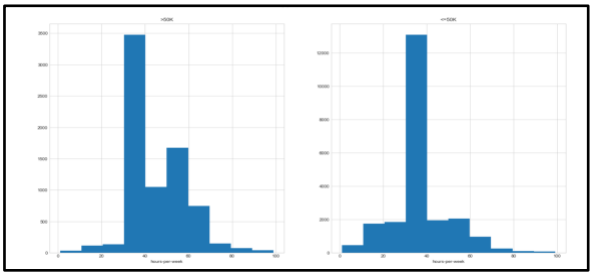
- The Bar graph shows that the data is almost normally distributed with the mean lying approximately between 40-50 for individuals with income above 50k. However, the data is rightly skewed for individuals with income less than 50k.
- The scatter plot between Capital Gain and Age shows us that the two are approximately well separated and also People with a higher capital gain seem to be the ones earning more than 50k. Thus, it should be a good feature to select for training the ML model.

Hours-per-week:

Hours-per-week was up on the list of features which had a high correlation with income, with a value of 0.23. We implemented a box-plot and a bar graph to further understand the relationship between the two features.

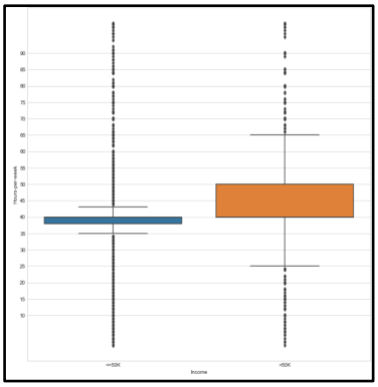
Bar Graph:

The bar graph here helps us to identify the range in which the number of hours each person works for, in both the <=50K and >50K category. Here we can see that for <=50K, the maximum count of people lies in the interval 30-40 hours, which is more than 6 times the count of the next highest interval count. In the >50K category, we can see that even though the 30-40 hours interval has the maximum count, the number of people who work for 40-70 hours is also closely behind.



Box plot:

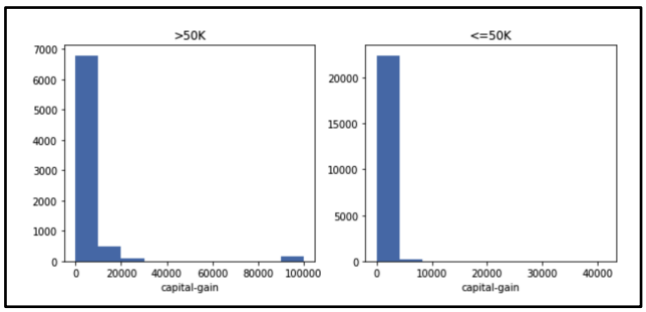
From the box plot here, we can see that for <=50K the 1st and 3rd quartile is very close and has a value of 38 and 40 respectively. For >50K, this is spread out and equal to 40 and 50 hours respectively. The median for both the plots lie at 40. The box plot also shows that people who earn more than 50K tend to work for about 7 hours a week more on average than those who earn less than 50K.



This feature alone is not capable enough to make a prediction that the income of an individual is above or below 50K. This can be used in conjunction with other relevant features.

Capital-gain:

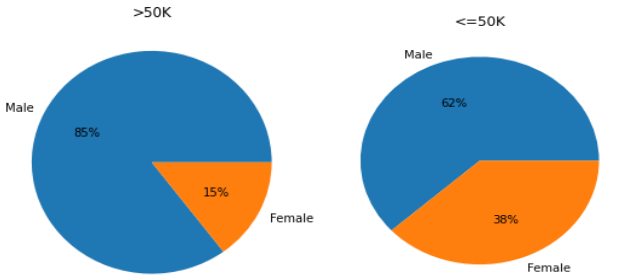
Capital-gain is another one of the features that has a higher correlation factor with income than most attributes with a value of 0.22. To further analyze the attribute we have a bar graph and a box plot. And it looks like if the capital gain is high it is more likely that the income of the person is above 50k. Since the pattern for both



the cases looks similar we can conclude that this feature might not be relevant for income prediction.

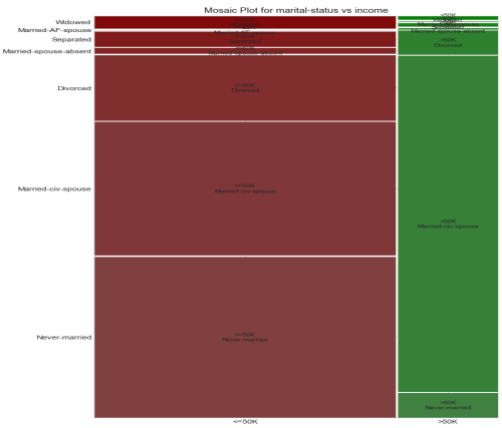
Sex:

Sex was up on the list of features which had a high correlation with income, with a value of 0.22. We implemented a pie chart to further understand the relationship between the two features.



Inference:
It is seen from above pie charts that males dominate both classes income>50k and income<=50k however, it is significantly higher among the salary range >50k. This feature alone is not sufficient to predict income.

Marital-status:



For Marital status the correlation factor with income was -0.19 and a mosaic plot was plotted to derive the inference.

Inference:

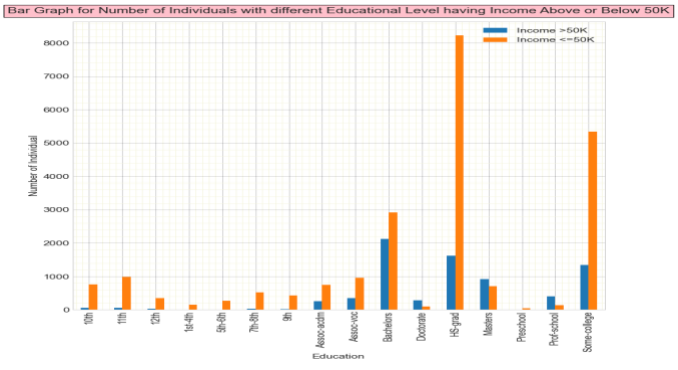
- Widowed individuals have very less section >50k
- Very Large section of Never-married people are <50k
- Married-AF-spouse have negligible contribution in both above and below 50k section
- Married-civ-spouse are the highest in number having income >50k

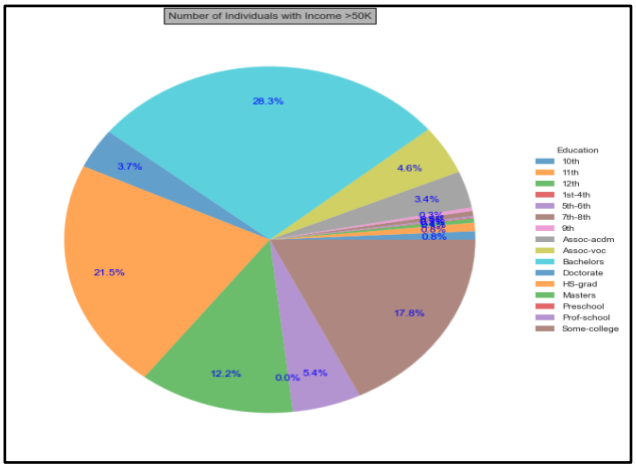
Education:

Education has the ninth highest correlation with income at 0.078987. Even though it had lower correlation in comparison to capital loss it was selected for analysis as the education-num showing similar attributes had a higher correlation factor. We wanted to check the output of a parameter with a lower correlation factor with the income of an individual.

Inference: The stacked bar graph shows that even though the number of individuals with income above and below 50K is not uniformly distributed, there are few education backgrounds for which there are more number of individuals with income above 50K. The degrees for which chances for an individual with same degree of education to have higher income are:

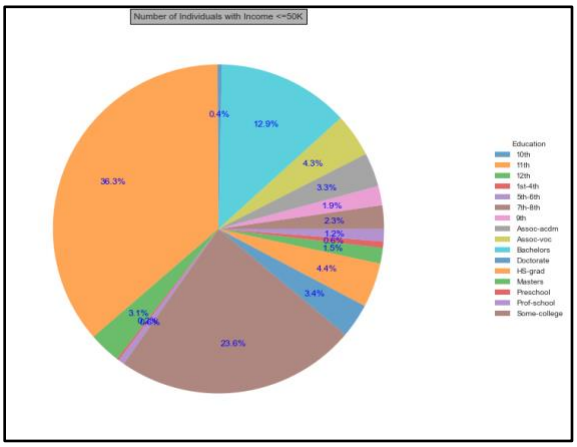
1. Doctorate
2. Masters
3. Prof-school





Inference: The pie chart on the left shows the distribution of education qualifications for individuals with income above 50K. The most common qualification for individuals with income above 50K(5% or above) are :

1. Bachelors - 28.3%
2. HS-grad - 21.5 %
3. Some-college - 17.8 %
4. Masters - 12.2 %
5. Prof-school - 5.4 %



Inference: The pie chart on the right shows the distribution of education qualifications for individuals with income below or equal to 50K. The most common qualification for individuals with income below or equal to 50K(5% or above) are :

1. HS-grad - 36.3%
2. Some-college - 23.6 %
3. Bachelors - 12.9 %

Pie chart for both the scenarios shows that people with education qualifications like Bachelors, HS-Grad, Some-college are higher in percentage than other qualifications. Thus making a marketing profile based on these results will be not much helpful. Thus, Education can not be selected as one of the attributes for the marketing profile model.Anyways, this feature has a high correlation with Education-num which is selected for marketing profile,thus using this feature will not be relevant.

Tools & Technologies

Tools and technologies used in this project are Python, Anaconda Python Distribution, Jupyter Notebook, Google Colab, GitHub and VS Code.

Questions

- **How to decide what features to pick?**
  - Since we decided to select 8 features out of the given 14, the first question that popped in our minds was which ones to choose and how to select them. Since the end goal is to pick the features which will help in training the ML model, we decided to pick correlation as a metric to help decide which features to pick.
  - Ideally for a ML model, we don't want the features to repeat or provide the same information as that will just add an overhead while training so we want to pick the ones with not too high correlation. If two features have a high correlation (>0.50), they are dependent and we can drop one of them. [1] Also, this means if we want to predict income, we should select features which have the highest correlation with income. [2]

    - In correlation, we should compare the absolute values as both 1 and -1 signify a strong correlation, a negative correlation just means that the features are inversely correlated, i.e. when one increases the other decreases. [3]
- **What kind of visualizations should we use for correlation?**
    - There are many plots available in python which help to determine the correlation between features.
    - Initially the analysis for correlation was done using a scatter matrix. Unfortunately we didn't quite get the results we expected. Since we only had binary values for income, correlation plots using scatter matrix didn't do any benefits.
    - The next best solution seemed to be heatmaps. By observing how cell colors change across each axis of a heatmap, one can simply observe if there are any patterns in value among the features of the dataset.
- **How to deal with categorical variables?**
    - To find the correlation between features it was important to convert categorical data to numeric values.
    - To use categorical data, we have to convert it into numerical values by enumerating them as 0,1,2,... and so on.
- **What visualization will be best for the features?**
    - Upon analysing the data, it quickly became clear that we can not use the same visualization for every feature. Some visualizations are better suited for univariate analysis while others are better suited for multivariate analysis. The same is valid for the type of data- categorical and numerical/ continuous data.
    - Visualizations like mosaic plots, donut plots & pie charts were used to represent categorical data and histogram, box plots & scatter plots for continuous data.
- **The visualizations are done but what about the skewness of the dataset?**
    - In our project, initially the visualization for income was done. The donut chart plotted for income clearly showed that the data is not distributed uniformly.
    - 75.1% of the entries in the dataset are for people who make less than 50k while there are almost 24.9% entries for people with income greater than 50K. Thus, proving that the data is highly skewed.
    - If the dataset is used as it is, then the model will be biased towards the less than 50k category. Thus, it is important to figure out a solution for this so that further analysis on the data can be proceeded with training the model. This will be considered as one of the future works for this project.

Future Work

We plan on doing the following in the future which is well within the scope of our project.

- We can use dimensionality reduction to find the best/new features for analysis , which will provide us with more than just the option of choosing features through the correlation matrix. Features with comparatively less correlation might have interesting relationships with Income, which are being overlooked right now.
- The relevance of a feature right now is determined by the correlation matrix. In future, we can work to use machine learning algorithms to determine the feature relevance, by training the models based on information from other datasets.
- We can also implement machine learning algorithms to predict salary groups. We can use these algorithms to determine if a person with various attributes listed will fall under the >50K category or <=50K category.

References

[1] R, Vishal. 2018. “Feature selection — Correlation and P-value.” towards data science. https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf.

[2] Charfaoui, Younes. 2014. “Hands-on with Feature Selection Techniques: Filter Methods.” heartbeat. https://heartbeat.fritz.ai/hands-on-with-feature-selection-techniques-filter-methods-f248e0436ce5.

[3] “Negatively correlated features.” 2020. https://datascience.stackexchange.com/questions/77511/negatively-correlated-features