SYSTEM DOCUMENTATION REPORT
For
CSE 578 SPRING 2021 GROUP PROJECT
GROUP 19:

Taylor Bart, Zuwei Guo, Pankaj Janakiraman,
Fatma Mokhtar, Ahmed Shabbir, Subhash Chandra Bose Vuppala

Roles and responsibilities

Task	**Assignee(s)**
Indicator Metrics	**Zuwei Guo, Fatma Mokhtar**
Education Level	**Taylor Bart**
Marital Status	**Taylor Bart, Vuppala Subhash Chandra Bose**
Hours Per Week	**Pankaj Janakiraman**
Sex	**Pankaj Janakiraman, Fatma Mokhtar,Vuppala Subhash Chandra Bose**
Relationship Status	**Taylor Bart**
Race	**Ahmed Shabbir**
Native-Country	**Ahmed Shabbir**
Age	**Fatma Mokhtar, Pankaj Janakiraman**
Occupation	**Vuppala Subhash Chandra Bose**
Capital Gain	**Pankaj Janakiraman**

Team goals and a business objective

The goal of this project is to help UVW College by developing targeted marketing profiles based on income as a key demographic in the decision making. We will be utilizing data from the United States Census Bureau to look at various features that can affect a person's income including age, sex, education and marital status etc. The key target provided by UVW College is determining if an individual has an income above or below $50,000. This will help establish the marketing profiles to be developed for the college. In order to achieve this we plan on using visualizations to find correlations between the individual features from the dataset and the income of a person.

Assumptions

- UVW college has chosen a salary as a key demographic to determine criteria for marketing its degree programs.
- The marketing team at UVW would like to develop an application to find factors that determine the individual's income.
- A dataset from the United States Census Bureau will be used to develop this application.
- The dataset used is accurate.

User Stories

1) As a member of the UVW marketing team, I want to know if the **education level** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

2) As a member of the UVW marketing team, I want to know if **marital status** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.
3) As a member of the UVW marketing team, I want to know if **age, hours per week and capital gain** are relevant factors in determining their income label when used together so that I can decide whether or not it should be integrated into our team's prediction tool and see if the results from each are coherent.
4) As a member of the UVW marketing team, I want to know if **sex** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.
5) As a member of the UVW marketing team, I want to know if the **relationship status** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.
6) As a member of the UVW marketing team, I want to know if **race** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.
7) As a member of the UVW marketing team, I want to know if **native-country** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.
8) As a member of the UVW marketing team, I want to know if **age** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.
9) As a member of the UVW marketing team, I want to know if the **occupation** is a relevant factor in determining their income label so that I can decide whether or not it should be integrated into our team's prediction tool.

Data Preprocessing

Data preprocessing is the process of fixing or removing incorrect data from the dataset. After the initial inspection of the dataset, we noticed that some fields contain the value '?' which is equivalent to null or undefined value. To fix this issue we:

- Identified which features contain these undefined values, and those were the columns: workclass, occupation, and native-country.
- Identified the mode or most frequent value of these features.
- Fill in the null values with the most frequent value for each feature.

We also pivoted on qualitative columns (education level, marital status, relationship status, etc.) and vectorized the new columns (i.e. Preschool, Doctorate, Husband, Wife, etc.) to 0 and 1 (1 if value was present in the original, pivoted column and 0 if not).

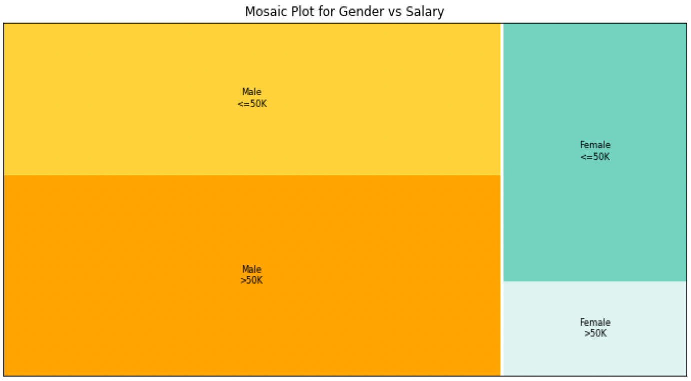
Visualizations

The use of visualizations is key to achieve the business objective. Visualizations are used to obtain knowledge and patterns from the dataset. We have used visualizations to figure out if the various features, namely, age, sex, occupation, etc, are dependent on a person's salary, or if multiple features are correlated with each other. This knowledge is key in deciding which features we will be using to implement machine learning models, and which features are redundant (i.e. not dependent on the person's salary).

The various visualizations we have implemented are presented in the following sections.

Mosaic Plot

A mosaic plot is used to check how categorical data are related to each other. The categorical feature from the dataset that has been used in this particular visualization is Gender, and its relationship with Salary (<=50K or >50K)

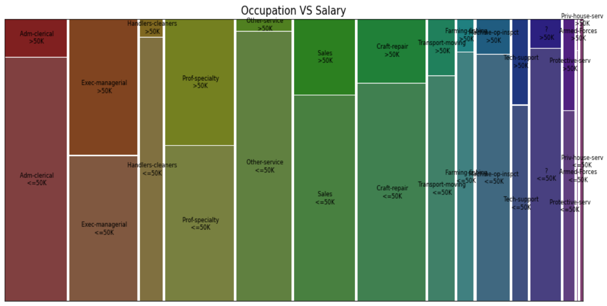


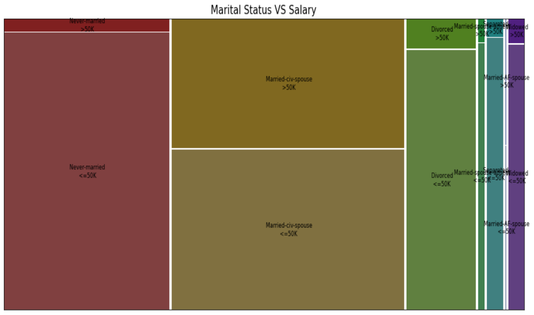
The color coding is done such that the smaller proportion has a lighter color and the larger proportion has a darker color respectively.

From the Mosaic plot above, we can conclude that:

1. A large proportion of males earn more than 50K salary.
2. A small proportion of females earn more than 50K salary.

Mosaic Plot for Occupation Vs Salary





The categorical feature from the dataset that has been used in the left visualization is Occupation, and its relationship with Salary (<=50K or >50K).

From the Mosaic plot above, we can conclude that:

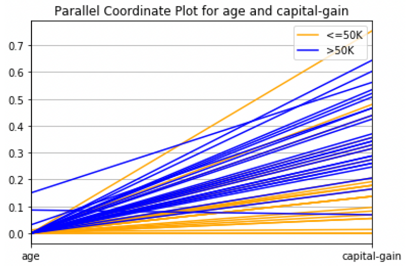
1. For most categorical data, the dispersion of the two classes are exceptionally skewed indicating that this component can be utilized to recognize among the two classes.
2. Individuals with occupations such as “Exe-managerial”, “Prof-speciality” are more likely to earn 50K income.

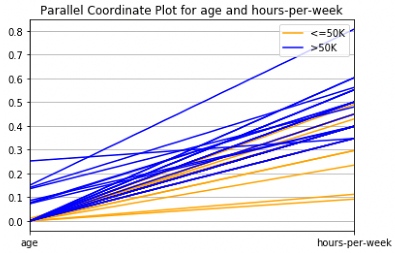
The categorical feature from the dataset that has been used in the right visualization is Marital Status, and its relationship with Salary (<=50K or >50K).

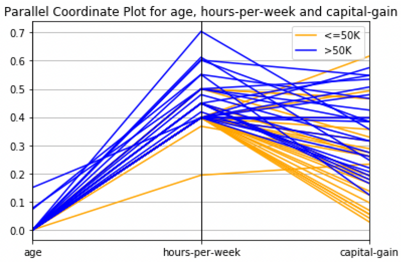
From the Mosaic Plot we can conclude that Married-civ-spouse are more likely to earn more than 50K income.

Parallel Coordinate Plot

A parallel coordinate plot (PCP) is a visualization used to plot multivariate data, in order to check the relationships between them, i.e. if they are correlated. For this scenario, PCP has been used to check the relationship between three variables, namely, hours-per-week, age, and capital-gain.







In the PCP plots above, the orange lines represent the salary values in the range <= 50K and the blue lines represent the salary values in the range > 50K.

From the PCP plots, we can conclude that:

1. Individuals at an older age are likely to earn a higher salary than individuals at a younger age.
2. There are always exceptions to every case, for example, an individual at a younger age can earn a salary greater than 50K as well.

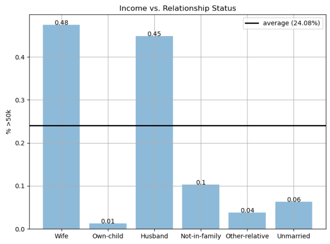
From the PCP plots, we can conclude that:

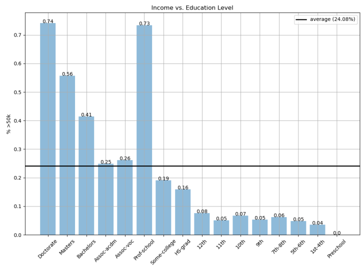
1. Individuals at an older age are more likely to work more hours per week than individuals at a younger age.
2. Individuals that work a large number of hours are more likely to earn a salary greater than 50K.

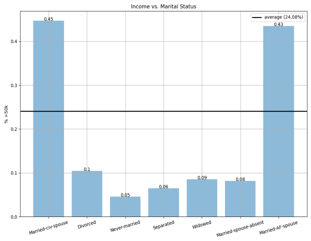
From the above PCP plots, we can see the combination of all three features together, i.e. how they are correlated to each other, and also how they are individually dependent on the class outcome, i.e. less than or equal to 50K or greater than 50K.

Bar Chart

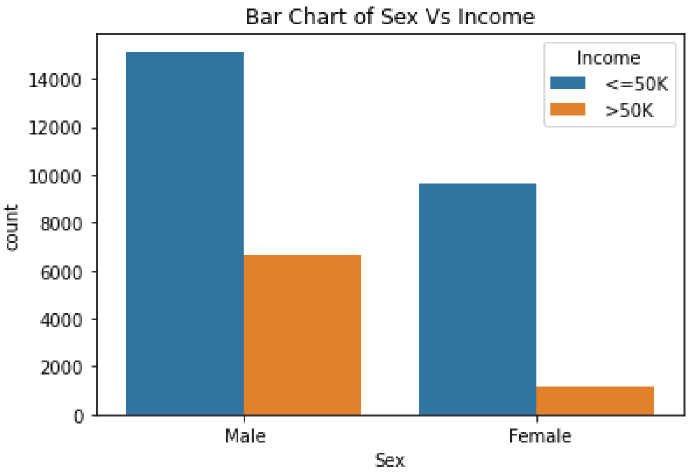
A bar chart can help compare values within the same feature and see if possible clustering or patterns emerge. A mean line was added to show how each sub-category performed against the average. The more polarized the values are, the more likely that the feature will be a good indicator of income level.







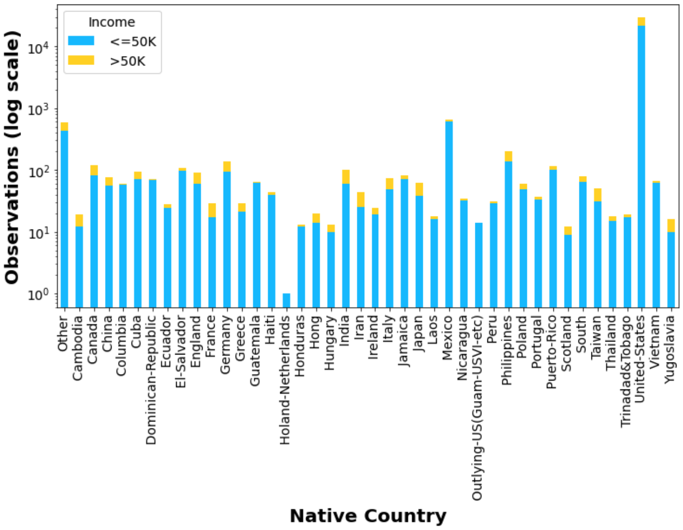
- From the leftmost chart, we can see that individuals that are Wife/Husband are far more likely to have an income of >50k.
- From the leftmost chart, individuals with Own-child, Not-in-family, other-relative, or unmarried as a relationship status are significantly likely to have <=50k salary (ranges from 90% to 99% likely to be <=50k depending on the label.
- From the middle chart, income >50k is highly correlated with having a specialized, secondary education (i.e. Doctorate, Masters, Prof-School)
- From the middle chart, those with degrees less than a Bachelors are statistically likely to have a salary of <50k.
- From the rightmost chart, people with Income >50k are more statistically probable to have a marital status of married and present.
- From the rightmost chart, Income <=50k is highly correlated with being unmarried or being married but not present.



- The above chart shows the relationship between Sex and Income.
- X-axis represent Sex and Y-axis represent count of Income(both greater than 50K and less than 50k)
- From the above bar chart we can infer that:
    1. Males with income greater than 50K are high in number when compared to females with income greater than 50K.
    2. Males with income less than 50K are high in number when compared to females with income less than 50K.

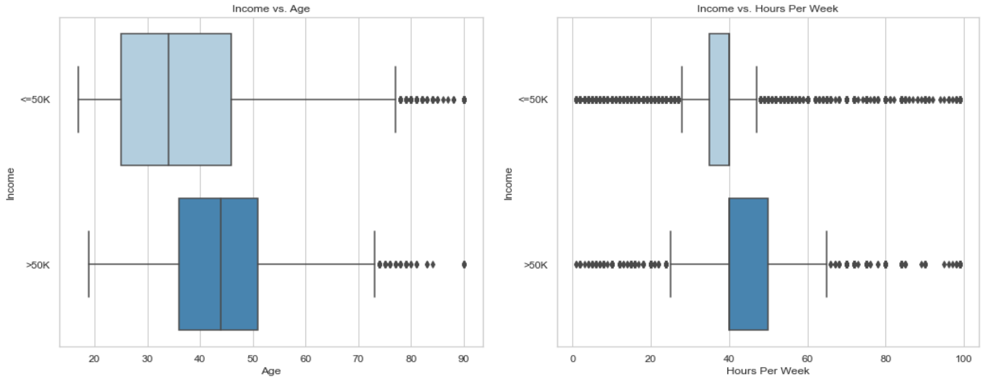
Stacked Bar Chart

A stacked bar chart can be used to compare numerical data across multiple categorical variables. Compared to a simple bar chart, more information can be condensed into a single visualization for multiple categories of data.



- Native Country feature is not a reliable indicator of income
- All countries have much more observations that have <=50K income as label
- Log scale used in visualization as data is highly skewed

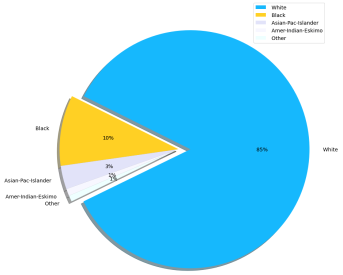
Box and Whisker Plot



- People whose age between 35 and 50 are likely to have a salary above 50K.
- However, there is a good portion of people whose age between 35 and 45 have a salary below 50K. In addition to those between the ages of 25 and 35 who are more likely to have income of 50K and below.
- There are some outliers between age 72 and 90 that have income above 50K.
- Most of the people who work 40 to 50 hours per week have income above 50K.
- Most of the people who work 35 to 40 hours per week have income below 50K.
- There are some outliers who work either less than 30 hpw or more than 60 hpw and have income above 50K.

Pie Chart

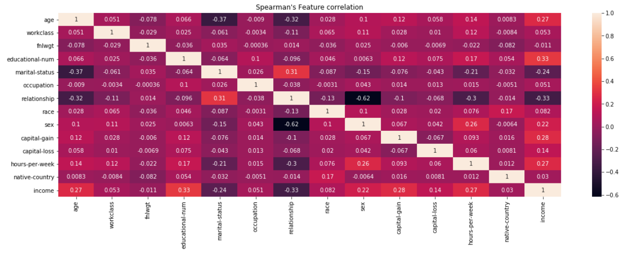
A **pie chart** takes **categorical data** from a statistical sample and breaks them down by group, showing the percentage of individuals that fall into each group.



- Race feature is highly skewed
- 85% of sample is 'White'
- 10% of sample is 'Black'
- 5% of observations for other races

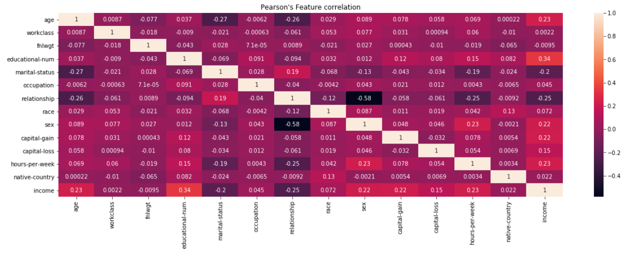
Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. In this study, the non numerical columns are converted into categorical variables. For example, in 'sex', 'Male' is converted to 1 and 'Female' is converted to 0. After this conversion, the correlation between all 14 variables are calculated.



- Spearman's correlation coefficient shows that income is positively correlated with education, capital-gain, age, hours-per-week, sex, captical-loss, race, workclass, occupation and native country.

	educational-num	capital-gain	age	hours-per-week	sex	capital-loss	race	workclass	occupation	native-country	fnlwgt	marital-status	relationship
income	0.33	0.28	0.27	0.27	0.22	0.14	0.08	0.05	0.05	0.03	-0.01	-0.24	-0.33



- Pearson's correlation coefficient shows the similar correlation that income is positively correlated with education, capital-gain, age, hours-per-week, sex, captical-loss, race, occupation and native country.

	educational-num	age	hours-per-week	capital-gain	sex	capital-loss	race	occupation	native-country	workclass	fnlwgt	marital-status	relationship
income	0.34	0.23	0.23	0.22	0.22	0.15	0.07	0.05	0.02	0.00	-0.01	-0.20	-0.25

Tools

- Data preprocessing and analysis was done in Python using Pandas and Numpy libraries.
- Matplotlib and Seaborn were used to create the visualizations for the project.
- Sklearn library was used to implement various machine learning models like decision tree, nearest neighbor, etc.

Questions

How to figure out which features to use in order to train the dataset?

Visualizations have been created in order to figure out how each feature affects the salary/income of an individual. Also, we have created a correlation matrix that determines which features are highly correlated with one another, and which features are redundant, i.e. unrelated to the salary/income class prediction.

Not doing

A couple neural networks were implemented with predictive powers in the mid 80s (i.e. ranging from 80-90%). However, these models were omitted from the project due to having no explanatory power for the features they used, therefore implying that they would not be useful for the end user.