CSE 578 Data Visualization
System Documentation Report

Introduction

Marketing plays a vital role in any organization. In growing universities marketing budgets are usually limited and the expected returns are high. To tackle this problem we propose analysis of key factors that influence the salary of an individual. UVW has chosen salary as a key demographic to determine criteria for marketing its degree programs.

Roles and responsibilities

Product Owner: XYZ Corporation.
Stakeholder: UVW College.
Team Members: Ajay Kannan, Aparokshith Rao, Raakesh SureshKumar, Rithvik Chokkam, Sanjay Arivazhagan, Vivek Bellalacharvu Srinivasa Rao.

The main objective of the project is to analyse the US census data with different visualisations. The work was divided into five tasks - Data preprocessing, Feature selection, Exploratory Data Analysis, MultiVariate analysis and Documentation.

Task	Description	Assigned to
Data preprocessing and cleaning	The given data was checked for null values and unknown values marked as '?' were cleaned. The distribution of the salary labels was also analysed. The test data was combined with the adult.data dataset. Data sampling was performed based on the requirements for visualization. For attributes such as the capital-gain and loss, the significance of 0s was analyzed.	Rithvik and Raakesh

Exploratory Data Analysis	Everyone in the team worked on analysing all the features relationship with salary using various visualisations.	All team members
MultiVariate analysis	Multivariate Analysis was done on the given data and plots were drawn to find further insights in the data.	Aparokshith and Ajay
Feature selection	Each feature was analysed to find their importance. For example, fnlwgt can increase the number of units in the particular population but it does not directly affect the salary of an individual.	Vivek and Sanjay
Documentation	System documentation and executive reports were prepared by everyone in the team.	All team members

Team goals and a business objective

This project's goal is to analyse the US census data and identify patterns to determine important factors that influence the salary of an individual. The analysis will be presented to the UVW executives. Using this analysis, a predictive model will be developed to predict the salaries. The UVW marketing team will use the predictive model and the data analysis to tailor their marketing efforts when reaching out to individuals.

Assumptions

- The feature "fnlwgt" represents the weight for a responding unit in a survey data set is an estimate of the number of units in the target population that the responding unit represents. This can complicate the dataset in a way that each record can be expanded using the final weight and this will change the record count. Even though this feature doesn't affect salary directly but combined with other rows it may make a big change but it will complicate the process. For this reason, fnlwgt is not considered in this application.

- The data contains unknown values masked as '?'. These unknown values can be found only in three features - workclass, occupation and native_country. The count of unknown values is recorded in Table 1. These unknown values are removed from the data during the data preprocessing.

Table 1 Count of unknowns in the dataset.

Features	**Unknown count**
Workclass	1836
Occupation	1843
Native_country	583

- After initial analysis, the relationship column was found to be positively correlated with the marital_status column.Their analysis was also found to be similar. Due to this we have assumed that relationship features can be excluded from the analysis.

User Stories

1. **Age -** After initial analysis, we can conclude that people in the age group of 31-60 are more likely to get a salary above $50k. On exploring the data, we can also infer that people below 30 and above 60 are more likely to earn below $50k.

2. **Final Weight -** According to the dataset description, The final weight for a corresponding record is an estimate on the number of persons with similar properties. This cannot be considered as an influencing factor to determine the salary of an individual.

3. **Education, Education-num -** Education and its numerical representation, Education-num represent the highest educational qualification obtained by an individual. There seems to be a positive correlation between salary and education/education-num.
   As people obtain higher valued education they seem to get paid more.

4. **Marital-status -** From the stacked barchart we are able to conclude the ratio of salary distribution for each marital status category. Married people are more likely to earn more than $50k.

5. **Occupation -** People who work in Armed forces and Private work services earn less than $50k compared and have no instances of people who earn more than $50k. This information can be used to narrow down occupations that pay more than $50k.

6. **Race -** In the given data, while analyzing the race dimension we find that people in the white race are more likely to earn more than $50k when compared to others.

7. **Sex -** Number of males is demographically more than females in the given dataset. Analysing them based on salary, we find that the data is skewed on male dominance. It becomes more apparent when analysing the > $50k class.

8. **Hours-per-week -** In the given data, people working 40 to 60 hours are more likely to receive a salary above $50k. In one of the analyses, the data is segregated into 4 categories as A (Over-time), B (Full-time), C (Part-time), D (Contract).

9. **Native-country -** 92% of the given sample is native to the United States. In the other 8% of the data, people from other countries are more likely to receive a salary less than $50k. This information can be used for classifying the data.

10. **Capital Loss/Capital Gain -** This column cannot not be considered as a valid factor because most of the data only contained zeros.

Visualizations

Exploratory data analysis was done by all the team members. All the features were individually analysed with the salary label. After individual analysis, some of the columns were combined for multivariate analysis. The visualizations used for the analysis were stacked barchart, barchart, mosaic plot, correlation matrix, violin plots, scatter plots and line charts. The analysis from both EDA and multivariate analysis are explained in this section along with necessary visualizations.

Correlation

Correlation is a statistic to measure the degree of relation between two variables. To measure the degree of correlation that exists among the dimensions of the data, we plot a correlation matrix (Fig 1). As we can observe from the visualization, there is a relatively strong positive correlation among age, sex, education and hours_per_week with salary.

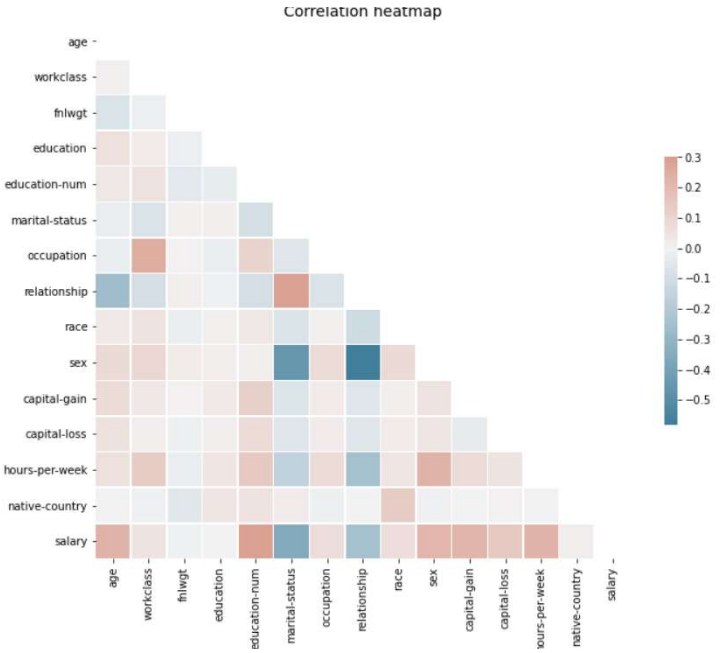


Fig. 1 Correlation matrix for all the columns in the dataset.

Occupation

When we compare age distribution with salary by occupation (Fig 2), we notice that the armed forces have the least age distribution. As expected, the armed forces have lots of employees in the age group 21-30. None of the armed forces personnel earn more than $50k. This might be a job characteristic. While the age distribution in private house services is significantly large, experience does not translate to higher salaries. Another observation from this chart is that the majority of the workforce are in the age group of 20-30. Apart from armed forces and private house services, other occupations seem to value experience and it can be observed with the increase in people receiving salaries greater than $50k.

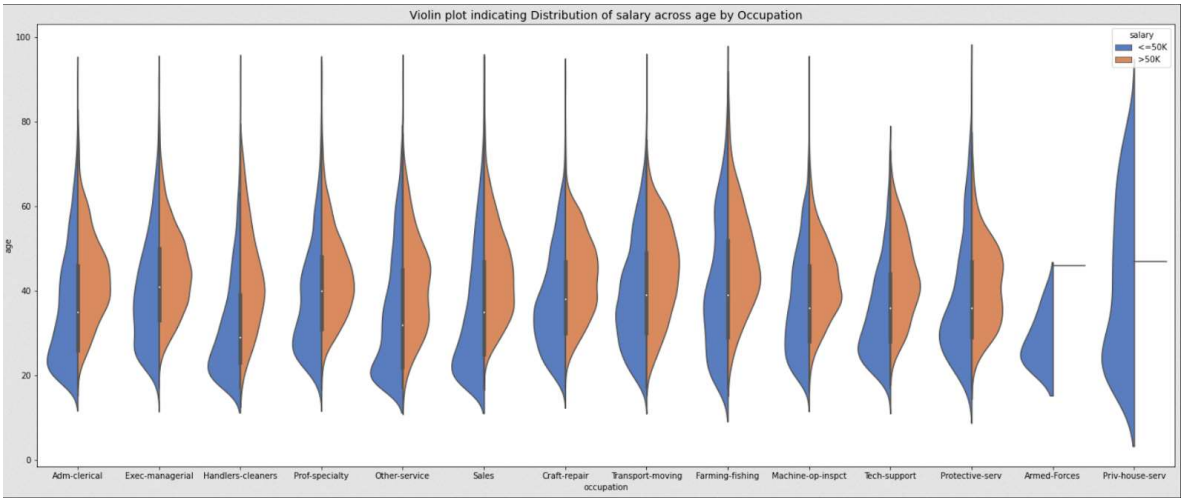


Fig. 2(a) A violin chart for age distribution with salary by occupation

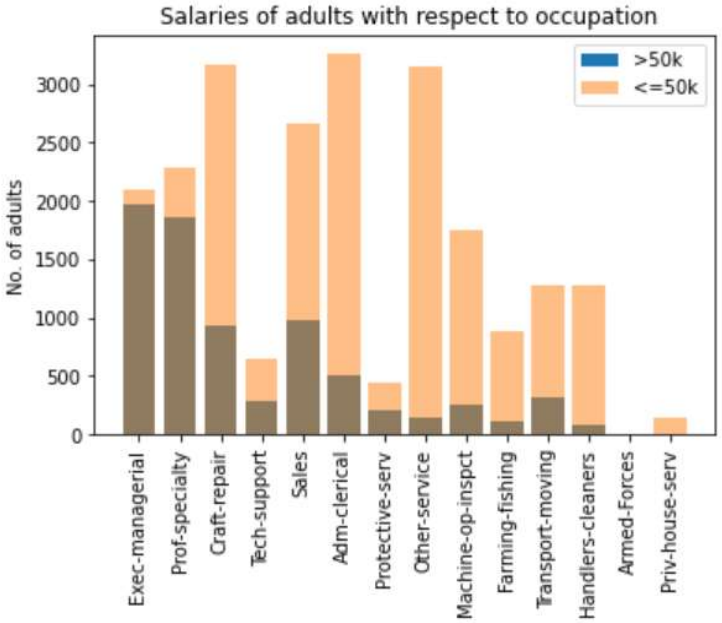


Fig. 2(b) Salary distribution of adults with respect to occupation.

Fig. 2. (b) shows that in the data, people with occupation “Exec-managerial” have the highest ratio of earning greater than $50k to earning less than $50k. We can conclude that holding a position in the “Exec-managerial” category increases the chances of earning greater than $50k compared to any other type of occupation.

Fig. 2. (c) and Fig. 2. (d) are mosaic plots segregated through Occupation and Sex based on Salary. The inference is that for people earning below $50K, the salary is distributed fairly among various occupations for both genders. But it can't be said the same for high paying jobs (>$50K). The data is skewed for people earning more than $50K towards males for most of the occupations.

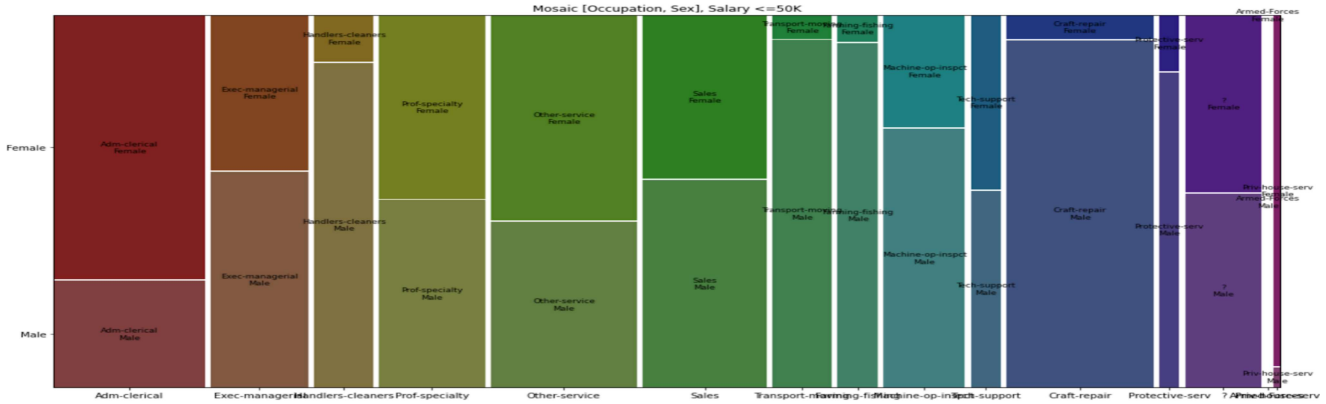


Fig. 2. (c) Occupation distribution based on sex for salary below $50K

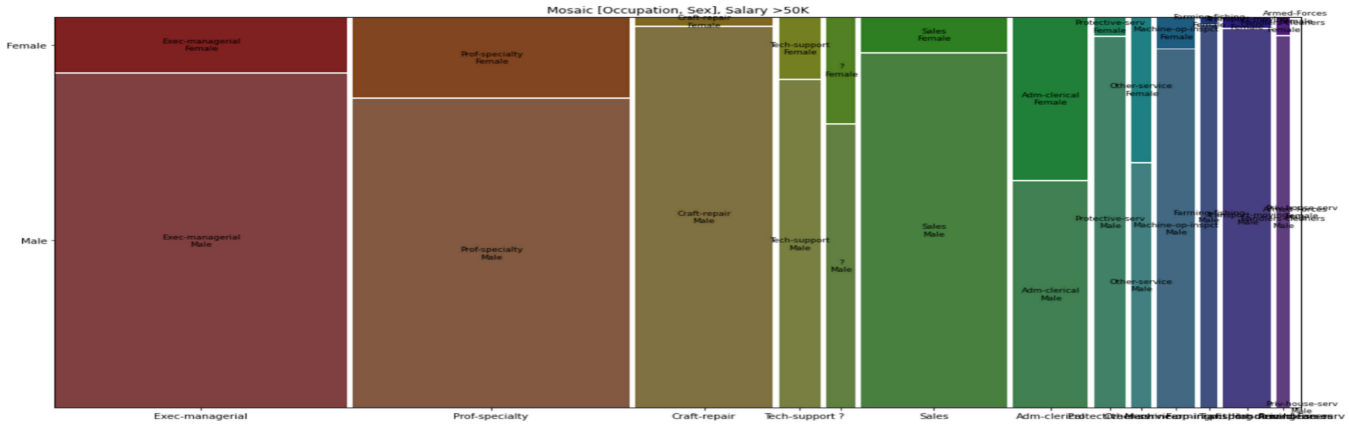


Fig. 2. (d) Occupation distribution based on sex for salary more than $50K

Marital-status

When marital-status is combined with the age factor, lots of non married people who are in their 20's earn less than $50k but when we compare the population who earn more than $50k and have never been married they seem to be in their late 30's and early 40's. While people in other categories such as married-separated and widowed, the age of the population who earn more than $50k are in their late 40's and above.

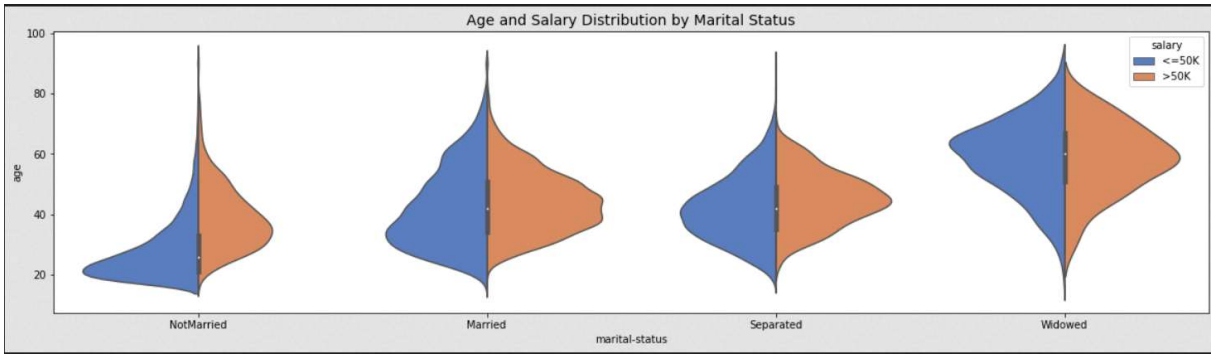


Fig. 3(a) Age and Salary distribution by Marital Status.

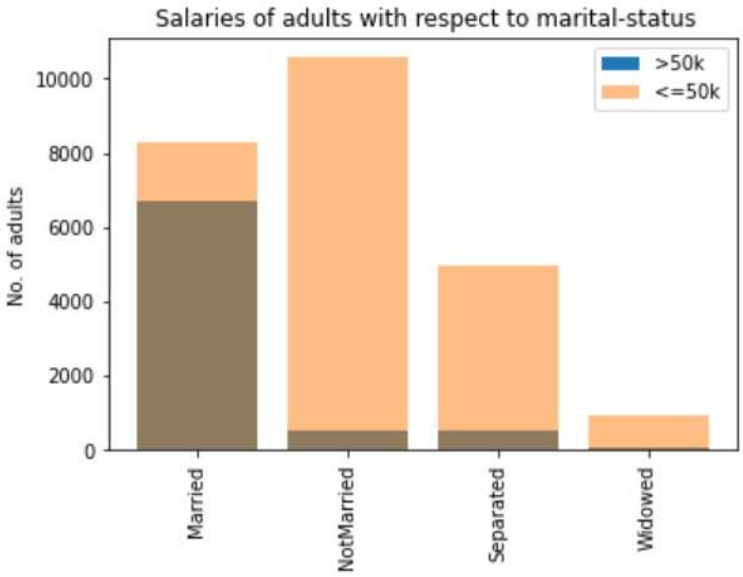


Fig. 3(b) Salary distribution with respect to marital status.

Fig. 3(b) shows that in the data, people who are married have the highest ratio of people earning greater than $50k compared to those earning less than $50k. So, we can infer from this that being married increases the chances of earning more than $50k.

Age

The dataset was filtered for people with age greater than 16. Therefore the minimum age in the dataset is 17. From fig 3, it can be observed that people in the age group of 17-25 are most likely to receive a salary less than or equal to $50k. There is a relatively higher chance that individuals in the age group 25 to 65 are more likely to receive a salary greater than $50k. There are fewer people above the age of 65 with a salary. This may be due to the fact that many people in skilled occupations retire around 65 years (as observed from fig. 4)

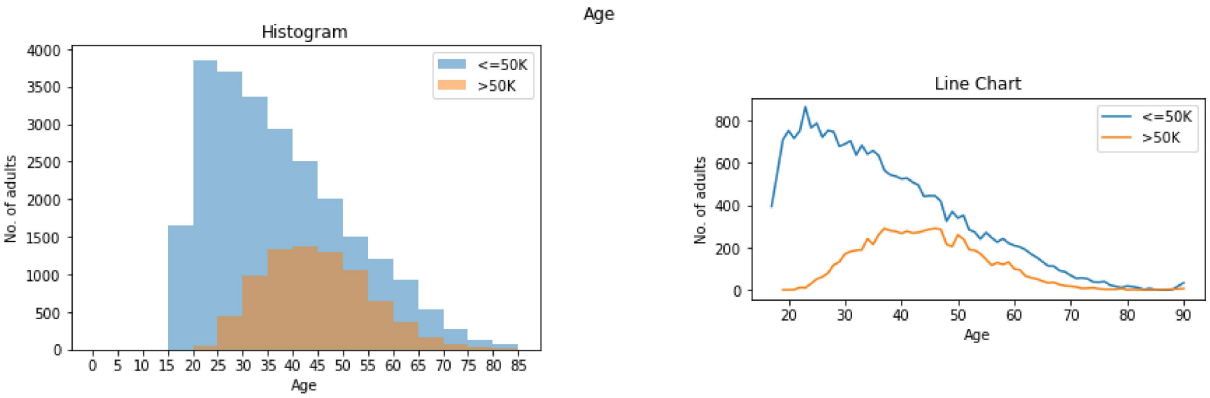


Fig. 4 Age distribution over salary

Education, education-num

Education-num is the numerical representation of the highest educational qualification attained by an individual. Table 2 shows the corresponding education level for each value in education-num.

Table 2 Education and education-num

education-num	education
1	Preschool
2	1st-4th
3	5th-6th
4	7th-8th
5	9th
6	10th
7	11th
8	12th
9	HS-grad
10	Some-college
11	Assoc-voc
12	Assoc-acdm
13	Bachelors
14	Masters
15	Prof-school
16	Doctorate

There seems to be a positive correlation between salary and education. It is observed from Fig. 5 that as people obtain higher valued education, they get paid more.

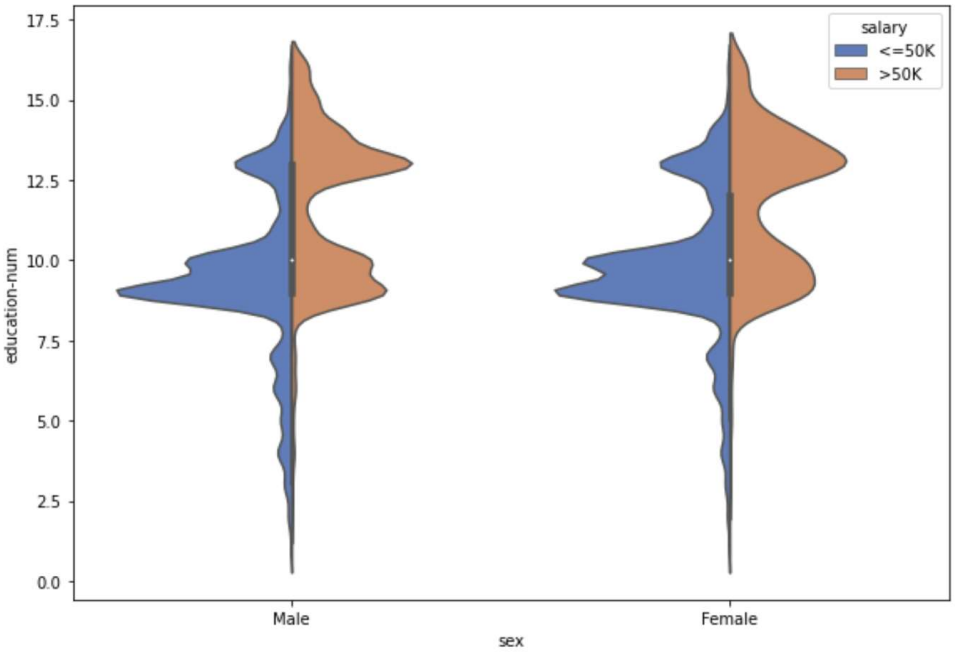


Fig. 5 A comparison between the salaries among the sexes at various education levels

Fig. 6(a) shows the salary among males and females with the maximum educational qualification as a High-school grad. The ratio of individuals earning more than $50k to less than $50k is quite less among both males and females.

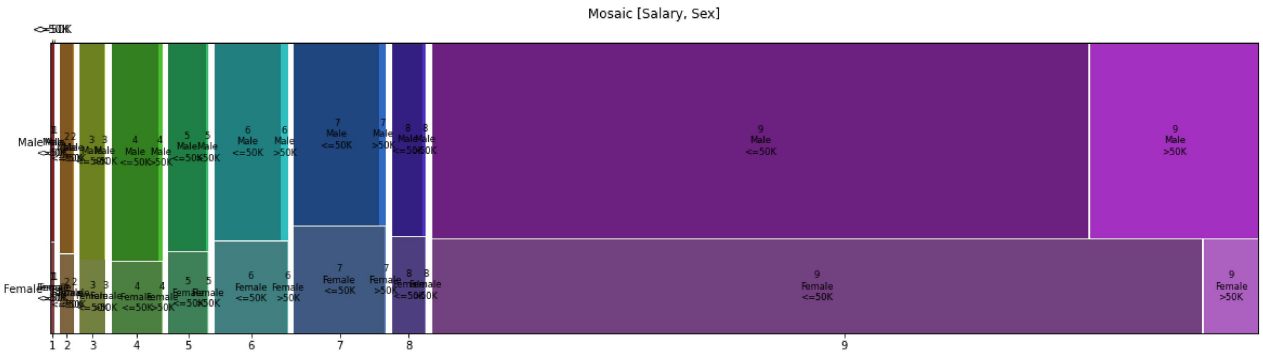


Fig. 6(a) Salary distribution among the sexes [Education upto HS-grad]

Fig. 6(b) shows the salary among males and females with an education higher than a High-school grad. There is a significant improvement in the ratio of individuals earning more than $50k to less than $50k in males. However, this improvement is not quite as evident in females with higher education.

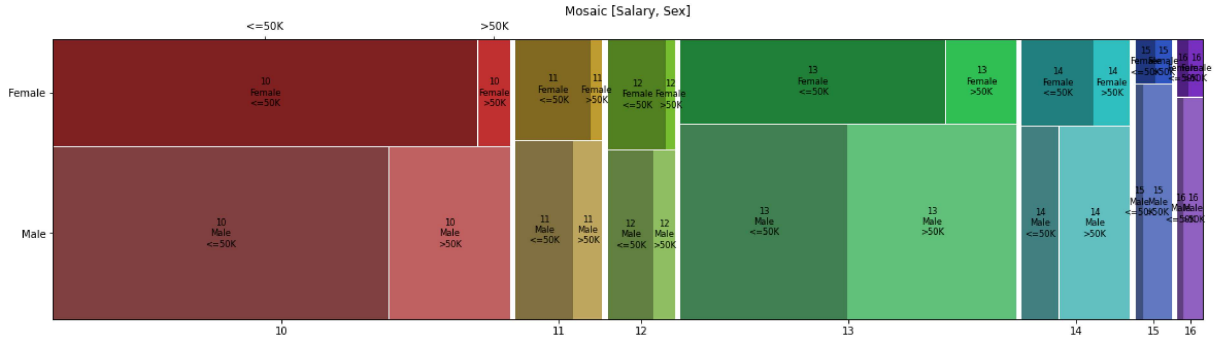


Fig. 6(b) Salary distribution among the sexes [Education more than HS-grad]

Fig. 6(c) shows the ratio of number of people earning more than 50k to less than 50k is only more in cases of masters and doctorate degree holders. For all other education levels, the ratio is lesser i.e more people are earning less than 50k compared to more. So, from the data, having a degree of master's or above increases the chances of earning over $50k.

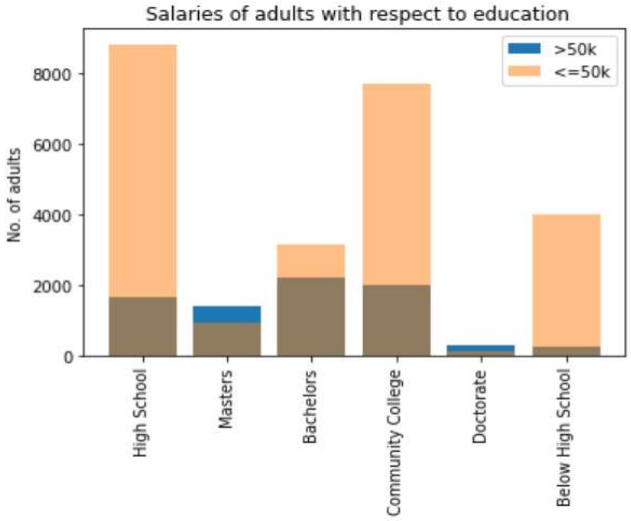
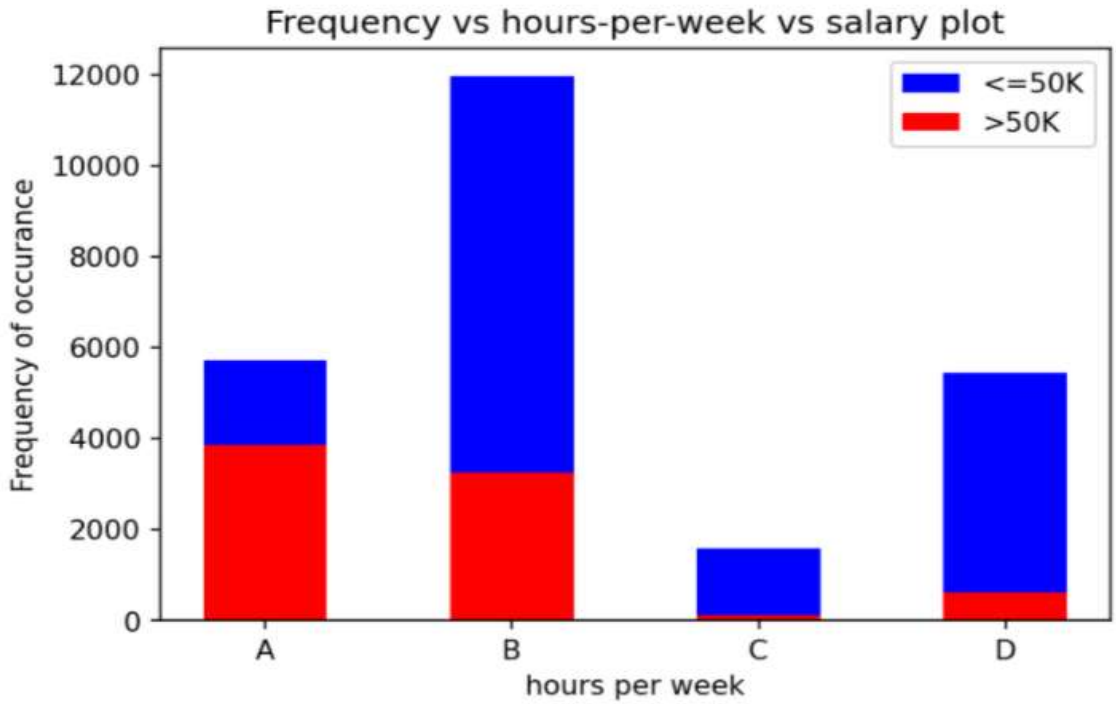


Fig. 6(c) Salary distribution with respect to education

Hours per week

In the overall perspective, the number of people doing full-time seems to be the highest. Unfortunately, the people doing full-time are also earning less than 50k which amounts to about (12000 - 4000) = 6000 people from the equally distributed sample of 25000 people. This amounts to about 25% of the sample.

Also, the percentage of people doing part-time seems to be significantly lower compared to the ones performing full-time and one could see equal distribution for Over-time and contract based work in the acquired sample.



- A = Greater than 40 hours (Over-time)

B = Equal to 40 hours (Full-time)

C = Between 20 and 40 hours (Part-time)

D = Less than 20 hours (Contract)

Fig. 7 Bar plot depicting the count of adults within categorized salary ranges.

Obviously, we could see the majority of the over-time sample earning greater than $50k with a certain percentage still finding it difficult to make ends meet.

Race and Native Country

In the dataset, we observe a similar pattern in Race and Native country features. There is a skewness in the data towards a single category. The other values were combined into a single category to compare with the dominating category. It is observed that even in the other categories, there are more people earning a salary less than $50K.

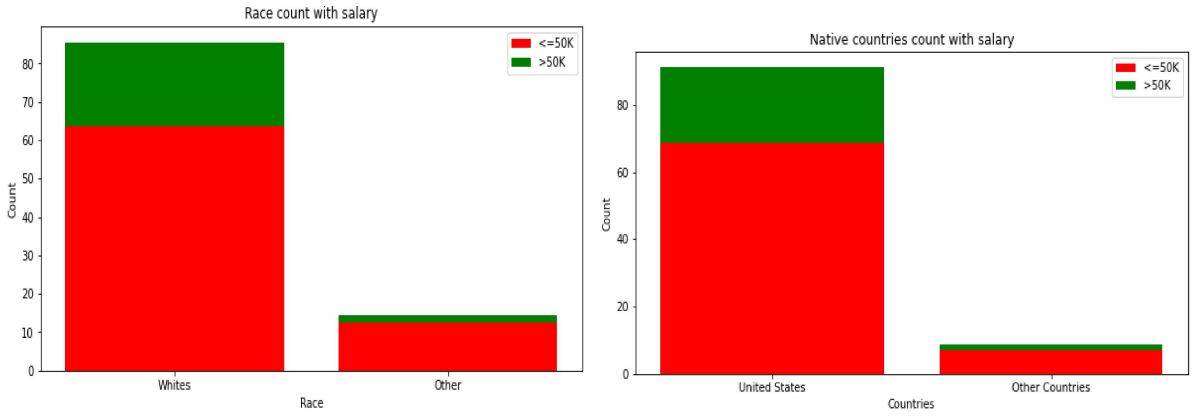


Fig. 8 Race and native country against salary.

Workclass

For workclass, the data distribution is very skewed. For the categories Never-worked and Without-pay, there is only very little data available (14 for without-pay and 7 for never-worked) and from the name it is understood that their salary is less than $50k. The private sector dominates the dataset in the workclass feature. It cannot be used properly to predict the salaries except when it comes to two categories: never-worked and without-pay.

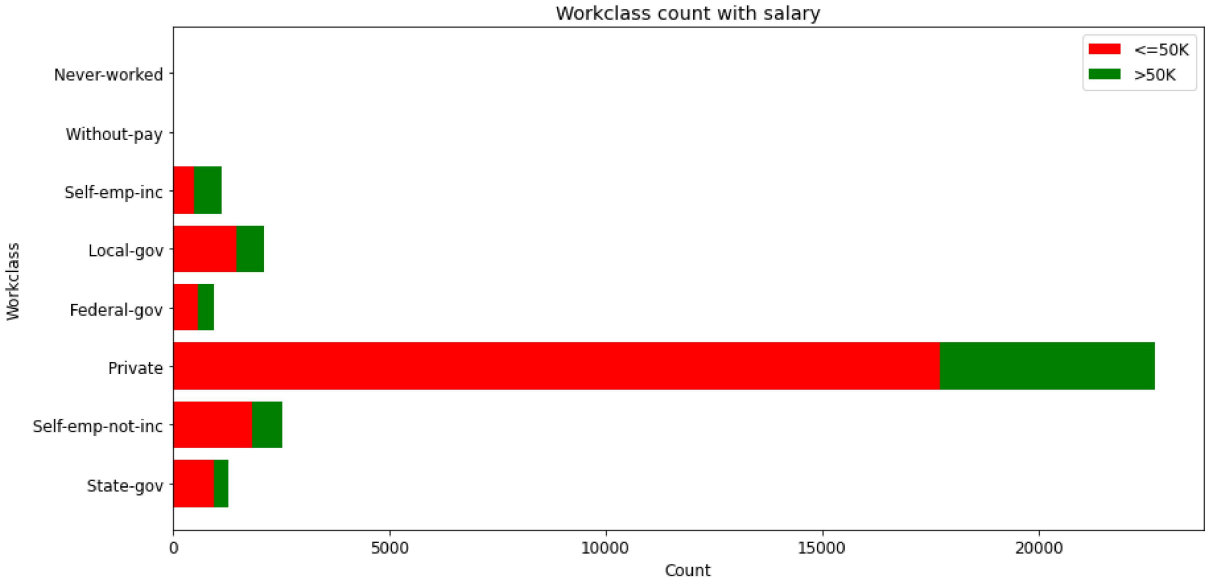


Fig. 9 Workclass count with salary.

Tools and Packages

Programming Language: Python
Libraries: Matplotlib, Pandas, Seaborn, Plotly, Statsmodels.

Questions

1. *How can we approximate the relationship between features in the dataset?*

   To find the relationship between variables a correlation matrix was created. Correlation is a statistic to measure the degree of relation between two variables. Using this we approximated the features that have a relatively higher influence on the salary label.

2. *How to infer useful information from features where there is more skewness?*

   We found two features where the data is more skewed towards one category. For analysing the data, the non-dominating categories were aggregated and compared against the dominating category.

Future Works

- Income prediction model

A machine learning model can be built to analyze the dataset for prediction of the salary above or below the $50,000 mark. Pipelines can be built using the pre-existing models like the Naive Bayes, Decision Tree, Ensemble, Logistic Regression and the like. This can be done by choosing the 5 best parameters influencing the trend by performing a Principal Component Analysis. The dataset has already been divided into train and test which is in the ratio 4:1.