GROUP ASSIGNMENT REPORT

Submitted By: Group 16

**Introduction:**

The US Adult Census dataset is a repository of 48,842 entries. This data was extracted by Barry Becker from the 1994 US Census database. We explored the data at face value in order to understand the trends and representations of certain demographics in the corpus.

The data has been extracted from UC Irvine Machine Learning Repository.

Link to the data: <http://www.cs.toronto.edu/~delve/data/adult/desc.html>

The Dataset**:** The Census Income dataset contains the following information about an individual:

1. **age**: The age of an individual.

* Integer greater than 0

1. **workclass**: A general term to represent the employment status of an individual.

* Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

1. **fnlwgt**: Final weight. In other words, this is the number of people the census believes the entry represents.

* Integer greater than 0

1. **education**: The highest level of education achieved by an individual.

* Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

1. **education-num**: The highest level of education achieved in numerical form.

* Integer greater than 0

1. **marital-status**: Marital status of an individual. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces.

* Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

1. **occupation**: The general type of occupation of an individual.

* Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

1. **relationship**: Represents what this individual is relative to others. For example, an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status.

* Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

1. **race**: Descriptions of an individual’s race.

* White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

1. **sex**: The biological sex of the individual.

* Female, Male.

1. **capital-gain**: Capital gains for an individual.

* Integer greater than or equal to 0

1. **capital-loss**: Capital loss for an individual.

* Integer greater than or equal to 0

1. **hours-per-week**: The hours an individual has reported to work per week.

* continuous

1. **native-country**: Country of origin for an individual.

* United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad Tobago, Peru, Hong, Holland-Netherlands.

1. **income**: Income earned by an individual per year.

* >50K, <=50K

**Topic/Objectives:**

**AIM:**

* What factors make an individual earn more than others?
* To predict an individual’s income per year (using the KNN Model) based on several attributes from the census data.
* **Sub-questions:**
* What is the relationship between income of an individual and age?
* Which work class is getting the highest income?
* Which gender has a better chance to get income > 50k?
* What combination of education and gender gives high income?
* Does marital status affect the income of an individual?

**APPROACH:** The approach to tackle the problem was to first clean and prepare the data for analysis. This was followed by a univariate analysis to study one attribute at a time and see its distribution in the data. Next, a bivariate analysis was done to study how different variables are related to an individual’s income. Finally, we used this information to form a model to predict whether an individual made more or less than $50,000 per year.

**Data Preparation and Cleaning:**

The data on the website was initially split into train and test files, so the two files were merged and income was changed to be formatted the same way. With our new data set, there were a total of 48842 rows and 15 columns.

Using data.info(), some of the columns showed ‘?’. We took these values out and replaced them with NaN and dropped all NaN values using dropna(). The new data frame has a total of 45222 rows.

Some columns were dropped: fnlwgt, education, race, relationship, capital-gain, capital-loss and native-country. Fnlwgt was removed due to it not being useful for the analysis. Education and education-num are similar, one being categorical and one being continuous. In our case, education-num was kept as it is an easier data to analyze.

Race and native-country were dropped, as we noticed that the count for white people was significantly higher compared to the other races. Thus, race and native-country were dropped. Capital-gain and capital-loss were dropped due to the data mostly showing zeros.

**Analysis:**

*Note - refer to Appendix for all Figures referenced in this section of report*

The variables included in the analysis are provided in (**Table 1)** broken by numerical and categorical attributes.

**Table 1-Variables in Dataset for Analysis**

| **Categorical** | **Numerical** |
| --- | --- |
| Workclass | Age |
| Marital-Status | Education-Num |
| Occupation | Capital-Gain |
| Relationship | Capital-Loss |
| Race | Hours-per-Week |
| Sex | Age Bins |
| Native-Country |  |

Univariate analysis was performed on each attribute in the dataset (df) to determine patterns within the raw data. For Income, the dataset distribution was 75.22% to 24.78% for <=50k to >50k respectively (**Table 2**). This distribution was taken into consideration in our predicted modeling.

**Table 2-Distribution of Income in Dataset**

| **Income** | **Total Number of Instances** | **Percentage of Sample Size (%)** |
| --- | --- | --- |
| **<=50k** | 34,014 | 75.2156 |
| **>50k** | 11,208 | 24.7844 |

Information gathered on the numerical features using the describe function indicates a high variance for Capital-Gain and Capital-Loss which is attributed to the large spread (min and max values). Distribution plot (**Figure 1,** **Figure 2**) of both show most values centering around 0. Given the distribution and the summary of information for these categories, their correlation to income was not further explored and dropped from the dataset.

The numerical attribute, Age, was grouped into intervals of ten years to reduce unique entries and improve pattern recognition. (**Figure 3)** illustrates a distribution of grouped ages (Age\_Bins), with the largest group sampled belonging to the group 30-40. Approximately 87% of individuals surveyed are within the range 20-60 with the mean age being 38 years old.

As seen in (**Figure 4)**, most of the sample size for years of education is between 9-12 years. Approximately 54% of the sample size fell within 9-10 years of education completed. For hours worked per week, (**Figure 5)** distribution plot indicates most people work around 40 hours per week.

Patterns in the surveyed data were also observed in the categorical univariate analysis. (**Figure 6)** depicts sample size based on Work Class, majority of the entries in the census were from the Private sector.

The skewness observed in Work Class was also noted when comparing Sex (**Figure 7**) with double the sample size of male compared to female. The distribution for Occupation is mostly uniform with no obvious bias to report (**Figure 8**).

The count plot for Marital Status (**Figure 9**) reveals most of the sample surveyed was made up of three categories: Never Married, Married-civ-spouse, and Divorced.

Bivariate analysis was conducted to infer the income split based on each feature to determine the feature’s level of influence on income. The percentage of population with income >50k across all features were plotted(**Figure 10 – Figure 16**). Of the Work Class that had an income >50k, the sector self-employed was highest. We noted that those married with spouse had a smaller difference between income (<50k vs >50k) whereas the gap between the two classes of income was larger for those that are not married (**Figure 11**). Different factors such as age, education, etc. may be linked to this logic.

For ratio of Occupation (**Figure 12**), an obvious correlation between Occupation and income >50 k was observed. A similar pattern was also observed for years of education where more education generated more income (**Figure 14**). From this we understand that individuals with lower education typically earn >50k.

The disparity between Sex (**Figure 13**) implies women are being paid less and while not overly surprising we explored it further by comparing Sex to Occupation and Education. In Occupation, Sex seemingly plays a role in Income for Protective-serv, Sales, and Tech-support (**Table 3**). For years of education, there is a higher ratio of Male that continue their education past the 9th year relative to Female. Based on this, we deduced gender has a large bias on income.

**Table 3-Gender Distribution of Income based on Similar Occupations**

|  | **Female (%)** | | **Male (%)** | |
| --- | --- | --- | --- | --- |
| **<=50k** | **>50k** | **<=50k** | **>50k** |
| **Prof-Spec** | 74 | 26 | 44 | 56 |
| **Protective-serv** | 88 | 12 | 66 | 34 |
| **Sales** | 93 | 7 | 62 | 38 |
| **Tech-Support** | 88 | 12 | 60 | 40 |

**(Figure 15)** displays the percentage of individuals making >50k based on hours per week worked. A positive correlation between hours worked and income was noted when considering the hours worked to be <40, 40, >40 hours where, as hours increased so did the ratio of income classes >50 k. This can be explained as being full time jobs typical equates to a stable career. Furthermore, working more hours directly translates into more income earned.

**(Figure 16)** shows how there is a lower percentage of individuals earning >50k under the age of 30 with a peak from ages 40 to 60. This can show how people’s careers do not get fully started until the late 20s while the lower percentage in the older population can represent how retirement allows them to stop working which in turn prevents further income.

**(Table 4)** shows the age bin along with the mean for education and hours worked per week. From this we can observe how the average age for people earning over 50k is higher than people earning less than 50k. We can also see that people earning over 50k went to school for longer and work more hours per week.

**Table 4-Education mean and Hours worked per week mean per Age Bin**

| **age\_bins** | **education-num** | **hours-per-week** |
| --- | --- | --- |
| 10-20 | 8.520553 | 28.096024 |
| 20-30 | 10.119288 | 39.727418 |
| 30-40 | 10.344317 | 43.485803 |
| 40-50 | 10.589489 | 43.658158 |
| 50-60 | 9.949254 | 42.749186 |
| 60-70 | 9.602622 | 36.563291 |
| 70-80 | 9.084291 | 28.275862 |

Based on the inferred relationships between the Attributes and Income, we hypothesize that Age, Sex, Occupation, Years of Education, and Marital Status have varying degrees of influence on Income.

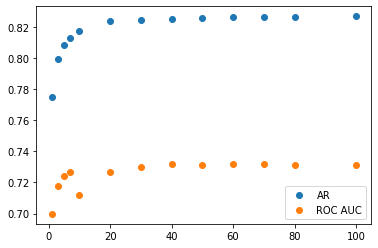
**KNN Model:**

We selected Age, Sex, Occupation, Work Class, Years of Education, Hours worked per week, and Marital Status to build KNN model to predict if an individual’s income>50K.

Below are the steps:

1. OneHotEncoder was used to transform categorical features into binary numeric values per unique feature value.
2. StandardScaler was used to standardize numeric features.
3. Transformed data frame was split into 70% train and 30% test datasets, train dataset used to train KNN model for a given hyperparameter n\_neighbors, and test dataset were used to predict results
4. Step 3 was repeated for 10 times for each hyperparameter n\_neighbors used in KNN model to get an average accuracy score and ROC AUC score as a measure for model performance
5. Hyperparameter n\_neighbors were fine-tuned by repeating Step 3 and 4

Our observation in the chart below was both average accuracy rate and ROC AUC score plateaued at 82.7% and 0.73 respectively when n\_neighbors was set to 60 or greater.



**Conclusion:**

The conclusion we reached from our analysis was that while there were a number of factors that appeared to impact an individual’s income, the factors that were the most prevalent to us were Age, Sex, Occupation, Work Class, Years of Education, Hours worked per week, and Marital Status. These factors were chosen to train our KNN Model.

When it came to answering our previous sub-questions, interesting observations were made on which factors saw the highest incomes.

Looking at the different working classes we see that the *self-emp-not-inc* shows the highest representation out of classes that earned >50k at 55.41% followed by *Federal-gov*. This is despite both being overshadowed by the *Private* working class when looking at the entire dataset population.

Marital Status showed an interesting story where individuals that were *Married-civ-spouse* and *Married-AF-spouse* were both nearly 4 times as prevalent as other marital statuses at 45.42% and 43.75% respectively when it came to individuals that earned over $50k a year.

Males were significantly overrepresented over Females out of individuals that earned over $50k a year at nearly a 3:1 ratio with 31.24% of the male population earning more than 50k while only 11.35% of the female population managed to do the same.

Age was also another interesting factor to look at as it showed a parabola starting from the age of 20 all the way to 80 with the peak being located from the 40-60 range.

Finally, there was a strong trend where having more years of education as well as higher levels of education increased an individual’s odds of earning over $50k a year with a noticeable increase after 9 years of study.

**Appendix:**

Figure 1 Figure 2

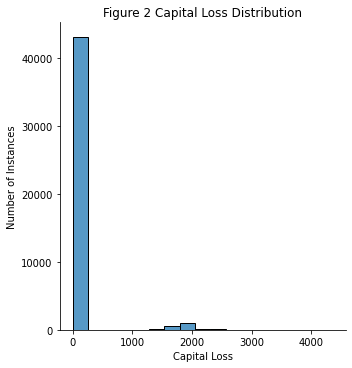
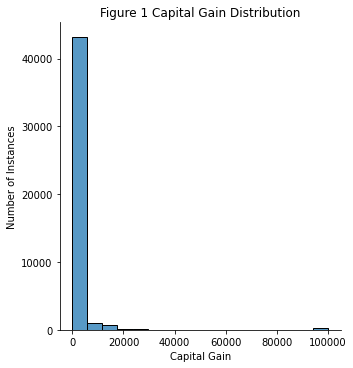


Figure 3

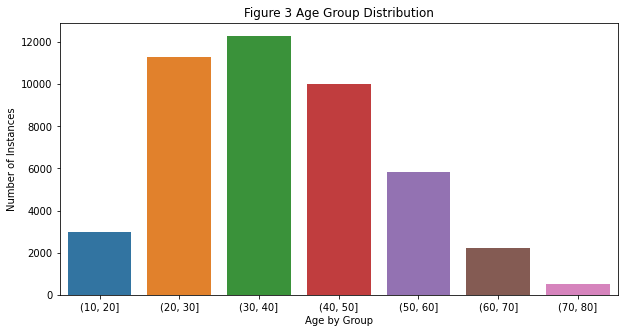


Figure 4

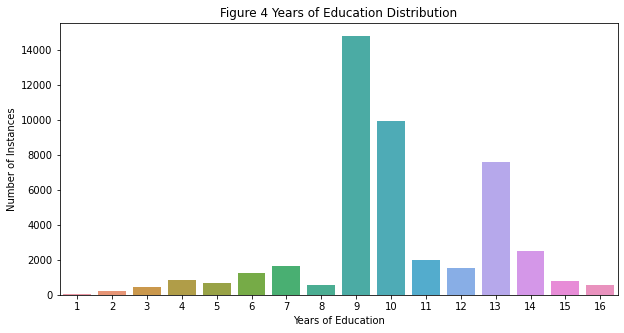


Figure 5

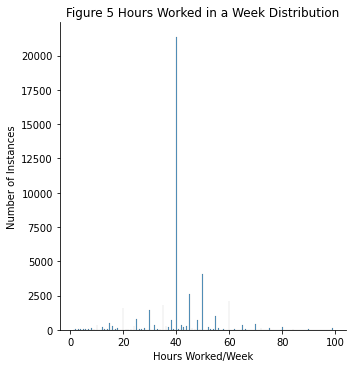


Figure 6

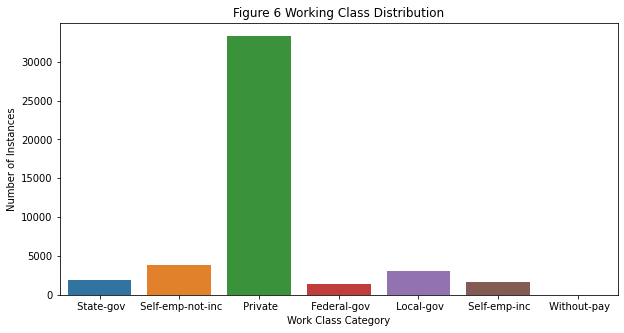


Figure 7 Figure 8

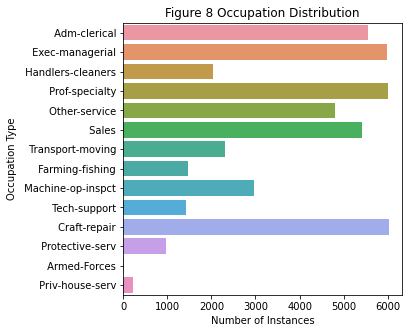
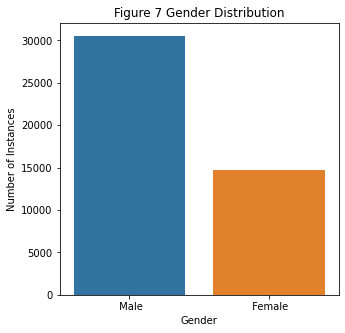


Figure 9

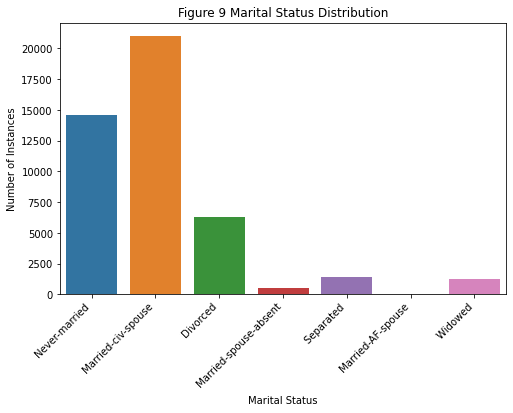


Figure 10

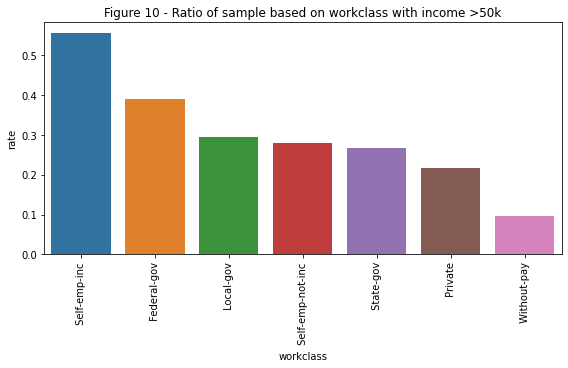


Figure 11

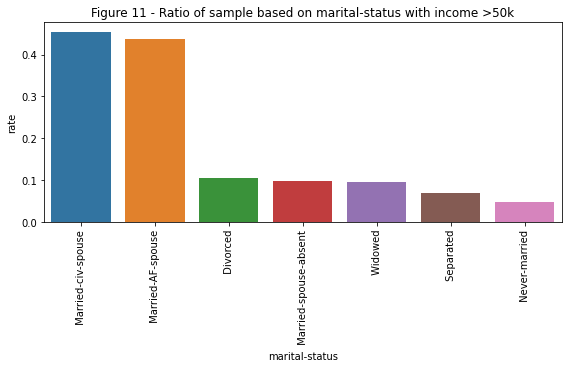


Figure 12

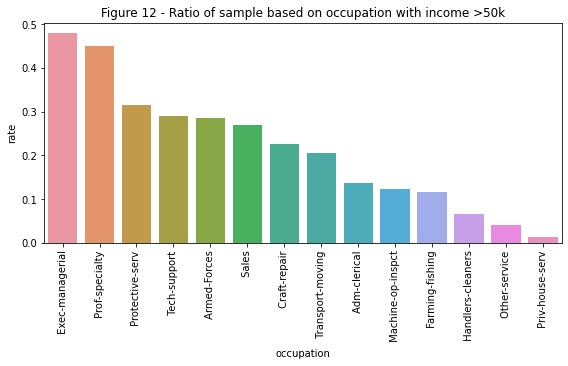


Figure 13

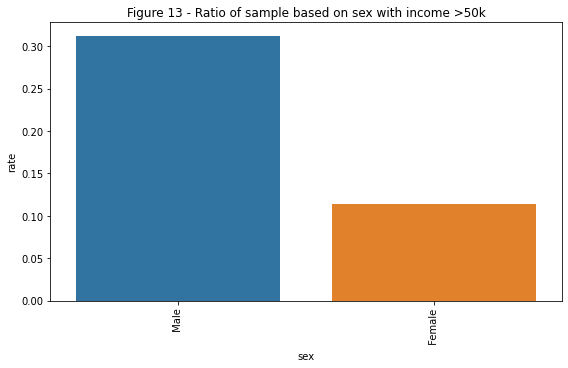


Figure 14

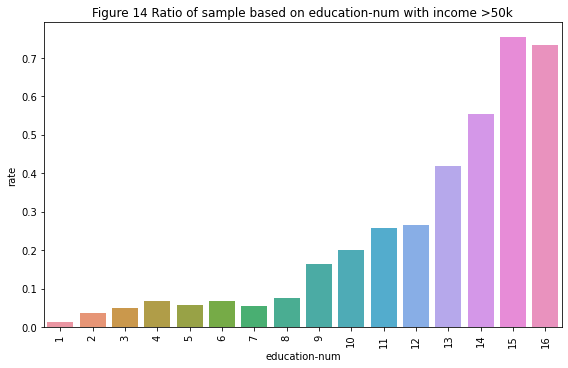


Figure 15

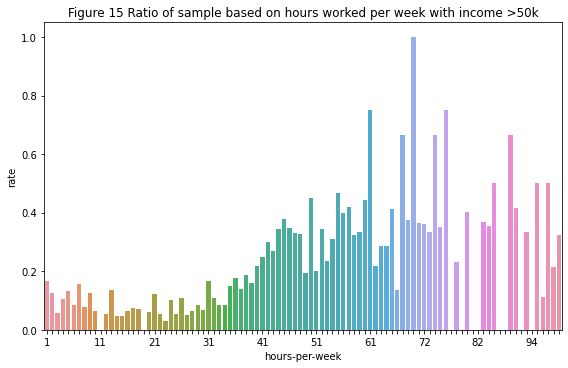


Figure 16

