EXPLORATORY DATA ANALYSIS

Visualization Project

Wajiha Zafar

Introduction:

In this data analytics, I am using the census income dataset, which contains weighted income data extracted from 1994 and 1995 current population survey conducted by the US census bureau and it was retrieved and used by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The clean set of records were extracted using the following criteria: Age > 16, Capital-loss > 100, Hours per week > 10 and final weight > 1. The main purpose for this data was to predict weather person is making income more than 50k per year.

Data source and Data collection:

This data set is downloaded from the UCI Machin learning open data repository: https://archive.ics.uci.edu/ml/datasets/adult. This data set contains 14 attributes related to demographics and employment. The weights of the current population survey indicate the number of people in each record of population survey are controlled to independent estimation of non-institutional US population. These records are prepared on monthly basis for the US population division in census bureau. For the real analysis, the final weight field must be used but its not the suitable attribute for the classifiers. The whole dataset consists of 199523 instances while the test file is comprised of 99762 instances. The data was split into train and test file with 2/3 and 1/3 approx. proportions.

The minimum estimated age of the population is 16+. The data is characterized or controlled based on race, age and sex. The data sample is collected from the 51 states, each with its own probability for the selection. Most of the predictive/independent variables are categorical with many levels while response/ in-depended variable is considered binary.

The order of tasks performed in this data analytics case is:

Acquire and read the data set: The dataset is read directly through the UCI Machine learning repository.

Clean the dataset: The real-world datasets are messy and noisy with a lot of missing and null values. To do a proper exploratory analysis we clean the data to make it in representable form.

Exploring the Predictive Variables: In analysing a dataset, the most important step is exploring the independent variables. The exploratory analysis helps in predicting and analysing the effect of independent variable to the response/ dependent variable. By looking at the distribution of variable, how the response variable is acting, which variable is affecting which way and what skew it has. The relation between independent and dependent variables and visualizing them.

Data Description:

This data set is downloaded from the UCI Machin learning open data repository: https://archive.ics.uci.edu/ml/datasets/adult. This data set contains 14 attributes related to demographics and employment. Most of the predictive/independent variables are categorical with many levels while response/ in-depended variable is considered binary.

Dependent Variable: The response/ dependent variable is the level of income which is binary variable with value of >50k mean adult income greater than 50000 and <=50k means adult income less than 50000.

Independent Variable:

Below are the independent variables (features or predictors) from the Census Data

Variable Name	Description	Type	Possible Values
Age	Age of the individual	Continuous	Numeric

Workclass	Class of Work	Categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt	Final Weight Determined by Census Org	Continuous	Numeric
Education	Education of the individual	Ordered Factor	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-num	Number of years of education	Continuous	Numeric
Marital-status	Marital status of the individual	Categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Occupation of the individual	Categorical	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
Relationship	Present relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	Race of the individual	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Sex of the individual	Categorical	Female, Male
Capital-gain	Capital gain made by the individual	Continuous	Numeric
Capital-loss	Capital loss made by the individual	Continuous	Numeric
Hours-per-week	Average number of hours spent by the individual on work	Continuous	Numeric
Native-country	Average number of hours spent by the individual on work	Categorical	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, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

Technologies Used and Preprocessing of Data:

To perform this analysis, I used these python library and packages: Pandas to download and read the dataset, seaborn and matplotlib for ploting/ visualizaing the data. Sk learn, numpy and scikit to perform the cleaning and preprocessing on the dataset.

Download and Read data: The data set is downloaded from the UCI repository by using panda's data frame library and defined the header for the data file which contains the comma separated columns.

Income_adult = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data', header = None, na_values="?")

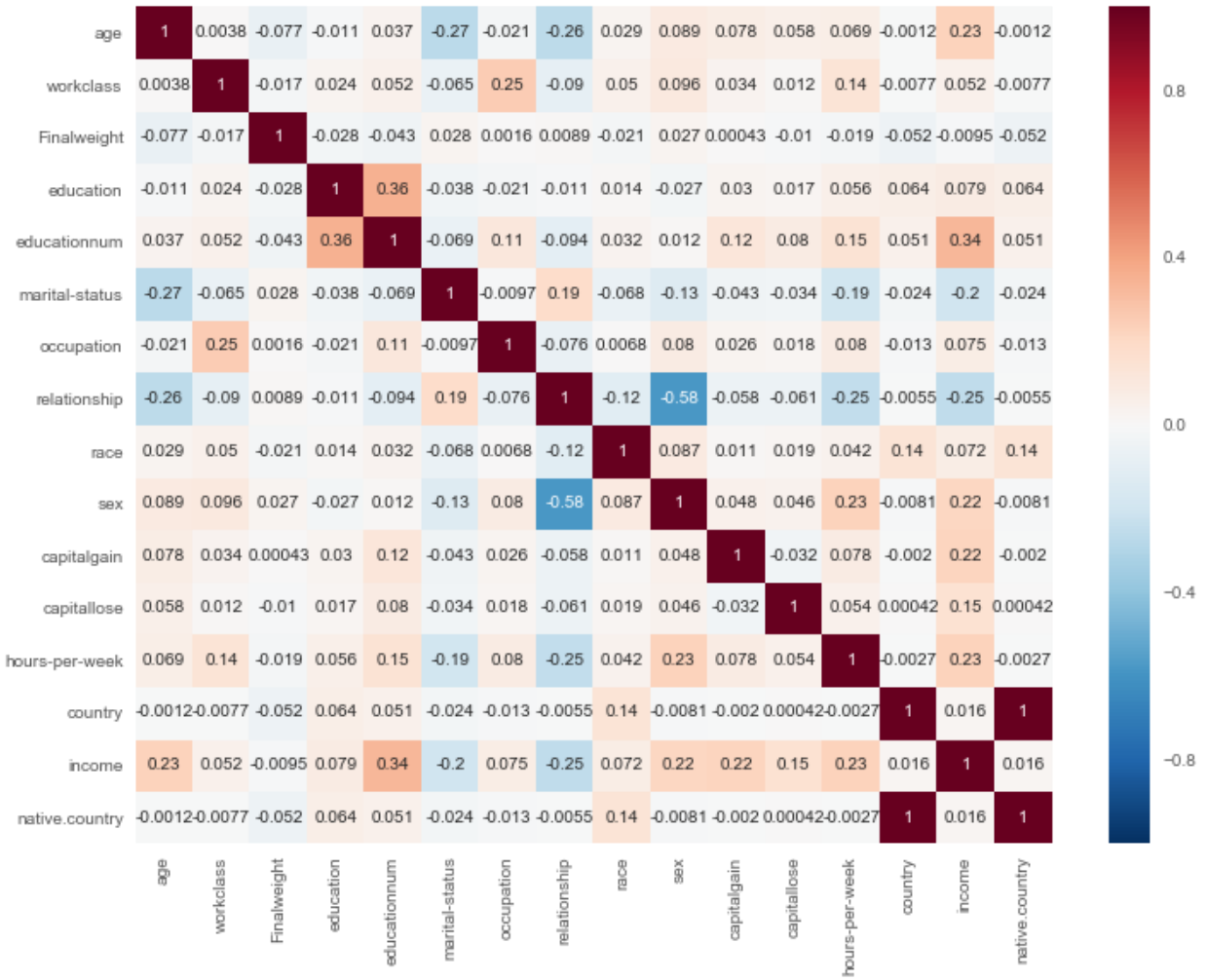
Income_adult.columns = ['age', 'workclass', 'Finalweight', 'education', 'educationnum', 'marital-status', 'occupation', 'relationship', 'race', 'sex', 'capitalgain', 'capitallose', 'hours-per-week', 'native-country', 'income']

Cleaning of Data: This dataset contains a lot of missing values and invalid values in Education, Occupation, Country, Work class columns. These invalid or missing values can be handled by imputing the value in data but this imputation can cause skew the data as, occupation, education and country can be considering as good variables for the income level study. I used the numpy package to replace and impute the value in dataset. The final weight is estimated total population which was calculated while conducting the survey. As this attribute have non-significant impact on income level so I remove this variable from the dataset.

Preprocessing of Data: This dataset has a lot of categorical variable. To identify the non-linear and linear relationship between the different features, we plot correlation between different features. Generally having too many correlations between the variable is not a good and it may indicate that the dataset is not very good. So, I preprocess the dataset by using the label encoder in sklearn, scikit learn package. I encode the categorical variables to numerical variable and calculated the correlation matrix by using corr() function and plot the heat map.

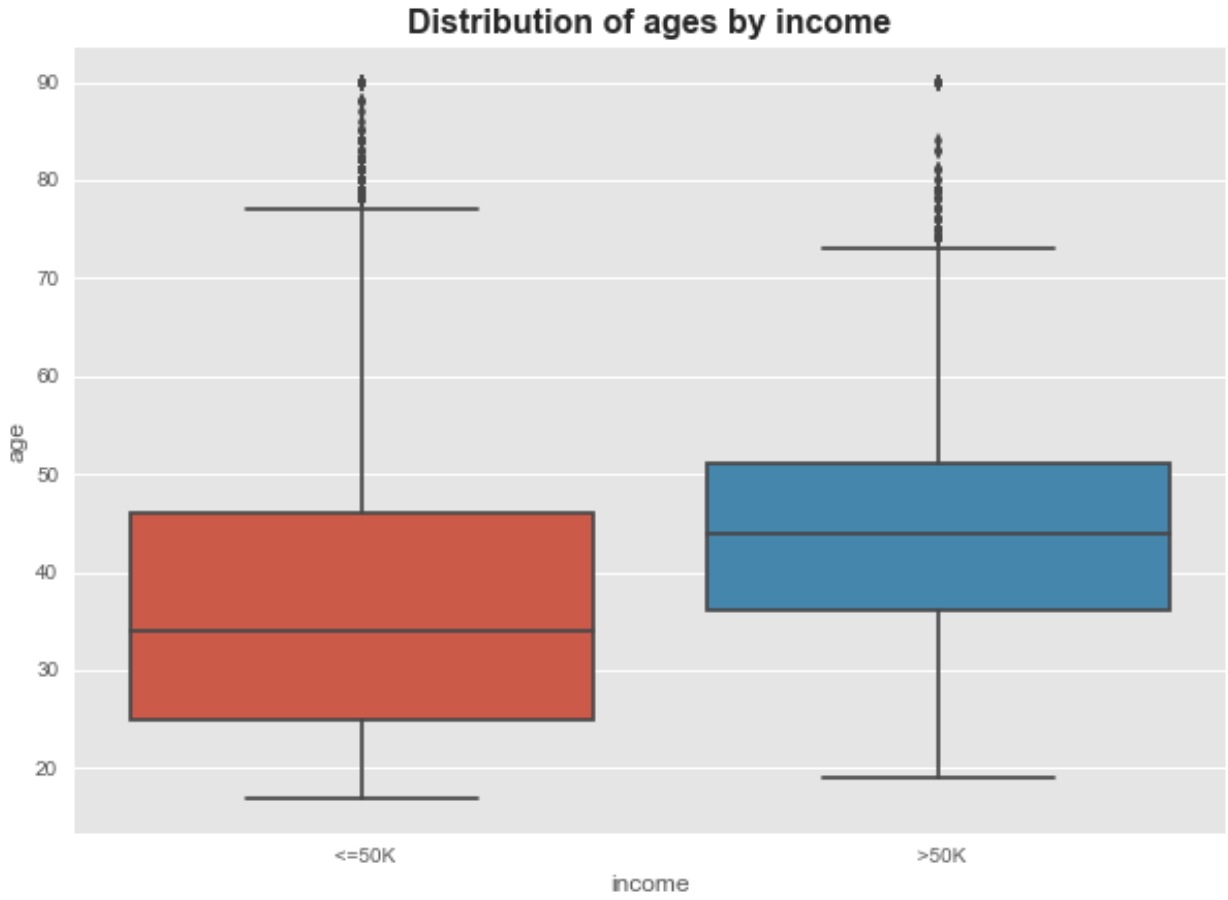
Exploratory analysis and Visualization:

Each of the variables are explored and studied the relationship between independent and dependent variables by calculating the correlation matrix and plotting the heatmap:



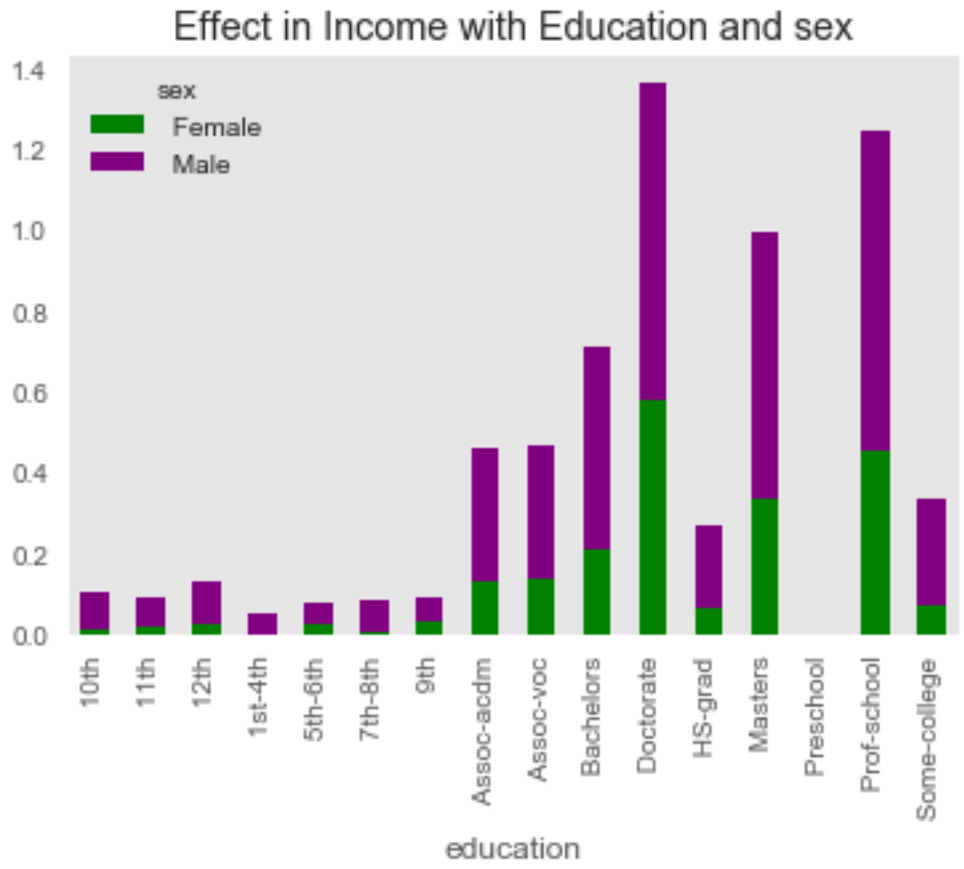
According to visualization, pair of variables that have correlation values closer to 1 and -1 are highly correlated. Correlation values that are closer to 0 indicates a non- linear relationship between variables. According to the correlation plot and matrix, there is slightly positive correlation between the dependent variable and educationnum of 0.34. There is also some correlation between the age and dependent variable of around 0.23. Where-as relationship, marital-status with dependent variable showed negative correlation of -0.25. The highest correlation is seen between the educationnum and income. This shows that the person with highest year of education have higher income. From the correlation plot we can see that education and education number have high correlation of 0.36, which means that the categorical variable education and the numerical variable education number both represents the same features.

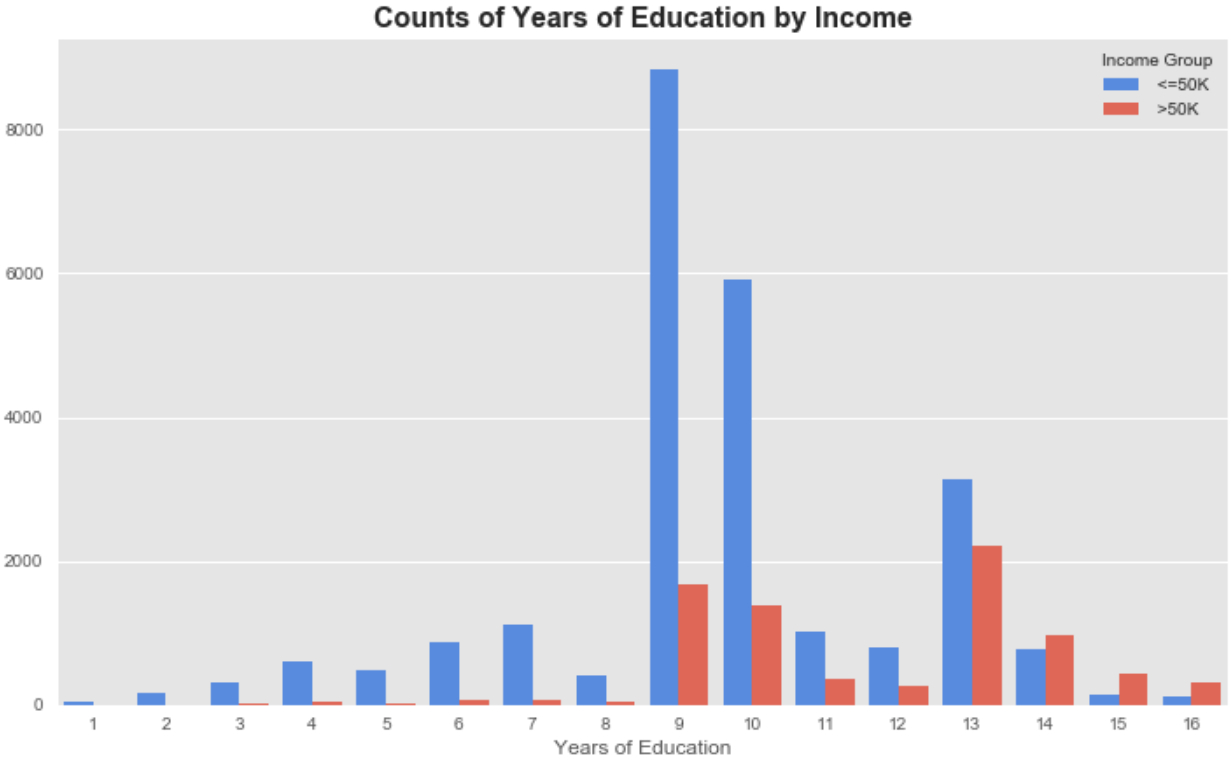
Exploring the AGE variable: The age variable has a wide range of variability. The correlation plot shows that the income and age are positively correlated to each other. So I plot the distribution of Age with the income group.



The box plot between income and age shows that the distribution and mean of ages are different for both income levels. The people count is high in income with less than 50000k group. This shows the age is a good predictor for the income level.

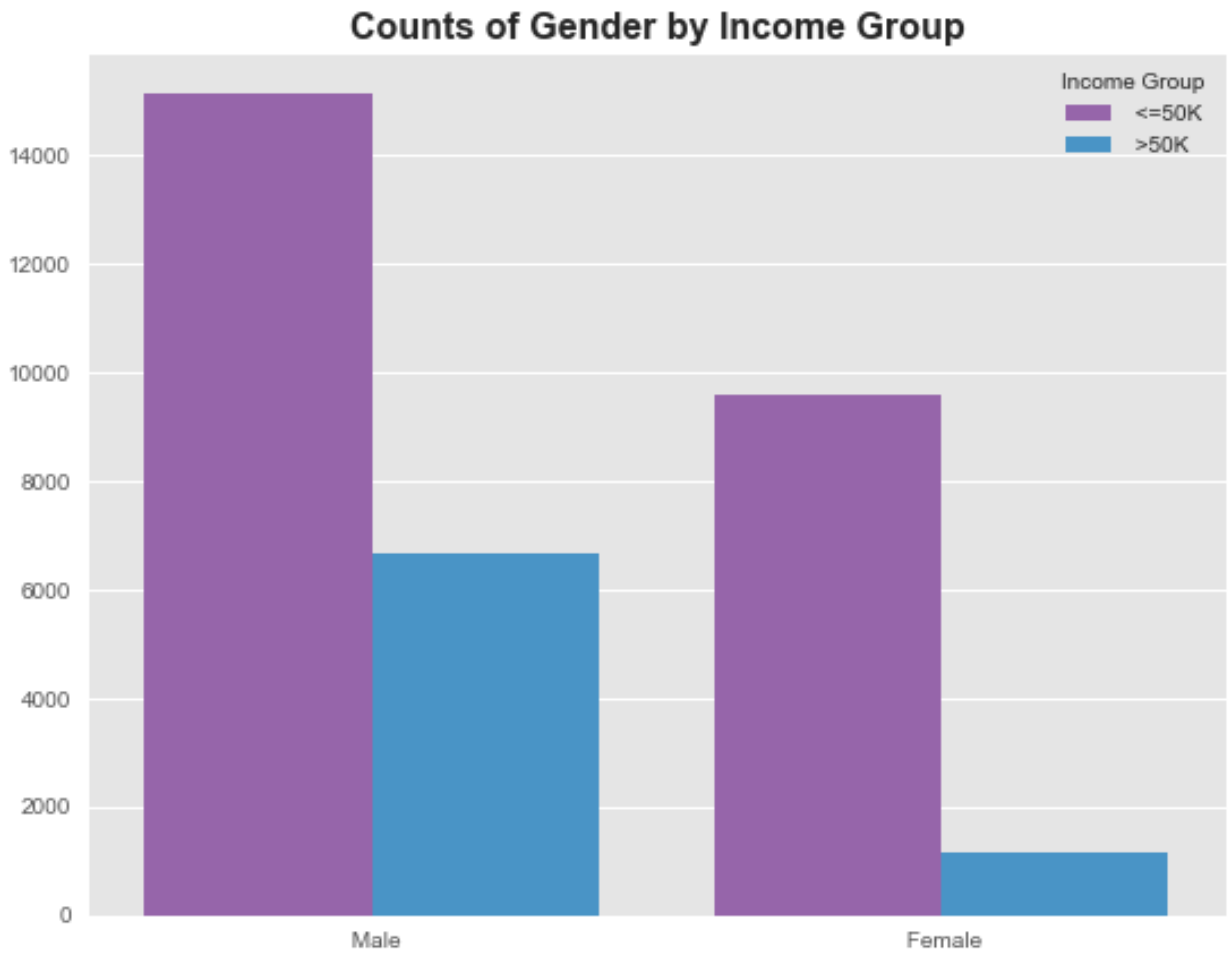
Exploring the Education variable: As per the correlation plot, the education number and income are highly correlated.


Effect in Income with Education and sex
sex
Female
Male
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
10th
11th
12th
1st-4th
5th-6th
7th-8th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Preschool
Prof-school
Some-college
education


Counts of Years of Education by Income
Income Group
<=50K
>50K
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Years of Education

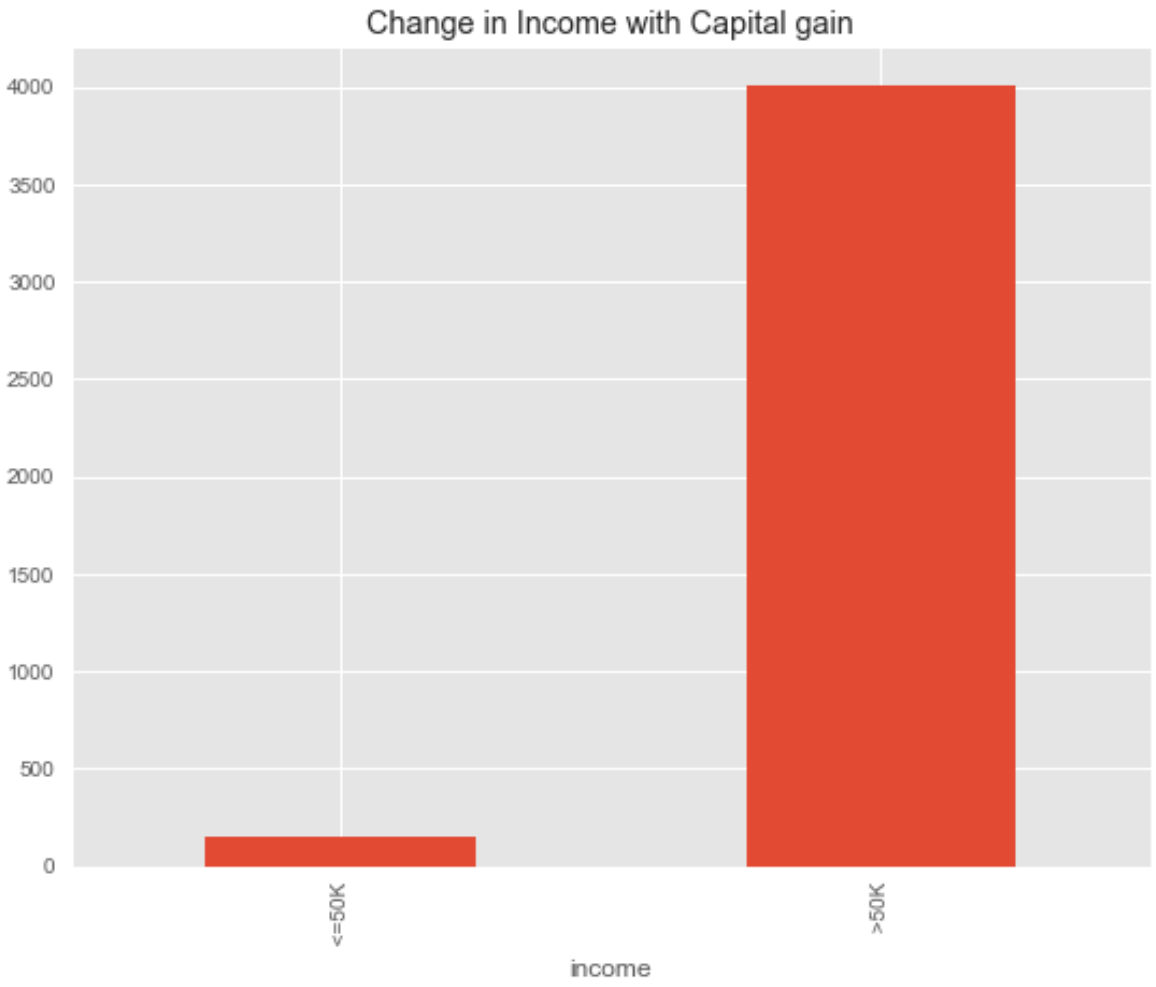
After visualization the education number and Education factors with income and number of people surveyed, it shows that the high number of people in High school and college graduates earn less than 50k. But the ratio of female to male gender looks little biased. The average Male adults earn more than female and males with doctorate and Professors degree have salary around or greater to 50k. This can be reason that the number of people participated in survey are male.

Exploring the Gender variable: Generally the sex variable is not consider a good predictor but the correlation plot shows that income and sex have positive correlation of 0.22.



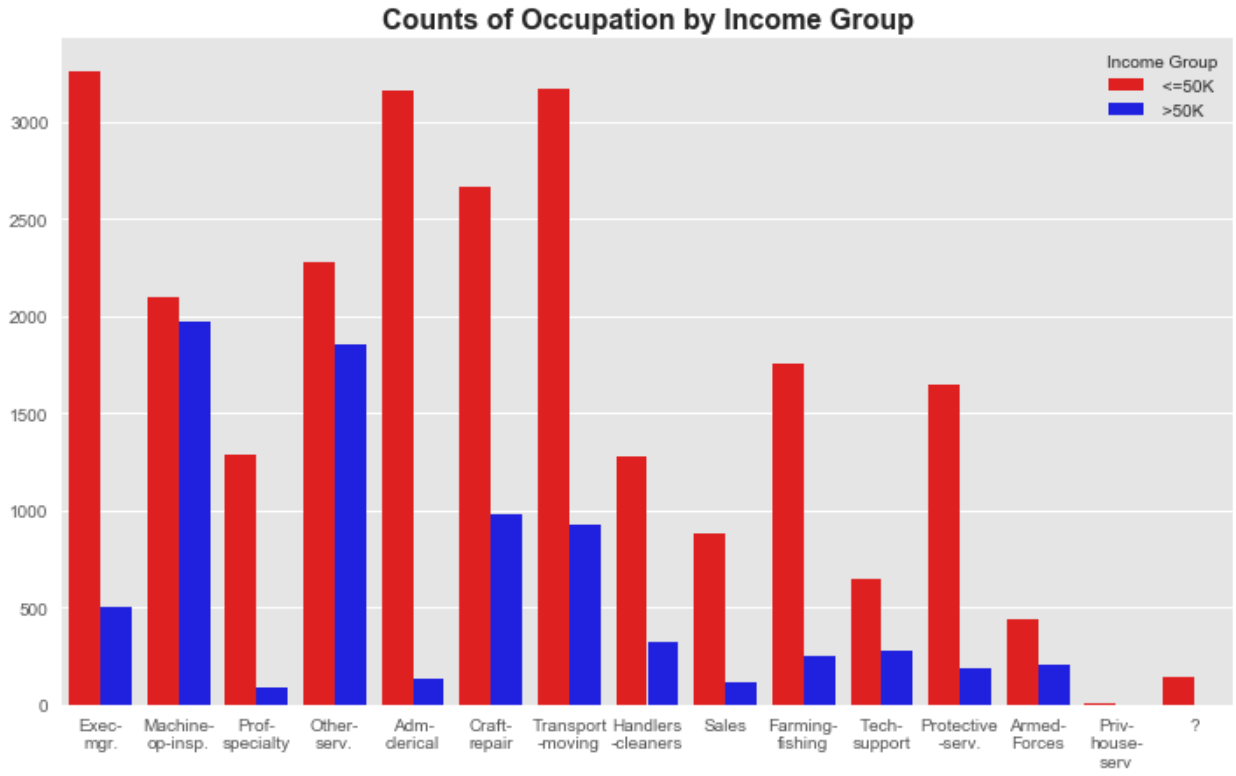
The number of male and female participants earning less than 50k are more than those earning more than 50k. The plot shows that male participants are more than the female participants. This can be possible that at the time of surveyed, male participants can be more than female.

Exploring the capital gain with income:

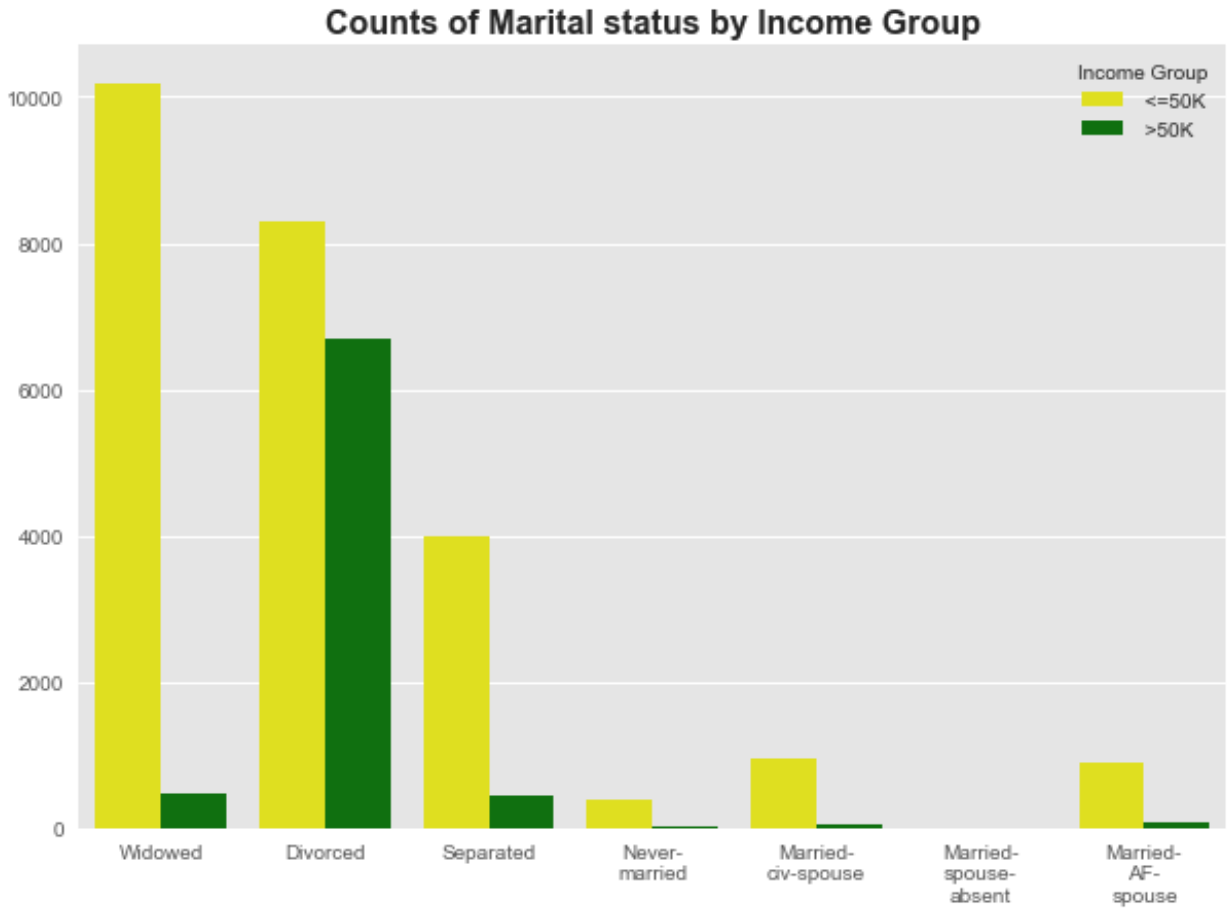


The plot show that average capital gain shows a different level for income. Which mean the capital gain variable can be a good predictor for the income level.

Exploring the Work class, Occupation, Marital Status: According to the correlation matrix and correlation plot the work class and occupation are slightly correlated but the marital status is showing negative correlation with the income.

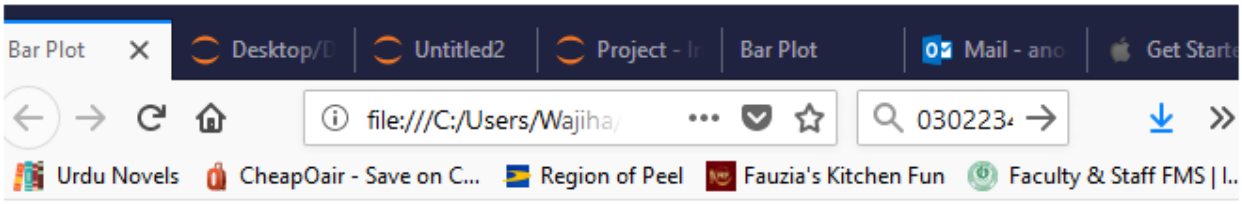


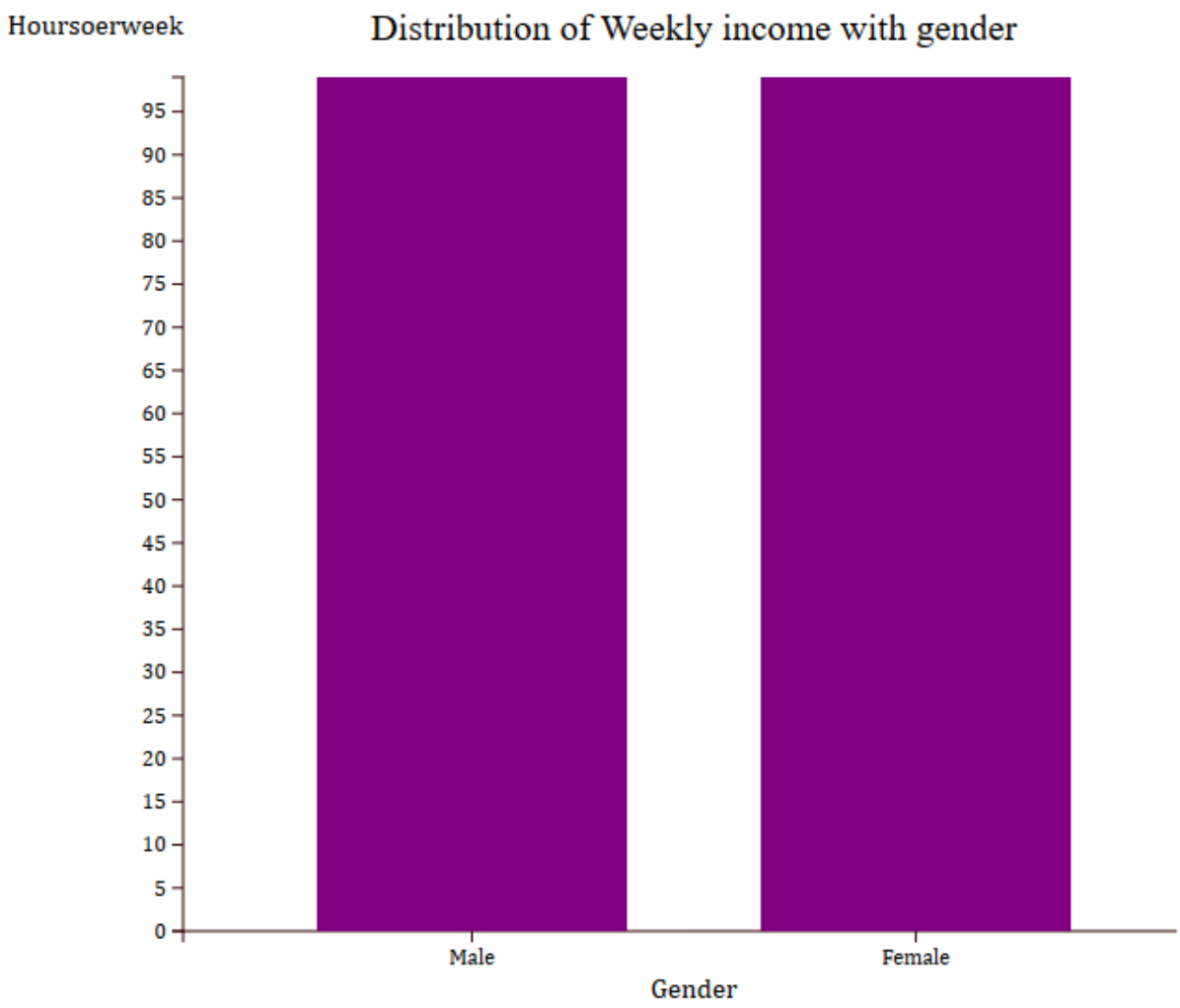
The number of people earning less than 50k are more than earning more than 50k in manager and admin clerk and prof-specialty level. The rest of occupation are showing slightly equal level in income. Now we look at the income level comparison with marital status:



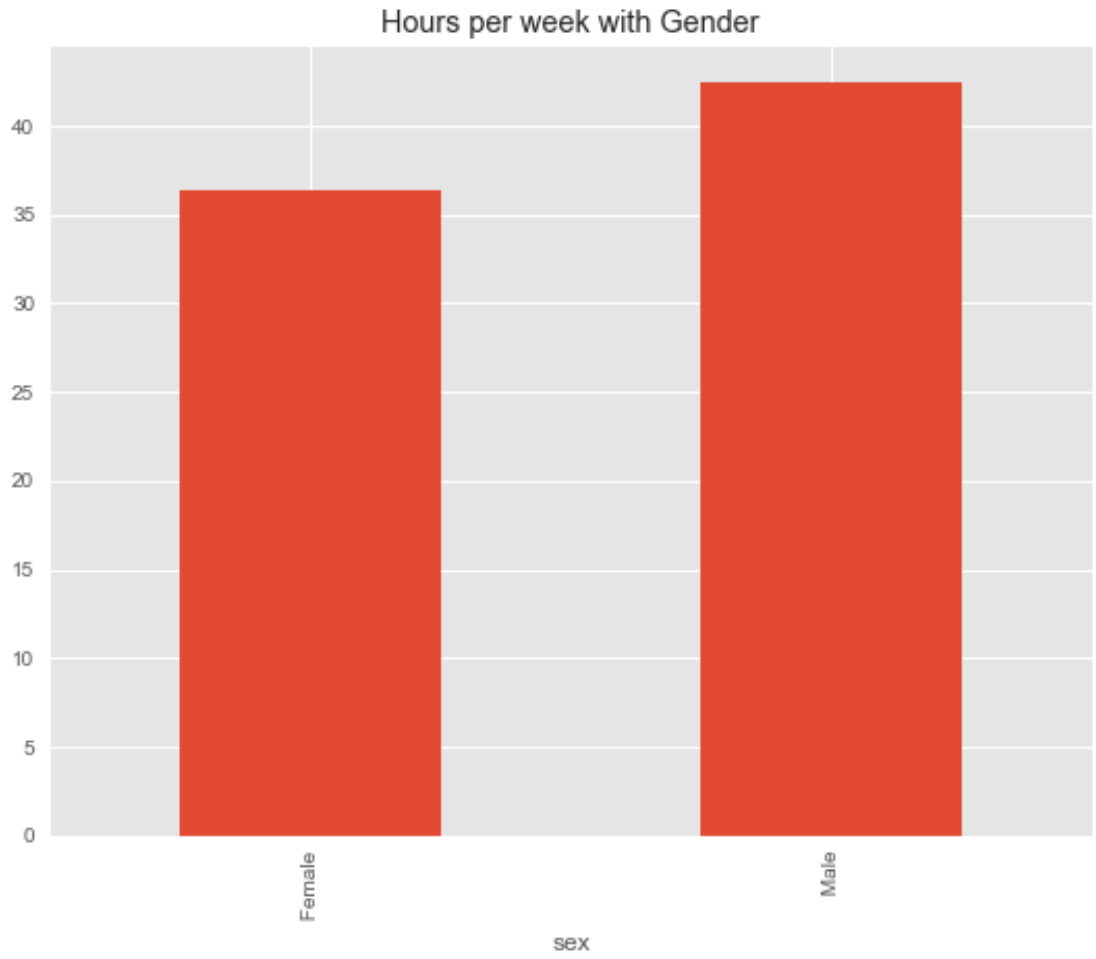
The widowed participants show a great increase in income level less than 50k, while the divorced participants have almost equal in income level. So, this variable is showing slightly negative and less accurate result for income level prediction.

Relationship gender an hours per week: The work class and gender is showing positive correlation in correlation plot, so I plot them in D3 to explore the variables.





But this didn't give any significant result. So, I plot the same plot in python, which shows that male participants earn more hours per week income than females. The reason can be that most of the females are involved in house hold chores and taking care of kids and work less hours than male due to the family and house hold care.



Conclusion and Lesson Learned:

From this data analysis case, it is inferred from the marital status, education, gender plots that Male on average earns more than female. But there is also a case that male participants in data collection can be more than female. The married citizens with spouse have higher chances of earning more than 50k than widowed and divorced. The participants with higher education earn more than the high school and graduate participants, so education leads to higher income in most cases. While looking at the race variable, it shows that the white and Asia-pacific have highest average income.

There is also imbalance in the class of outcome variable, this can be due to the less weight on more than 50k income group. If we increase the weight on more than 50k income group, it will help in getting more accurate and need to involved equal level of male and female participants to predict the accurate income level. The education and education number appear to measure the same outcome and the work class and occupation shows approx. same result with the income level. Thus, they both are correlated to each other more than the outcome variable. In short always designed the question related to the dataset or identify the problem and then answer these questions with visualization.

References:

- https://www.kaggle.com/martinpsz/predicting-income-group-with-logistic-regression/notebook information extracted on April 9 2018
- https://cloudxlab.com/blog/predicting-income-level-case-study-r/ information retrieved on April 8
- https://www.kaggle.com/mshaurya/adult-income-predictor-python-tableau information retrieved on April 9 2018
- https://www.valentinmihov.com/2015/04/17/adult-income-data-set/ information retrieved on April 8 2018
- Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996
- http://www.dataminingmasters.com/uploads/studentProjects/Earning_potential_report.pdf information retrieved on April 9 2018.
- https://bl.ocks.org/mbostock information retrieved on April 9 2018.