



BRACT'S
Vishwakarma Institute of Information Technology

Department of Computer Engineering
2020-2021

Business Intelligence and Data Analytics
Assignment 7(Mini Project)

Submitted By

Name	Roll No	Gr No
Raturaj Tambe	323058	21810362
Amit Zope	323066	21810714
Vedant Padalkar	323071	21920123

Class : TY-Comp
Div : C
Batch : C3
BIDA Practical Batch : E8

URL : <http://vedant7101.pythonanywhere.com/>

Aim

Perform a Mini project for any Financial, Societal, Education problem. Use methodologies, Models, Algorithms, visuals applied in assignments 1 - 6.

Objectives

- Understanding problem statement
- Finding suitable dataset
- Analyzing data
- Finding outcomes

Theory

Classification Algorithms

1. Classification is the process of recognizing, understanding, and grouping ideas and objects into preset categories or “sub-populations.”
2. Using pre-categorized training datasets, machine learning programs use a variety of algorithms to classify future datasets into categories.
3. Classification algorithms in machine learning use input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories.
4. One of the most common uses of classification is filtering emails into “spam” or “non-spam.”
5. In short, classification is a form of “pattern recognition,” with classification algorithms applied to the training data to find the same pattern (similar words or sentiments, number sequences, etc.) in future sets of data.

Examples of Classification Algorithm

1. Logistic Regression
2. Naive Bayes Classifier
3. K-Nearest Neighbors
4. Decision Tree
5. Random Forest
6. Support Vector Machines

We are going to use Random Forest Classifiers.

Random Forest Classifiers

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

Census Income

US Adult Census data relating income to social factors such as Age, Education, race etc. The US Adult income dataset was extracted by Barry Becker from the 1994 US Census Database. The data set consists of anonymous information such as occupation, age, native country, race, capital gain, capital loss, education, work class and more. Each row is labelled as either having a salary greater than ">50K" or "<=50K".

.

This Data set is in **dataset.csv** The goal here is to train a binary classifier on the training dataset to predict the column income_bracket which has two possible values ">50K" and "<=50K" and evaluate the accuracy of the classifier with the test dataset.

Dataset

This dataset contains income of people in United States. Each row represents particular number of people and the number is described in fnlwgt (Final Weight).

Features

1. age - Age of group
2. education.num - Number of years spent on education
3. workclass - Type of job
4. occupation - Occupation of group

- I) Adm Clerical
- II) Executive Managerial
- III) Handlers-Cleaners
- IV) Pro-speciality
- V) Other service
- VI) Sales
- VII) Craft repair
- VIII) Transport Moving
- IX) Farming-Fishing
- X) Machine-op-inspect
- XI) Tech support
- XII) Protective service
- XIII) Armed Forecs
- XIV) Private House Service

5. Workclass

- I) State Government
- II) Self Employed not Income
- III) Private
- IV) Fedaral Government
- V) Local Government
- VI) Self Employed Income
- VII) Without pay
- VIII) Never worked

6. Sex - Sex of group

- I) Male
- II) Female

7. capital.gain

8. capital.loss

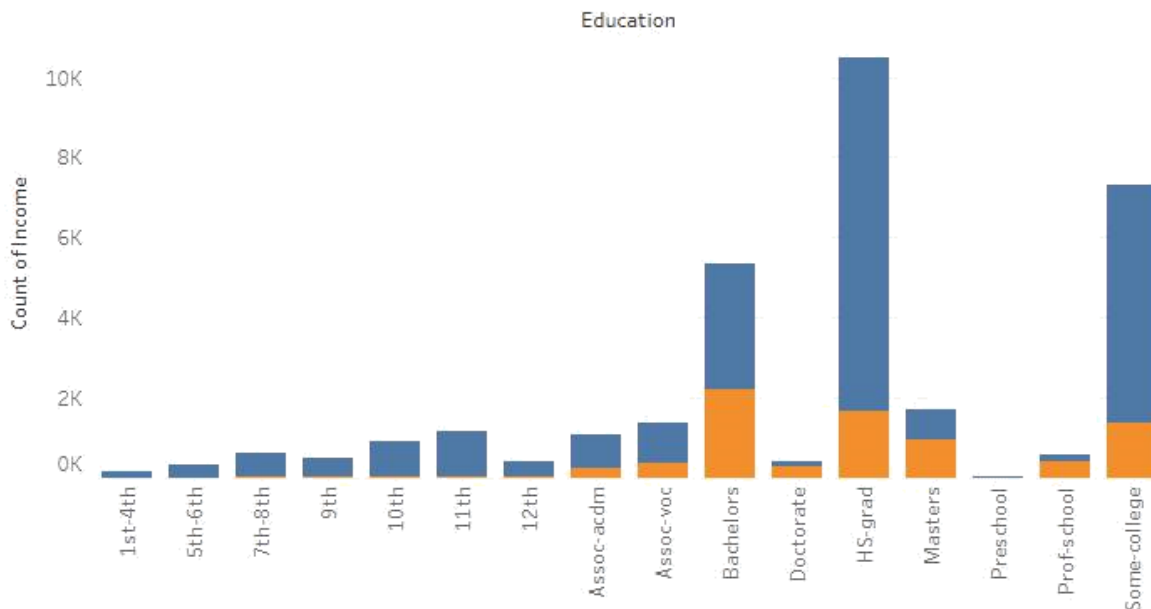
9. hours.per.week - Number of hours spent on work

10. income - It is categorical features. (1 for income > 50K and 0 for income <= 50K)

Visuals and their Inferences :

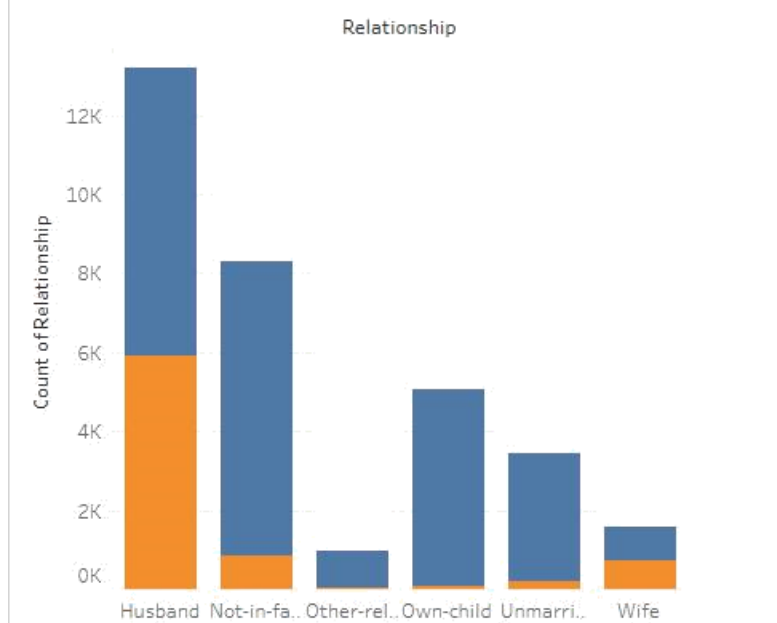
Blue – Indicates salary $\leq 50k$ and **Orange**- Indicates salary $> 50K$

Income by Education

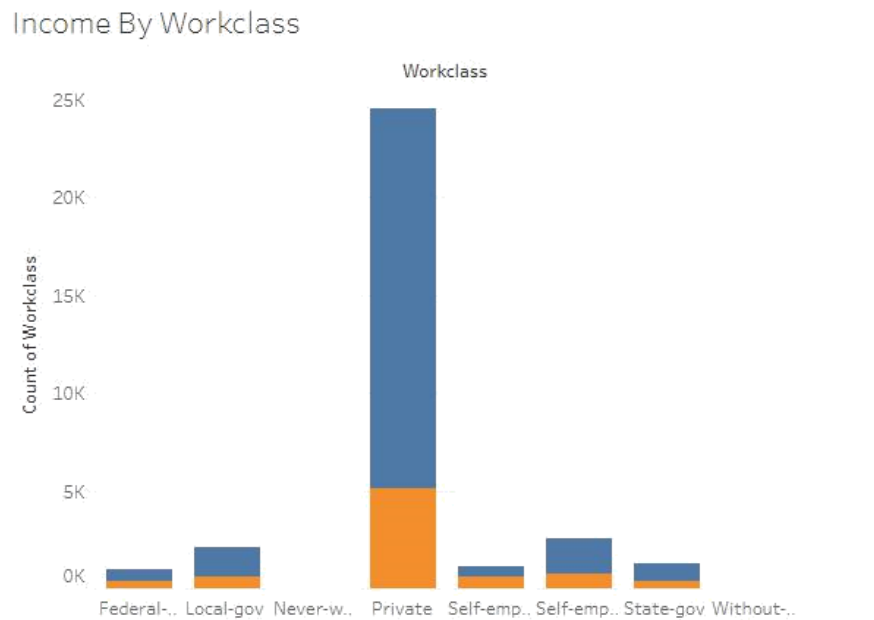


This plot shows the number of people falling in the category having income $> 50K$ and $\leq 50K$ on the basis of education they have. It shows that HS-Grads have highest number of people having income $\leq 50K$.

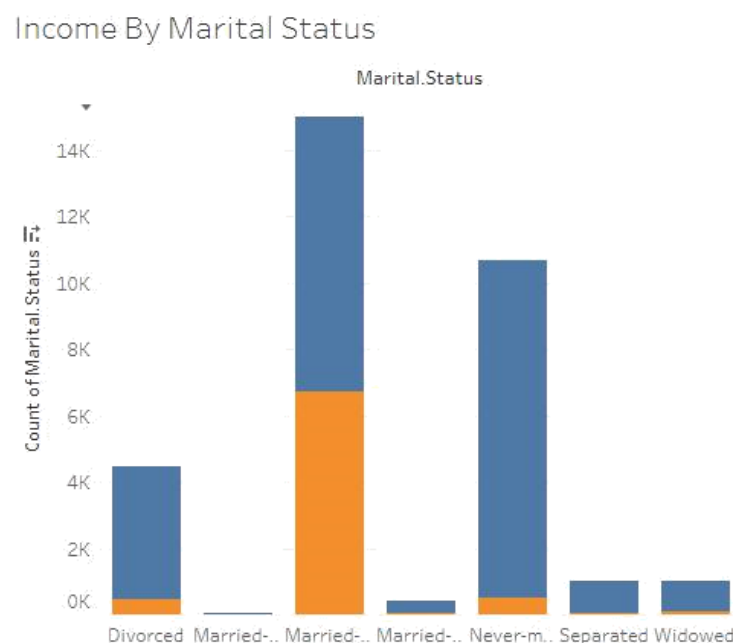
Income By Relationship



This plot shows the number of people falling in the category having income $>50K$ and $\leq 50K$ on the basis of relationship they have. It tells that the persons with relationship husband are likely to be engaged in work and thus have salary $>50K$ or $\leq 50K$.

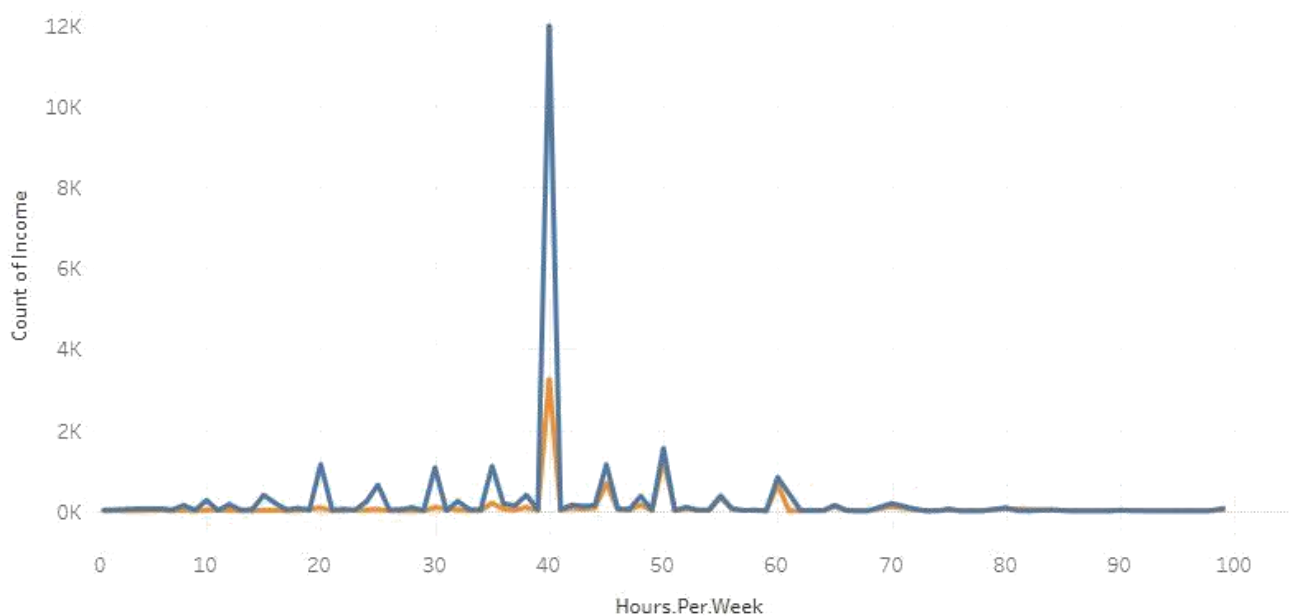


This plot shows the number of people falling in the category having income $>50K$ and $\leq 50K$ on the basis of Workclass they have. Lots of People are doing job in private sector and in private sector majority of people have income $\leq 50K$.

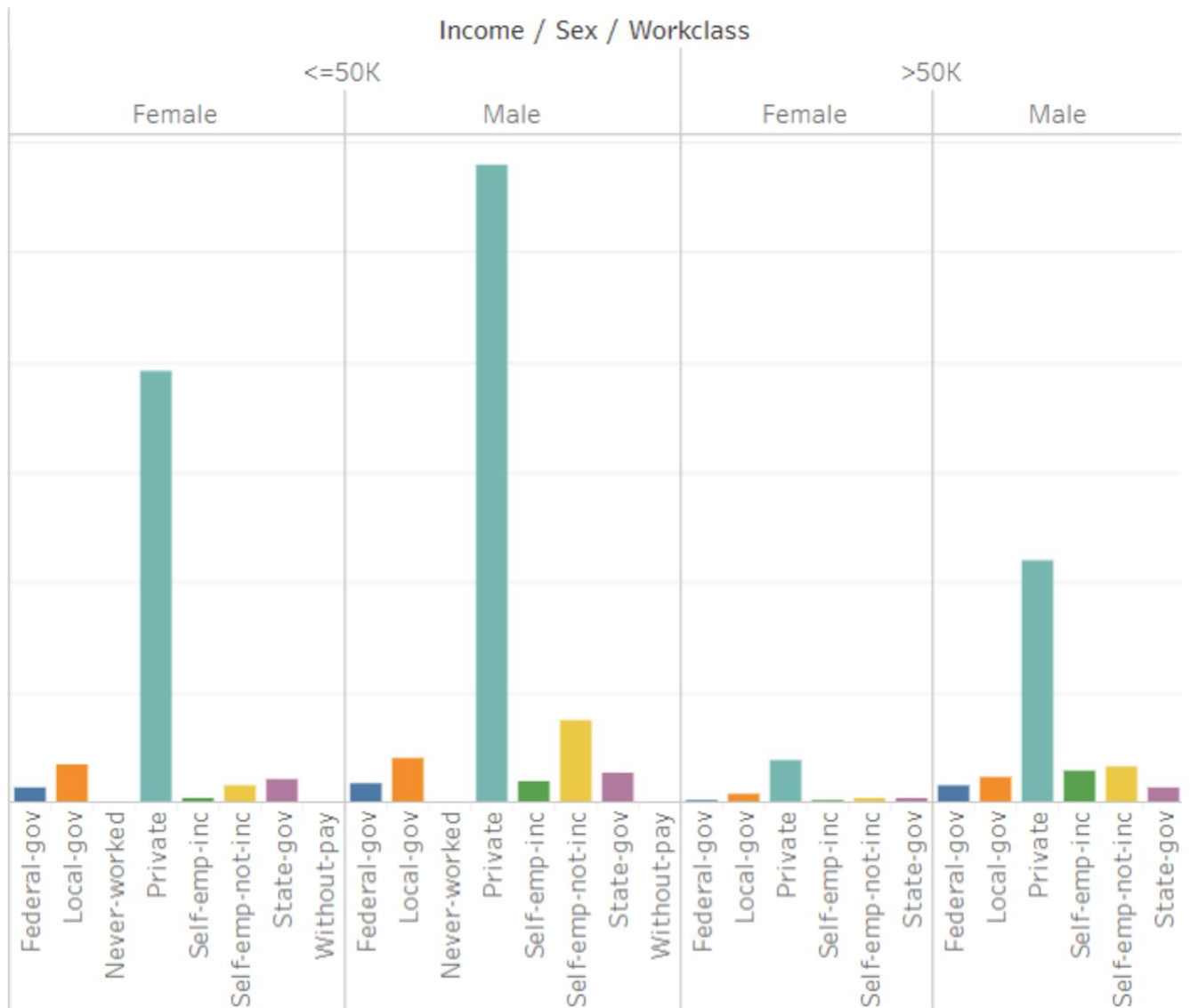


This plot shows the number of people falling in the category having income $>50K$ and $\leq 50K$ on the basis of Marital Status they have. It is seen that 8284 persons who are married and have civilian spouse have income $\leq 50K$ and 6692 persons have income $>50K$.

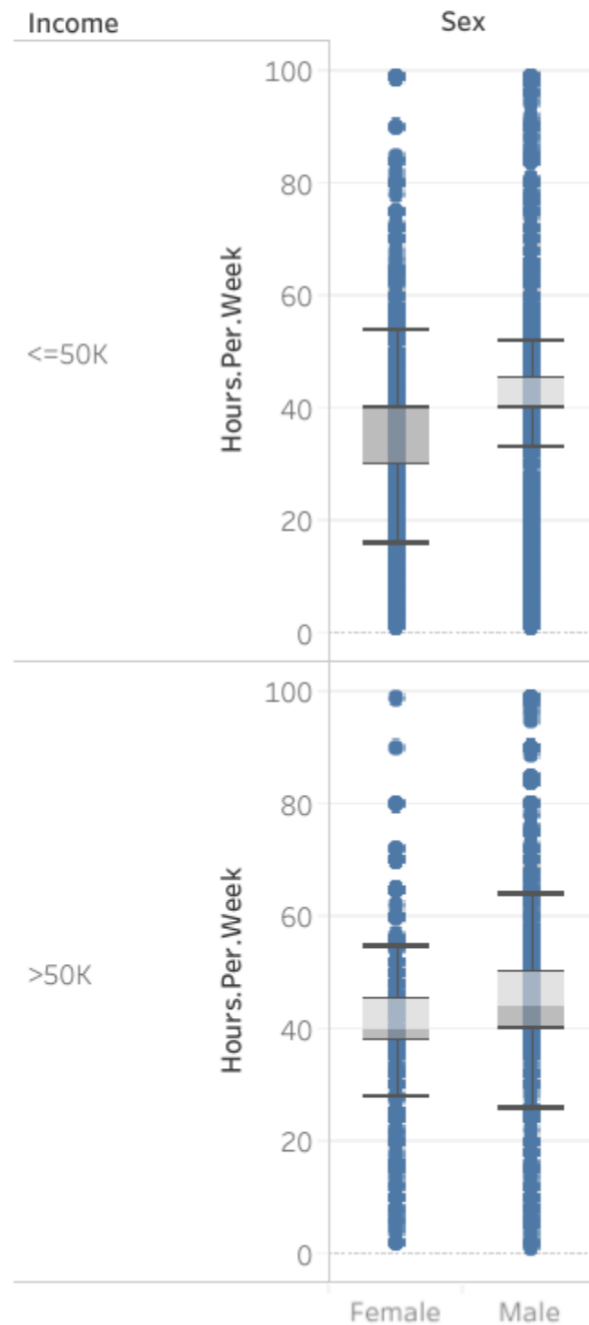
Income By Hour/Week Line Plot



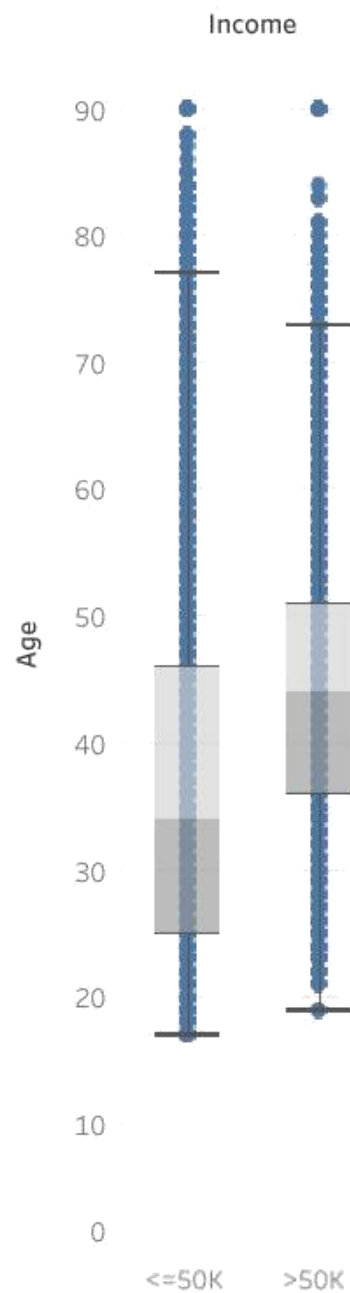
This plot shows the hours of work done by the people per week and for both categories $\leq 50K$ and $>50K$ maximum number of people are working for 40 hrs per week although number of people having income $\leq 50K$ is more than number of people having income $>50K$.



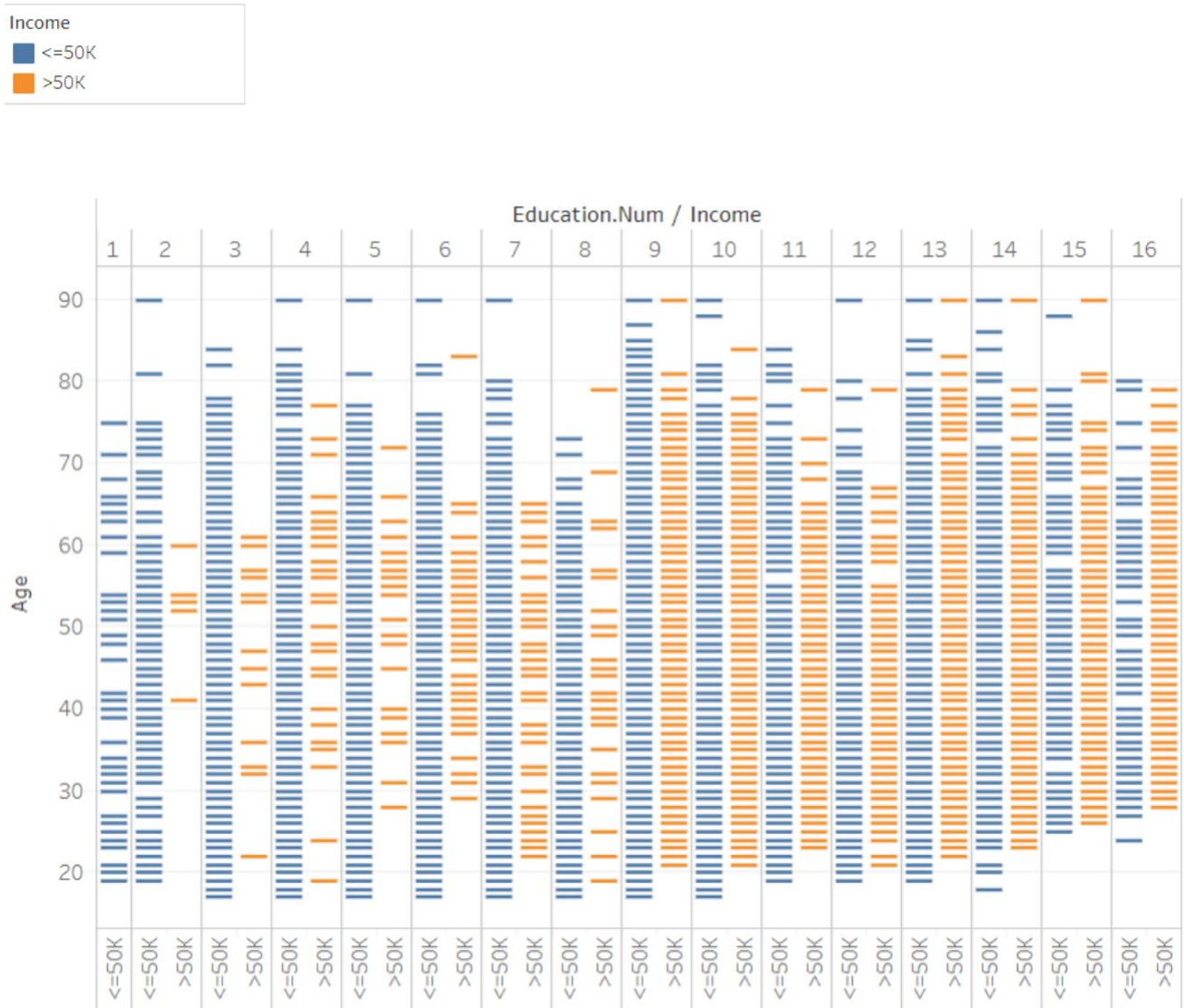
The above plot displays the distribution according to gender and workclass in the income category of >50K and <=50K. One thing to note here is that , in the income category of >50K majority of the males and females work in the private sector. Also the number of females is less as compared to the males in the income category of >50K.



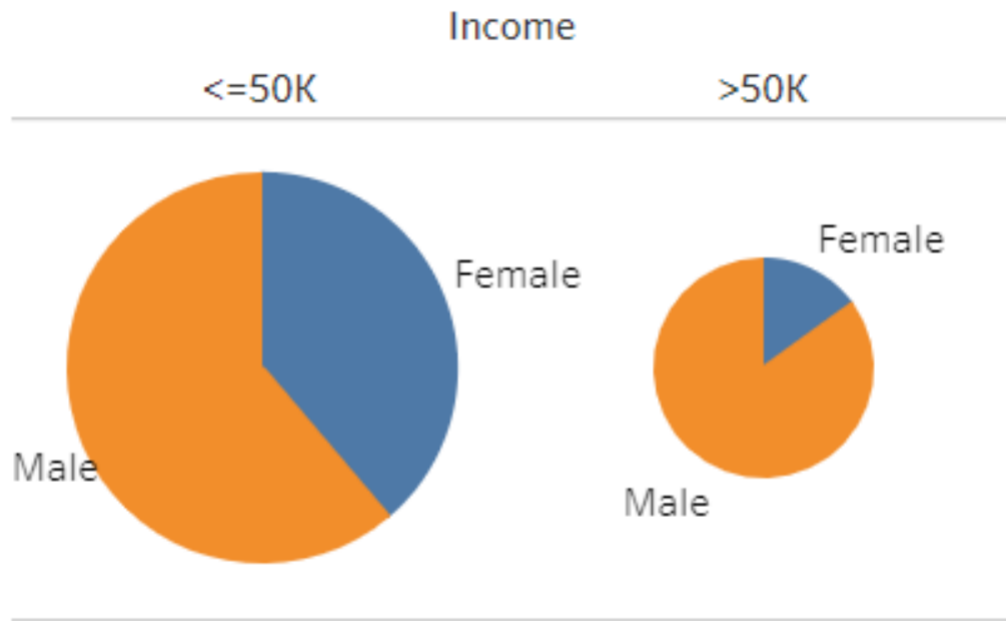
The above plot is a box plot of number of hours per week separated by income category and gender. The median value of number of hours per week for females in the income category $\leq 50K$ is around 36 and it is around 42 for males. In the $>50K$ category the median value of number of hours per week for both the genders is almost the same.



The above plot is a box plot of age for both the income categories . For the people earning $\leq 50K$ 75% of them have their age below 46 , whereas for the people earning $> 50K$ 75% of them have their age below 52.



The above plot gives an idea of the trend of income by education number and age. We observe here that for the education number upto 3 very few orange ticks are visible that is the number of people having income >50K for education number upto 3 is very less. As the education number increases the orange ticks also increase , that is the number of people earning >50K also increases . We can also observe that out of the total number of people earning >50K most of them have high education number and are above 30 years of age.



The above pie chart shows the number of males and females in both the income categories . In the category of <=50K the number of males is higher , but the number of females is also significant . Similarly , in the >50K category the number of males is higher but that of the females is very less as compared to that of the males.

Dataset Example

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Female	0	0	40	Cuba	<=50K
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	Female	0	0	40	United-States	<=50K
6	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Female	0	0	16	Jamaica	<=50K
7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	45	United-States	>50K
8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	Female	14084	0	50	United-States	>50K
9	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Male	5178	0	40	United-States	>50K

Accuracy of models on dataset

	Method	Training Accuracy	Testing Accuracy Score
0	Decision Tree Classifier	97.315352	81.515712
1	Logistic Regression	79.610957	78.927911
2	Random Forest Classifier	97.315352	84.674845

Census Income Dashboard

Census Income Analysis

Income by Education

Education: All, Sex: All

Income By Relationship

Relationship: All

Income By Workclass

Workclass: All

Income By Marital Status

Marital.Status: All

Income By Sex

Sex: Female, Male

Income By Hour/Week Line Plot

Hours.Per.Week: 1 to 99 and Null values

Income Category by States

Income: <=50K, >50K

Income by Occupation sex and age

Occupation: All, Sex: All, Age: All

Income Category By Sex and Workclass

Income / Sex / Workclass

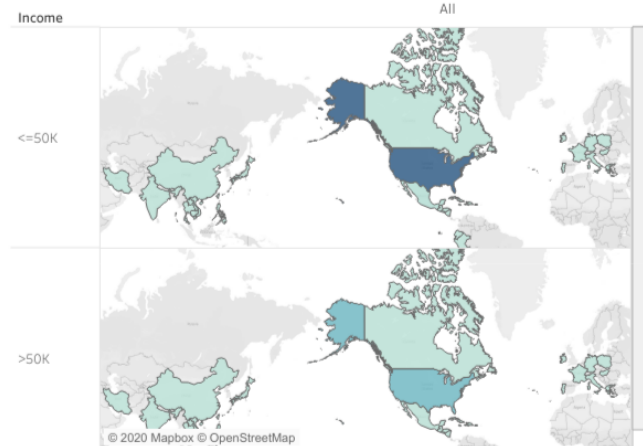
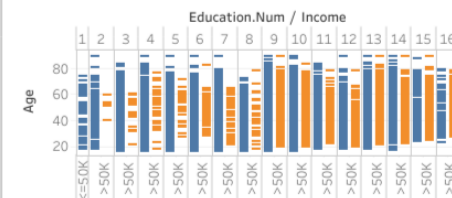
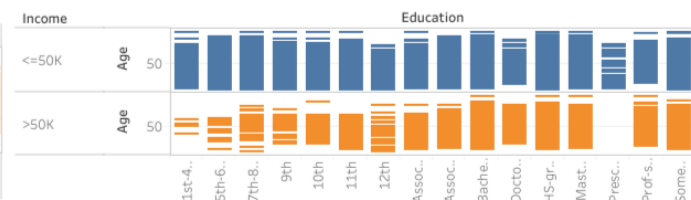
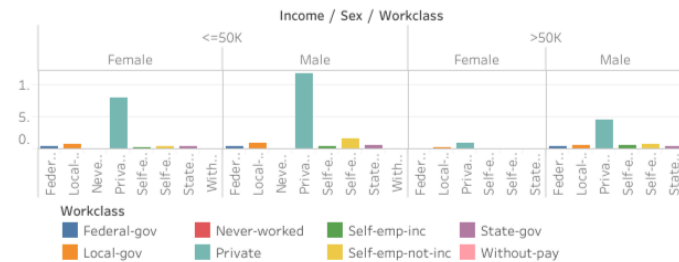
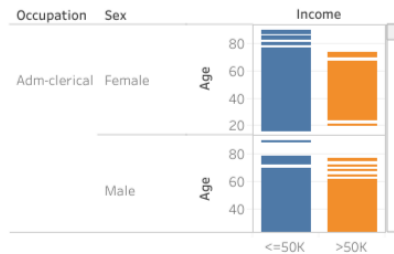
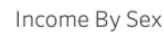
Income Trend by Education No. and Age

Education.Num / Income

Income Trend By Education and Age

Income: <=50K, >50K

Income by Education



Prediction Portal

Dashboard

Prediction

Prediction

Name

Vedant Padalkar

Number of educational years

16

Hours per week

50

Age

25

Occupation

Executive Managerial

Capital Gain

0

Workclass

State Government

Gender


Male

Capital Loss

0

Predict

You income may be <=50K



Conclusion

1. Decision Tree Classifier and Random Forest Classifier works best on training data.
2. Gradient Boosting Classifier and Random Forest Classifier works best on testing data.
3. We have performed classification using Random Forest Classifier.
4. We have got 87% accuracy on training data and 84% accuracy on testing data.
5. Age, Qualification and Occupation are most important factor in this classification.
6. If age and qualification is high of person, then he/she will more likely to have income > 50K.