Bhargav Pabbisetti
COEN 281
10/12/2015

HW #1

1)
nominal, binary, ordinal, interval, and ratio.
Age: age of the person
In this particular dataset age is ordinal attribute with discrete values.

Age can be classified as nominal, binary, ordinal, or ratio. It really depends on the application. Sometimes numerical variables are categorical due to lack of values. Age can also be binary in some particular surveys where age represents two categories, like 6 year old or 7 year old. Age can also be ordinal so it can represent ranks in a survey. Lastly age can be a ratio where it is a continuous or discrete variable. For example, age measured in days of a specific seed growing into a plant.

Work: type of employer
In this particular dataset, work is nominal.

Work can be classified as nominal(categorical), it can also be ordinal like CEO, CTO, Project Managers, technical engineers and etc.

edu: level of education
In this dataset, Edu can be classified as nominal, and ordinal. Ranks like Phd, Masters, Bachelors, high school diploma.

marital: marital status
In this dataset, marital status can be classified as nominal. There are different categories.

It can also be considered as binary if dataset only accepts like married or not married.

occupation: type of occupation
In this dataset, occupation status can be classified as nominal. There are different categories

race:
In this dataset, race can be classified as nominal. There are different categories

sex: Male and Female
In this dataset, sex can be classified as nominal, binary. There are different categories and only two categories generally speaking so it can be considered binary.

hrs_per_week: hours worked per week.

Bhargav Pabbisetti
COEN 281
10/12/2015

Hrs_per_week is ordinal because ranks can be given for the amount of hrs one has worked.

Income earned:
This can be ordinal or binary. You can rank the income levels and in this particular data set

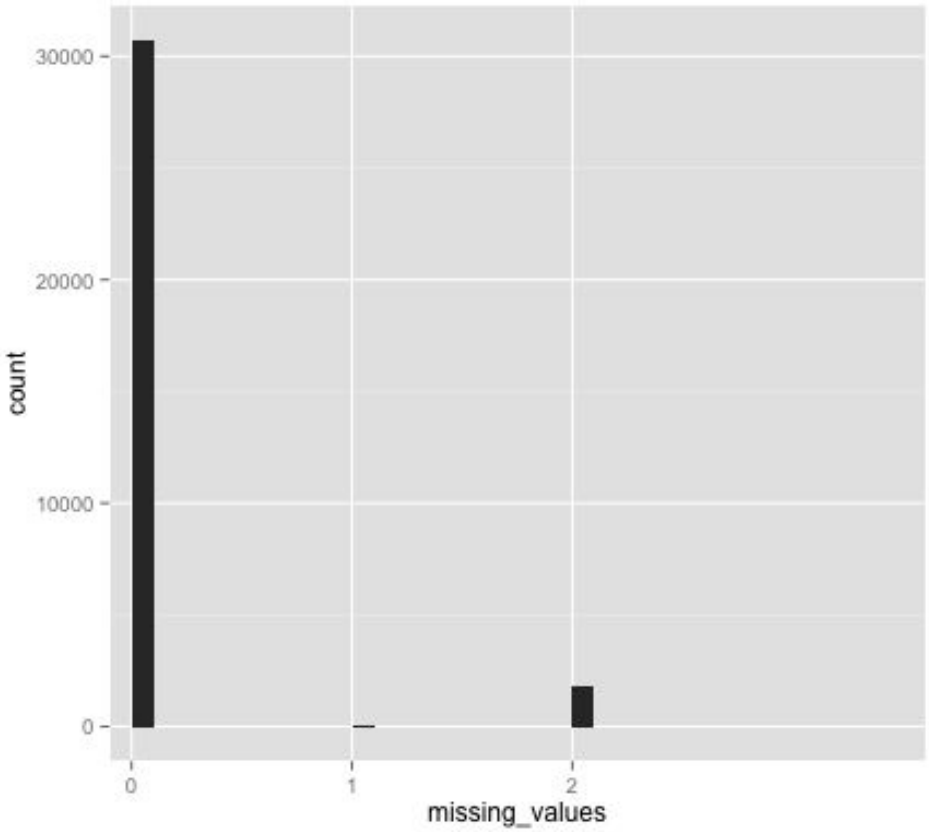
2a)

Percentage of missing values for age: 0.00
Percentage of missing values for work: 5.64
Percentage of missing values for edu: 0.00
Percentage of missing values for marital: 0.00
Percentage of missing values for occupation: 5.66
Percentage of missing values for race: 0.00
Percentage of missing values for sex: 0.00
Percentage of missing values for hrs_per_week: 0.00

Only attributes work and occupation have missing values. Only about 6% of the data is missing within each of these attributes.

Bhargav Pabbisetti
COEN 281
10/12/2015

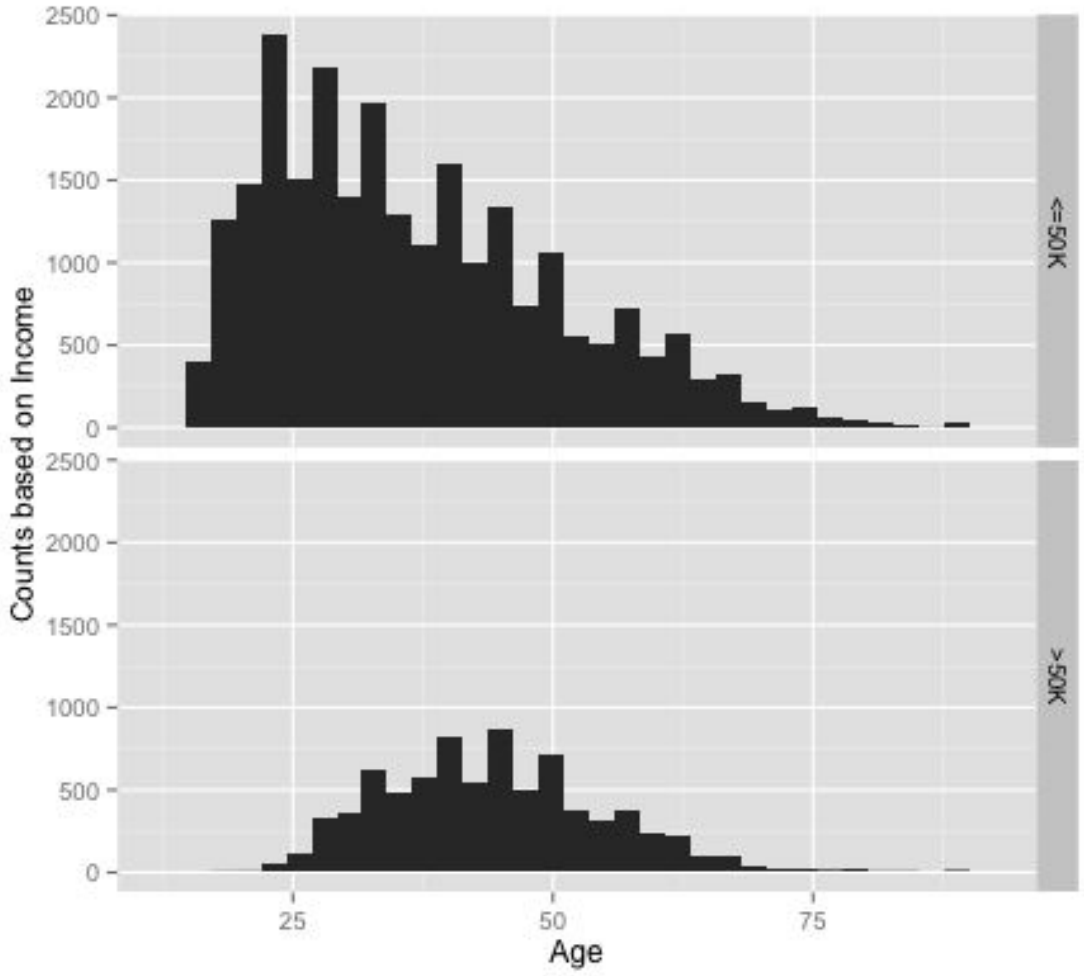
2b)



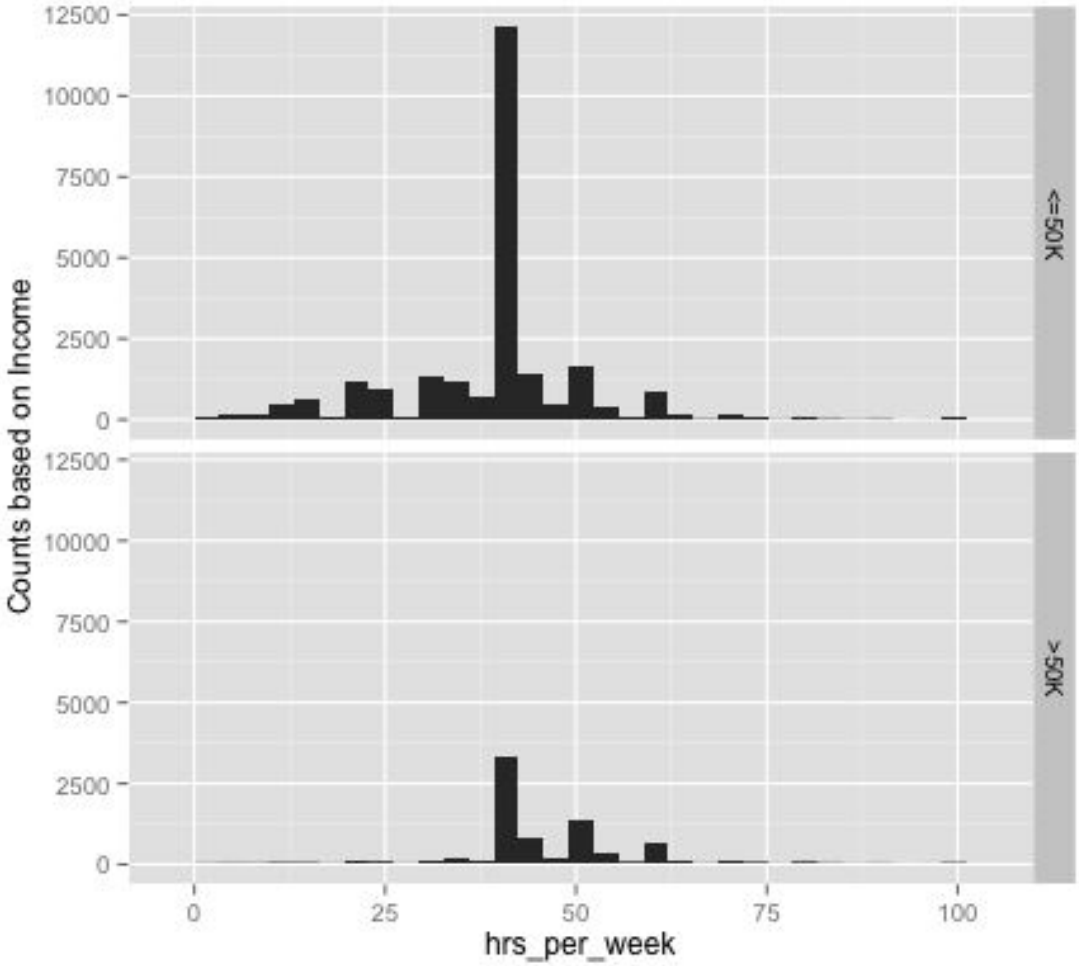
Most of the rows are filled with complete data corresponding to missing_value = 0.
Some rows are missing two elements of a single data point/row. (missing_value =2)
Very few rows are missing only one element of a single data point/row.(missing_value =1)

Bhargav Pabbisetti
COEN 281
10/12/2015

3a)
This graph below represents many young people under 35 or 40 making less than 50K. There are less number of people who make above 50K compared to people who make less than 50K. Majority of people who make above 50K tend to be little bit older. Starting age for people greater than 50K is around 25years, compared to approximately 15 for <= 50K.



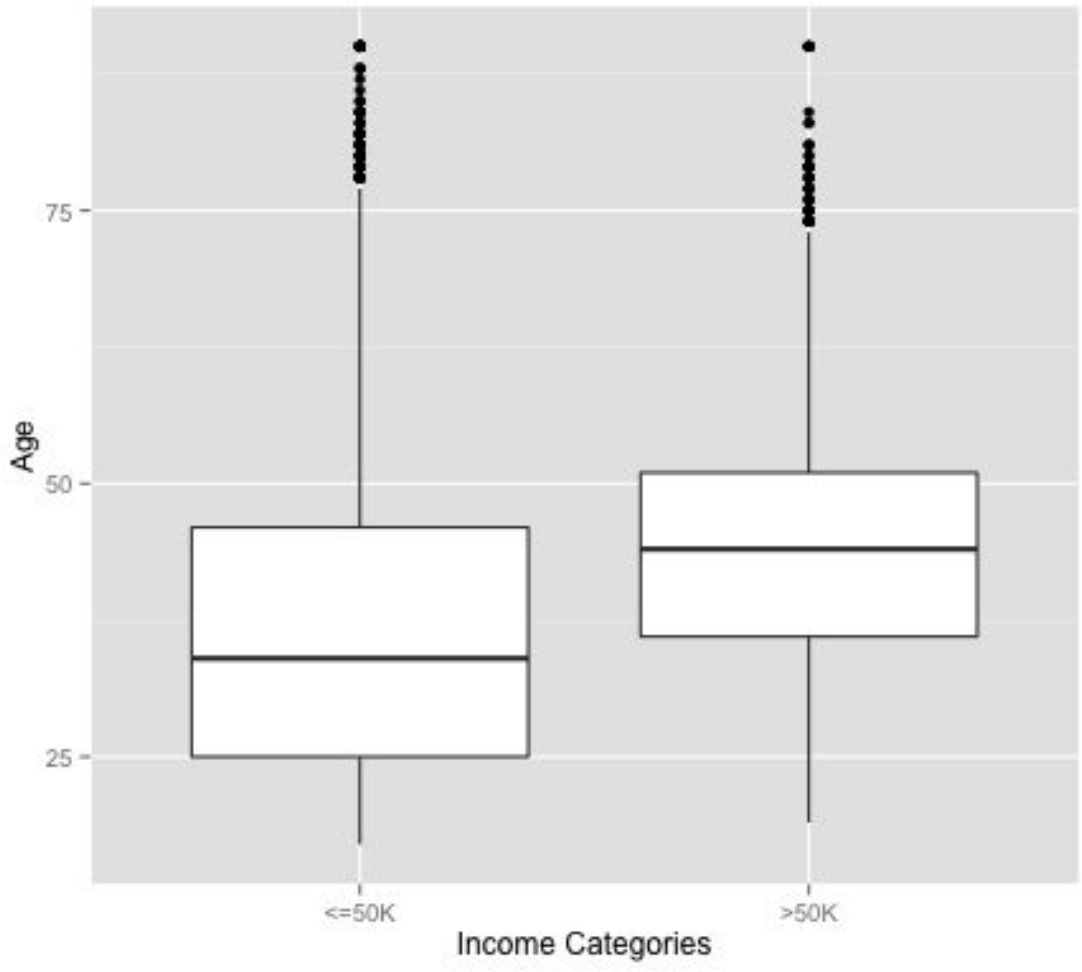
Bhargav Pabbisetti
COEN 281
10/12/2015



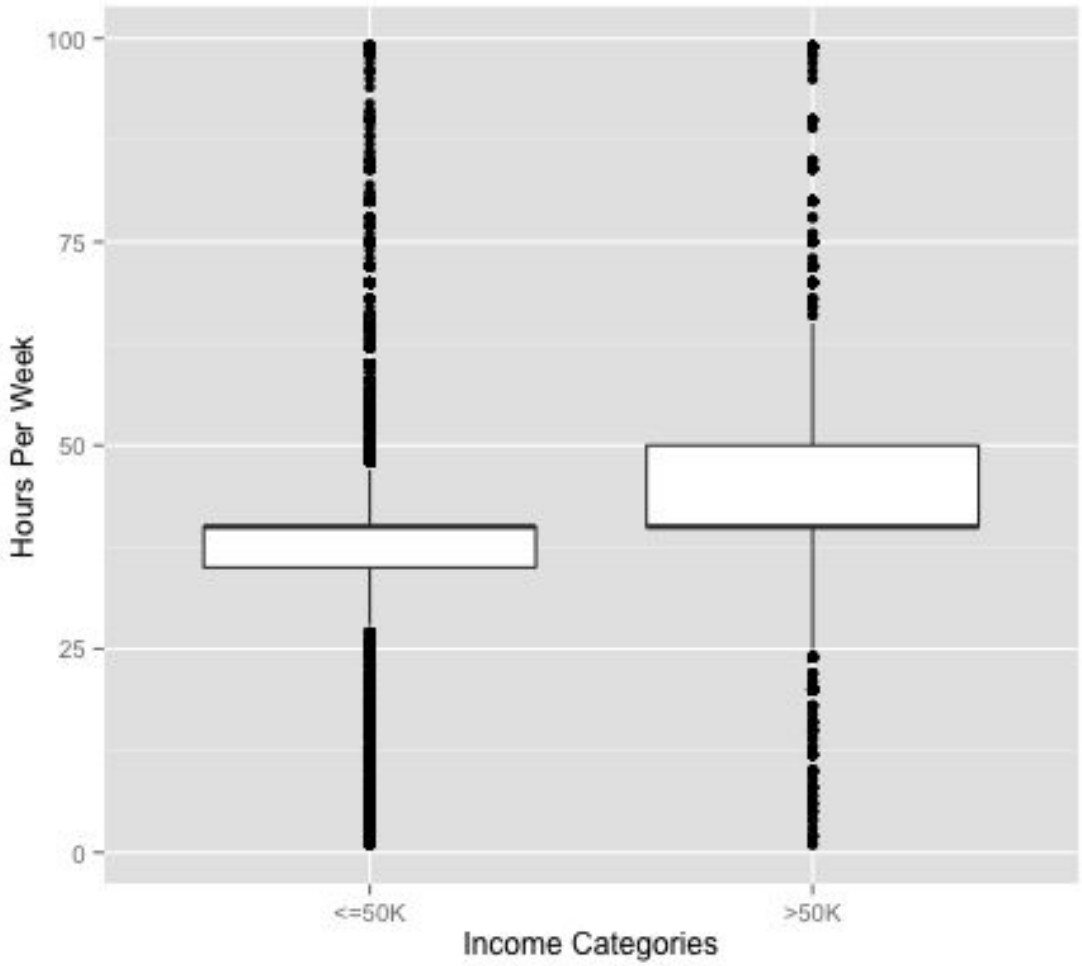
The above graph seems like overall people who make more than 50K tend to be between 35-65 hrs per week. For people who make less than 50K, most of them tend to work around 40 hrs per week. You can see huge bar around 40. Other age groups for this category tend to be widespread and very few people work more than 65 hrs per week for both categories.

Bhargav Pabbisetti
COEN 281
10/12/2015

3b)
Below graph, gives better picture about different quartiles for each category of income. For people who make less than 50K, median is around 33. First quartile is 25. Third quartile is around 44. Inter quartile range approximately goes from age group 25-44. Maximum is around 90. Minimum is around 17. For people who make more than 50K, median is around 40. First quartile is around 35 and third quartile is around 52. Inter quartile range is approximately from ages 35-52. Maximum is around 90 and minimum is around 19. Looks like people who make more than 50K tend to be more older according to interquartile range.



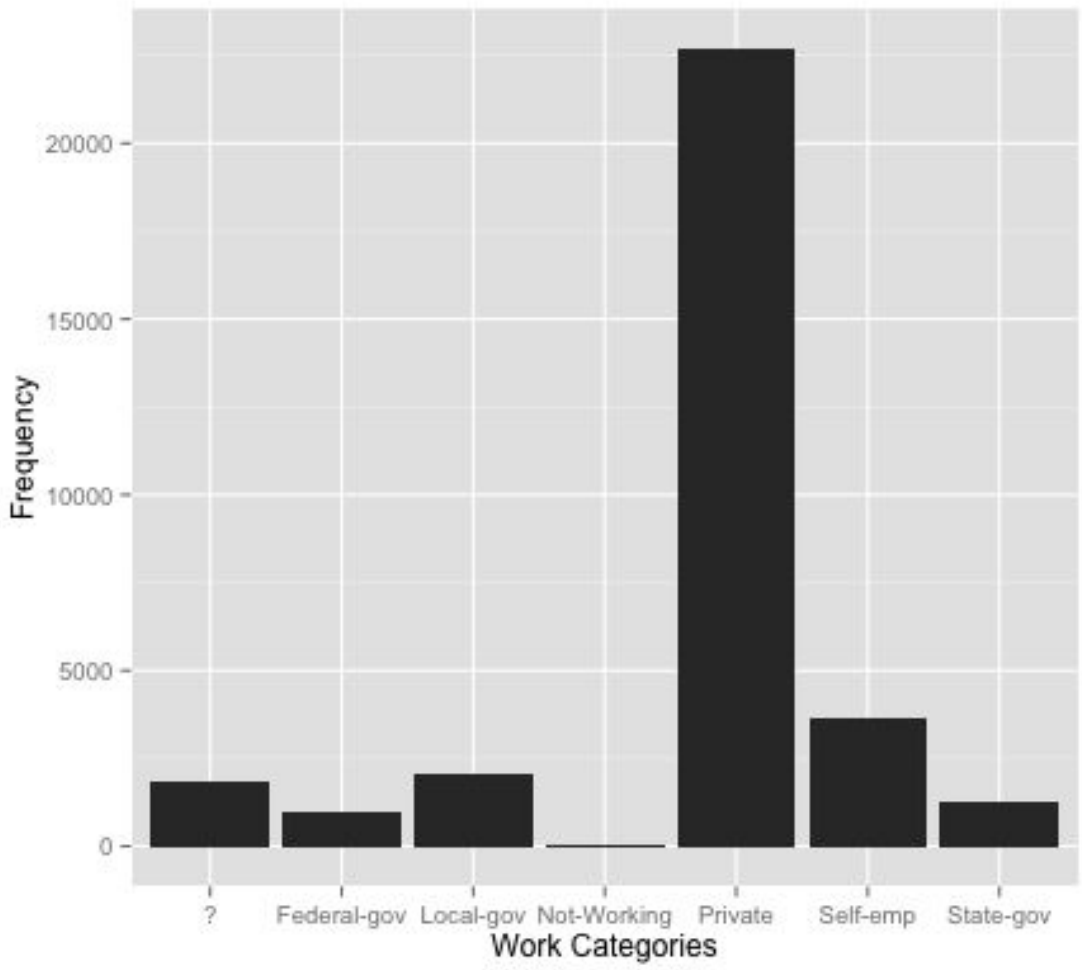
Bhargav Pabbisetti
COEN 281
10/12/2015



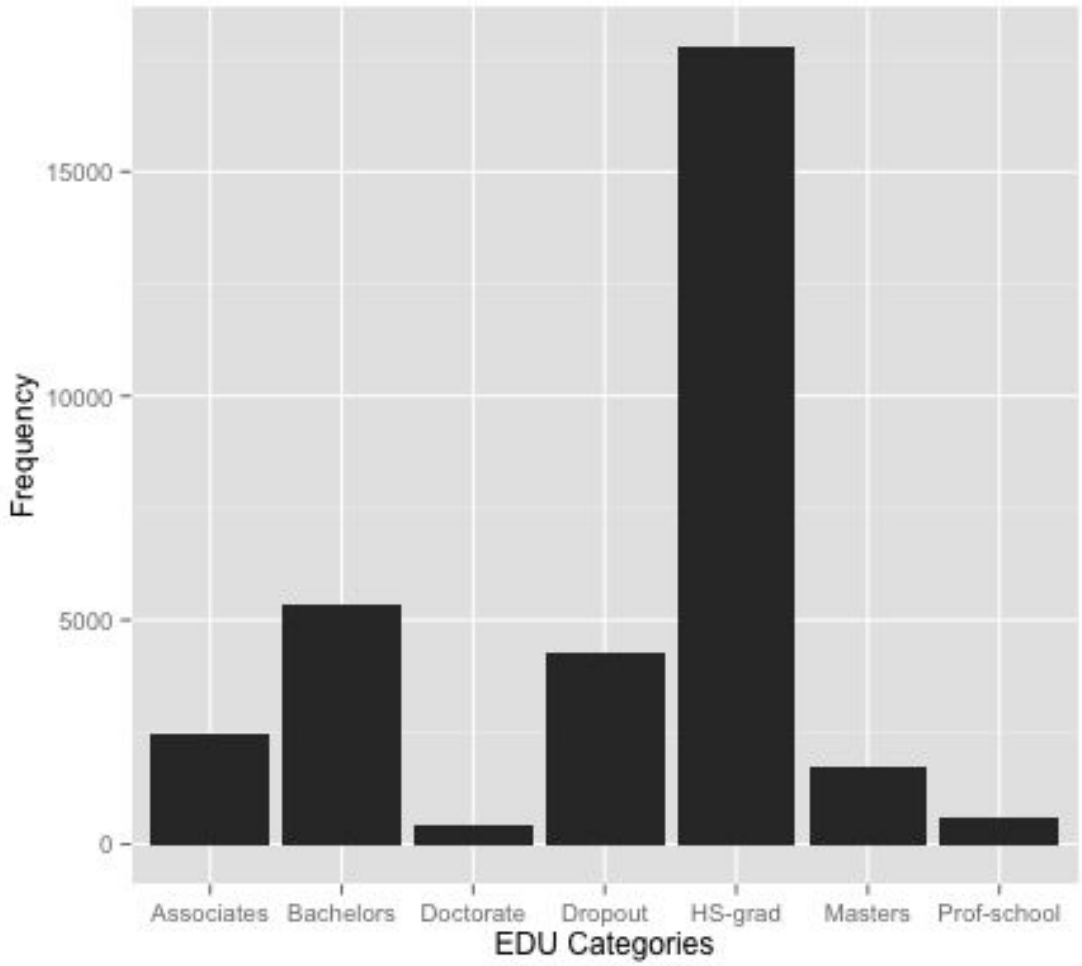
Above graph, gives better picture about different quartiles for each category of income with hrs_per_week. For people who make less than 50K, median is around 40 hrs_per_week. First quartile is 35. Third quartile is same as median. Inter quartile range approximately goes from age group 35-40. Maximum is around 100 hrs_per_week. Minimum is around 2. For people who make more than 50K, median is around 40. First quartile is same as median and second quartile is around 50. Inter quartile range is approximately from ages 40-50. Maximum is around 100 and minimum is around 2. Looks like people who make more than 50K work less hours than people who make less than 50K, according to Interquartile range.

Bhargav Pabbisetti
COEN 281
10/12/2015

4a) Below graph, work has seven unique categories including "?" which represents missing information. There tends to be many people who work privately due to his rise in the bar. There are very few people who are "not working"

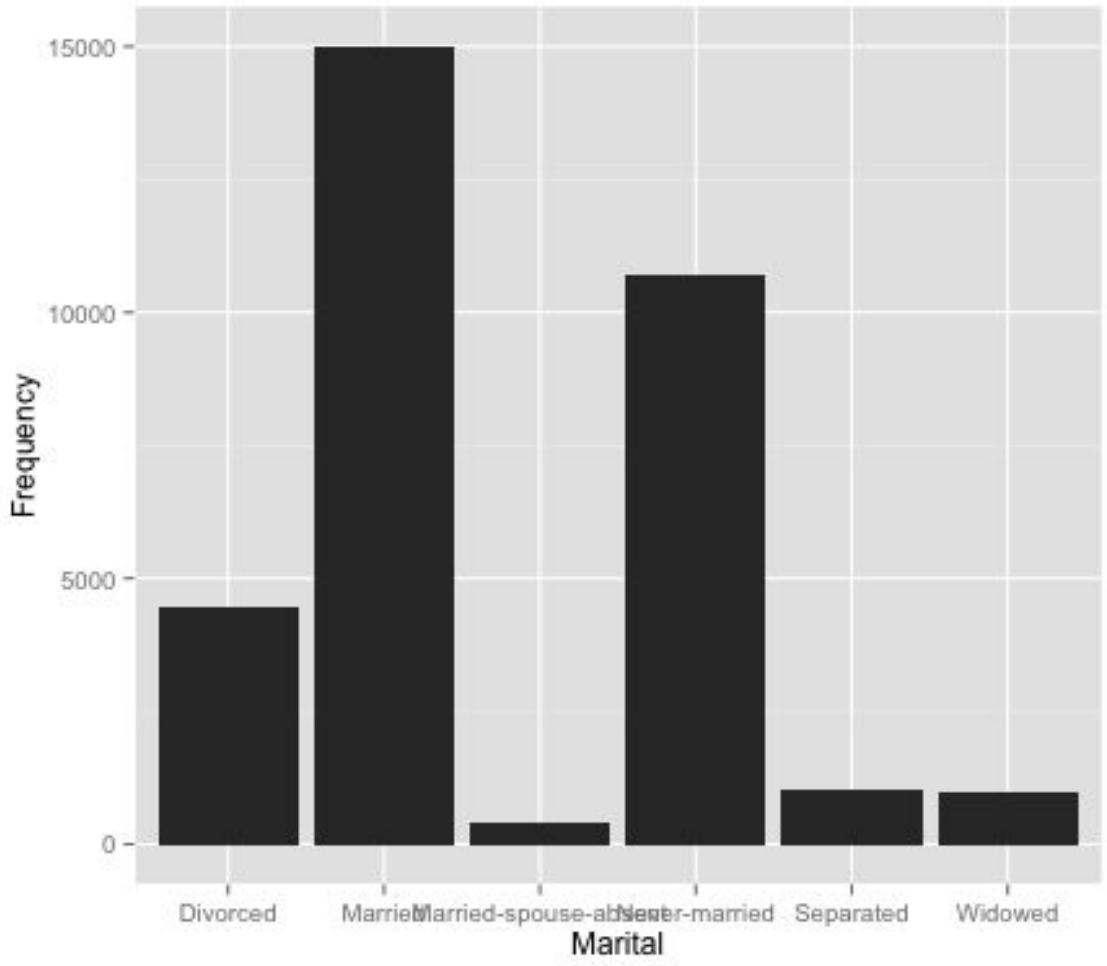


Bhargav Pabbisetti
COEN 281
10/12/2015



There are seven unique categories for EDU attribute. There tends to be many High School grads. Bachelors, Dropouts, Associates, Masters, Prof-school, and Doctorates arranged in decreasing order. Few Doctorates.

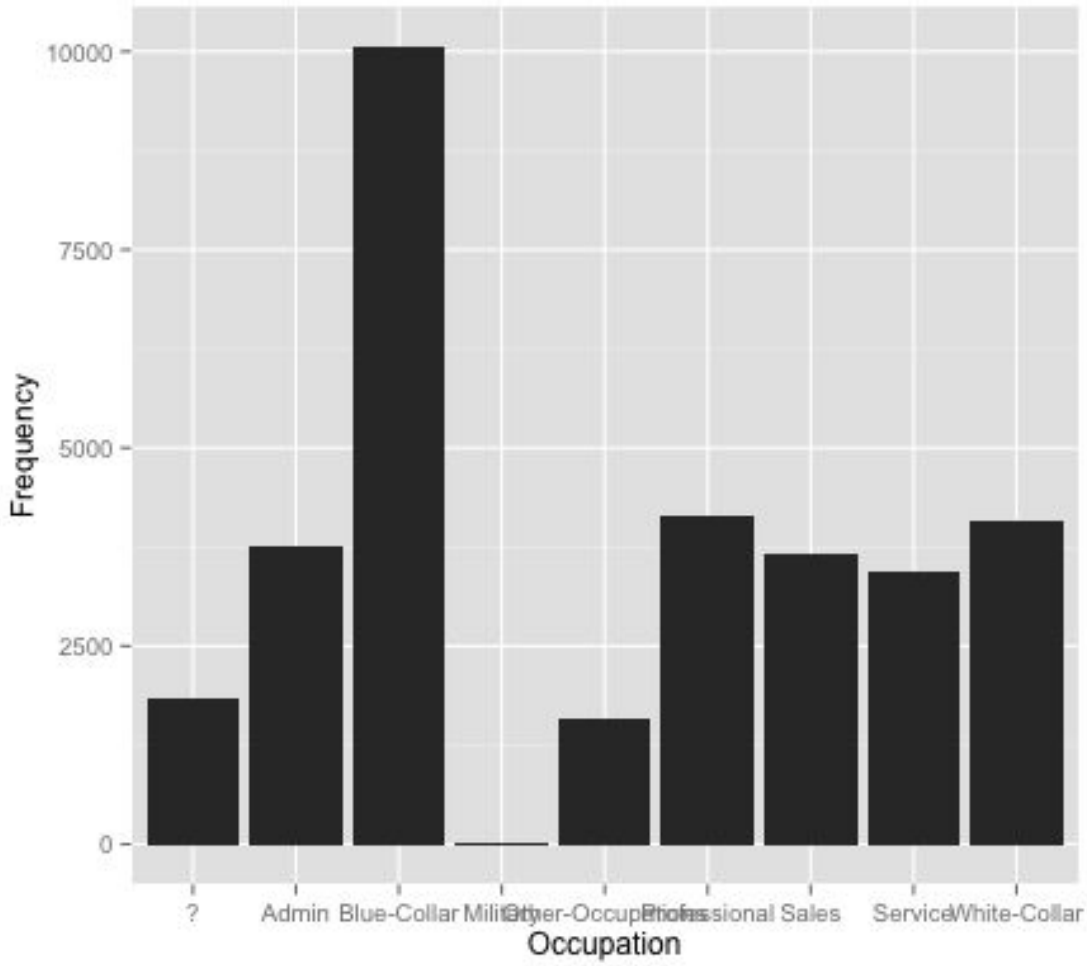
Bhargav Pabbisetti
COEN 281
10/12/2015



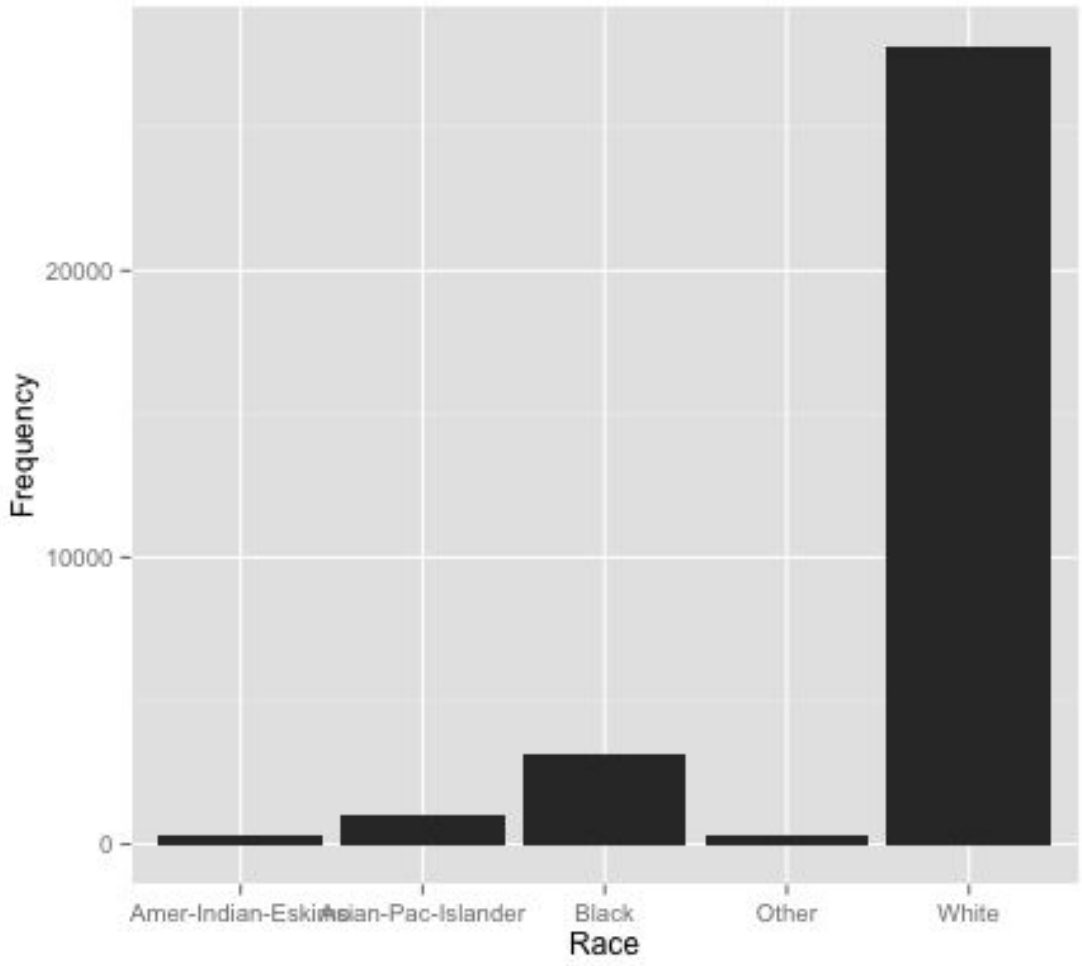
There are six unique categories associated with attribute Marital. Looks like majority of them are married and second highest is never married. Lowest is Married spouse absent.

Bhargav Pabbisetti
COEN 281
10/12/2015

Below graph, occupation attribute contains nine unique attribute including missing values which is represented by "?". Majority are blue collar workers and very few are in Military.

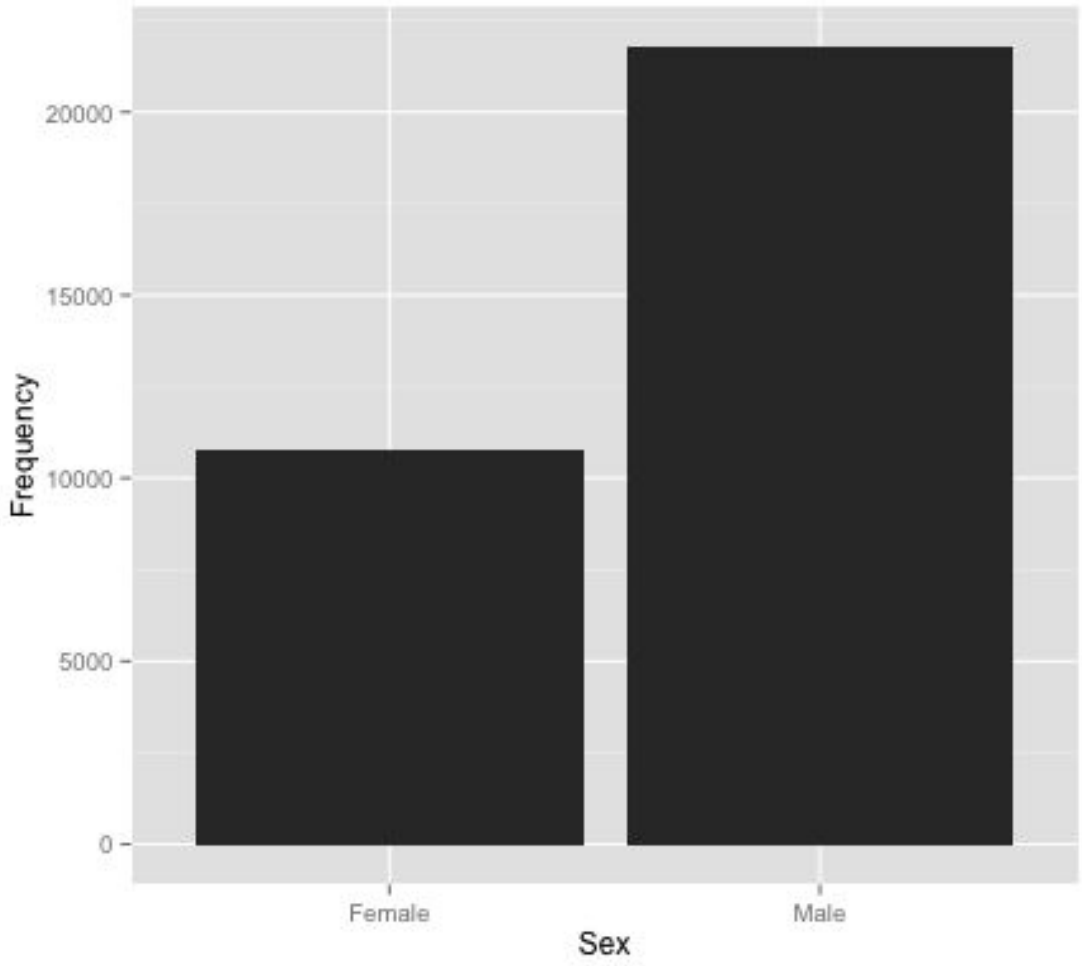


Bhargav Pabbisetti
COEN 281
10/12/2015



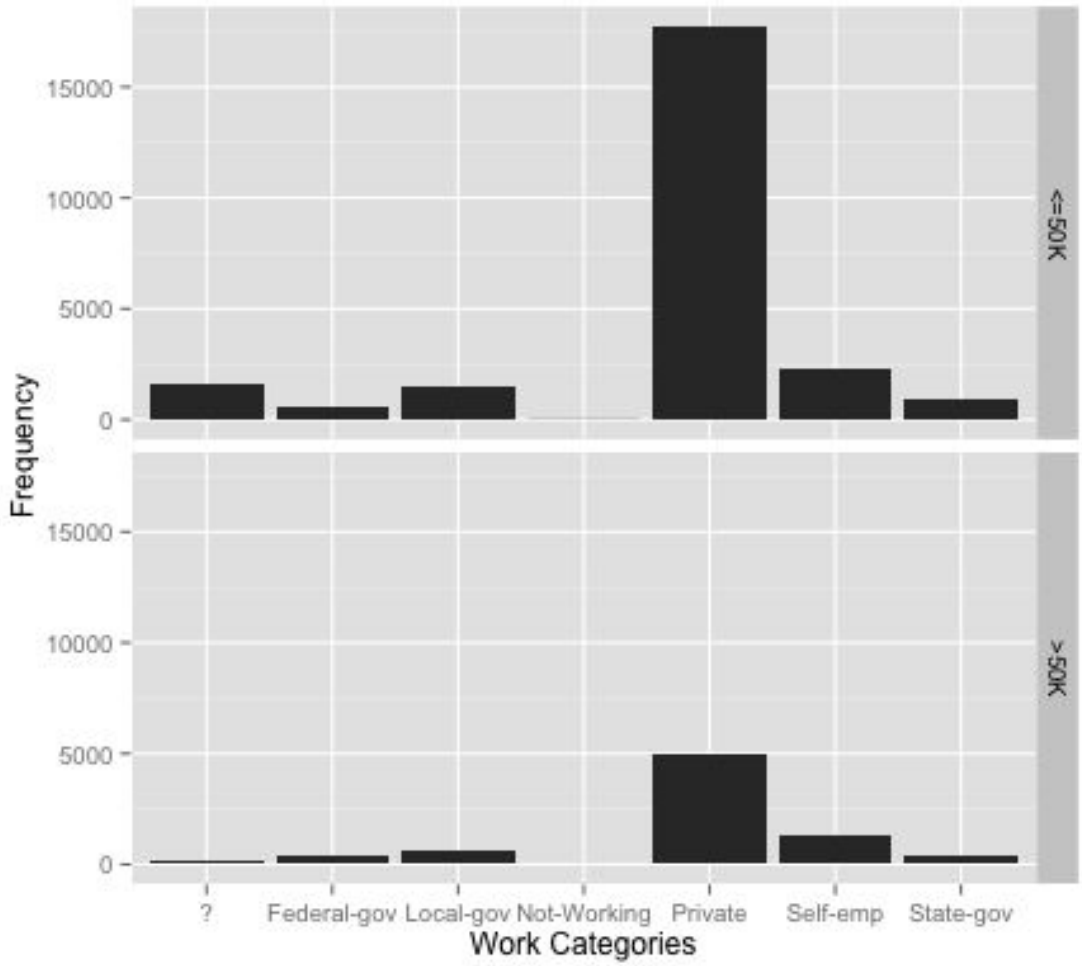
This graph shows that Race has five unique categories and majority of the people tend to be white and next highest number is black.

Bhargav Pabbisetti
COEN 281
10/12/2015



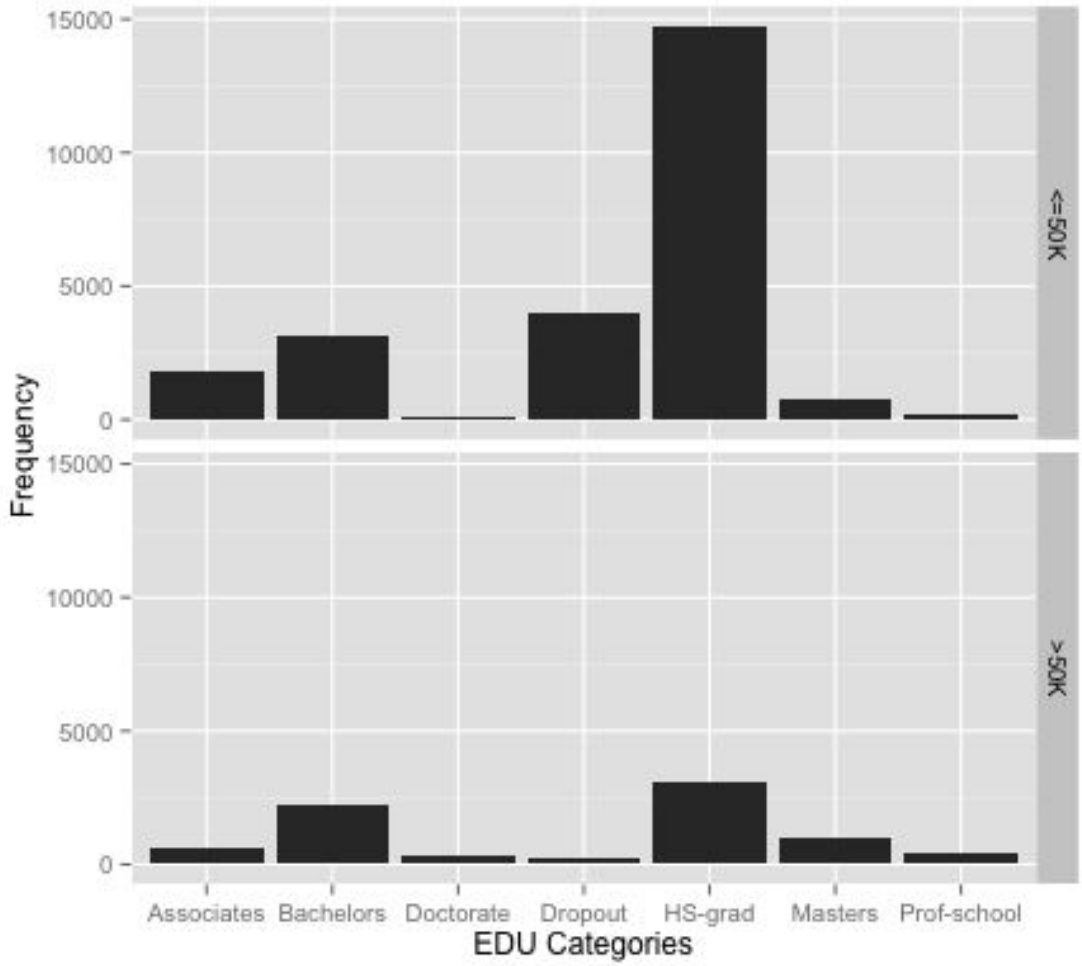
This graph shows that there are more males than females in this data set.

Bhargav Pabbisetti
COEN 281
10/12/2015



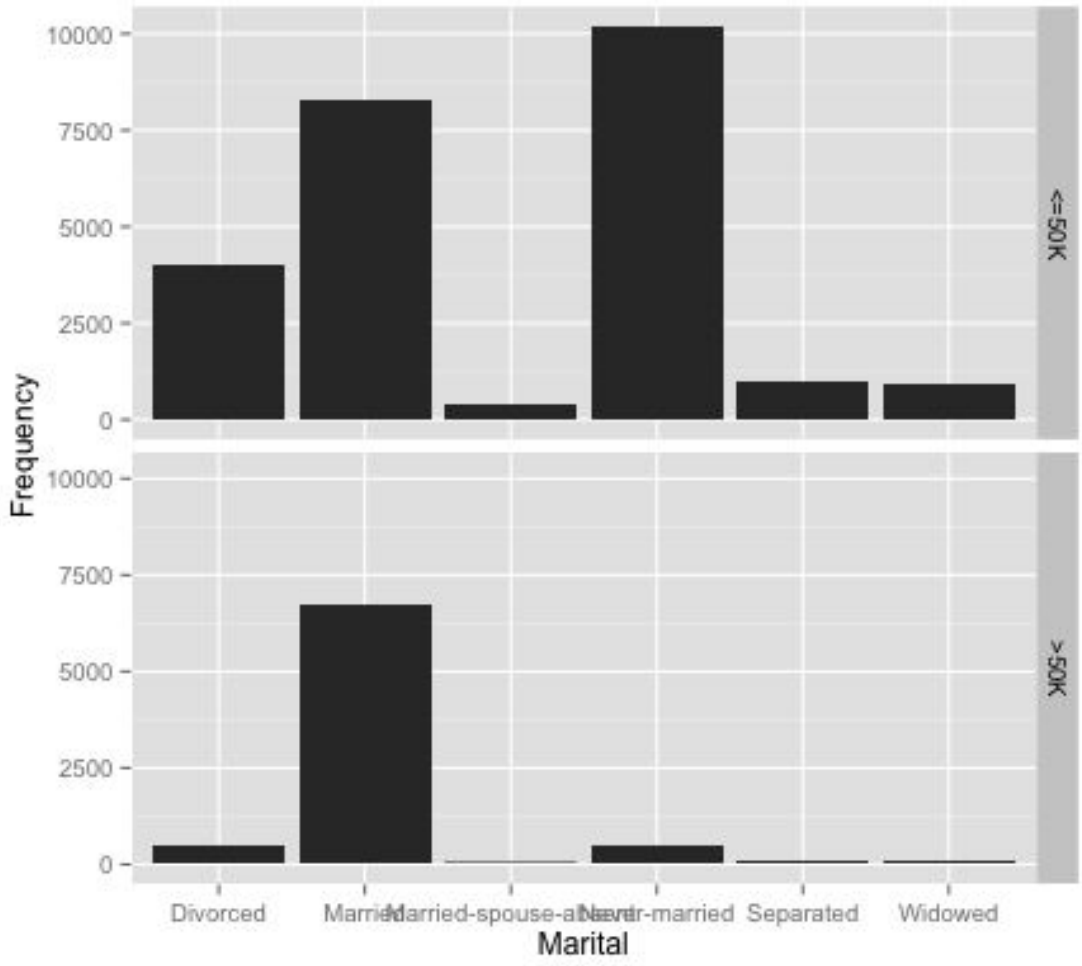
This graph shows more people who make less than 50K work privately. There are more missing values for people who make less than 50K. More self employed, state-gov, local gov people in less than 50K. Almost no one is not working in more than 50K category.

Bhargav Pabbisetti
COEN 281
10/12/2015



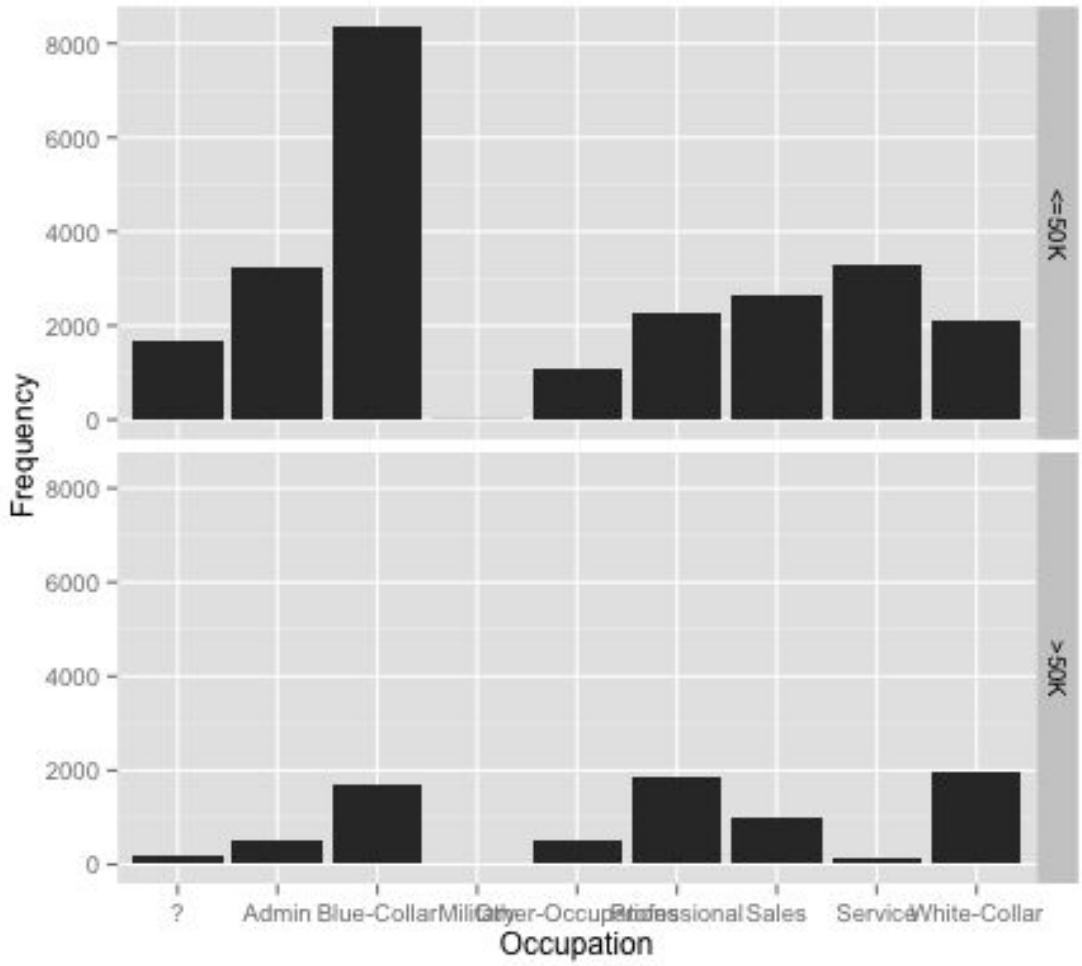
More High school grads in less than 50K category. More Bachelors, Doctorate, Masters, Prof school in less than 50K category. More associates in less than 50K category.

Bhargav Pabbisetti
COEN 281
10/12/2015



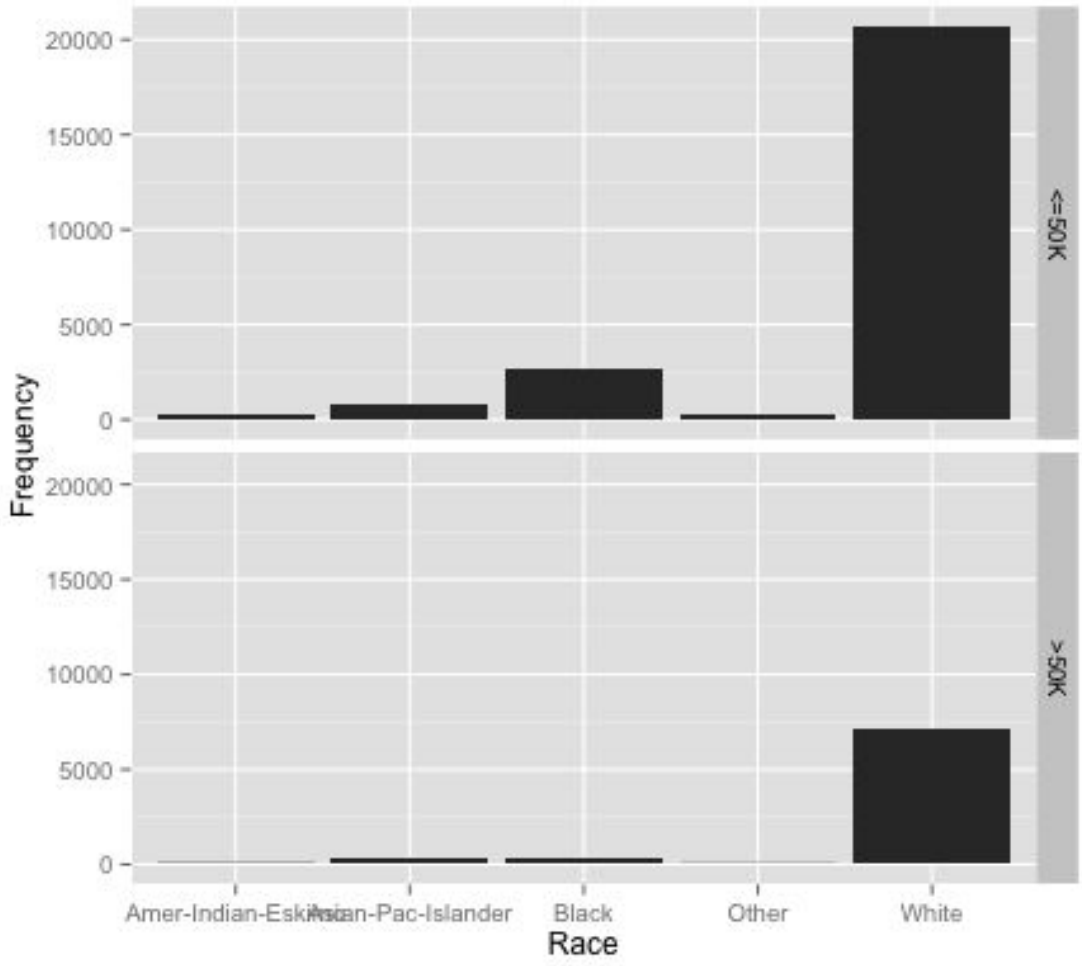
More people are divorced, married, never married, separated, married spouse absent and widowed in less than 50K category.

Bhargav Pabbisetti
COEN 281
10/12/2015



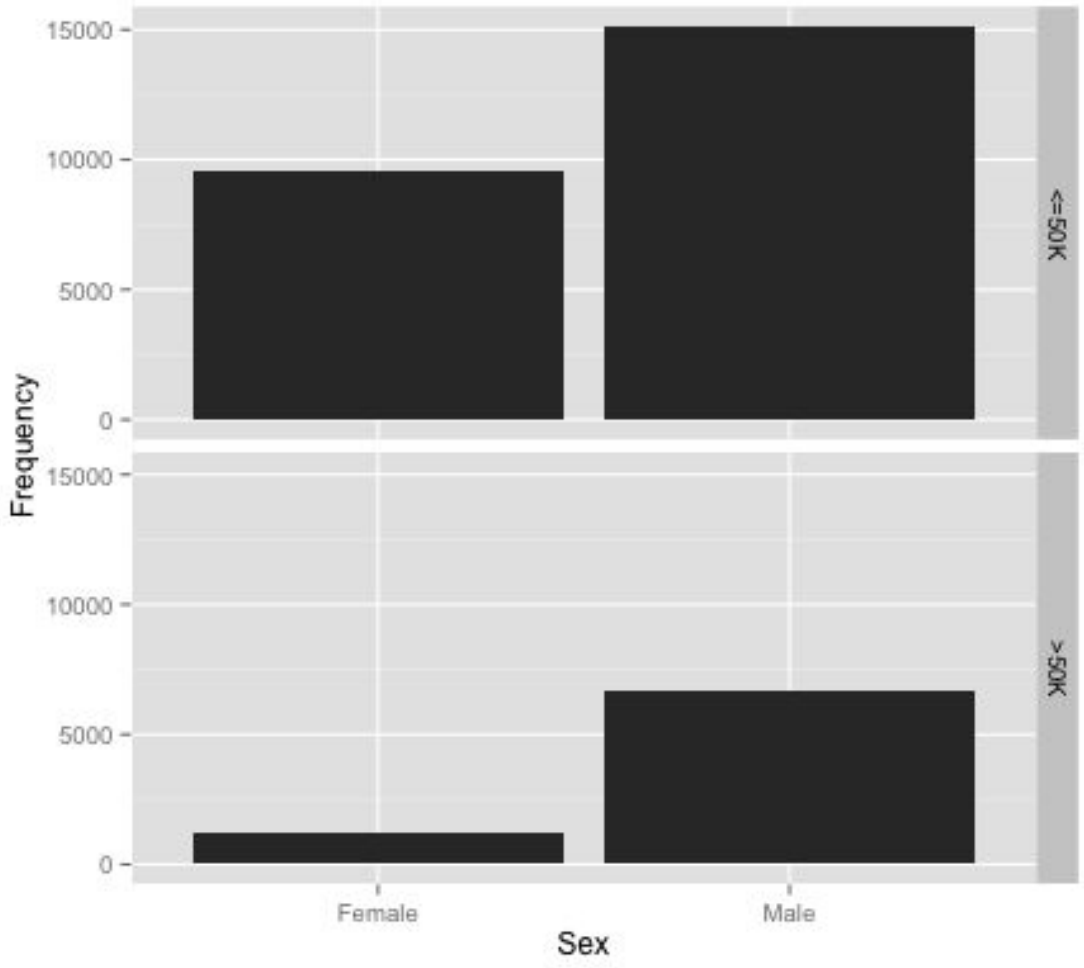
Every category is more in less than 50K.

Bhargav Pabbisetti
COEN 281
10/12/2015



There are more people who earn less than 50K in every category of race.

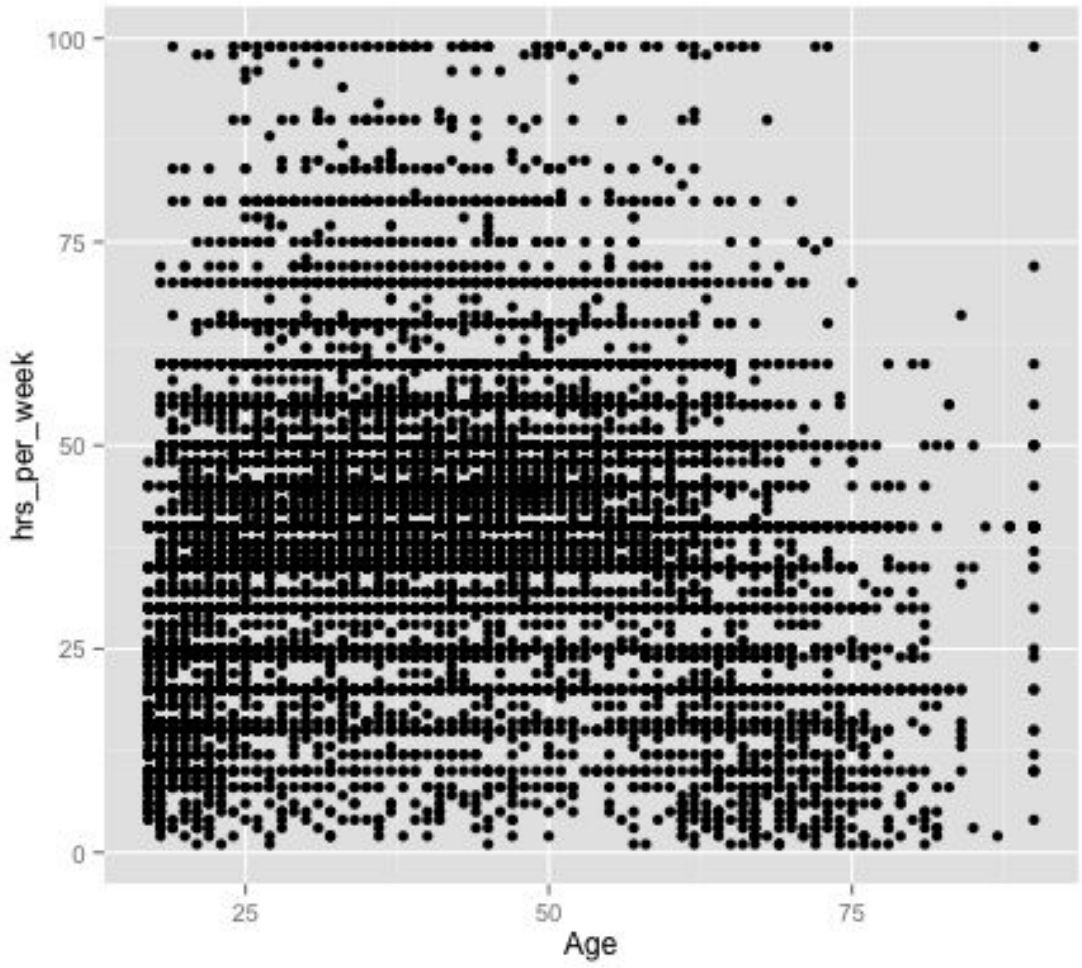
Bhargav Pabbisetti
COEN 281
10/12/2015



More males in both categories of income but ratio of male to female seems to be higher in people who earn more than 50K. Much more male domination in 50K above.

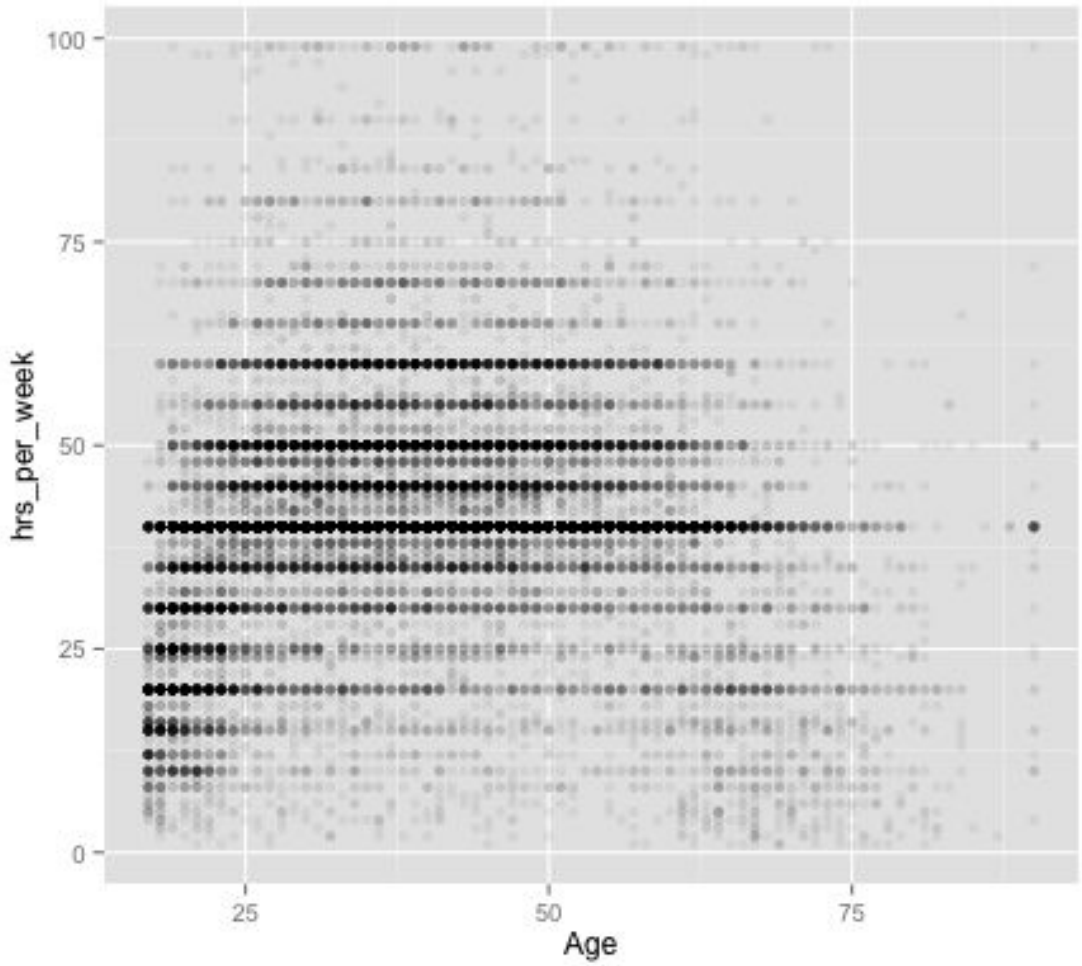
Bhargav Pabbisetti
COEN 281
10/12/2015

5)



Original plot, there is a lot of overplotting in the middle. This plot look like points are all over the place and majority in the middle and below area.

Bhargav Pabbisetti
COEN 281
10/12/2015



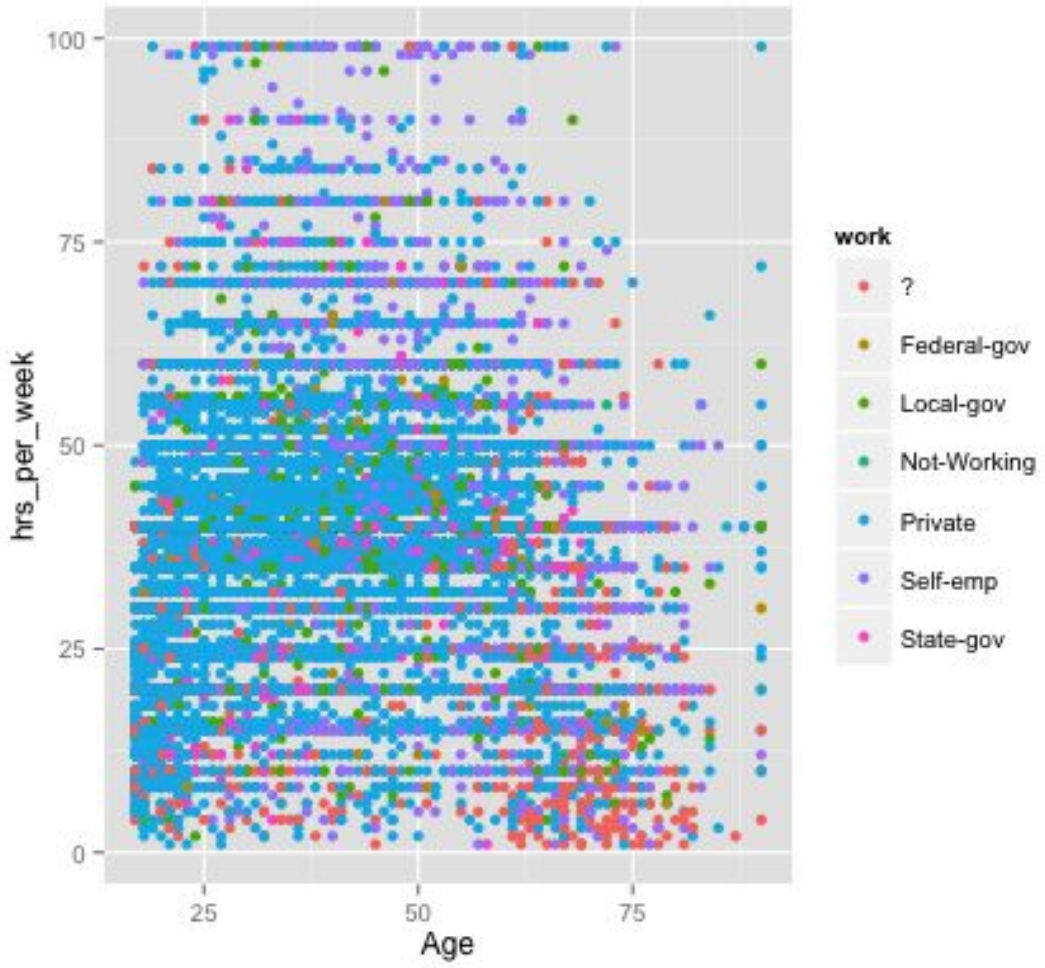
Fixed overplotting.

Pearson's product-moment correlation

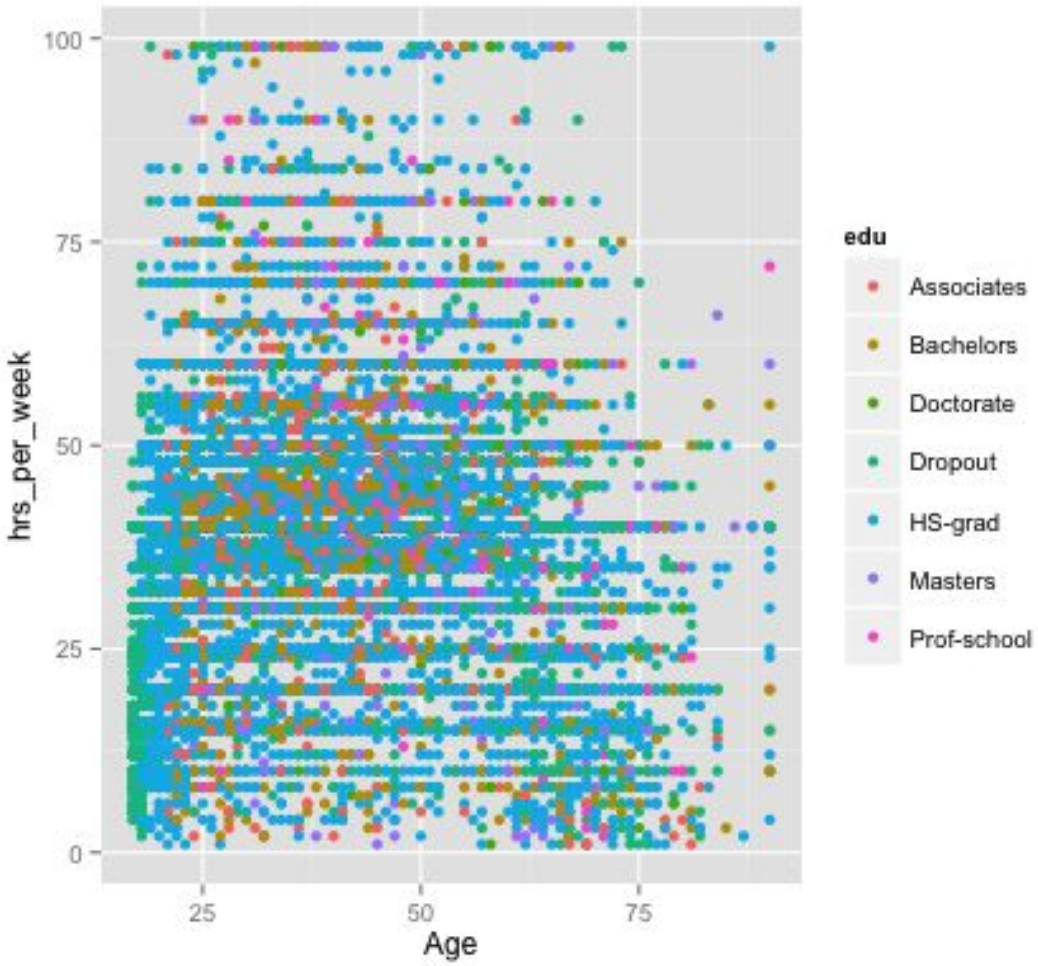
data: data2$age and data2$hrs_per_week
t = 12.436, df = 32559, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.05793717 0.07955810
sample estimates:
cor
0.06875571, This shows that there isn't much linear correlation between hrs_per_week and age.

Bhargav Pabbisetti
COEN 281
10/12/2015

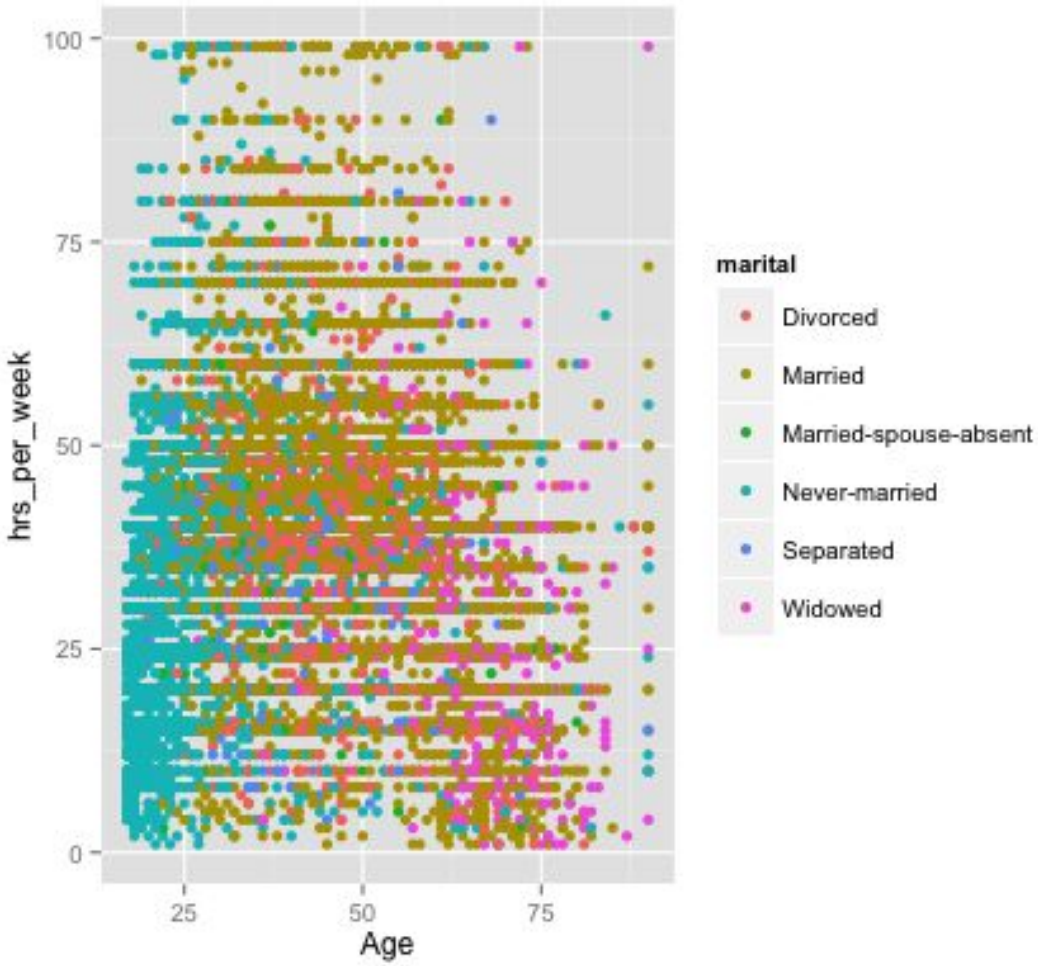
6)
All of these plots show data points of different element in a particular category. You can observe correlation between bar graphs previously and scatter plots.



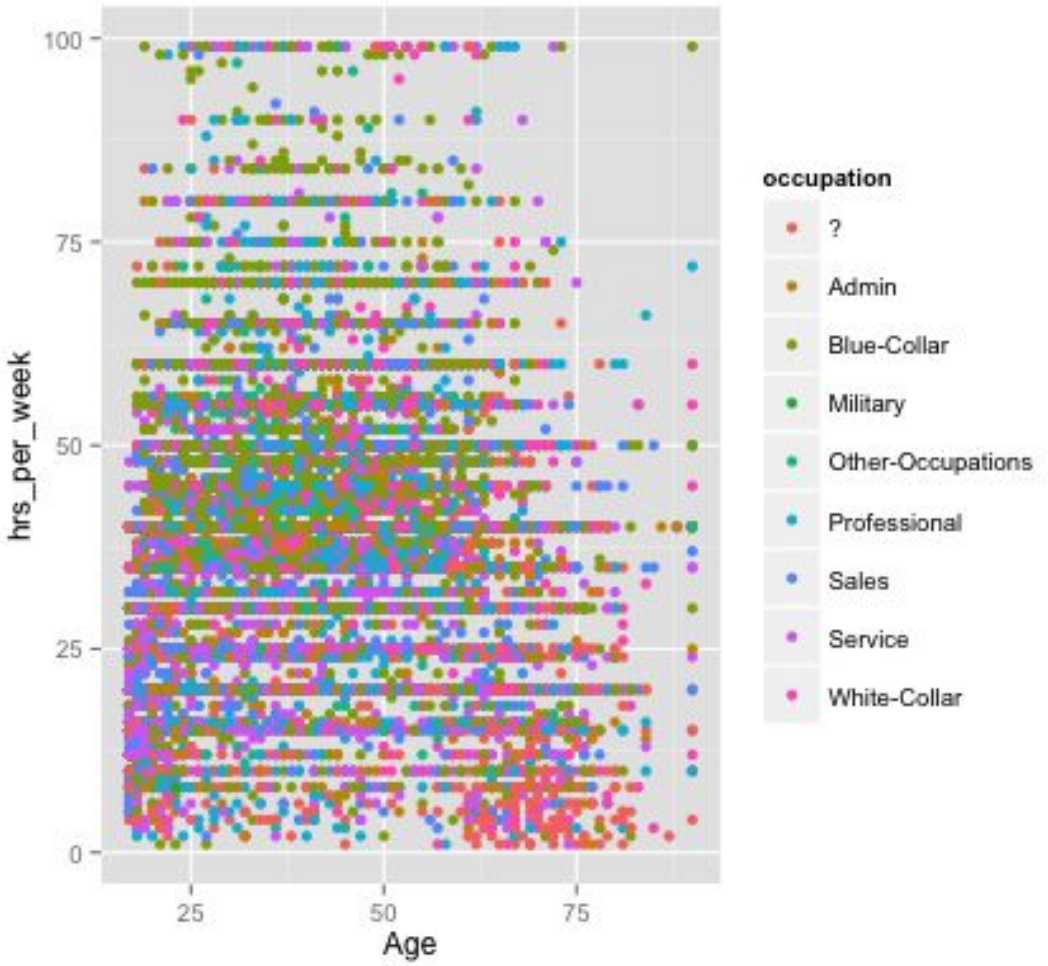
Bhargav Pabbisetti
COEN 281
10/12/2015



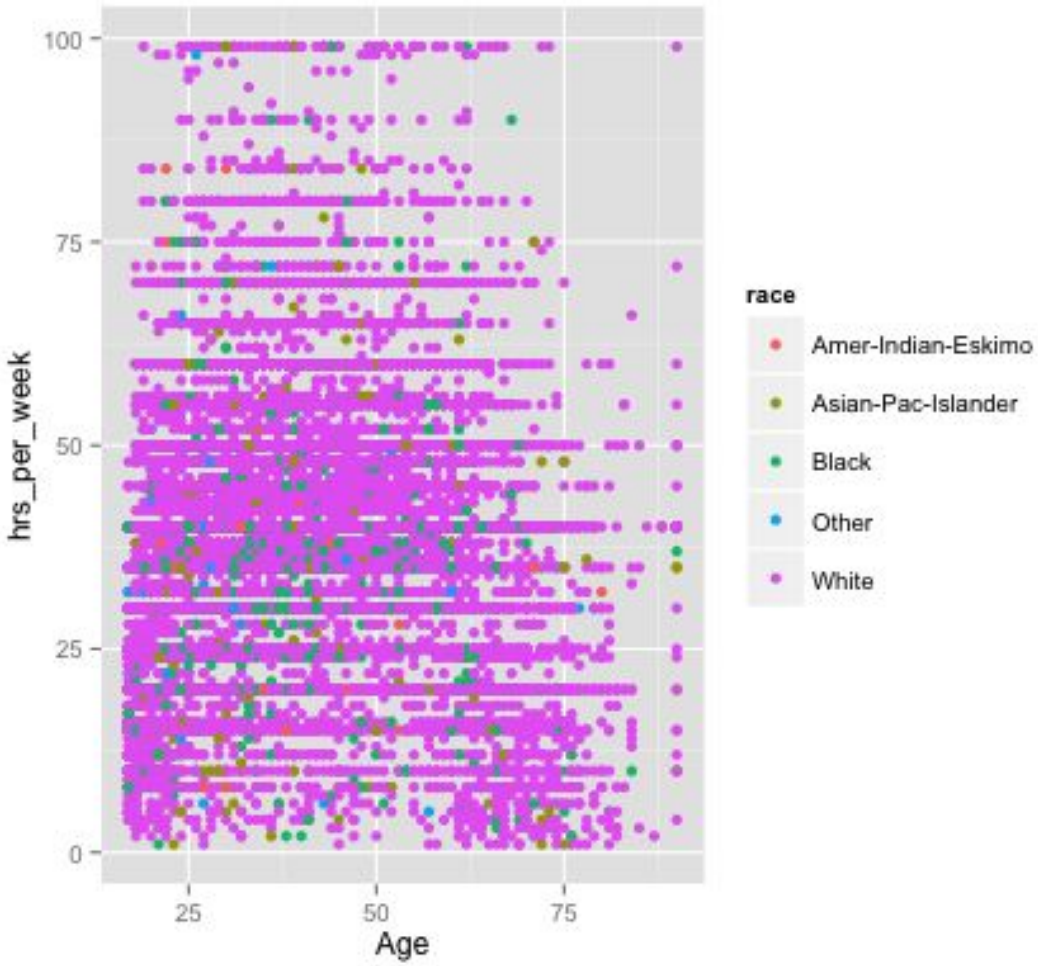
Bhargav Pabbisetti
COEN 281
10/12/2015



Bhargav Pabbisetti
COEN 281
10/12/2015



Bhargav Pabbisetti
COEN 281
10/12/2015



Bhargav Pabbisetti
COEN 281
10/12/2015