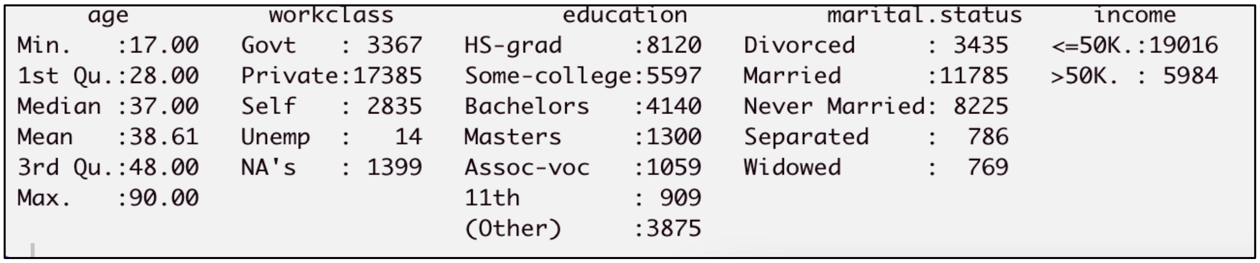
MAT 342

DIFFERENTFACTORS EFFECTING INCOME LEVELS

Kassegnin,Aretha O.(Student)

9-30-22

1. **Report of a summary of all the variables**

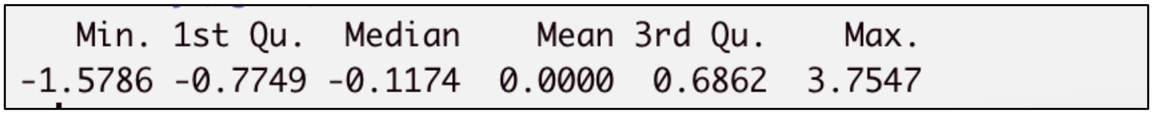


```
      age          workclass           education          marital.status     income
 Min.   :17.00   Govt   : 3367   HS-grad     :8120   Divorced     : 3435   <=50K.:19016
 1st Qu.:28.00   Private:17385   Some-college:5597   Married      :11785   >50K. : 5984
 Median :37.00   Self   : 2835   Bachelors   :4140   Never Married: 8225
 Mean   :38.61   Unemp  :   14   Masters     :1300   Separated    :  786
 3rd Qu.:48.00   NA's   : 1399   Assoc-voc   :1059   Widowed      :  769
 Max.   :90.00                   11th        : 909
                                 (Other)     :3875
```

Figure 1: 5 number summary of all the variables in dataset.

a. Yes, there are missing values in the dataset. There are 1399 missing variables in the workclass"variable. There are no missing variables in education, marital status, and income and age.
b. The only numerical variable is "age" and it is skewed right.
c. For the categorical variables the modes for them were "workclass" most people worked in the private sector, 17386. For the "education "variable, there was 8120 people who choose HS - grad. For "marital status 11785 people said they were married. Lastly, for income most people selected that they make less then or 50K per year is 19016.
d. The "typical" person would be around their 30's and, work in the private sector, be a high school graduate, married and make less then 50K per year.
e. Yes there are typical people in the data set because what I choose most people selected

2. **Z-Scores**

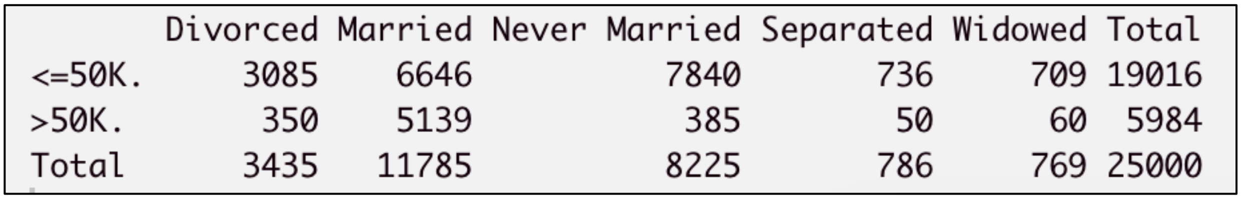


```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-1.5786 -0.7749 -0.1174  0.0000  0.6862  3.7547 
```

Figure 2: The Z score of Age variable 5 number summary.

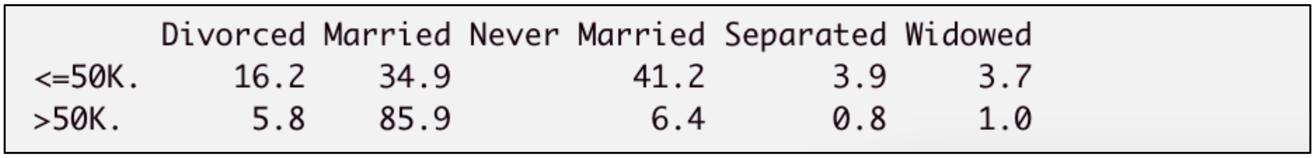
Yes, there are outliers because the max is greater is then 3. To find the outliers max outliers (because there was no min outliers) I used the IQR. Anyone over the age of 78 is an outlier.

3. **Tables**



```
        Divorced Married Never Married Separated Widowed Total
 <=50K.     3085    6646          7840       736     709 19016
 >50K.       350    5139           385        50      60  5984
 Total      3435   11785          8225       786     769 25000
```

Figure 3: Contingency table of Income and Marital status.



```
        Divorced Married Never Married Separated Widowed
<=50K.      16.2    34.9          41.2       3.9     3.7
>50K.        5.8    85.9           6.4       0.8     1.0
```

Figure 4: Percent Table of Income and Marital Status.

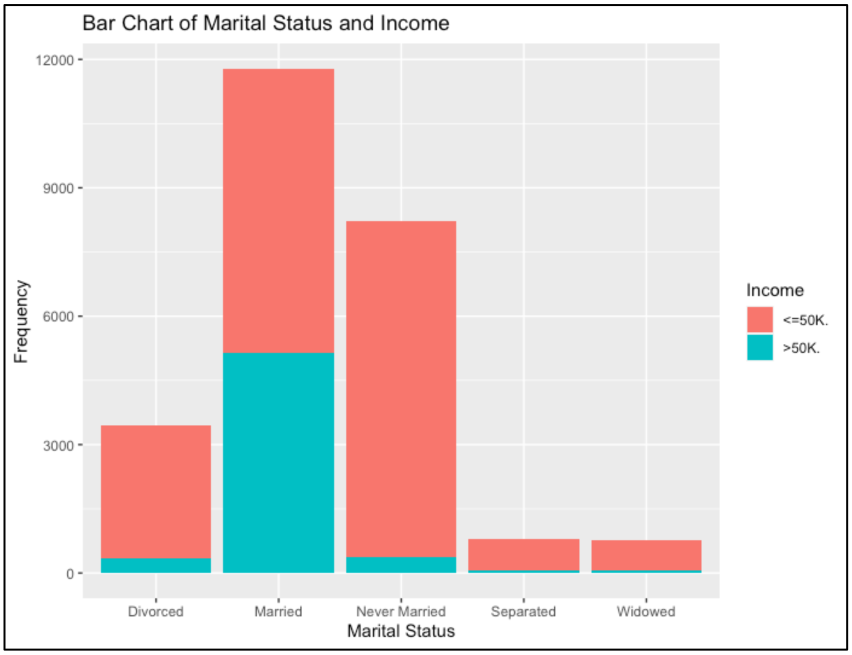


Figure 5: Bar Chart Stacked

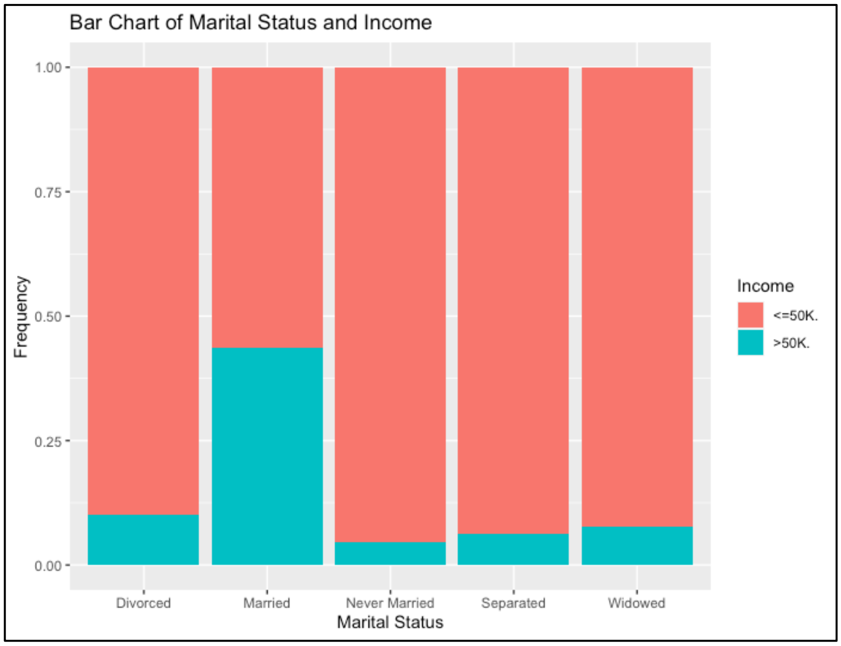
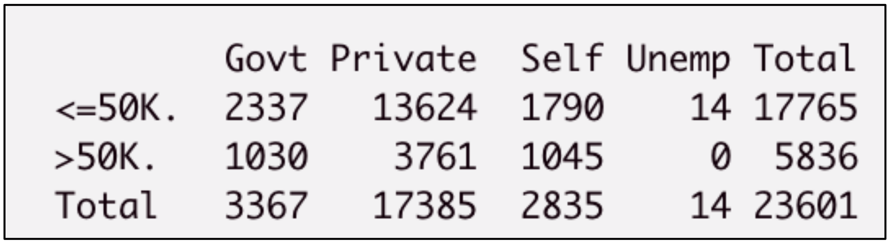


Figure 6: Bar Chart Normalized

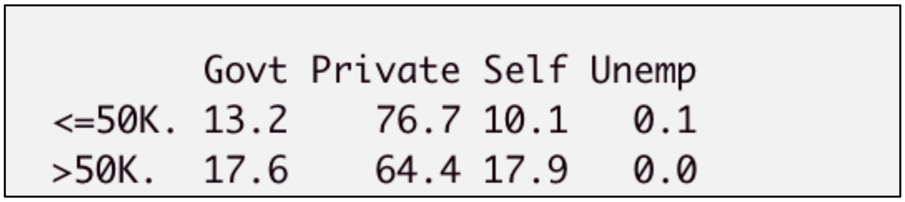
According to the data people who are married make more then 500K per year, 85.9%. Only 34.9% of married people make less then 50K per year. People who are divorced are more likely to make less then or equal to 50K per year with 3.9%. People who were never married are more likely to make less then or equal to 50K per year 41.2% while only 6.4% make more then 50K per year.

Yes, I believe that marital status is a good predictor of income because being married means that you have responsibly and would need more money to take care of things. I also believe that people who are not married don’t make as much because they just take care of themselves in most cases and don’t have to worry about anyone else. Using the data, I would fall into not married and they are most likely to make less then or equal to 50K which is also my circumstance.



```
        Govt Private  Self Unemp Total
<=50K.  2337   13624  1790    14 17765
>50K.   1030    3761  1045     0  5836
Total   3367   17385  2835    14 23601
```

Figure 7: Table Workclass and Income



```
        Govt Private Self Unemp
<=50K.  13.2    76.7 10.1   0.1
>50K.   17.6    64.4 17.9   0.0
```

Figure 8: Precent Table Workclass and Income

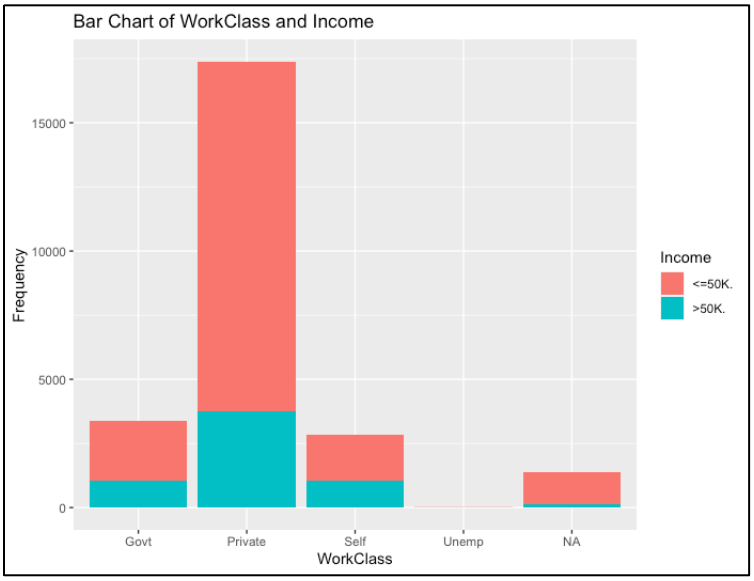


Figure 8: Bar Graph Workclass and Income Stacked

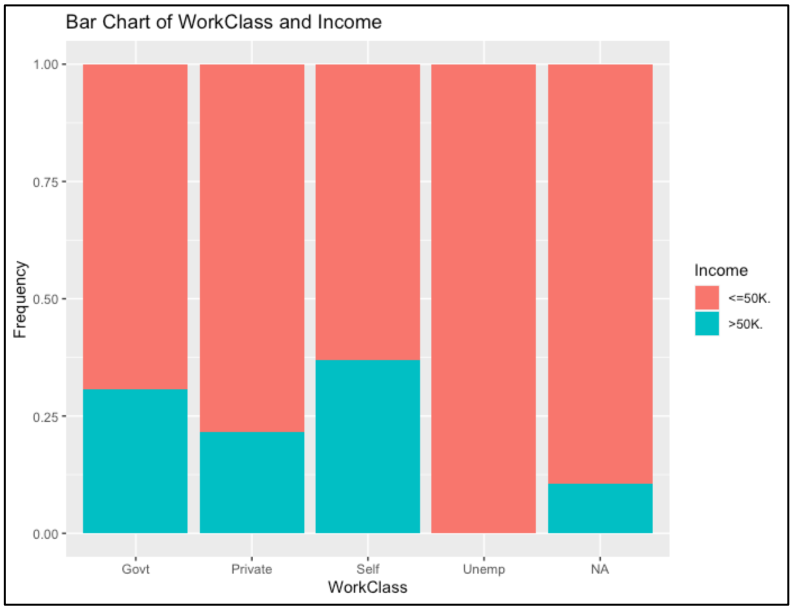


Figure 9: Bar Graph Workclass and Income Normalized

9. According to the data people who work in the private sector are more likely to make more then 50K per year, about 67%. The second workclass is the government sector with about 17% of people making more then 50K per year. It is to be noted that about 76% of people that work in the private sector make less than or equal to 50K. People who are self-employed are more likely to make more then 50K per year.

10. I don't believe that work class is a good inductor of income because for example when you look at the private sector there is no difference (a big one). When you look at the other sectors there is a difference but as much as it was with marital status.

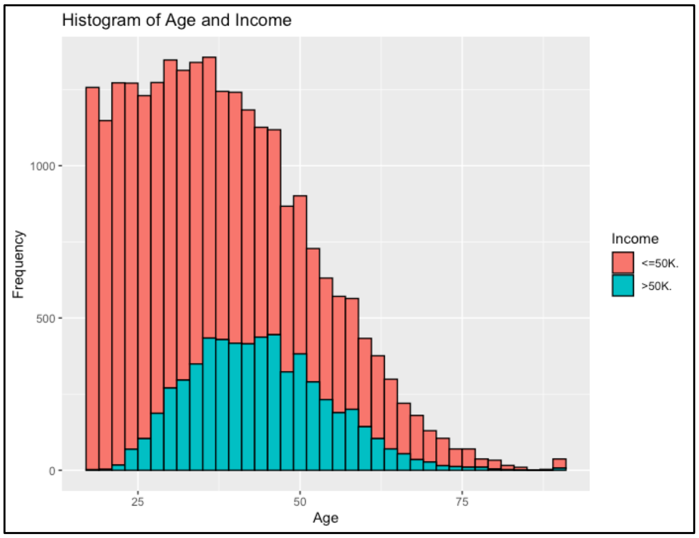


Figure 10: Stacked Histogram of Age and Income

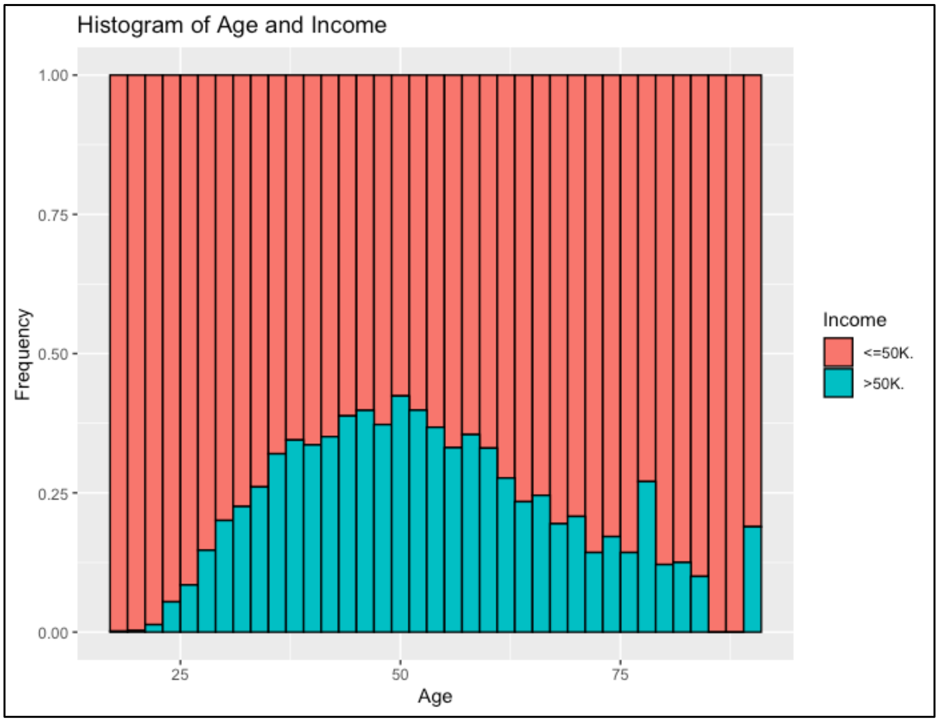


Figure 11: Normalized Histogram of Age and Income

The histograms of age and income are unimodal mostly with one main hump and everything else is part of the graph. In Figure 11 as you see there are some outliers in this dataset when you look towards the age of 75. For the most part most people below the age of 25 make less than or equal to 50K. From the age 30 to 50 we see that more people are making more then 50K for income. Lastly from 50 to 90 we see the income is decreasing and less and less people those ages make less than or equal to 50K.

Going off the data the typical person with a higher income may be like around their 30's – 50's with age and most likely work in the private sector and are married would have a higher income.

According to the data, the typical person with a lower income age would be under of 25. They would most likely work in the private sector and are not married.

According to my answers I would like to focus on workclass and the private sector I would like to break the age groups into generations to see what people older than us are doing and if that is making an income more then 50K. I said before in this report that workclass is a good variable because it looks similar and there are no differences. I would isolate the private sector and see how that effects age and income and I would like to know their education as well to see maybe that's why there is about the same people, and they make drastically.

The Code

```
library(psych)
library(plyr)
library(ggplot2)
head(adult01)
summary(adult01)
is.na(adult01)# there are some missing variables
age_skew <- adult01$age
boxplot(age_skew, horizontal = TRUE, xlab ="Age",
        main = "Boxplot of Age Variable") # to get the skewness

mean_of_age <- mean(adult01$age)
sd_age <- sd(adult01$age)
age_z<- (adult01$age - mean_of_age)/sd_age
age_z
summary(age_z)
IQR(adult01$age) # to get max outliers
#task 2
table1 <- table(adult01$income, adult01$marital.status)
table1
table1_1 <-  addmargins(A = table1, FUN = list(Total = sum),
                   quiet = TRUE)
table1_1
```

```
table1_precent<-round(prop.table(table1, margin = 1)*100, 1)
table1_precent

ggplot(adult01, aes(marital.status)) +
  geom_bar(aes(fill = adult01$income)) +
  xlab("Marital Status") +
  ylab("Frequency") +
  ggtitle("Bar Chart of Marital Status and Income") + labs(fill="Income")

ggplot(adult01, aes(marital.status)) +
  geom_bar(aes(fill = adult01$income),position = "fill") +
  xlab("Marital Status") +
  ylab("Frequency") +
  ggtitle("Bar Chart of Marital Status and Income")+labs(fill='Income')



ggplot(adult01, aes(marital.status)) +
  geom_bar(aes(fill = income)) +
  xlab("Marital Status") +
  ylab("Frequency") +
  ggtitle("Bar Chart of Marital Status and Income")

# income and workclass
```

```
table2 <- table(adult01$income, adult01$workclass)
table2
table2_1 <- addmargins(A = table2, FUN = list(Total = sum),
                       quiet = TRUE)
table2_1
table2_precent<-round(prop.table(table2, margin = 1)*100, 1)
table2_precent


ggplot(adult01, aes(workclass)) +
  geom_bar(aes(fill = adult01$income)) +
  xlab("WorkClass") +
  ylab("Frequency") +
  ggtitle("Bar Chart of WorkClass and Income")+labs(fill='Income')


ggplot(adult01, aes(workclass)) +
  geom_bar(aes(fill = adult01$income),position = "fill") +
  xlab("WorkClass") +
  ylab("Frequency") +
  ggtitle("Bar Chart of WorkClass and Income")+labs(fill='Income')
#task 3
# hist of age and income
```

```
ggplot(adult01, aes(age))+
  geom_histogram(aes(fill = adult01$income),
                 color = "black", binwidth= 2) +
  labs(fill='Income')+ xlab("Age") + ylab("Frequency")+
  ggtitle("Histogram of Age and Income")
```

```
ggplot(adult01, aes(age))+
  geom_histogram(aes(fill = adult01$income),
                 color = "black", binwidth= 2,
                 position = "fill")+labs(fill='Income')+xlab("Age") + ylab("Frequency")+
  ggtitle("Histogram of Age and Income")
```