An Exploration in Income Levels from Across the Country

Report 01

Brendan Cubberly
30 September 2022
Explorations in Data Science

Task 1: Describe and Standardize the Data

1. **Report 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
```

Summary of all the Variables from the Data Set adult01.

Comment on

a. **Any missing values in the variables.**
   In the variable workclass, there are 1399 missing values. This means out of the 25,000 responses, 1399 did not express which member of the working class they are.
b. **Any skewness in the numeric variable(s).**
   The variable age is slightly skewed right. The minimum is 17, the maximum is 90, the median is 37, and the mean is 38.41. In this distribution, since the mean is closer to the minimum value and the mean is greater than the median, the distribution is skewed right.
c. **The modes of categorical variables(s).**
   The categorical variables are workclass, education, and marital.status. The mode of each of these variables is Private for workclass, HS-grad for education, and Married for marital.status.
d. **What a "typical" person has for a value in each of the variables.**
   A "typical" person is around 38 years old, is a member of the Private working class, only graduated high school, is married, and makes less than or equal to $50k a year.
e. **Are there any "typical" people in this data set?**
   Yes, there are "typical" people in this data set.

2. **Obtain the Z-scores of the Age variable and save it as *age_z* in the data set. Are there any outliers (with Z-scores larger than 3 or less than -3) in the Age variable? If so, report the original variable values (e.g., the actual Age or Income values) that are outliers.**

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

Z-Scores of the Variable Age

Yes, there are outliers. The maximum is 3.7547 which is greater than 3. There are no minimum outliers.

	age	workclass	education	marital.status	income
223	90	Private	HS-grad	Never Married	<=50K.
431	80	<NA>	HS-grad	Widowed	<=50K.
919	81	Self	HS-grad	Married	<=50K.
1041	90	Private	HS-grad	Never Married	<=50K.
1169	88	Self	Prof-school	Married	<=50K.
1936	90	Private	Bachelors	Married	<=50K.

Outliers of the Z-Scores (First 6 using head() command)

<u>Task 2: Visualize the Categorical Variables with respect to Income</u>

Variable: Marital Status

3. **Make a contingency table of Income to Marital Status.**

	Divorced	Married	Never Married	Separated	Widowed
<=50K.	3085	6646	7840	736	709
>50K.	350	5139	385	50	60

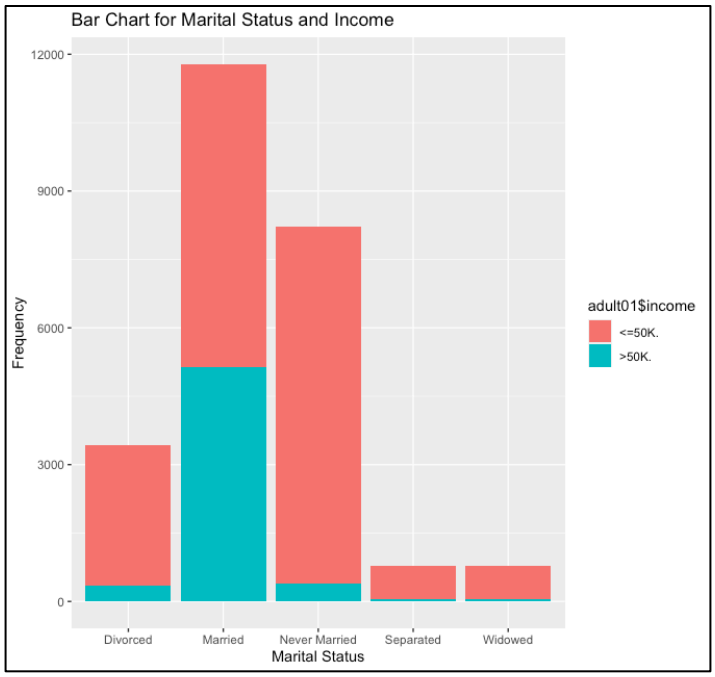
Table of Income to Marital Status

Also make a table showing the percent of Income within each level of Marital status

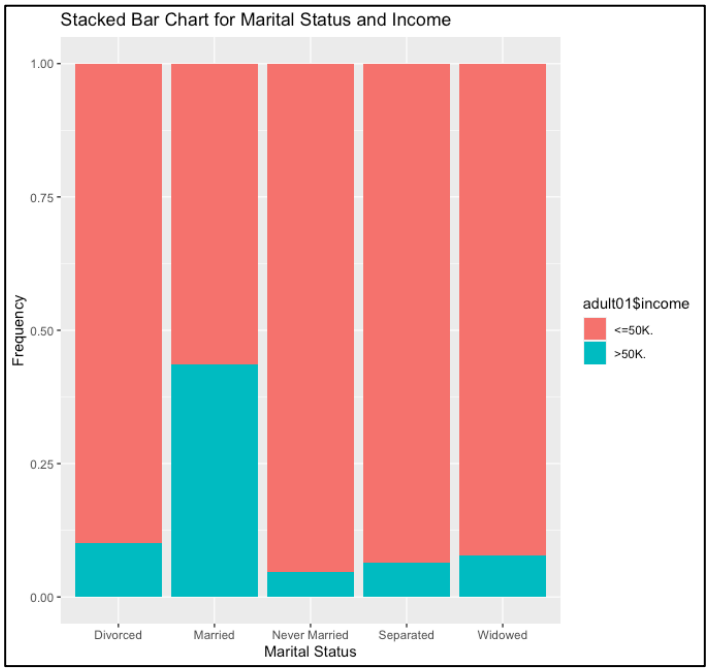
	Divorced	Married	Never Married	Separated	Widowed
<=50K.	12.340	26.584	31.360	2.944	2.836
>50K.	1.400	20.556	1.540	0.200	0.240

Percentages Table of Income to Marital Status

4. **Create a bar chart of Marital Status with Income overlay, in both stacked and normalized forms.**



Bar Chart for Marital Status on Income (Normal)



Stacked Bar Chart for Marital Status and Income

5. **Use the tables and the graphs from the previous two problems to discuss which marital statuses have higher or lower Incomes. Include specific counts and precents in your answer. You do not have to cover all marital statuses, only those with note-worthy observations.**

Of all divorced people in our data set, 12.34% (3085) have an income of less than or equal to $50k, and only 1.4% (350) had an income of greater than $50k. Those who were married had both higher percentages in income compared to the rest of the data. 26.584% (6646) had an income of less than or equal to $50k, and 20.556% (5139) made more than $50k. Those who had never been married, 31.360% (7840) made less than or equal to $50k, and only 1.54% (385) had more than $50k in income.

6. **Use the knowledge gleaned from the past few questions to discuss whether Marital status seems to be a good predictor of Income, and why.**
   No, marital status is a not good predictor of income because in each marital status level, most people made less than or equal to $50k. While the marital status Married had a the most responses for an income higher than $50k, most responses in this level made less than $50k in income. This means that regardless of marital status, most people will likely have an income of no more than $50k.

7. **Make a contingency table of Income and Workclass.**

	Govt	Private	Self	Unemp
<=50K.	2337	13624	1790	14
>50K.	1030	3761	1045	0

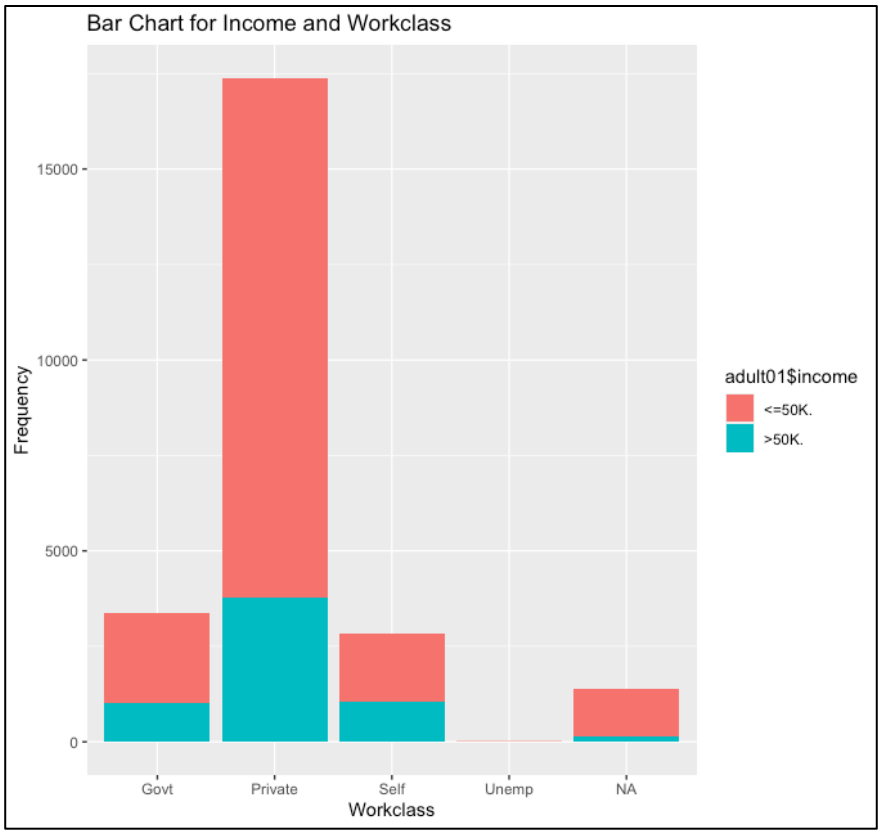
Table of Income and Workclass

Also make a table showing the percent of Income within each level of Workclass.

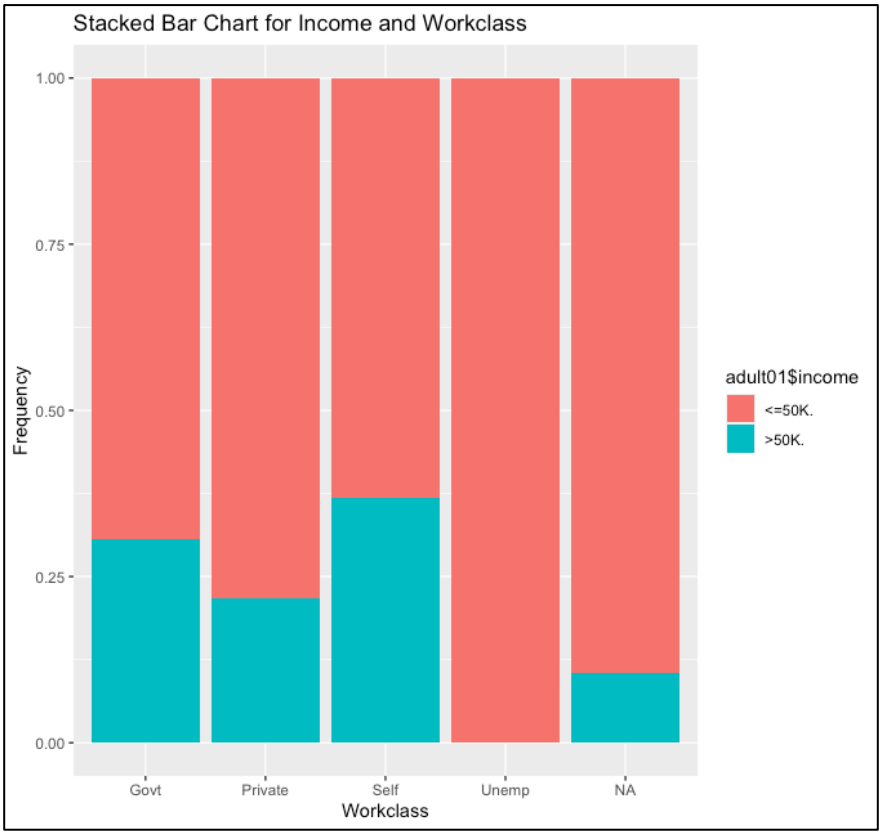
	Govt	Private	Self	Unemp
<=50K.	9.90212279	57.72636753	7.58442439	0.05931952
>50K.	4.36422186	15.93576543	4.42777848	0.00000000

Percentages Table of Income and Workclass

8. **Create a bar chart of Workclass with Income overlay, in both stacked and normalized forms. Don't worry about a messy X-axis for now.**



Bar Chart for Income and Workclass



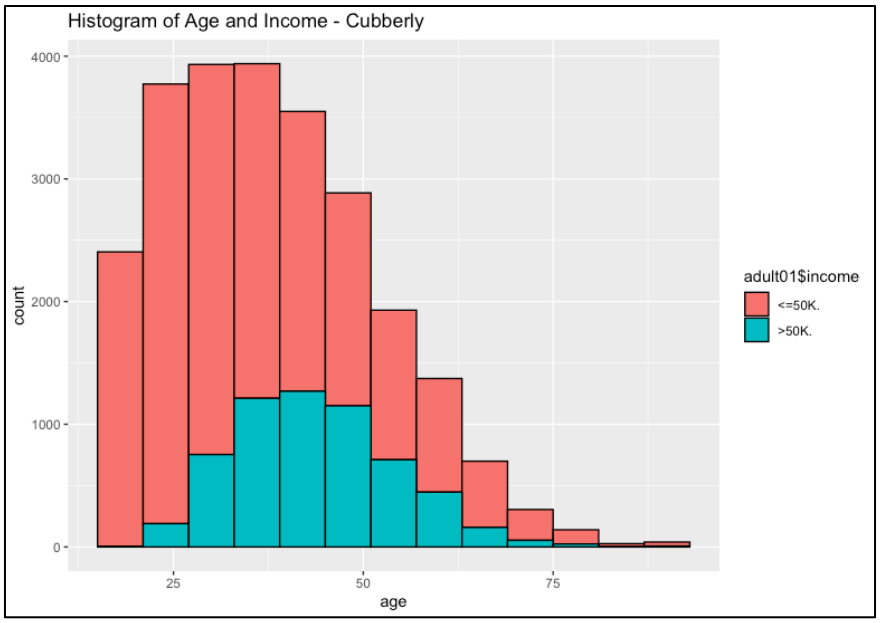
Stacked Bar Chart for Income and Workclass

9. **Use the tables and the graphs from the previous two problems to discuss which work classes have higher or lower Incomes. Include specific counts and percents in your answer. You do not have to cover all work classes, only those with note-worthy observations.**
   Most people of our data set are privately employed, of which 57.73% (13624) have an income of no more than $50k, while only 15.94% of those who made more were privately employed. For the unemployed, 0% made over $50k in income.

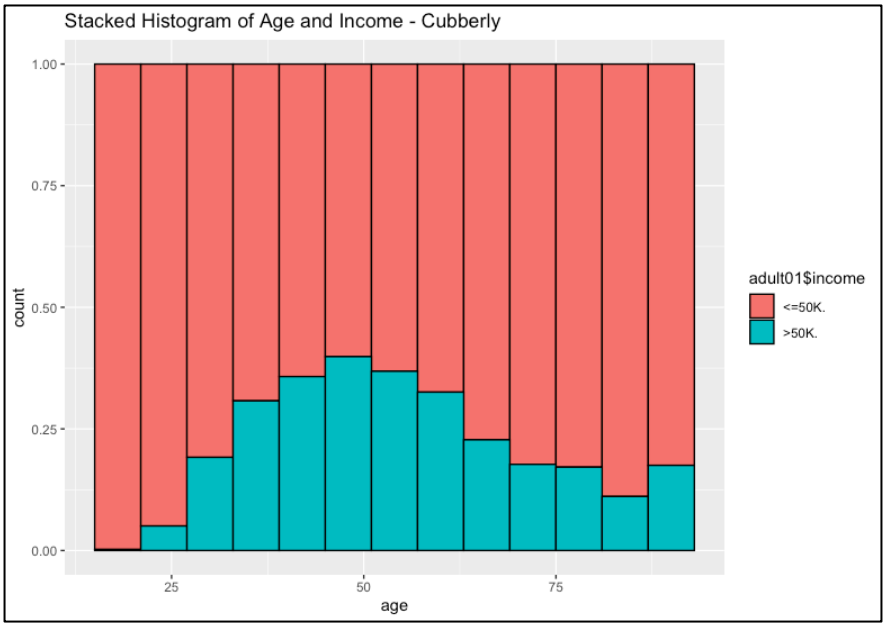
10. **Use the knowledge gleaned from the past few questions to discuss whether Workclass seems to be a good predictor of Income, and why.**
    Yes, Workclass seems to be a good indicator of income because those who are unemployed make no more than $50k in income. Those who work for private employers have a higher percentage of making more than $50k in income (15.94%), but they may also make no more than that (57.35%). Those who are self-employed have a higher percentage of making less than or equal to $50k in income, but the difference between making either more than $50k and no more than $50k in income are not too far off unlike those who work privately. For self-employers, 7.58% (1790) make no more than $50k, but only 4.43% (1045) make more than $50k in income. The difference between 7.58% and 4.43% is only 3.15%, which is smaller than the difference between those who work privately.

Task 3: Visualize the Numeric Variables with respect to Income

11. **Make a stacked histogram of Age with an overlay of Income. Also make a normalized histogram of Age with an overlay of Income, with the same ticks on the X axis. For both, add titles with the variables' names and your last name.**



Histogram of Age and Income



Stacked Histogram of Age and Income

12. Describe the pattern of income across the ages. Is income unimodal, bimodal, or trimodal (three modes) across the Age variable? What ages have the highest or lowest income?

The histogram depicts that as age increases, income increases up until the age of 50. After 50, there is a decrease in income amongst both income levels. Income is a unimodal distribution across the age variable. Ages around 25 through 50 have the highest incomes, with those younger than 25 who have the lowest income.

Task 4: Bring it All Together

13. **Describe what the typical person with higher income may be like, in terms of age range, marital status, and work class.**
    A typical person with higher income may be between 40 and 60 years old, they may work privately, and they may be married. These are all the variables that have the highest percentages and values for having an income of more than $50k.

14. **Describe what the typical person with lower income may be like, in terms of age range, marital status, and work class.**
    A typical person with lower income may be younger than 25, they may be unemployed, and they may have never been married. These are all the variables that have the highest percentages and values for having an income of no more than $50k.

15. **Your findings in the previous two questions suggest that you have very different subsets of people in your data set, based on what income they make. If you were to split the data into sub-groups to try and isolate pockets of higher income and lower income people, how would you do it? What variables would you split the data set using, and what values or levels of those variables would create the splits? Would it benefit future analyses to have more than two splits? Support your answer using the work done in this report.**
    To split the data into sub-groups, I would split it up into 4 different groups: $0k - $24k, $25k - $49k, $50k - $74k, and $74k+. This could help give variability rather than just two subgroups in which the individual makes $0k - $50k and $51k+. We could potentially see more numbers in other categories and may see how specific ages make $75k+ or even less than $25k in income. From this report, the data was either “one or the other” which seems restrictive to the data to give a true analysis of if these observed variables from across the country impact income.

Appendix

```
install.packages("psych")
library("psych")
install.packages("fansi")
library(fansi)
install.packages("ggplot2")
library(ggplot2)
install.packages("plyr")
library(plyr)
install.packages("caret")
library(caret)
summary(adult01)
is.na(adult01)
hist(adult01$income)
hist
meanage <- mean(adult01$age)
meanage
sdage <- sd(adult01$age)
sdage
age_z <- (adult01$age - meanage) / sdage
age_z
summary(age_z)
age_outliers <- adult01[ which(age_z < -3 | age_z > 3), ]
age_outliers
head(age_outliers)

tab1 <- table(adult01$income, adult01$marital.status)
tab1

tab1per <- prop.table(tab1)*100
tab1per
ggplot(adult01, aes(income)) + geom_bar()
ggplot(adult01$marital.status, aes(adult01$income)) + geom_bar()
ggplot(adult01, aes(marital.status)) +
  geom_bar(aes(fill = adult01$income)) +
  xlab("Marital Status") +
  ylab("Frequency") +
  ggtitle("Bar Chart for Marital Status and Income")
ggplot(adult01, aes(marital.status)) +
  geom_bar(aes(fill = adult01$income), position = "fill") +
  xlab("Marital Status") +
  ylab("Frequency") +
  ggtitle("Stacked Bar Chart for Marital Status and Income")

tab2 <- table(adult01$income, adult01$workclass)
tab2
tab2per <- prop.table(tab2)*100
tab2per

ggplot(adult01, aes(income)) +
  geom_bar(aes(fill = adult01$workclass)) +
  xlab("Income") +
  ylab("Frequency") +
  ggtitle("Bar Chart for Income and Workclass")
ggplot(adult01, aes(workclass)) +
  geom_bar(aes(fill = adult01$income), position = "fill") +
  xlab("Workclass") +
  ylab("Frequency") +
  ggtitle("Stacked Bar Chart for Workclass and Income")

tab3 <- table(adult01$income, adult01$age)
tab3

ggplot(adult01, aes(age)) +
  geom_histogram(color = "black", binwidth = 20)

ggplot(adult01, aes(age)) +
  geom_histogram(aes(fill = adult01$income),
                 color = "black", binwidth = 6) + ggtitle("Histogram of Age and Income - Cubberly")

ggplot(adult01, aes(age)) +
  geom_histogram(aes(fill = adult01$income),
                 color = "black", binwidth = 6,
                 position = "fill") + ggtitle("Stacked Histogram of Age and Income - Cubberly")
```