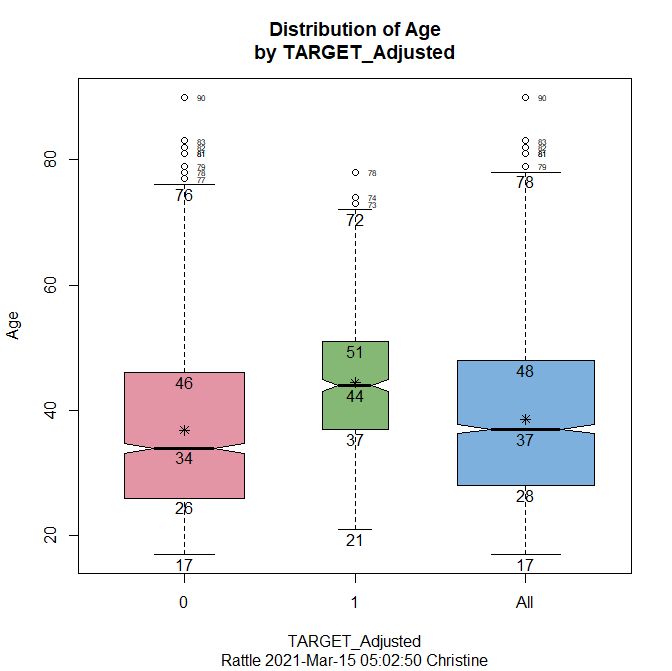
**BIAM510 Week 2 iLab Visual Summarizations**

Name: \_\_\_Christine Baxter\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_March 14, 2021\_\_

**Step 5: Numeric Data (Video 2.2.1)**

# Age/Box Plot

Copy/paste box and whiskers plot of Age below:



Answer the following question:

Q1: Using the Box and Whiskers Plot describe how Age might influence the likelihood of receiving a tax penalty. Is it statistically significant? If so at what level?  
HINT: Compare the alignment of the "0" and "1" boxes for Age on the box and whiskers plot.

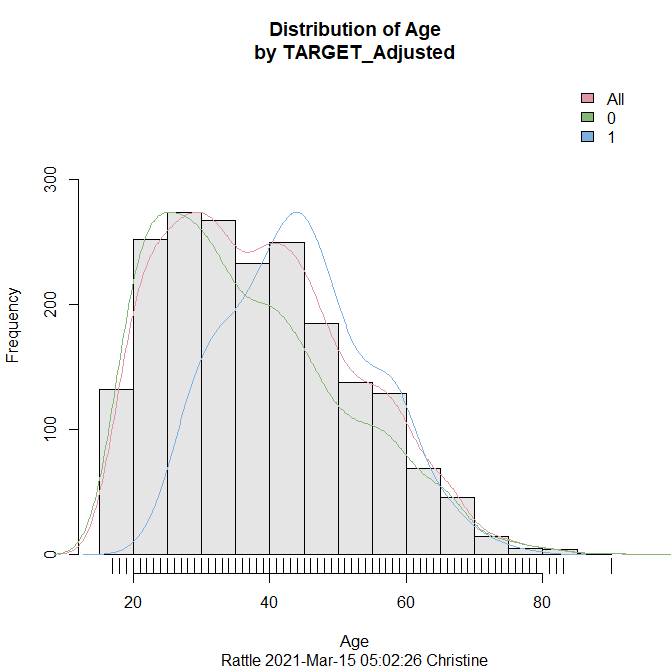
As depicted by the size of the boxes, people that did not receive additional penalties from the audit is a much larger group than those that had additional payment due to penalties. Adults between the ages of 37 and 51 with an average age of 44 were more likely to receive penalties; whereas those that did not have additional payments ranged from 26 to 46 years of age with an average age of 34. In addition, less than 25% of the people under the age of 37 received penalties but less than 25% of people under the age of 26 did not. When focusing on the Category Zero and Category 1 groups, the 95% confidence interval around the median does not overlap which means there is a statistically significant difference between the values of these two groups.

Your notes:

* TFC target adjusted = comparing the distribution of entire data set to the distribution for people who did not get their taxes adjusted because of the audit and people who did get their taxes adjusted because of the audit
* Category Zero (0) = the people that did not to have to have any sort of additional payments
* Category One (1) = those that had to have additional payments

# Age/Histogram

Copy/paste histogram of Age below:



Answer the following question:

Q2: Using the Histogram plot what information is provided that might support a difference in ages for the two groups.  
HINT: Consider the shape of the distribution you see for the two groups combined.

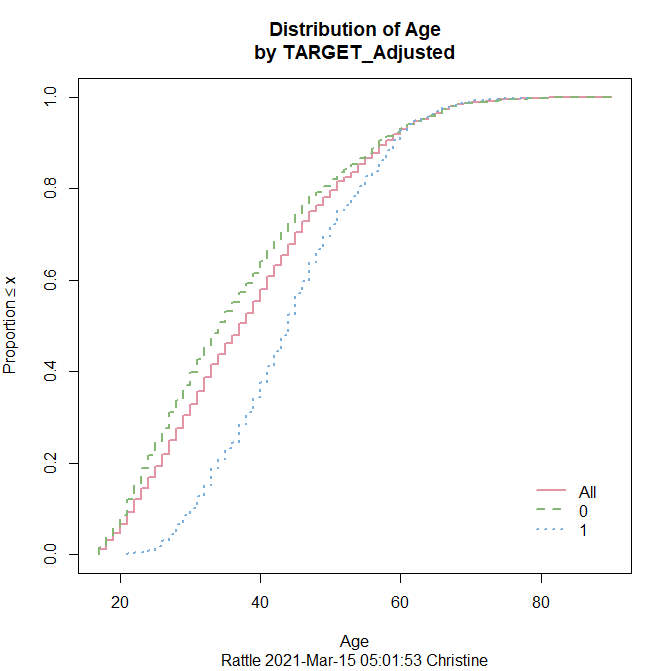
When viewing the distribution shape for the two different groups, there are two almost equal peaks where the most frequently occurring values are close to each other, producing a bi-modal distribution. In addition, Category 1 (people with penalties) becomes more normalized with a median of 44 years of age. However, when reviewing the combined groups, the frequency peaks around age 30 and the number of people with penalties to pay begins decreasing with a valley occurring between 35 and 40 years of age. There is a significant drop in frequency for adults between the ages of 45 and 50 and a significant increase in frequency for those in the 20- to 25-year-old range. When people reach the 55 to 60 age range, the frequency numbers return to those encountered for the 17- to 20-year-olds.

Your notes:

* TFC target adjusted = comparing the distribution of entire data set to the distribution for people who did not get their taxes adjusted because of the audit and people who did get their taxes adjusted because of the audit
* Category Zero (0) = the people that did not to have to have any sort of additional payments
* Category One (1) = those that had to have additional payments.

# Age/Cumulative Distribution

Copy/paste cumulative distribution of Age below:



Answer the following question:

Q3: Using the Cumulative Distribution plot describe where the differences in the two groups might be the greatest.  
HINT: Examine for what age range the "0" and "1" lines diverge.

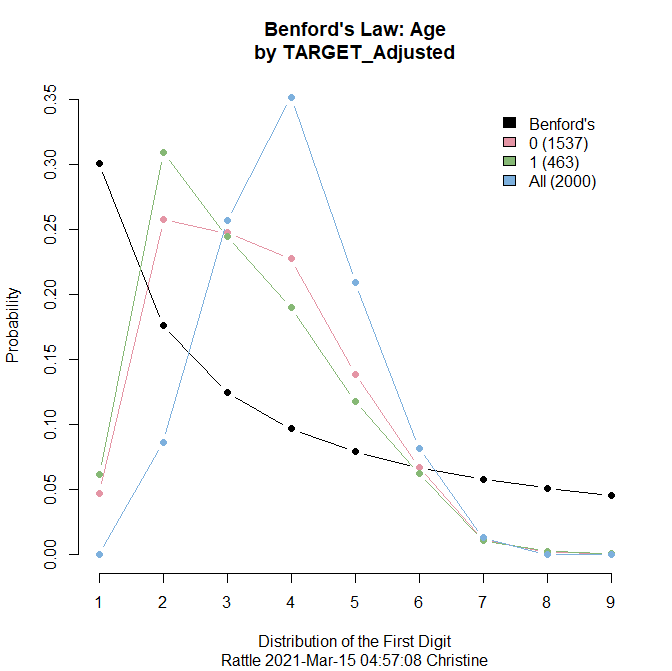
The greatest differences between Category 0 and Category 1 occur between ages 21 and 37. At age 37, a person is 25% more likely to have penalties versus a person in their twenties.

Your notes:

Enter your notes on these results in a green font.

# Age/Benford

Copy/paste Benford's Law plot of Age below:



Answer the following question:

Q4: Does Age follow Benford's Law?

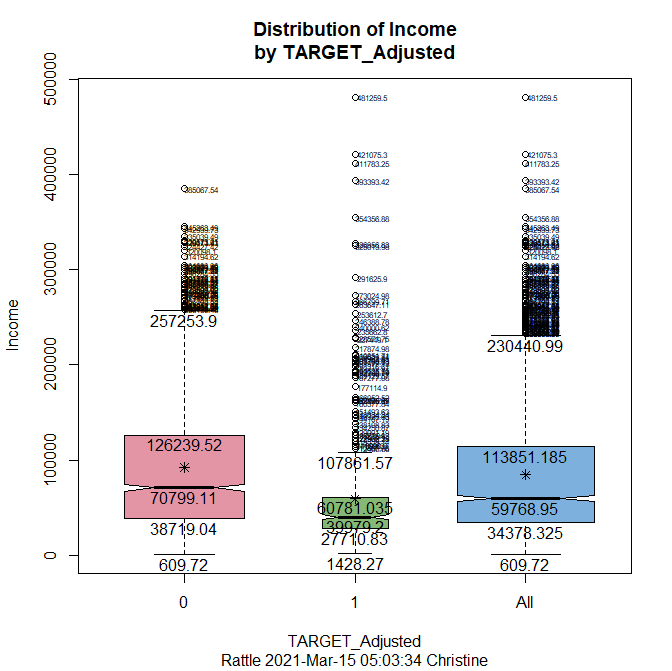
No.

Your notes:

Enter your notes on these results in a green font.

# Income/Box Plot

Copy/paste box and whiskers plot of Income below:

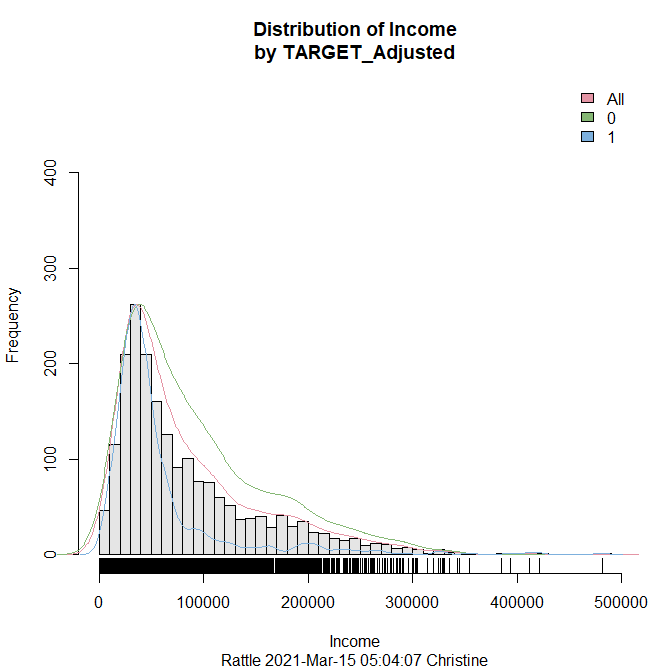


Your notes:

* No overlap of 95% confidence interval around the median, showing there is a statistically significant difference between the values of these two groups
* Category 1 has many outliers
* Size of boxes indicate people that did not receive additional penalties from the audit had larger incomes than those that had additional payments due to penalties

# Income/Histogram

Copy/paste histogram of Income below:



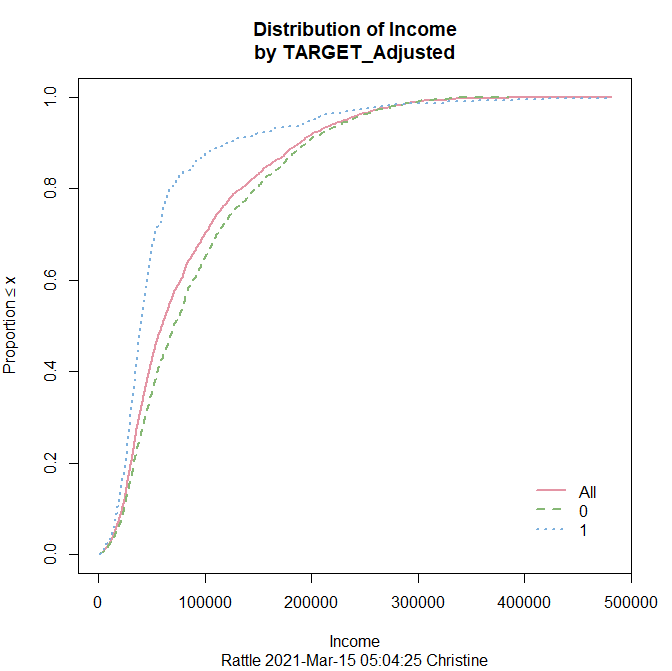
Your notes:

Enter your notes on these results in a green font

* TFC target adjusted = comparing the distribution of entire data set to the distribution for people who did not get their taxes adjusted because of the audit and people who did get their taxes adjusted because of the audit
* Category Zero (0) = the people that did not to have to have any sort of additional payments
* Category One (1) = those that had to have additional payments
* Combined group (All): the frequency peaks around 275 with an income range of $30,000 to $40,000
* Significant drop in frequency when income is > $40,000
* Significant increase in frequency when income is > $20,000

# Income/Cumulative Distribution

Copy/paste cumulative distribution of Income below:

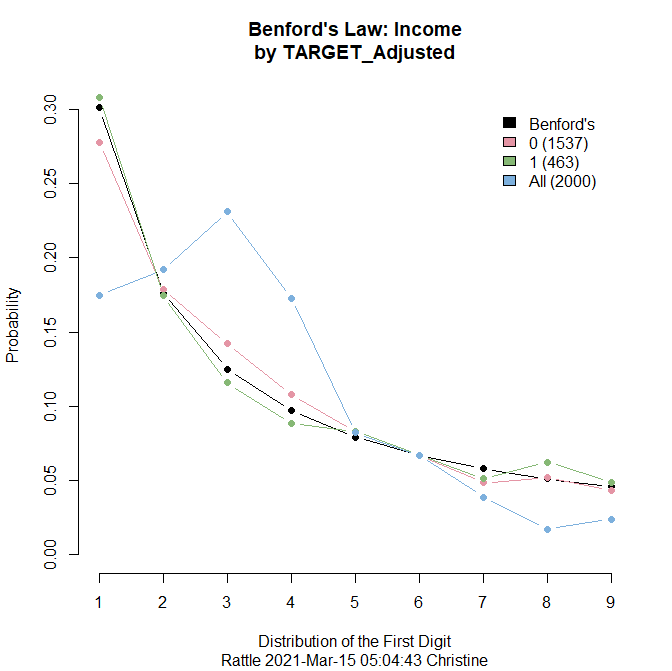


Your notes:

* Approximately 30% increase in income from $50,000 to $125,000
* 95% of people with income > $200,000 pay penalties

# Income/Benford

Copy/paste Benford's Law plot of Income below:



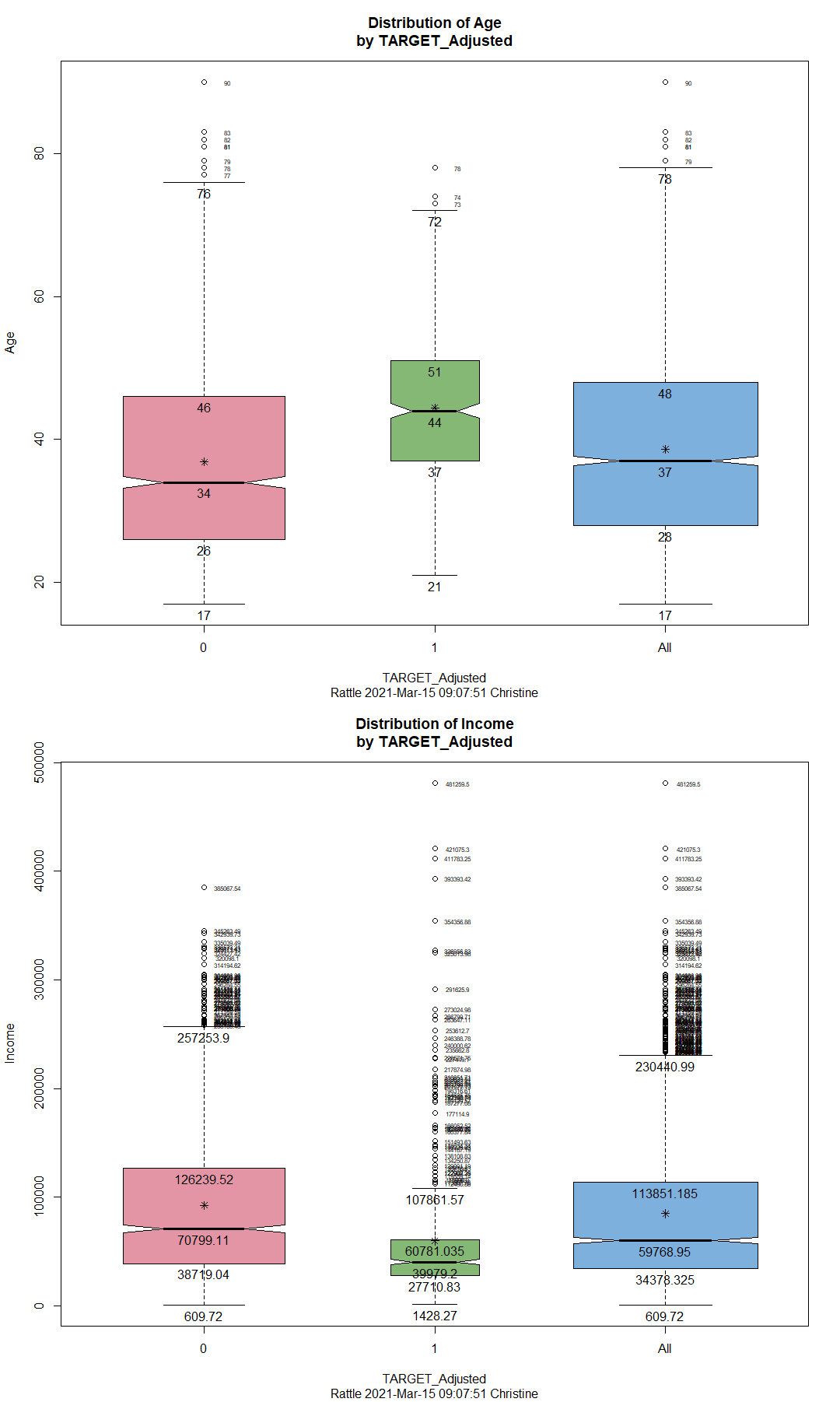
Your notes:

Income for Category 0 and Category 1 follows Benford’s Law; however, income levels for combined group (All) does not seem to completely follow Benford’s Law.

Answer **any one** (your choice) of the following 4 questions:

Q5: Comparing the Box and Whiskers plots of Age and Income which variable should we be most concerned about regarding the existence of outliers? Why?

We should be most concerned with the Income variable as the group with penalties to pay (Category 1) has a large number of outliers which then carries over to the combined group (All). Due to the large number of outliers, summarization/analysis results could be skewed or biased.



Q6: What do you think the relationship is between Income and the assessment of additional taxes owed by the IRS? Support your answer.

Enter your answer in a red font.

Q7: How are the distributions of Age and Income different?

Enter your answer in a red font.

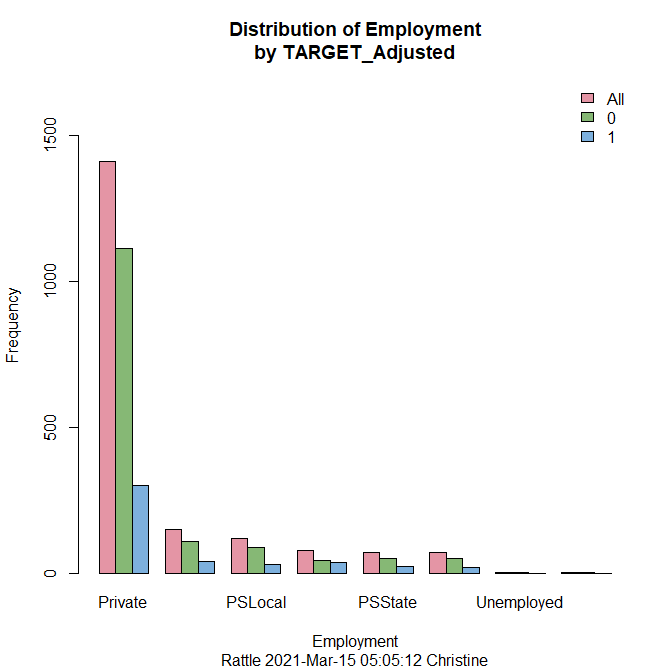
Q8: Which group does not seem to follow Benford's Law for the Income variable?

Enter your answer in a red font.

**Step 6: Categorical Data (Video 2.2.2)**

# Employment/Bar Plot

Copy/paste bar plot of Employment below:

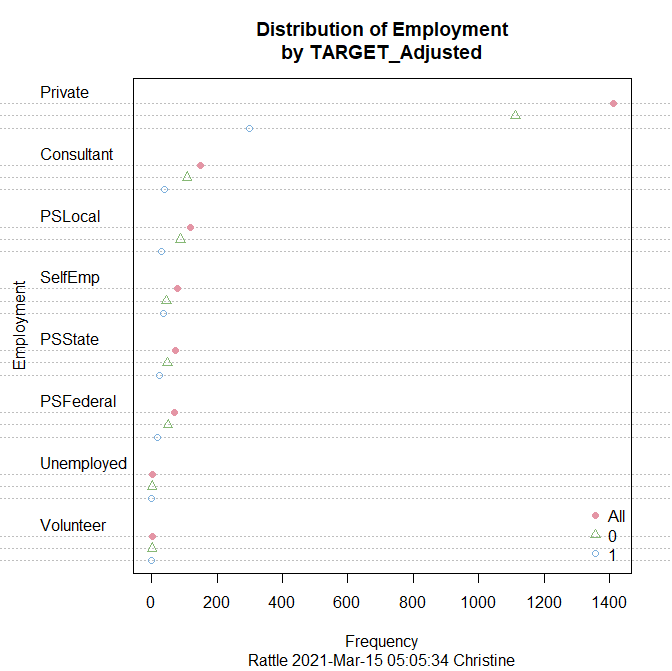


Your notes:

* Large frequency number for Private employment within Category 0 versus Category 1
* Category Zero’s frequency number for Private employment is approximately 77% more than Category One’s

# Employment/Dot Plot

Copy/paste dot plot of Employment below:

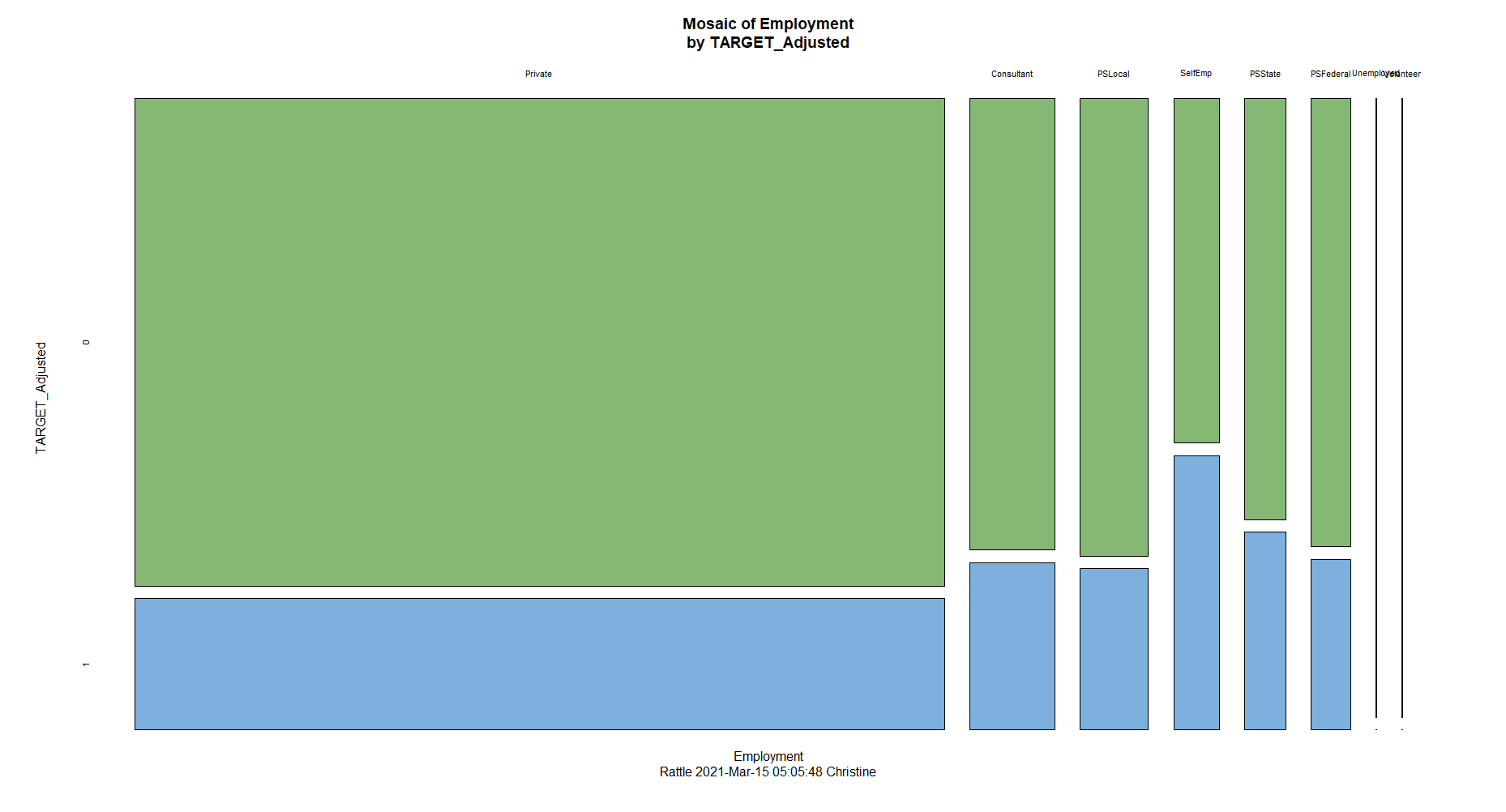


Your notes:

* All employment categories other than Private have a frequency number <200
* Majority of Private employment category did not have to pay penalties

# Employment/Mosaic Plot

Copy/paste mosaic plot of Employment below:



Your notes:

Higher number of Self Employed (SelfEmp) workers paid penalties versus the other employment categories

Answer the following question:

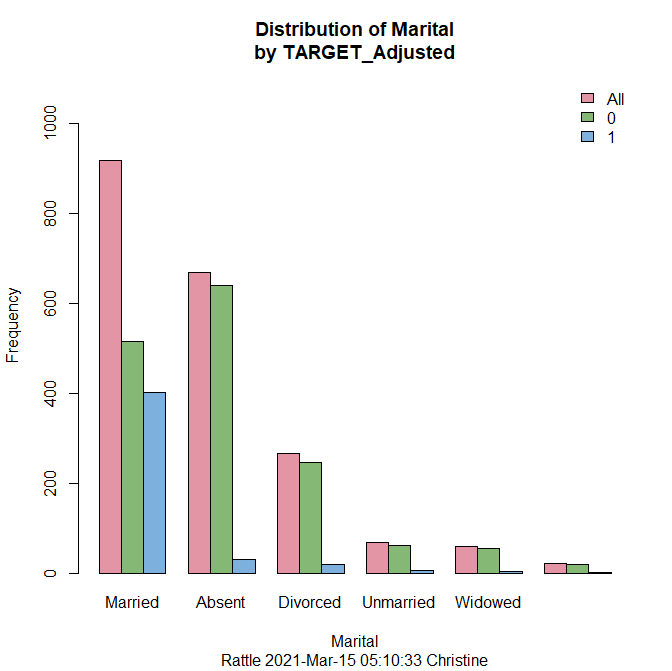
Q9: What should you keep in mind regarding the possible relationship between Employment and audit results based on the three plots?  
HINT: Consider whether any one Employment category seems to have a greater or lesser chance of having an audit adjustment than do other categories.

Apart from Self Employed (SelfEmp) workers, there seems to be a fair balance between Employment and audit results. As depicted in the Mosaic chart, SelfEmp workers have a greater chance of having an audit adjustment.

Select **one** of the remaining categorical variables (Education, Marital, Occupation, or Gender-- your choice) and generate the bar, dot, and mosaic plots for this variable. Paste each plot under the corresponding heading below, changing "Your Variable" in each heading to the variable you selected.

# Marital Status/Bar Plot

Copy/paste bar plot of your chosen variable below:

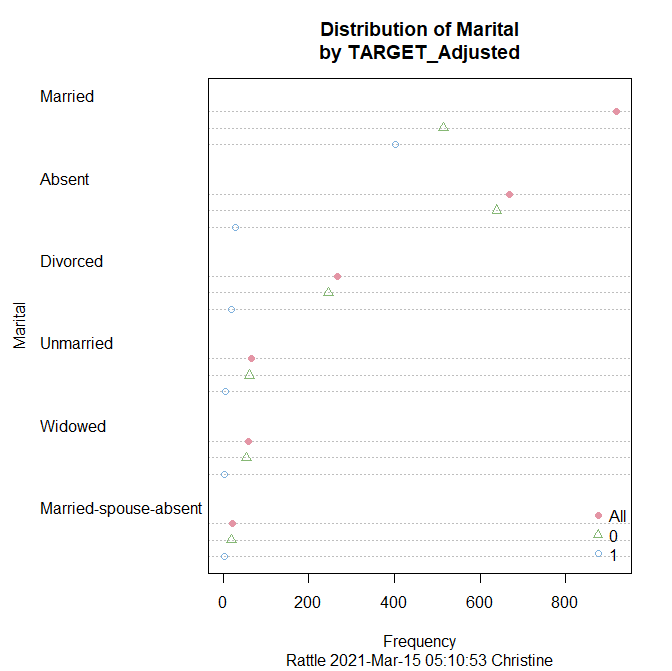


Your notes:

Enter your notes on these results in a green font.

# Marital Status/Dot Plot

Copy/paste dot plot of your chosen variable below:

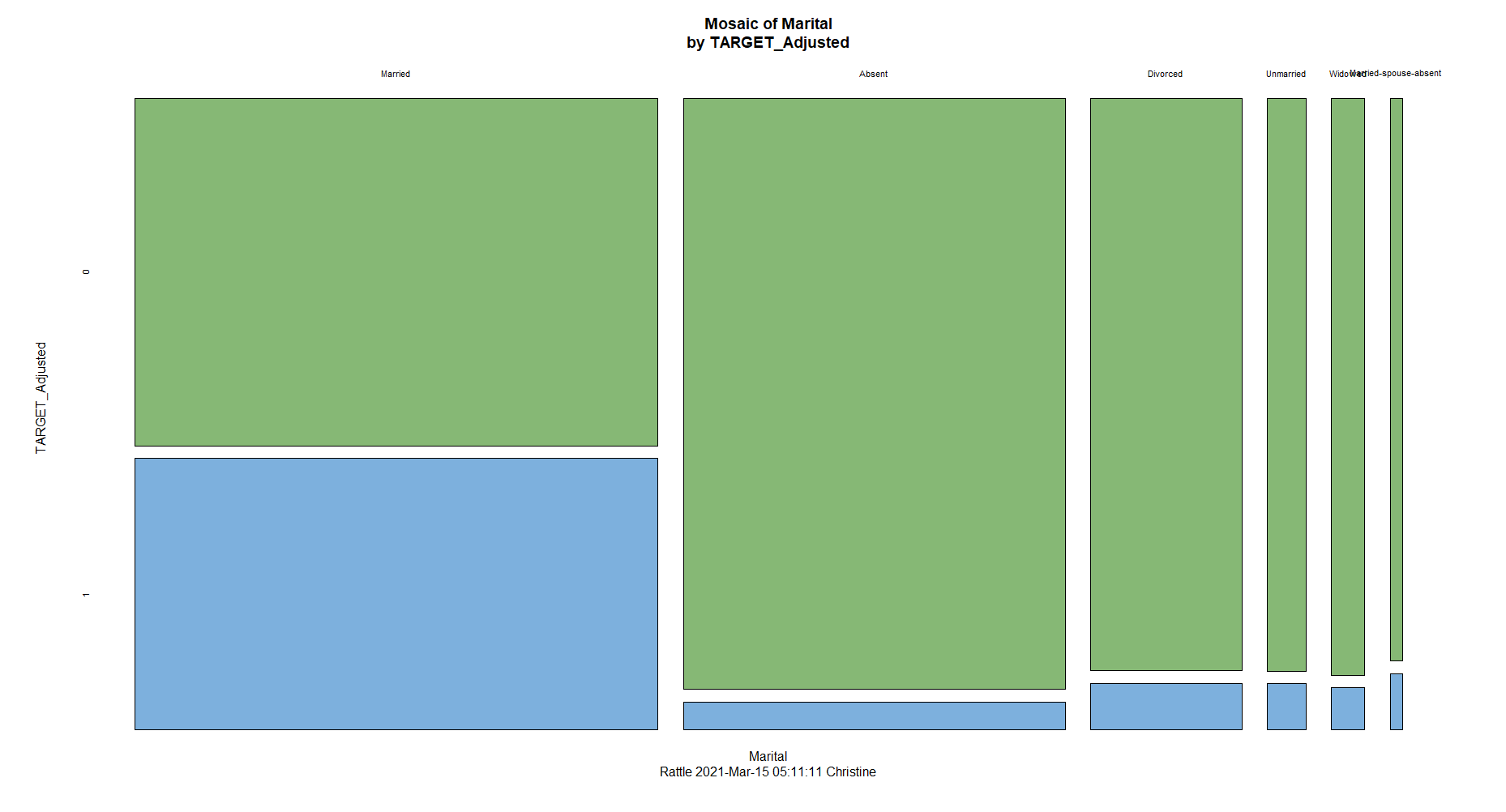


Your notes:

Enter your notes on these results in a green font.

# Marital Status/Mosaic Plot

Copy/paste mosaic plot of your chosen variable below:



Your notes:

* Married people seem more likely to pay penalties
* All Marital Status categories with the exception of ‘Married’ are less likely to pay penalties

Answer **only** the question from the list below that pertains to the variable for which you generated the plots above.

Q10: What should you keep in mind regarding the possible relationship between Education and audit results based on the mosaic plot?

Enter your answer in a red font.

Q11: Often the dataset will contain variables with ambiguous values. In this case the definitions of the levels of Marital are not clear. Find the official marital status definitions that the IRS uses for filing status. Discuss how you would map the IRS values to the ones provided. (b)What should you keep in mind regarding the possible relationship between Marital Status and audit results based on the mosaic plot?

* The official marital status definitions from the IRS are as follows:

**filing status**

Determines the rate at which income is taxed. The five filing statuses are: single, married filing jointly, married filing separately, head of household, and qualifying widow(er) with dependent child.

**Single filing status**

If on the last day of the year, you are unmarried or legally separated from your spouse under a divorce or separate maintenance decree and you do not qualify for another filing status.

**Married Filing Jointly filing status**

You are married and both you and your spouse agree to file a joint return. (On a joint return, you report your combined income and deduct your combined allowable expenses.)

**Married Filing Separately filing status**

You must be married. This method may benefit you if you want to be responsible only for your own tax or if this method results in less tax than a joint return. If you and your spouse do not agree to file a joint return, you may have to use this filing status.

**Head of Household filing status**

You must meet the following requirements: 1. You are unmarried or considered unmarried on the last day of the year. 2. You paid more than half the cost of keeping up a home for the year. 3. A qualifying person lived with you in the home for more than half the year (except temporary absences, such as school). However, your dependent parent does not have to live with you.

**Qualifying Widow(er) with Dependent Child filing status**

If your spouse died in 2014, you can use married filing jointly as your filing status for 2014 if you otherwise qualify to use that status. The year of death is the last year for which you can file jointly with your deceased spouse. You may be eligible to use qualifying widow(er) with dependent child as your filing status for two years following the year of death of your spouse. For example, if your spouse died in 2012, and you have not remarried, you may be able to use this filing status for 2013 and 2014. This filing status entitles you to use joint return tax rates and the highest standard deduction amount (if you do not itemize deductions). This status does not entitle you to file a joint return.

* To reduce the ambiguity of the Marital Status variable, I would reduce the six statuses into 2 statuses. One status for “Unmarried” and one for “Married”. The Unmarried category would include Divorced, Widowed, and Unmarried. The Married category would include Married-spouse-absent, Absent, and Married. For simplicity and clarity purposes, I would not include or use Head of Household or Qualifying Widow(er) with Dependent Child.
* Based on the Mosaic, married people seem more likely to pay penalties and all Marital Status categories except for ‘Married’ are less likely to pay penalties.

Q12: What should you keep in mind regarding the possible relationship between Occupation and audit results based on the mosaic plot?

Enter your answer in a red font.

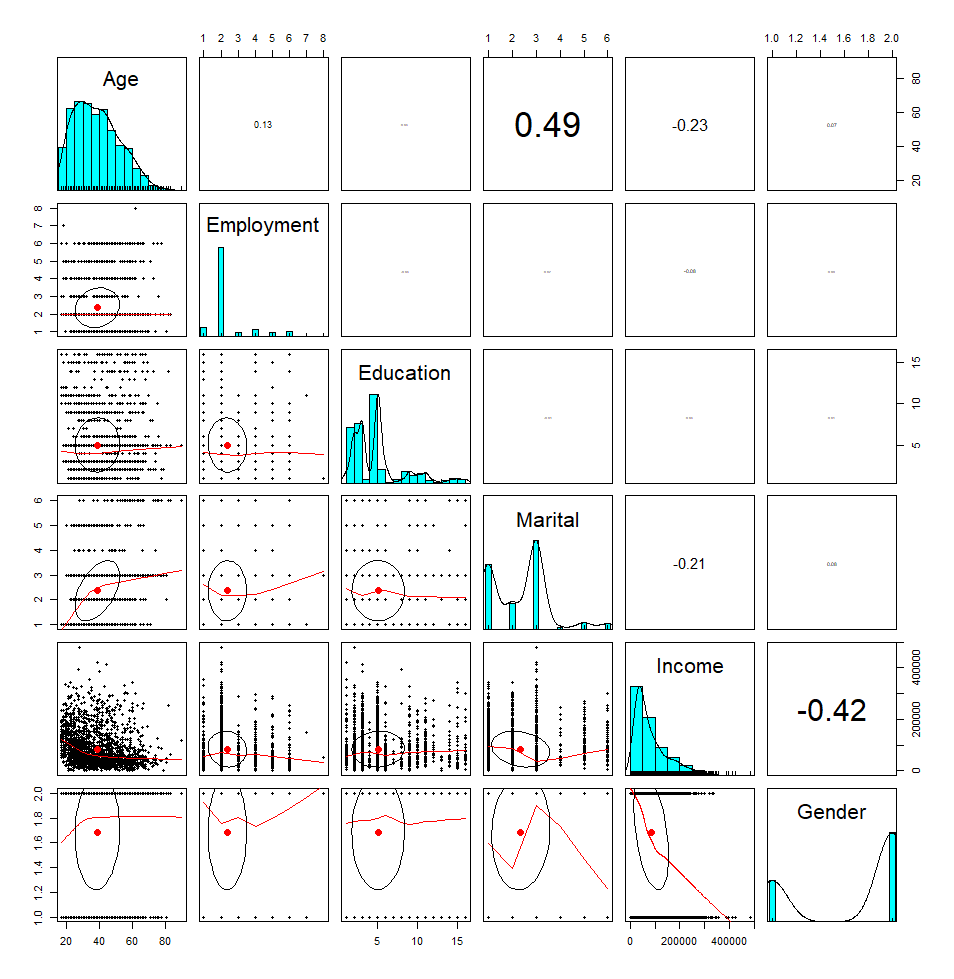
Q13: What should you keep in mind regarding the possible relationship between Gender and audit results based on the mosaic plot?

Enter your answer in a red font.

**Step 7: Pairwise Comparisons (Video 2.2.3)**

# Scatter Plot Matrix

Copy/paste at least one scatter plot matrix below:

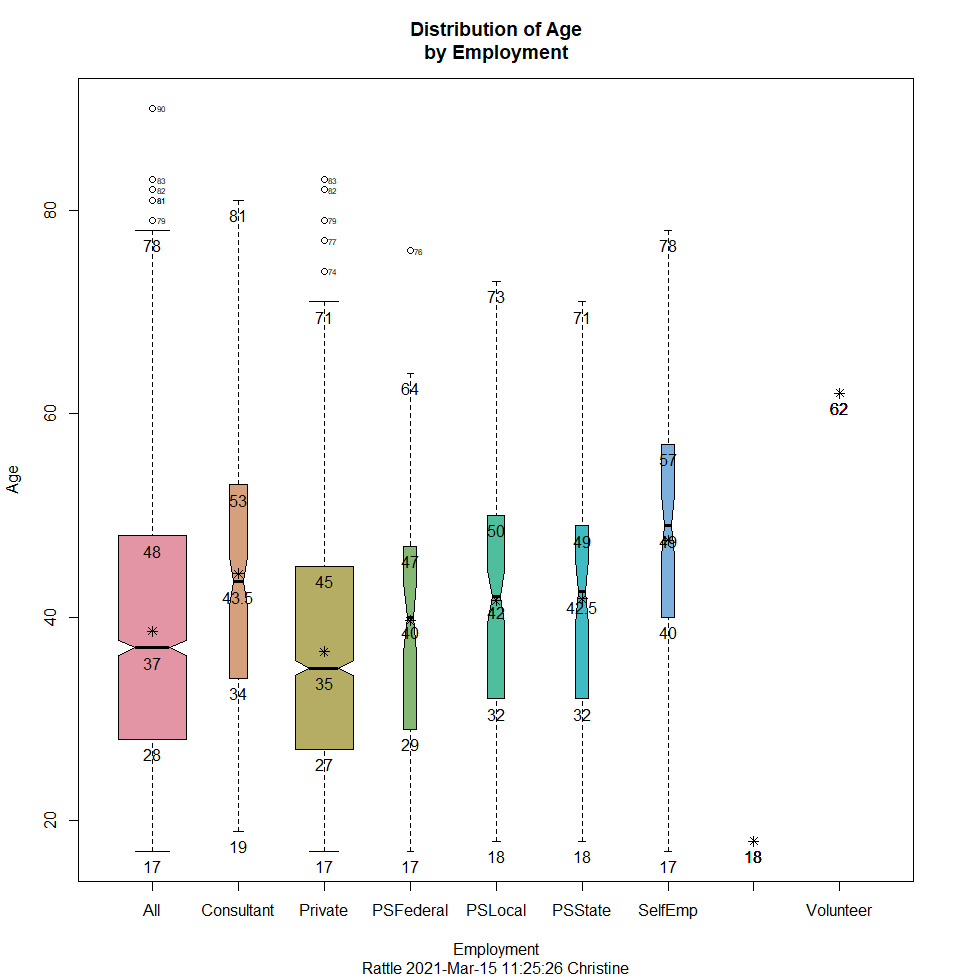


Your notes:

* Age and Marital Status
  + Positive correlation between variables
  + Potential correlation/relation between independent variables
  + .49 = .52 = .25 = 25% of variance
* Age and Income
  + Continuous variables
  + Negative correlation
  + -.212 = .04 = 4% of variance
  + Two numeric variables are not correlated, no violation of assumption
* Age and Gender
  + Negative correlation
  + -.42 = .1764 = 17.6% of variance

# Employment x Age Pairwise Comparison

Copy/paste the box plot showing distribution of Age by Employment below:

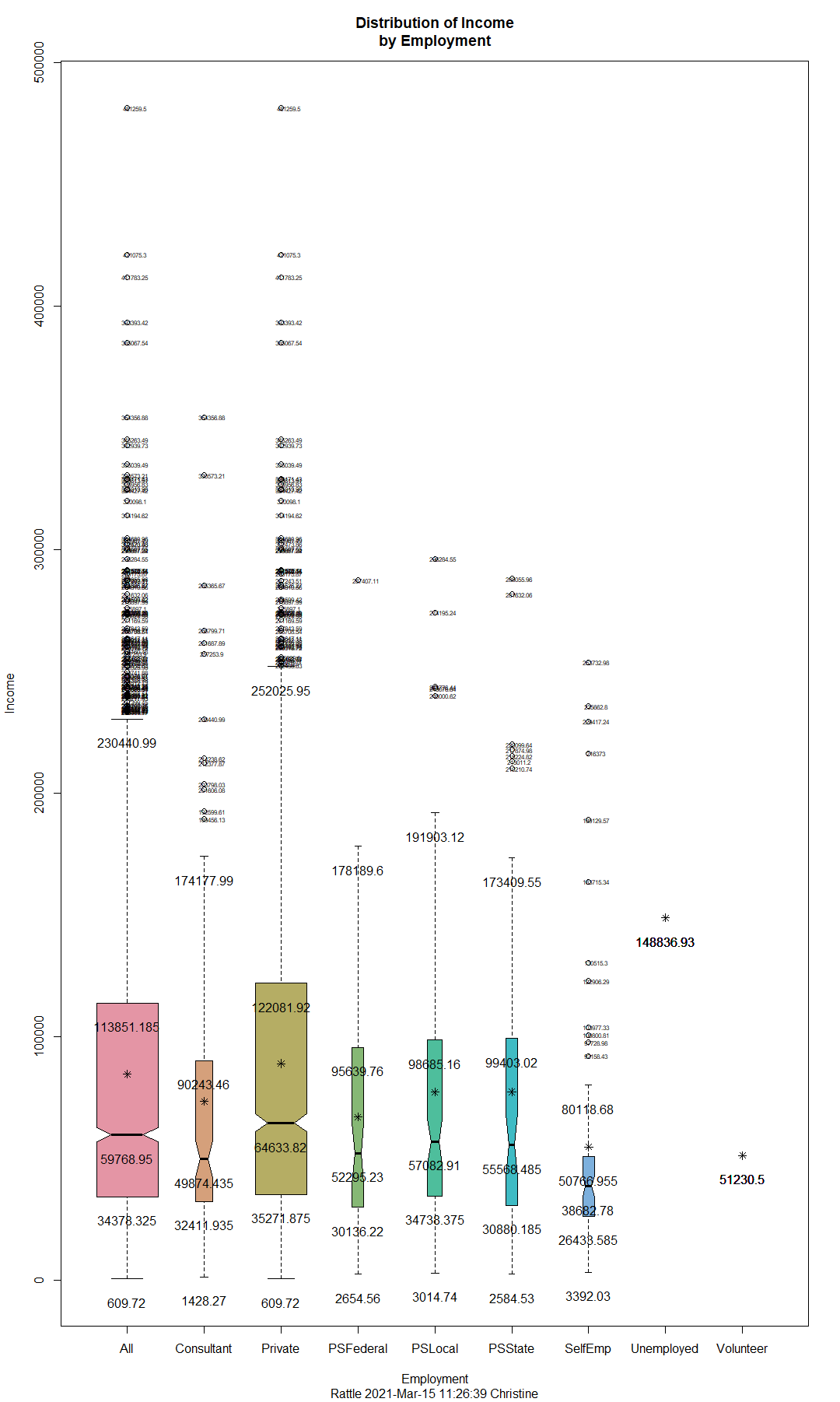


Your notes:

* Target = Employment Classification
* Outliers exist for…
  + Federal sector (1 outlier = 76 years old)
  + Private section (5 outliers = 74- to 83-years-old)
* Larger sized boxes indicates more people within that sector for the specified age range
* Adults between the ages of 53 and 57 are more likely to be self employed
* Adults between the ages of 50 and 53 are more likely to be consultants
* Federal sector…
  + Employs adults between 29 and 47 years of age
  + <25% within the 17 to 27
* Local and State sectors employ adults between the ages of 18 and 73 with half of the values falling in the 32 to 50 range
* Private sector seems more likely to hire younger workers

# Employment x Income Pairwise Comparison

Copy/paste the box plot showing distribution of Income by Employment below:



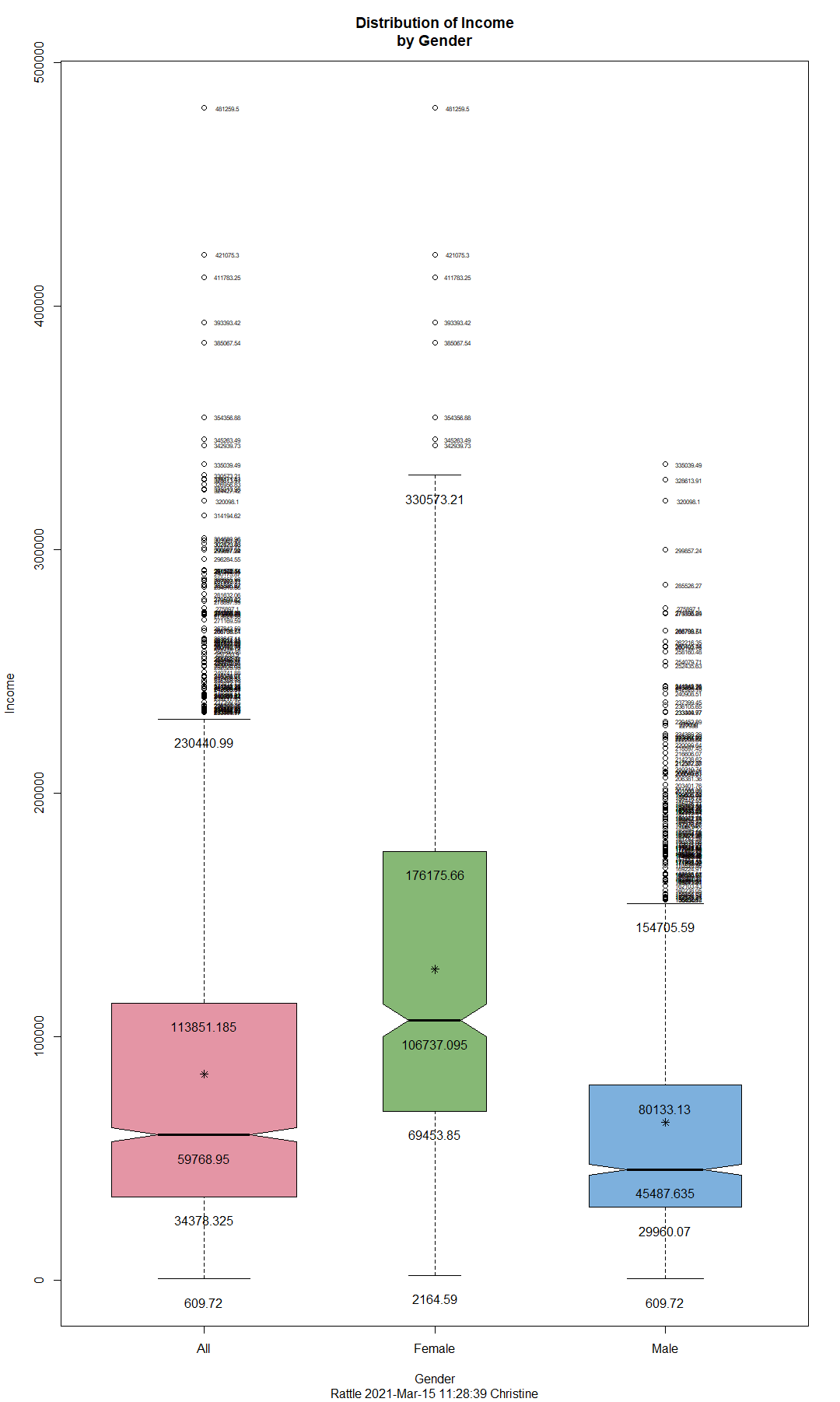
Your notes:

* Target = Employment Classification
* Larger sized box indicates more people with specified income by Employment
* Outliers
  + Potential issue with number of outliers for Consultants, Private sector, and Self-Employed workers
  + Federal sector has two outliers
* Income for Private Sector workers within the interquartile range from $35,271.88 to $122,081.92 with a median value of $64,633.82.

Generate at least **one other** pairwise comparison plot described in the video, such as Marital x Age, Gender x Income, etc. (your choice) and copy/paste it under the heading below. Change "Your Variable 1" and "Your Variable 2" in the heading to reflect your choice of variables.

# Gender x Income Pairwise Comparison

Copy/paste the box plot showing distribution of Income by Gender below:



Your notes:

* The 95% confidence interval around the median does not overlap which means there is a statistically significant difference between the values of these two groups.
* Outliers exist for…
  + Males = large number of outliers
  + Females = 8 outliers ranging from $330,573 to $481,255
* Income range for…
  + Males = $29,960.07 to $80,133.13
  + Females = $69,453.85 to $176,175.66
* Income ranges for Females are distinctly higher than male’s income ranges.

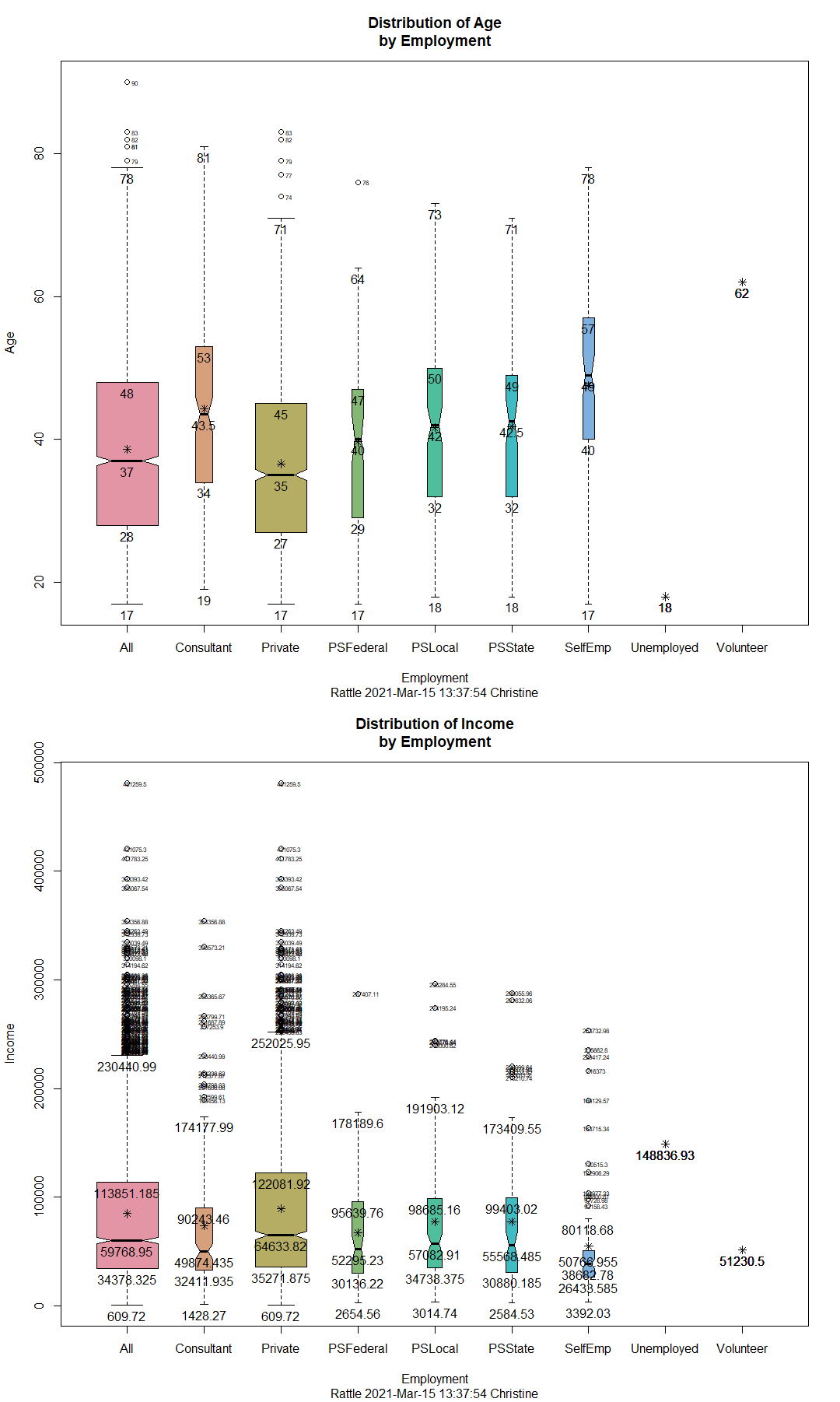
Answer **any two** (your choice) of the following 4 questions:

Q14: Based on the Pairwise Comparison plots is there a meaningful relationship between Age and Income? Explain why.  
HINT: Consider the value of the correlation coefficient for these two variables.

There is a negative correlation between the Age and Income variables with a 4% variance. Both are continuous variables and are not related to each other. As a result, there is no violation of the assumption.

Q15: Based on the two box plots of Employment X Age and Employment X Income which numerical variable is more likely to be related to Employment?  
HINT: Consider the variability in the position of the notches in the box plots within each set.

The income variable seems more likely to be related to Employment as the 95% confidence interval around the median has some minor overlapping which indicates there is not a statistically significant difference between the values of these two scenarios. There is less variability in the position of the notches within the Employment X Income set and higher variability in the position of the notches within the Employment X Age set.



Q16: If it is determined that the Marital level entitled Absent is in fact a missing observation what problems might we have with it and Age?  
HINT: Consider whether there is a relationship between Age and the Absent category in Marital.

Enter your answer in a red font.

Q17: Looking at the Gender X Income box plot describe why females are making more money than males. Use the numerical characteristics of the distribution as your answer.  
HINT: We are not looking for a causal explanation (societal, economic, etc.), but merely for what is it about the distribution that leads us to conclude this. Consider whether the difference appears to be due to outliers or is more fundamental.

Enter your answer in a red font.