STA130 Week 3 Tutorial

Kaixuan (Bryan) Hu

Today’s agenda:

- Brief overview of boxplots – when to use them, what types of information they display.
- Short group discussion relating to this week’s homework.
- Quick writing activity.
- 2nd half of tutorial reserved for meeting mentors.

Boxplot

- *When to use?*
- When you want to summarize the distribution of a quantitative (numerical) variable.
- Boxplots visualize five statistics, (Q: what are they?)
- minimum, maximum, median, 1st quartile and 3rd quartile, while also plotting unusual observations (e.g. outliers).
- You can also use a boxplot to summarize these values according to a categorical variable of interest.

Elements of a boxplot

- Line in the middle of the box: median: middle data value (50% of data values above, 50% below)
- Edges of the box:
- 1st quartile: value such that 25% of the data values less than it ($\mathbf{Q1}$)
- 3rd quartile: value such that 75% of the data values less than it ($\mathbf{Q3}$)

Example

```
x
```

```
##  [1] 0.14 0.15 0.15 0.44 0.54 0.76 0.96 1.18 1.23 2.89
```

```
data_frame(x) %>%
  ggplot(aes(x = "", y = x)) +
  geom_boxplot()
```

in ggplot

X-axis

Whiskers on the box extend to the most extreme value that is outside of the box (i.e. greater than Q3 or smaller than Q1) but within 1.5 x IQR

They are also some of the data points in your data set, but is not considered as outliers.

Interquartile Range (IQR)

- Defined as:
- IQR = 3rd quartile (Q3) - 1st quartile (Q1)
- Gives an indication of how spread out the data are.
- Boxplot can be used to **identify outliers** by using the 1.5 × IQR rule:
- Outliers are points that farther than 1.5 x IQR from the box.
- i.e., less than Q1 – 1.5 x IQR OR more than Q3 + 1.5 x IQR
- They are represented by dots on the Boxplot. (i.e. all the dots you see on the boxplot can be considered as outliers.)

What information can we get from Boxplot
(marriage dataset)

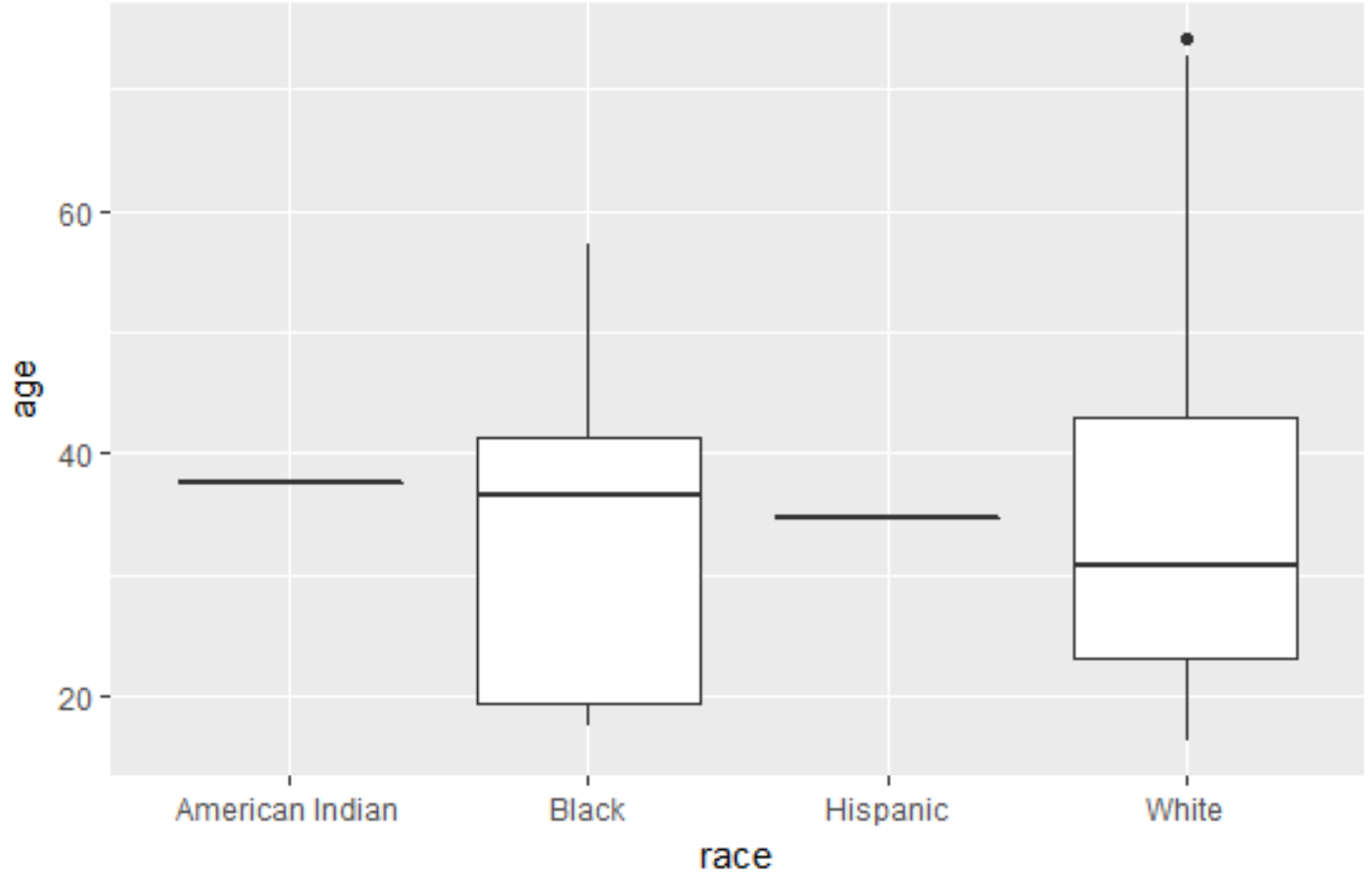
True or false?

Age of marriage does not differ between white and black people:

True

Race does not seem to be a contributing factor in the age of marriage:

True

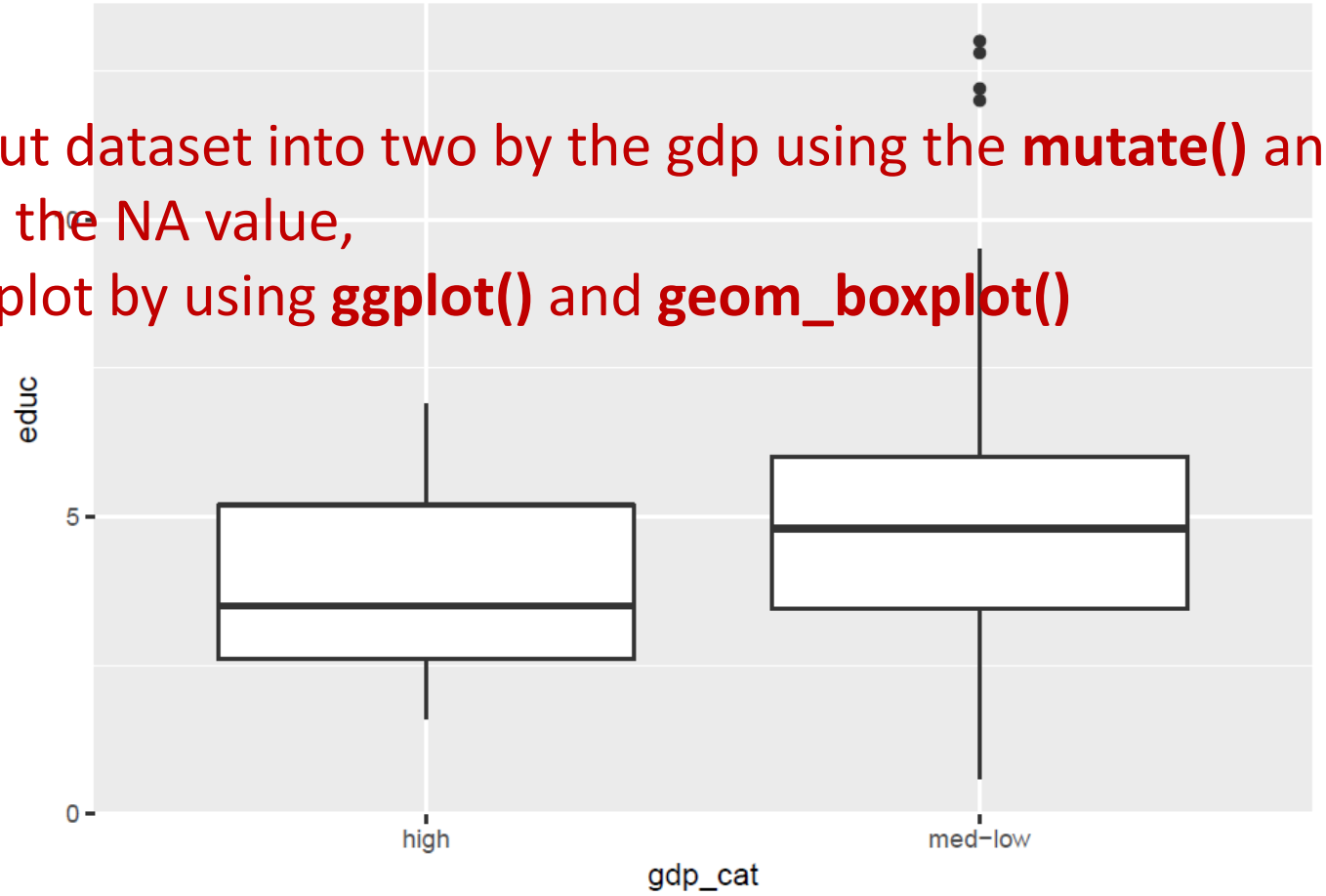


Homework3 question 2

- Use boxplots to compare the distribution the proportions of GDP spent on education for countries with a GDP of at least $50,000 compared to countries with a GDP of less than $50,000

```r
CIACountries %>%
  mutate(gdp_cat = ifelse(gdp >= 50000, "high","med-low")) %>%
  filter(is.na(gdp_cat) == F) %>%
  ggplot(aes(x = gdp_cat, y= educ)) + geom_boxplot()
```

We first sperate out dataset into two by the gdp using the **mutate()** and **ifelse()** function,
Then we filter out the NA value,
Finally we do the plot by using **ggplot()** and **geom_boxplot()**



- Example from previous data used in this class:
- (Homework 1, Q1) babies' birth weight, a continuous variable, can be visualized using a boxplot.
- Boxplot is particularly useful here if you want to show how this distribution varies by important categorical variables, such as: country of birth, month of birth, etc.
- It is a good practice to go back to homework 1 and do some boxplot on the dataset **ncbirths** by using the **geom_boxplot()** function in R.

Vocabulary for this week's material

- Outlier
- Interquartile range (IQR)
- Boxplot
- Proportion

Group Discussion

- *For Question 1D, you used both histograms and boxplots to visualize your data.*
- *Which features were easier/harder to observe from each of the visualizations?*
- *In what situations may you want to choose a boxplot over a histogram, or vice versa? Explain.*
- You may also discuss about any parts of the homework question, help each other if you cannot figure out how to solve some question or how to write code regarding those question.

Writing activity

- *Self-Reflection:*
  What questions, if any, do you have so far regarding the course materials?
- *What is one of your favorite things about tutorial? Least favorite?*
- *Any comment about the course in general.*