Variables

Continuous (numeric/quantitative): is a variable that has an **infinite number** of possible values.

Categorical: is a variable that can take on one of a **limited number** of possible values, assigning each observation to a particular group or category on the basis of some qualitative property.

Age: **Continuous** Sex: **Categorical** Race: **Categorical** Height: **Continuous**

Choosing the right graph to describe your data

- Scatter plot
- Histogram
- Box Plot
- Bar Plot

Scatter plot

- When to use:
- To see the relationship between **two continuous variables**.

- True or false Question
- We can use scatter plot to show the relationship between age and sex:
- False
- We can use scatter plot to show the distribution of people's age in Toronto:
- False

Histogram

- When to use:
- A histogram is used to plot **one continuous** **variable**. It helps to break the data into bins and shows frequency distribution of these bins. Thus, we can always change the bin size and see the effect it has on visualization.

- True or false
- We can use a histogram to plot the race of students in U of T:
- False
- We can use a histogram to show the relationship between age and heights of students:
- False

Box Plot

- When to use:
- Box Plots are used to plot a combination of **categorical and continuous variables**. Also, used for visualizing the **spread of the data** and detect outliers.

- True or false
- We can make a boxplot of age of people in this building:
- True
- We can make a boxplot of age versus sex of people in this building:
- True

Bar Plot

- When to use:
- We use Bar charts to plot a **categorical variable**.

- True or false
- We can use a bar plot to show the height of people in this building:
- False
- We can make a bar plot of race of people in this building:
- True

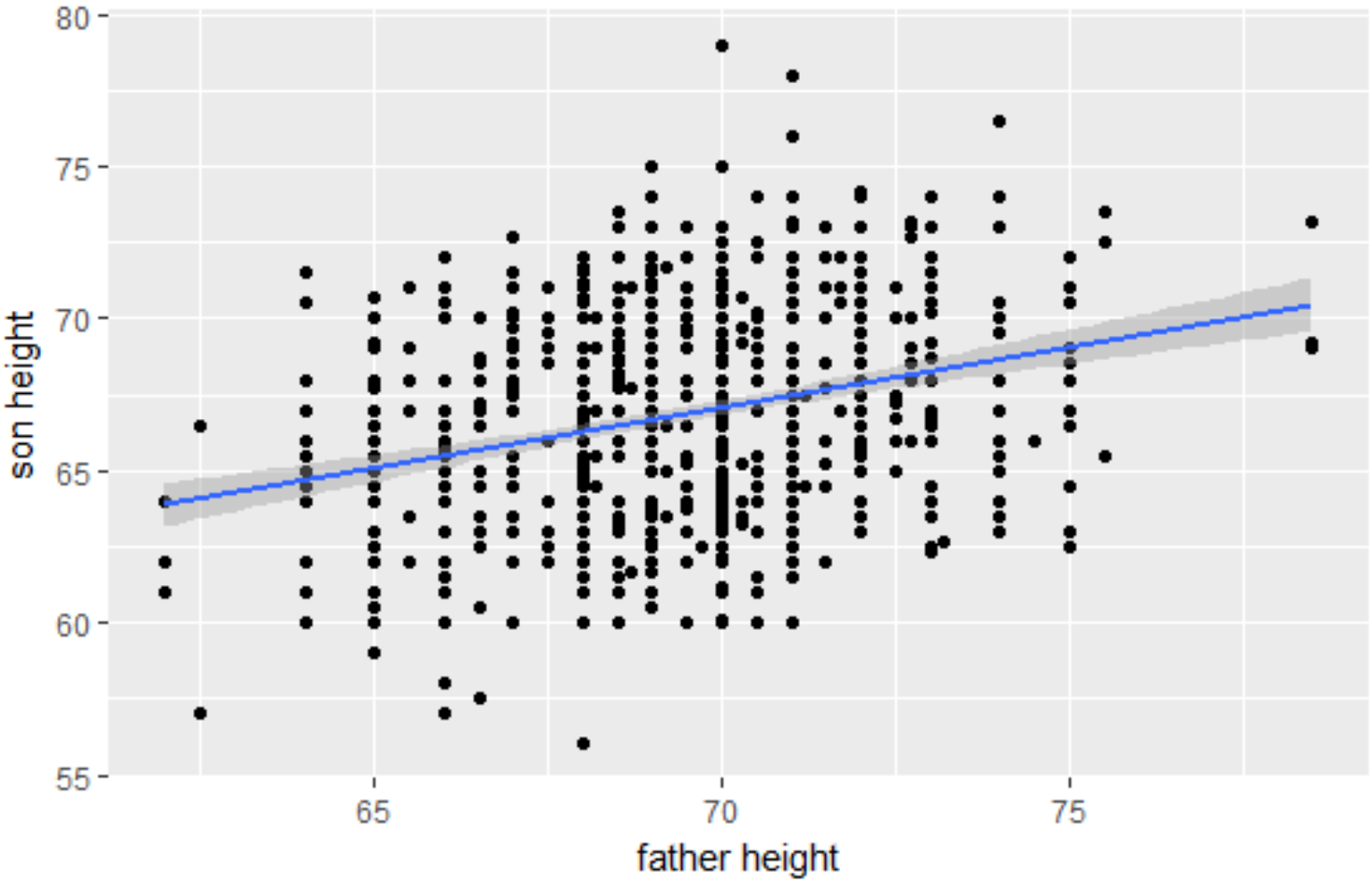
Histogram vs Bar plot

- They look similar. Pay attention to choosing the right one
- For continuous Random variable use: ??
- Histogram
- For categorical Random variable use: ??
- Bar plot

True or false?

We observe a positive relationship between son height and father height in this dataset:

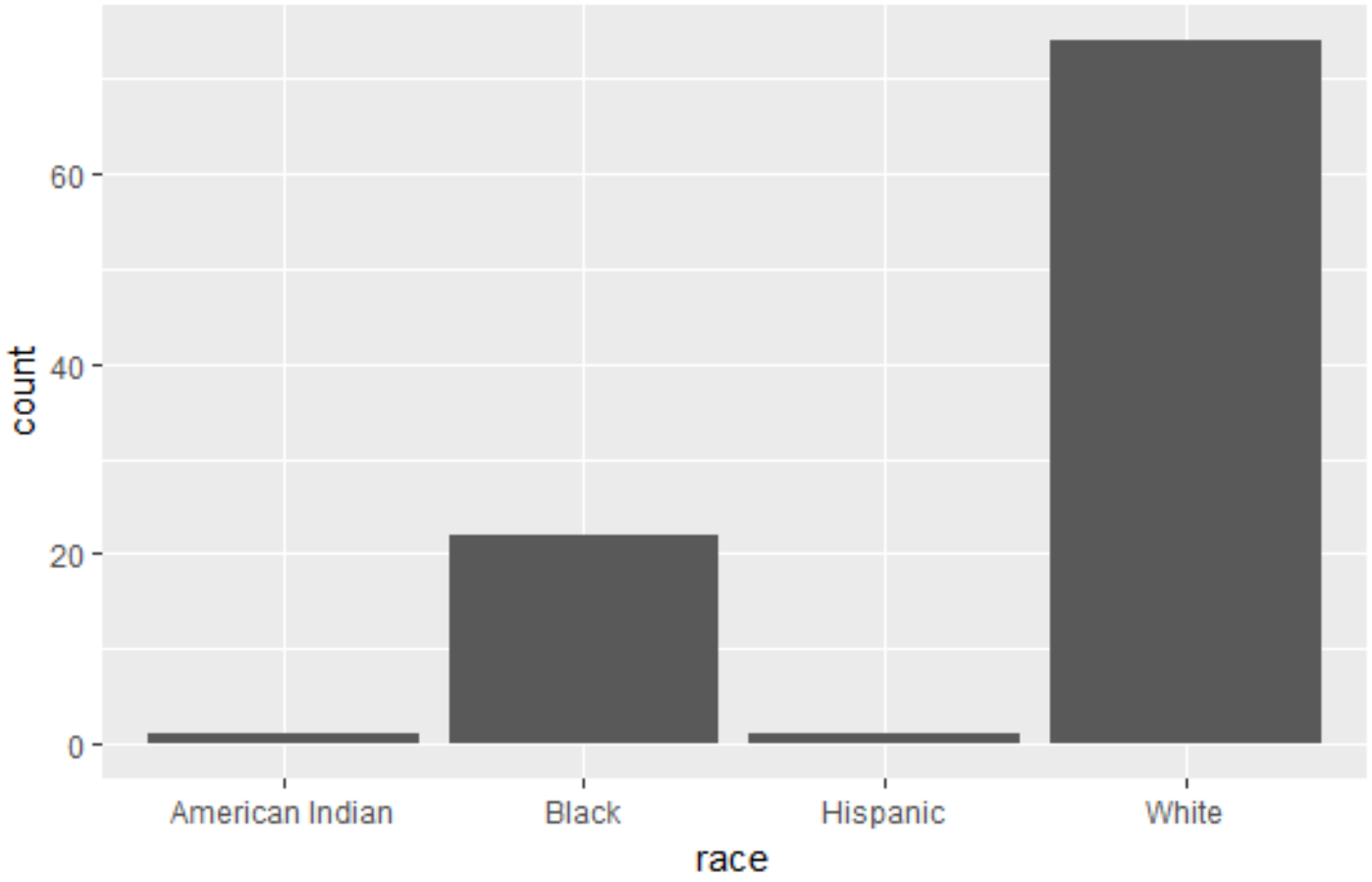
True



marriage dataset

True or false?

- White people tend to marry more than other races:
- False
- White people have more representatives in the marriage dataset compared to other races:
- True

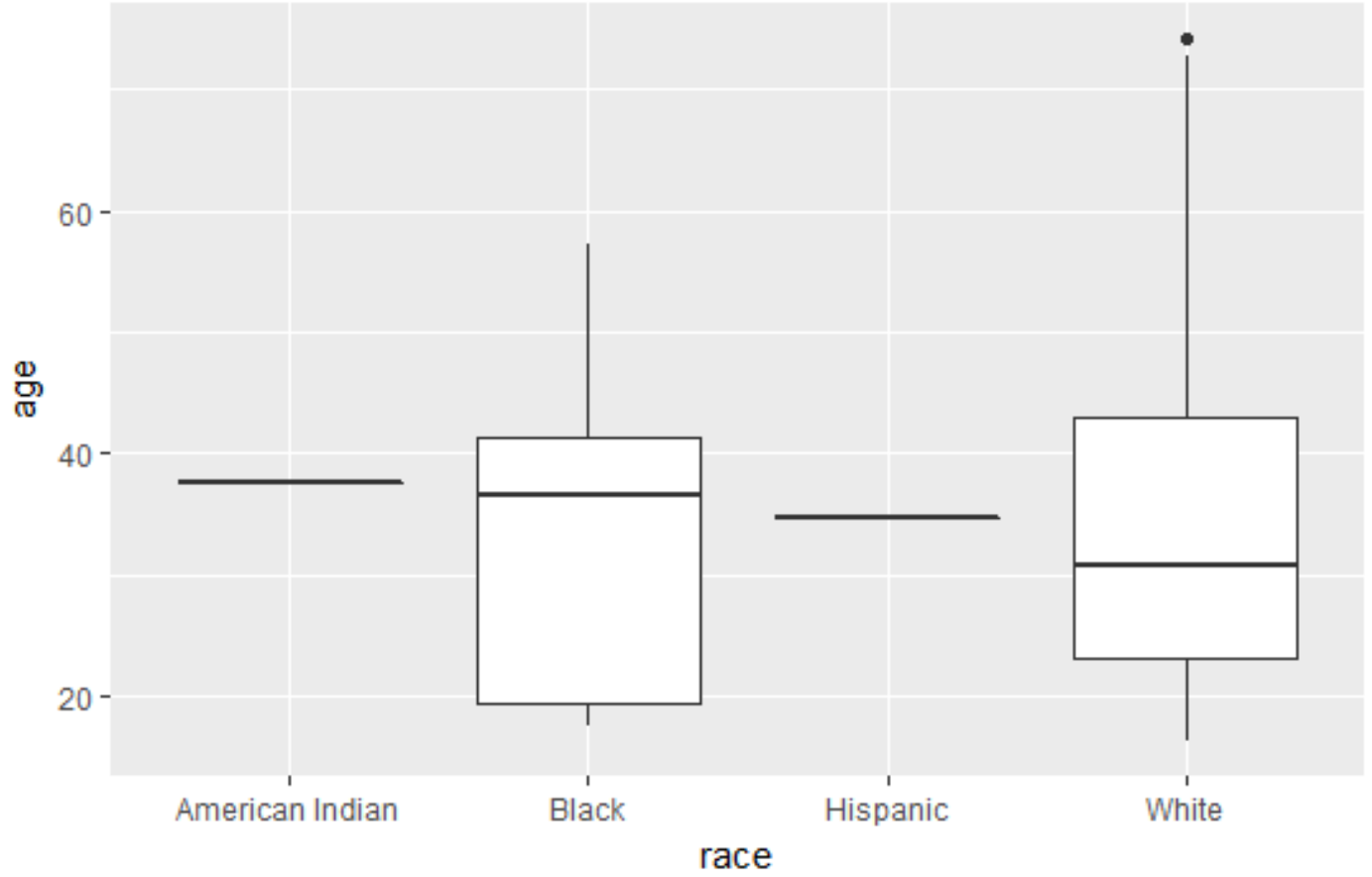


Type of plot? (marriage dataset)

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

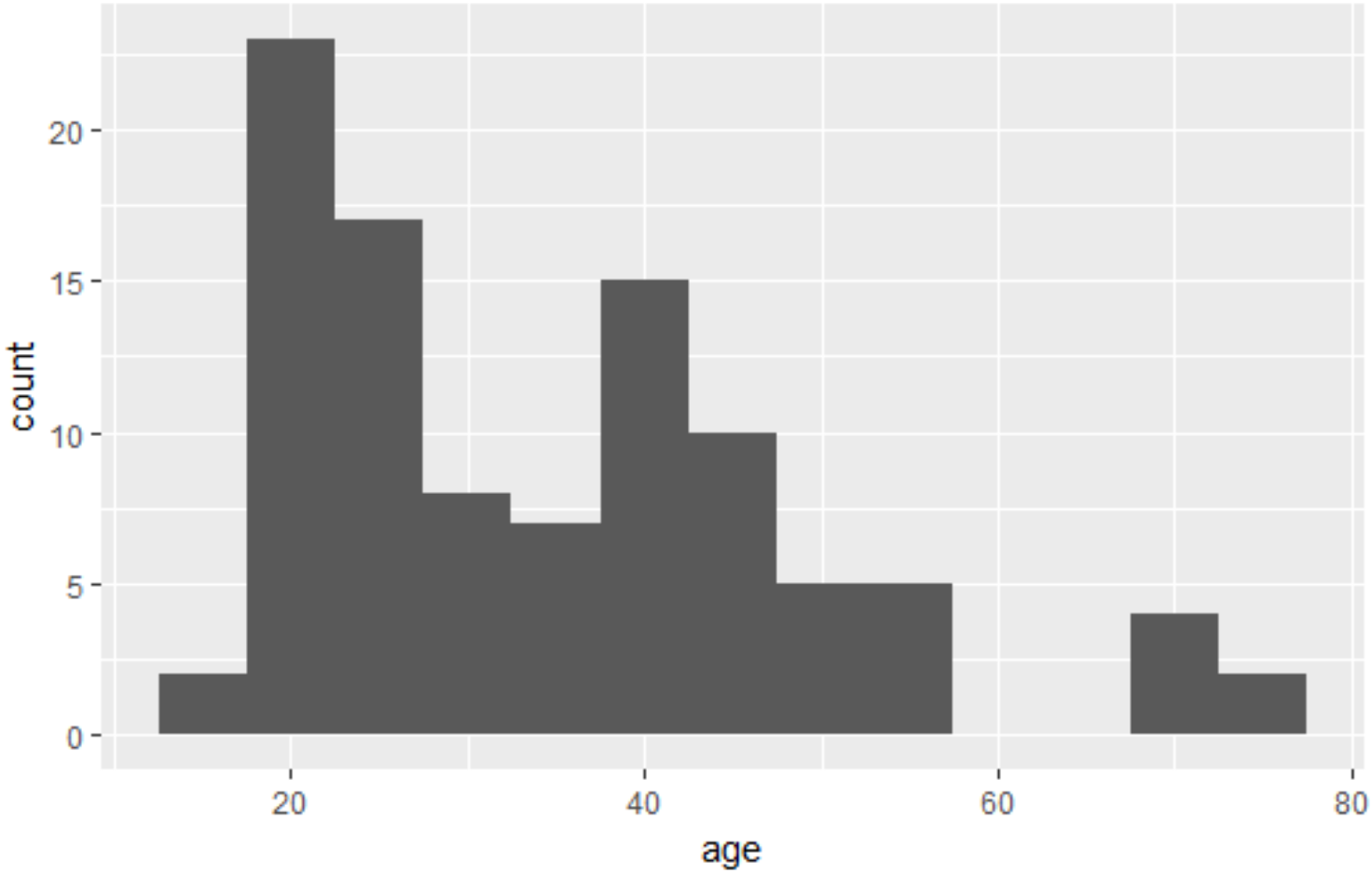


Type of plot? (marriage dataset)

True or false?

The distribution of age in the marriage dataset is bimodal with one peak ~20 and one peak ~40 years:

True



- Look at this plot from the happiness dataset and write what does it show (one or two sentences)?

Only tells you info about how many samples you have for each continent.

This shows there are more representatives in Europe, Africa and Asia compared to others.
Does not give you any information about happiness.

```
ggplot(data=happinessdata2016, aes(x=continent)) +
  geom_bar() +
  coord_flip()
```