STA130 Week 1 Tutorial

Kaixuan (Bryan) Hu

September 14, 2018

Outline for today

- Introduction
- Topic for today: **Data Visualization**.

We will go over what is Data visualization and providing some example.

- Go over question 1 and 3 with discussion.
- Attendance and homework checking & Group assignment for discussion of question 2 (approx. 30 mins).
- Go through the process of synthesizing information from visualization.
- Writing exercise: last 30 minutes; HAND IN for evaluation.

Write a short paragraph to describe the graphs you produced from question 2 and tell a story based on these graphs.

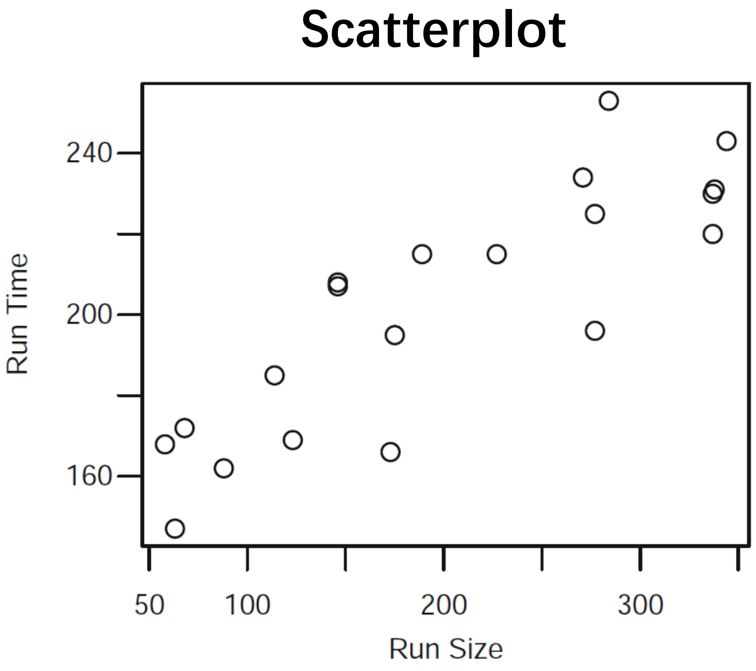
Data Visualization

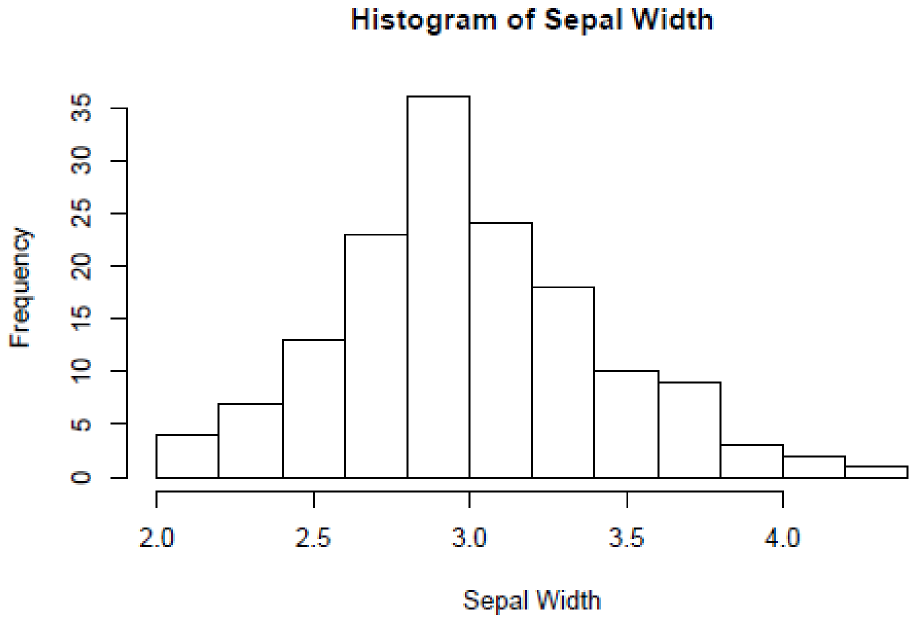
- What a visualization does and how to describe a visualization?

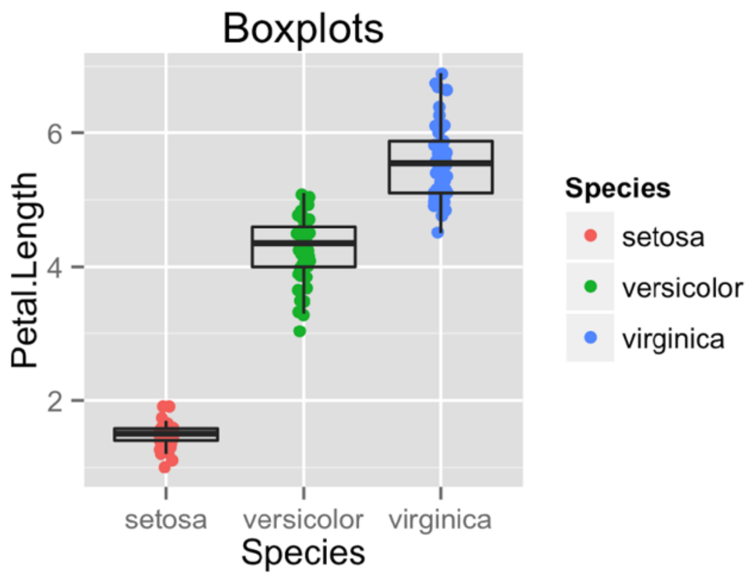
Basically, Exploring a dataset using graphs and come up with a story to describe the dataset.

- What are the most effective types of graphs to summarize information in categorical or quantitative variables?

Typically, **histogram**, **scatterplot, Boxplots** are the most commonly used graphs.







How do you describe them?

Key things to look at:

- Look for **distribution**

What does the distribution tell you about for each types of data?

e.g. Normally distributed; Symmetric; Skewed right; Skewed left;
Bimodal (i.e. double peak) distributed; Multimodal (i.e. multiple peak) distributed.

Use Statistics terminology! (Vocabularies when describing distributions of variables or relationships between two variables)

- Look for **relationships between two variables (e.g. trend)**

e.g. Linear (positive or negative) / non-linear relationship;

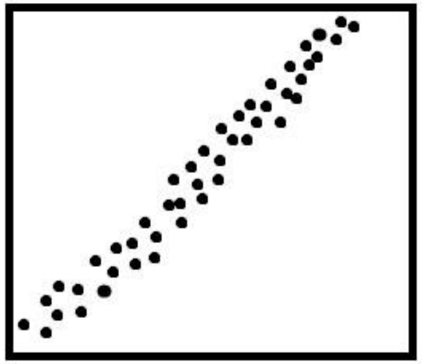
- Boxplots, histograms:

1. Where it is centred (towards the left, right, middle)
2. How much spread?
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed & thin-tailed)
4. Modes (i.e. peak): where, how many? (unimodal, bimodal, multimodal, uniform)
5. Symmetric; Skewness (left-skewed, right-skewed)
6. Outliers; Extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

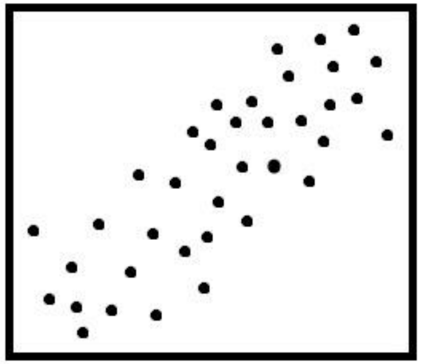
- Scatterplots:

Used to compare the relationship between two variables

1. Strong / weak relationship
2. Linear (positive or negative) / non-linear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming.



strong positive linear association

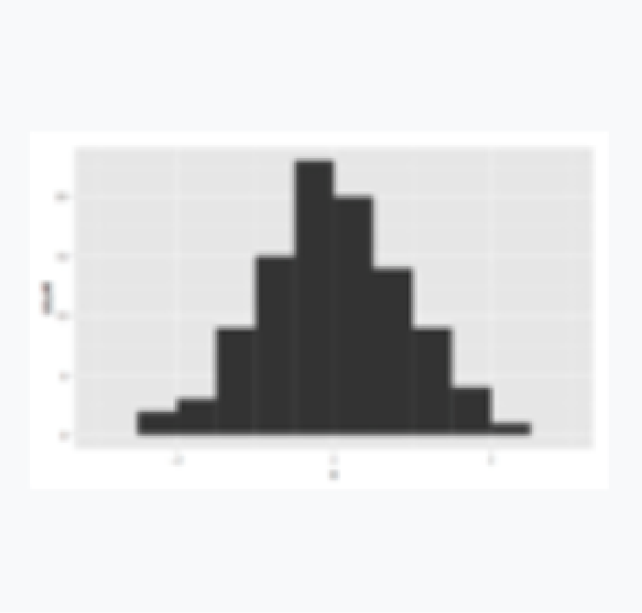


weak positive linear association

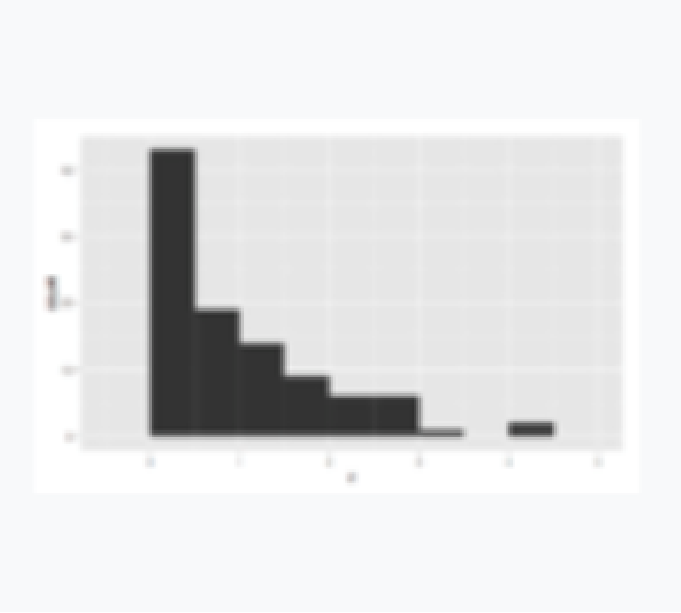
One of the **objective** of this course is to get you familiar with the statistical language, so it is helpful to understand and memorize the word or phrase shown above.

Try it by yourself!

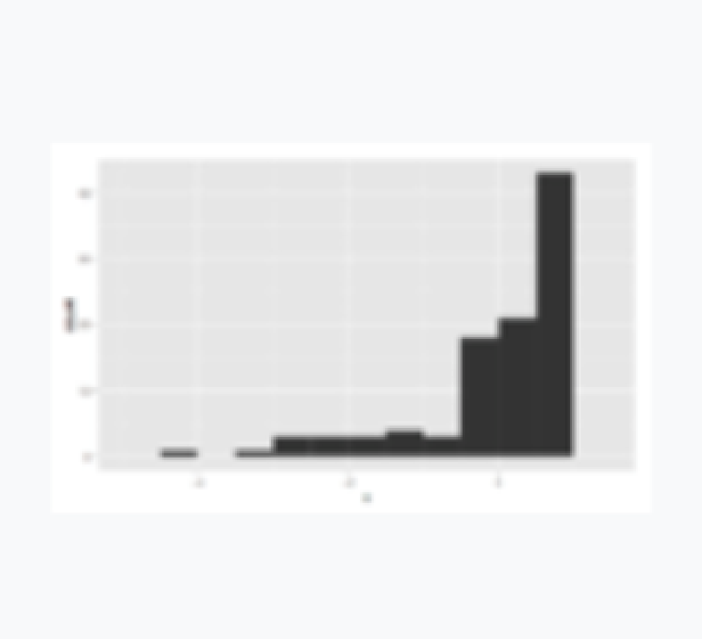
- Describe the following graphs using several correct (and precise) vocabularies.



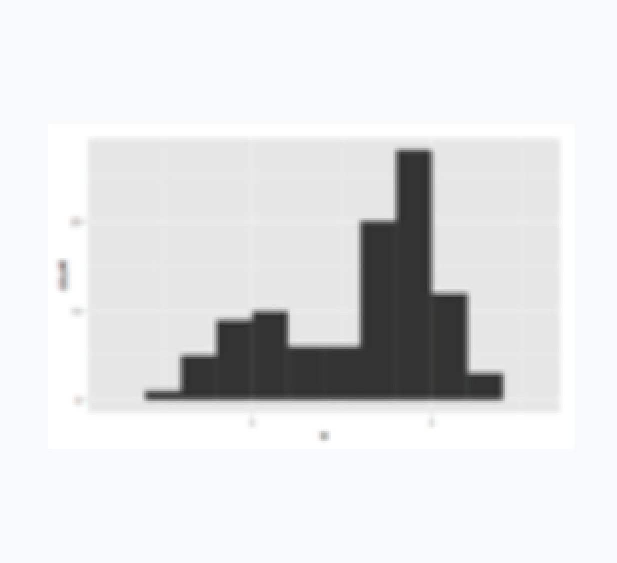
Symmetric or unimodal



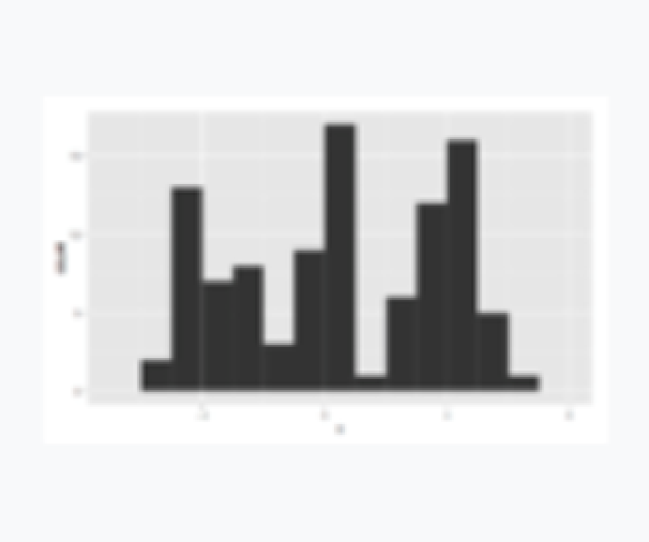
Skewed right



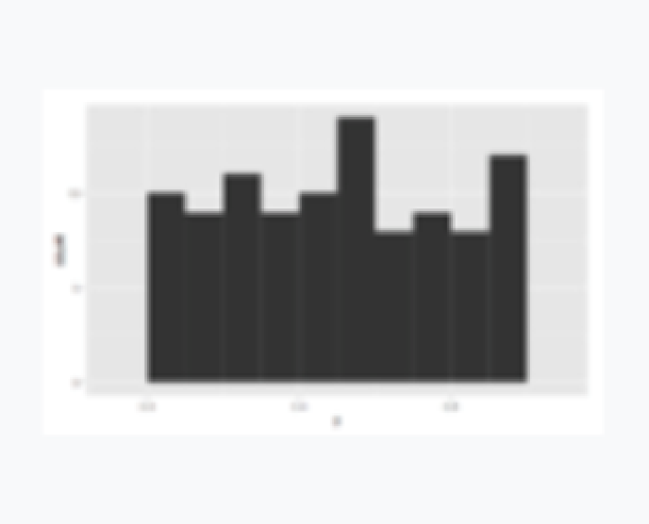
Skewed left



Bimodal



multimodal

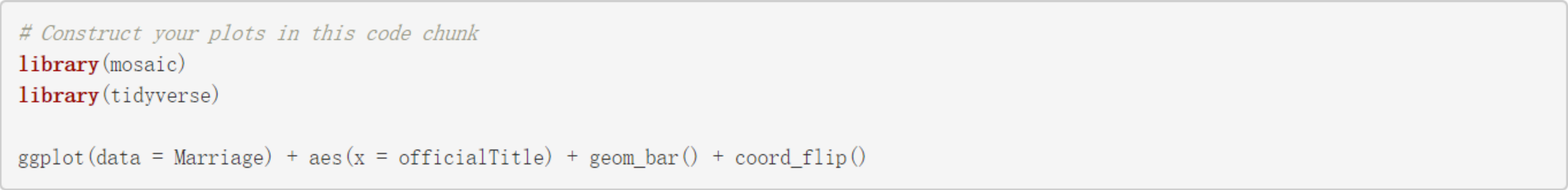


Symmetric

Question 1 from the Practice Problems

- Data Set: Marriage
- There are several variables of the data set.
- For question a: We will look at: **officialTitle** & **sign** (the 12 astrological sign).

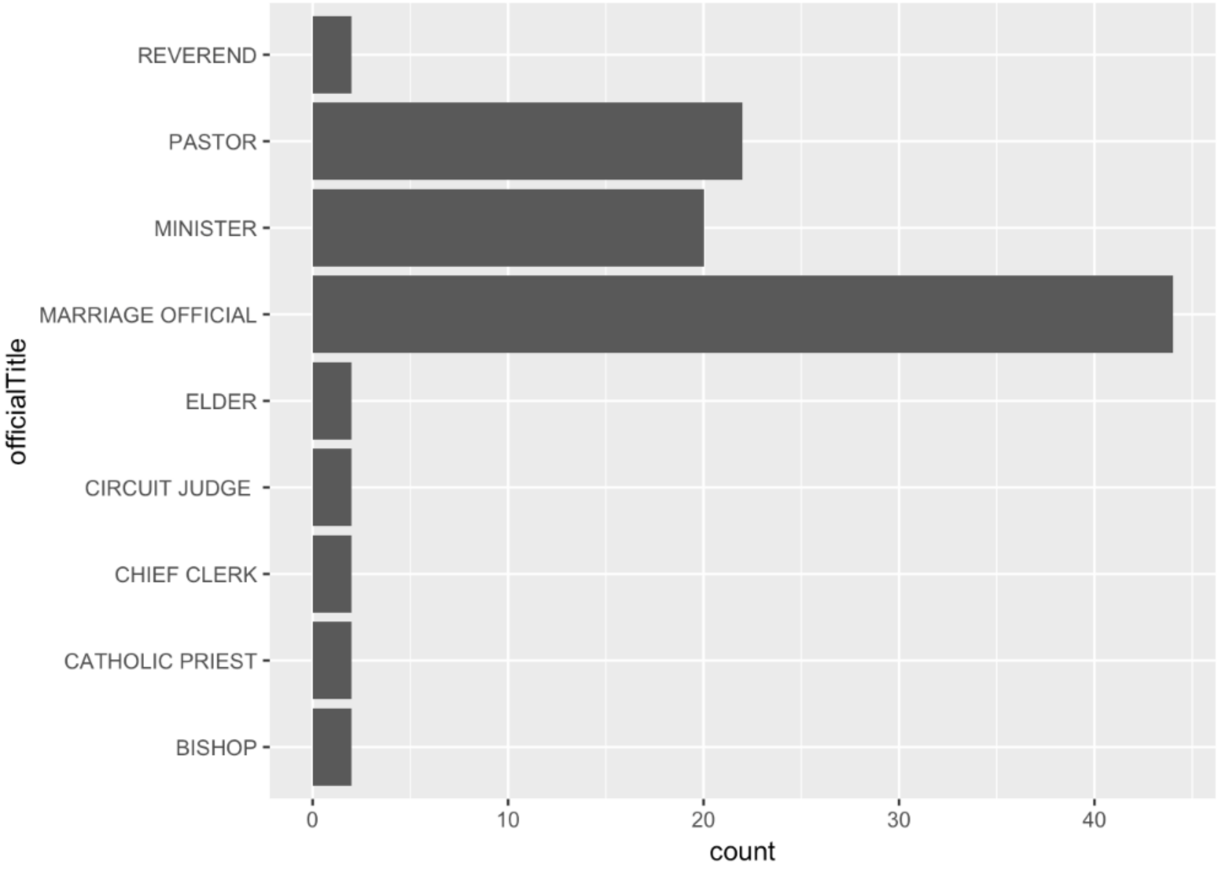
a. Choose two categorical variables and plot their distributions. Interpret the plots.



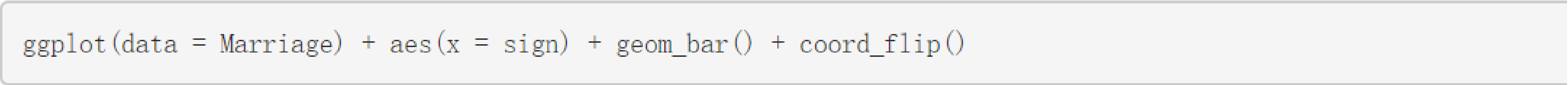
```
# Construct your plots in this code chunk
library(mosaic)
library(tidyverse)

ggplot(data = Marriage) + aes(x = officialTitle) + geom_bar() + coord_flip()
```

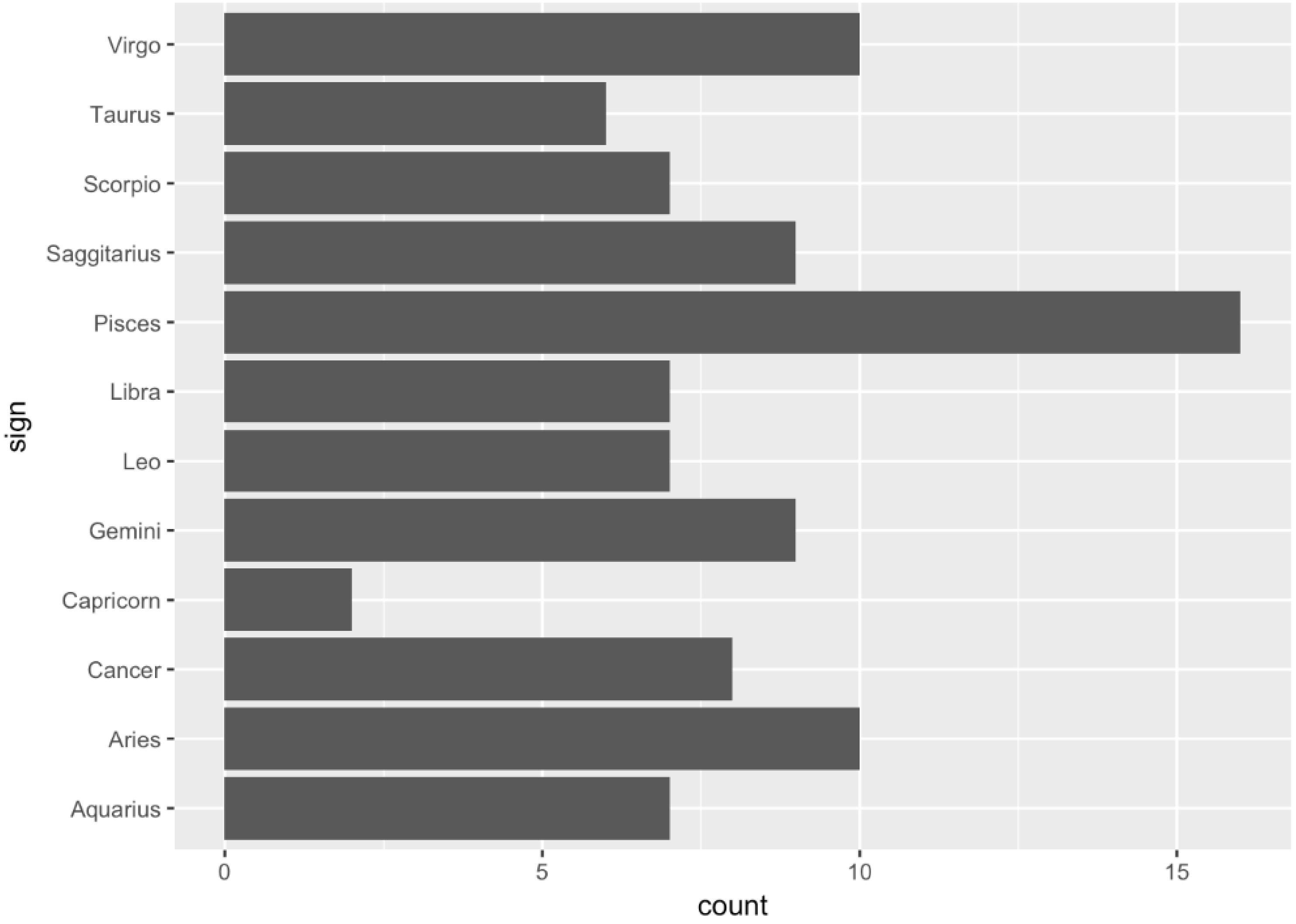
For variable: officialTittle



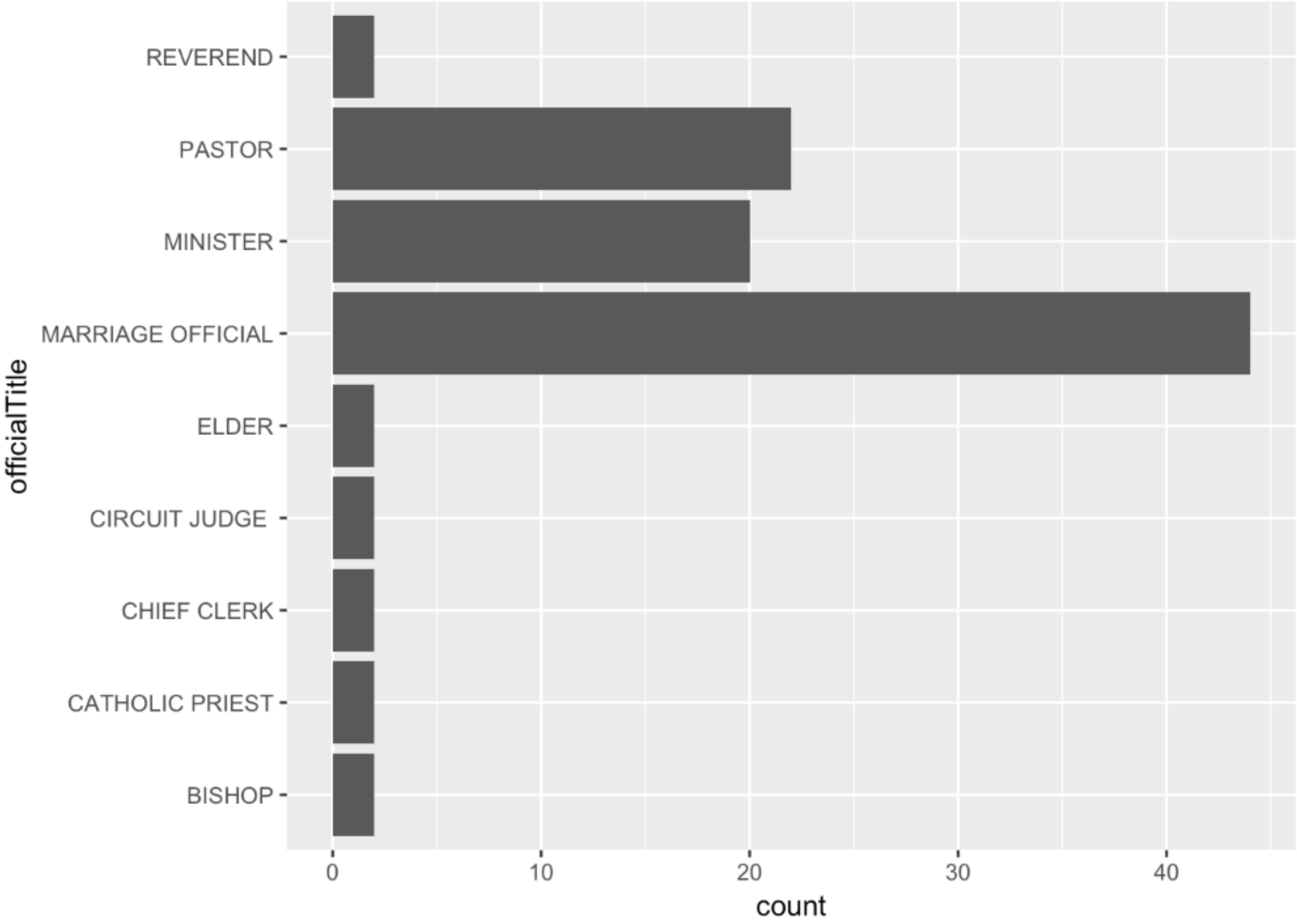
For variable: sign



```
ggplot(data = Marriage) + aes(x = sign) + geom_bar() + coord_flip()
```

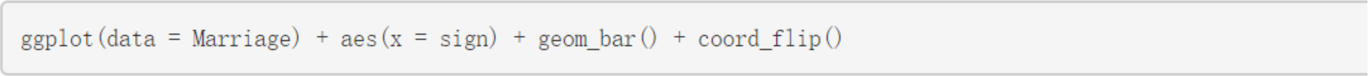


How do you interpret it (using one sentence)? Discuss it with the people seat next to you!

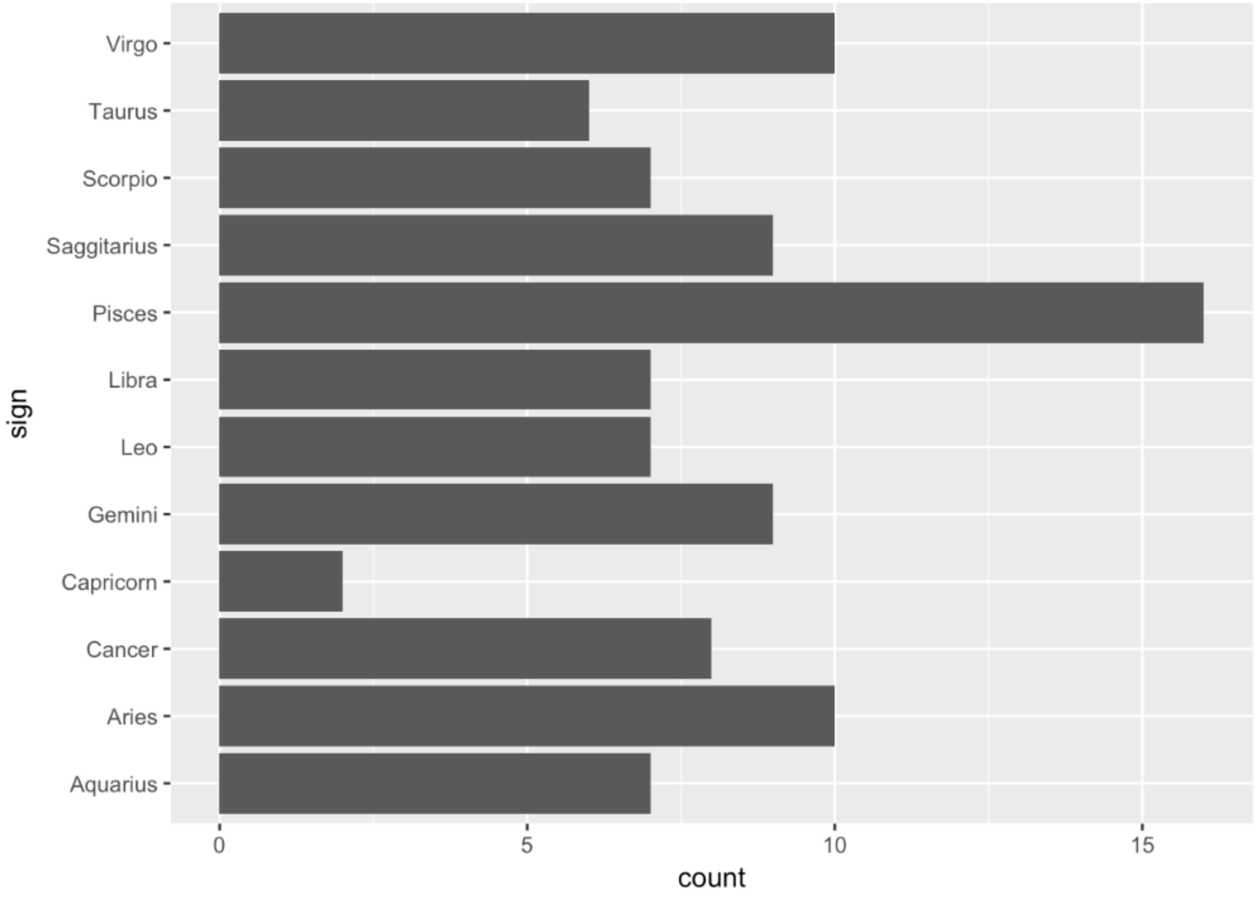


This plot shows that the majority of marriages were performed by a marriage official, pastor, or minister.

How do you interpret them (using one sentence)? Discuss it with the people seat next to you!



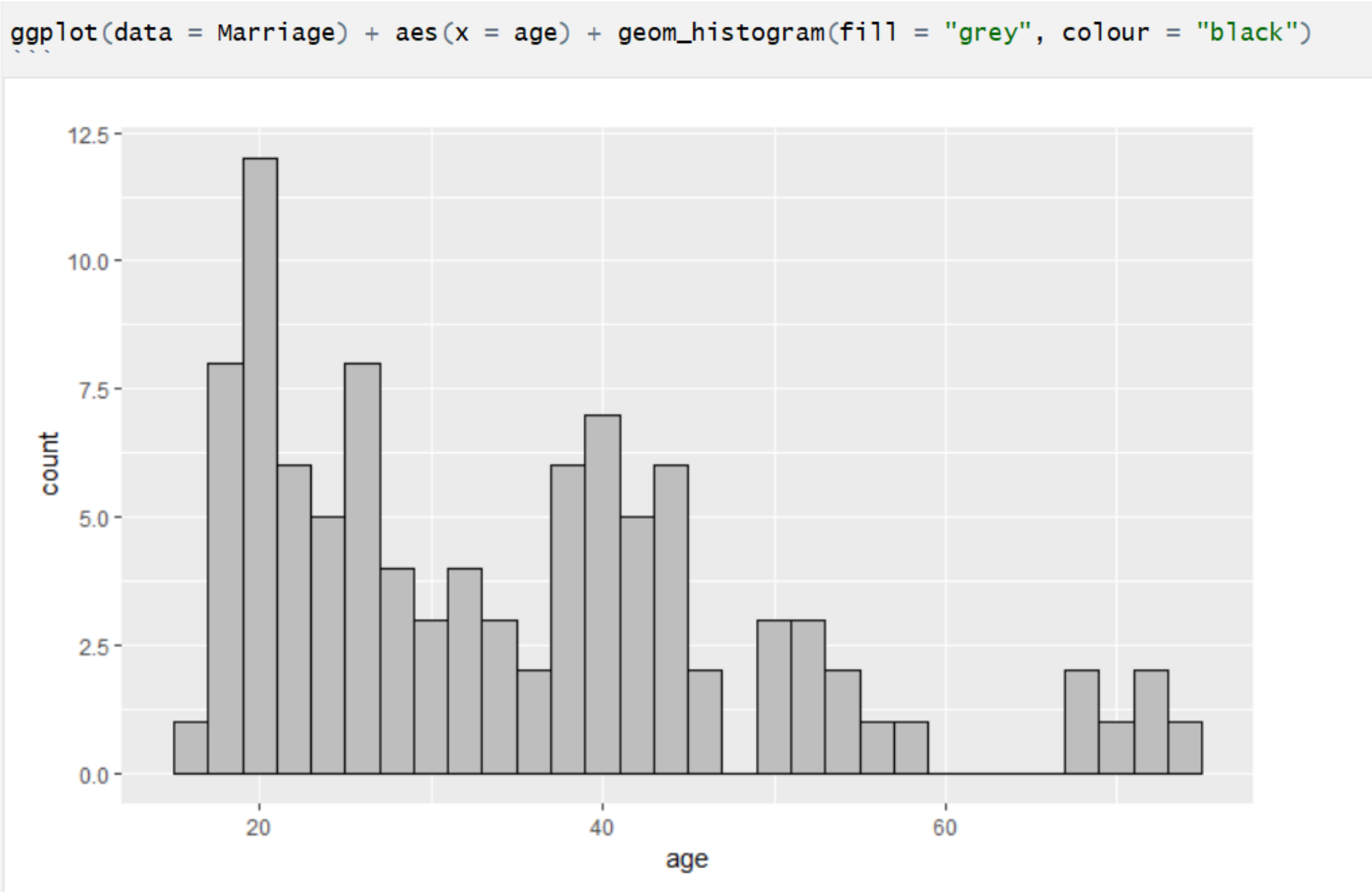
```
ggplot(data = Marriage) + aes(x = sign) + geom_bar() + coord_flip()
```



This plot shows that the majority of people that filled out the application are Pisces, and the next most common sign is Virgo and Aries.

b. Choose a quantitative variables and plot it's distributions. Interpret the plot.

We will look at the variable **age**



```
ggplot(data = Marriage) + aes(x = age) + geom_histogram(fill = "grey", colour = "black")
```

How do you interpret it (using one sentence)? Discuss it with the people seat next to you!

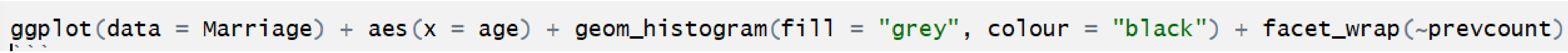
The distribution of applicant age is right skewed since the data trail off to the right. There are two prominent peaks so the distribution is bimodal. The modes appear around 20 and 40. This means that the two largest groups of marriage license applicants are in their twenties and forties.

c. Construct a plot that shows the relationship between two variables. What can you say about the relationship?

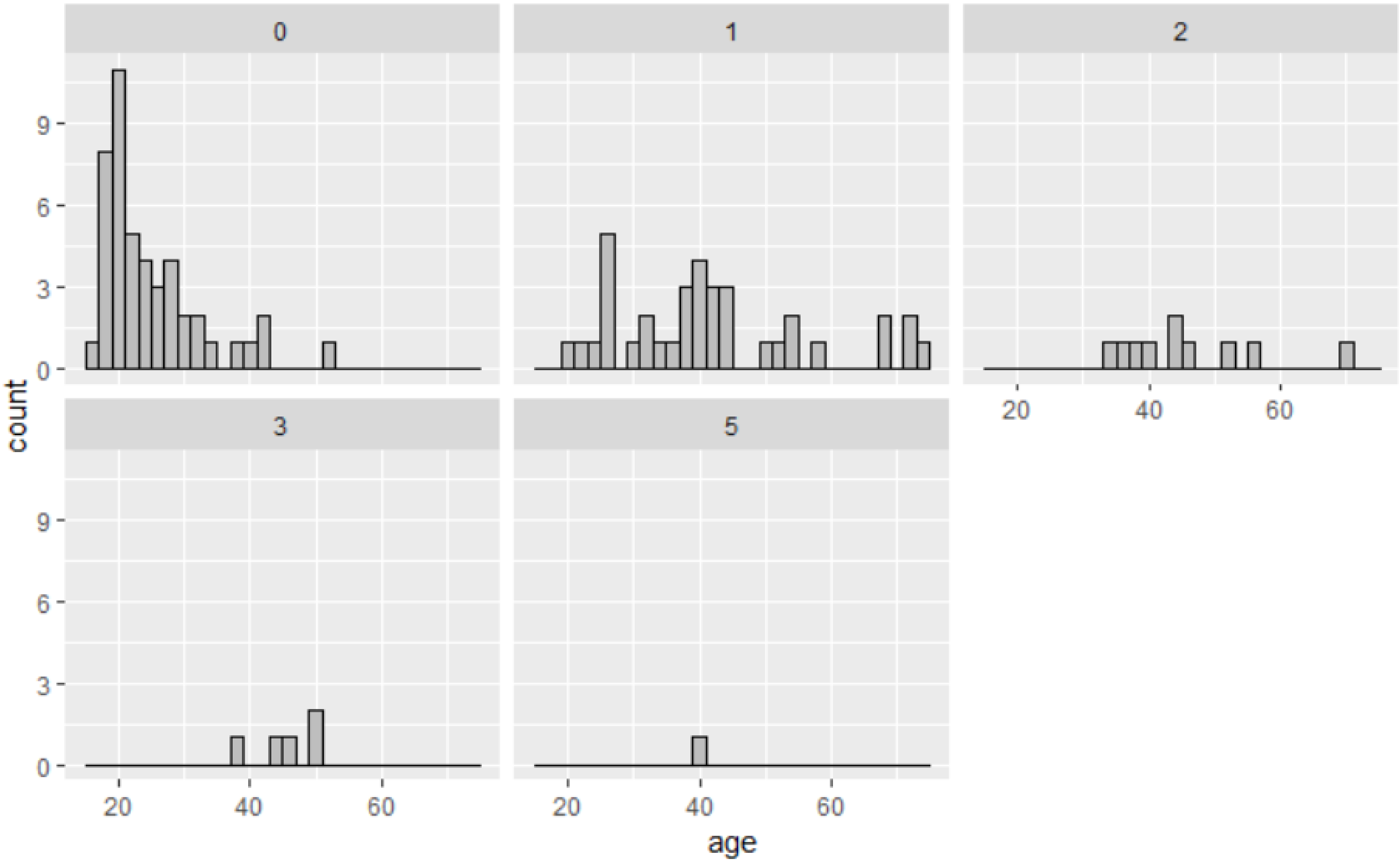
We will look at the variable **age** & **prevcount**: the number of previous marriages of the person.

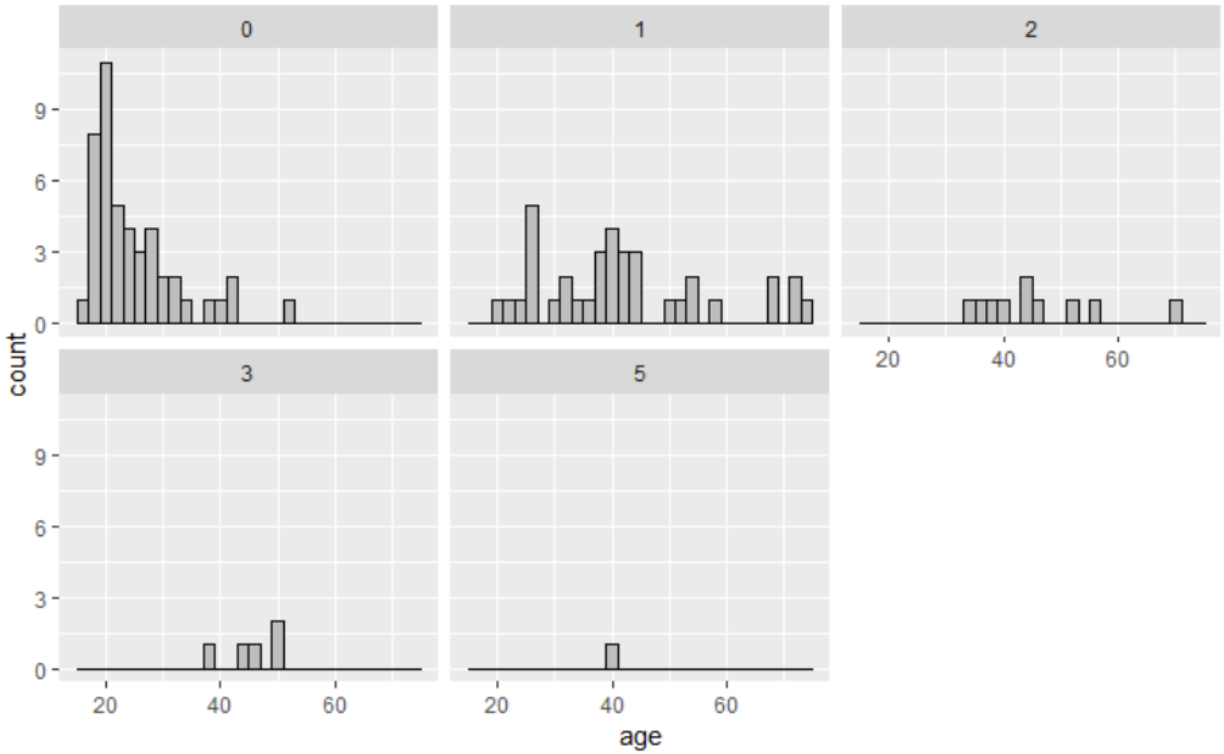
We use facet_wrap() function here!

As you learned in class, facet_wrap() produces a sequence of rectangular plot.



```
ggplot(data = Marriage) + aes(x = age) + geom_histogram(fill = "grey", colour = "black") + facet_wrap(~prevcount)
```



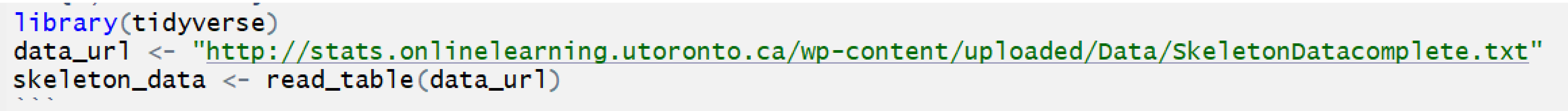


What can you say about the relationship between **age** and **prevcount**? Discuss it with the people seat next to you!

The distribution of age is shown by the number of previous marriages of the applicant. Applicants with at least two marriages tend to be older compared to applicants with fewer previous marriages.

Question 3

a. Loading data in R:

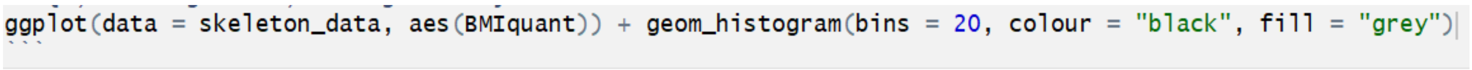


```
library(tidyverse)
data_url <- "http://stats.onlinelearning.utoronto.ca/wp-content/uploaded/Data/SkeletonDatacomplete.txt"
skeleton_data <- read_table(data_url)
```

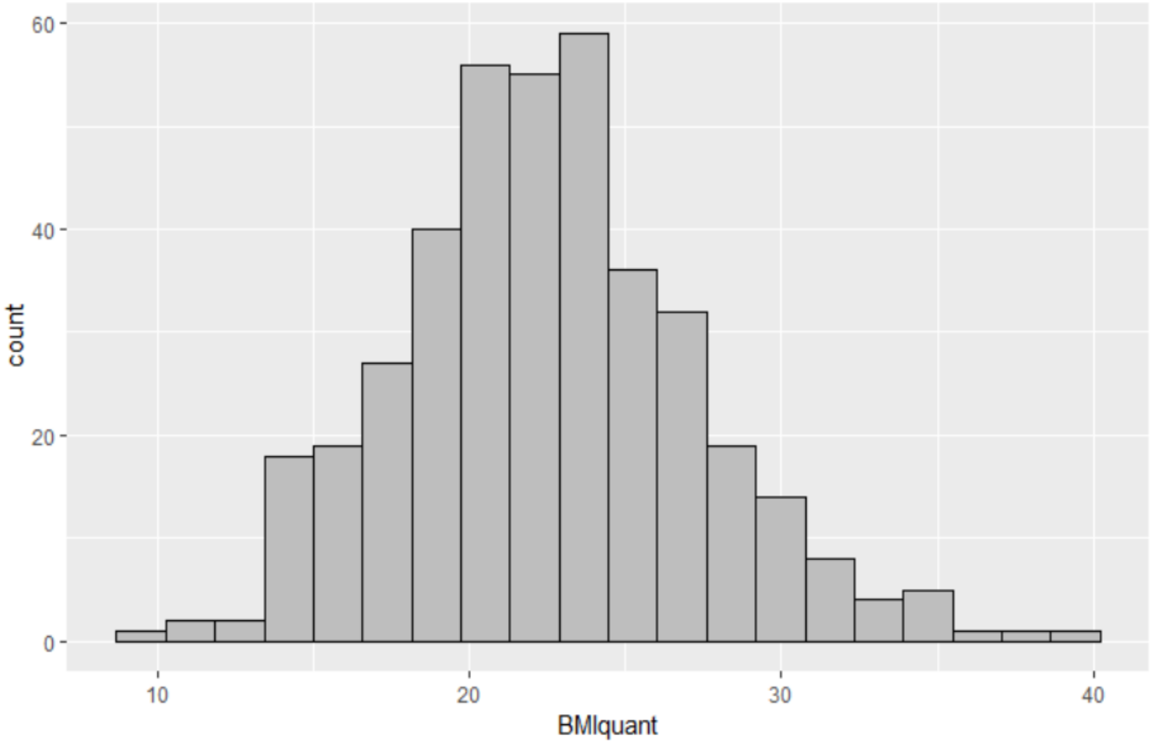
b. Construct at least four interesting graphs with the data, including: a graph of one categorical variable, a graph of one quantitative variable, a graph with at least two variables, a graph with at least three variables.

c. Describe what you learned about the data from your graphs.

Here we plot the histogram for the variable **BMIquant**



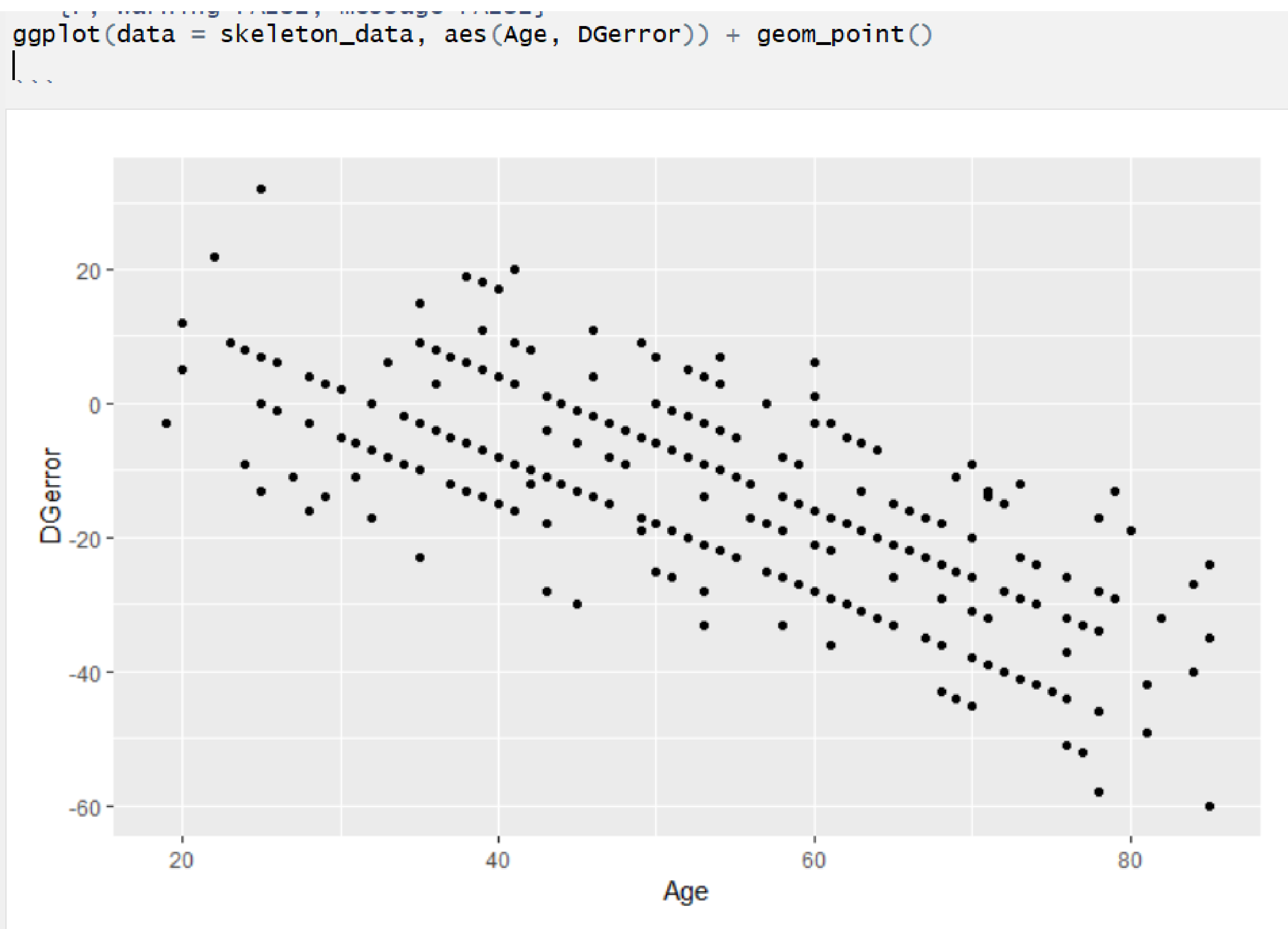
```
ggplot(data = skeleton_data, aes(BMIquant)) + geom_histogram(bins = 20, colour = "black", fill = "grey")
```



What you learned about the data from your graphs? Discuss it with the people near you!

The distribution of the quantitative measurement of BMI is symmetric and unimodal. The mode is between 21 and 25 kg/m^2.

The scatterplot for the variable **Age** and **DGerror**



~~~
{r, warning=FALSE, message=FALSE}
ggplot(data = skeleton_data, aes(Age, DGerror)) + geom_point()
```
~~~

What you learned about the data from your graphs? Discuss it with the people near you!

There is a strong, negative, linear relationship between **Age** and **DGerror**!

Group discussion!

- Show the graphs you produced from question 2 to each group members and describe them.
- Think about the message you are sending.

Are there any noticeable deviation or subgroups? What is the frequency of each variables? Are there correlation between them?

- Come up with a story based on a few graphs.
- Pay attention to the **logical order** to tell the stories.

What is the most effective logical order?

Question 2 Description

- Birth weight, date, and gestational period collected as part of the Child Health and Development Studies in 1961 and 1962. Information about the baby's parents — age, education, height, weight, and whether the mother smoked is also recorded.
- Key variable:

Process of synthesizing (e.g. getting) information from visualizations

- Think about the most logical order in which lead the reader through the visualization.
- Possible writing template:

- Give some context to the variables you are graphing based on what you know about the dataset (units and types of variables involved should be clear)

Either:

- Give the most striking features of the graphs (contrast or similarity).
- Synthesize these features and make a conclusion based on these features.

Or:

- Make a statement or conclusion based on your impression.

- Explain each of the features of the graphs (contrast or similarity) that support your statement or conclusion.

Writing exercise!

- Write a short paragraph to describe the graphs you produced from question 2 and tell a story based on these graphs.
- Use at least 3 graphs.
- You will have 30 minutes to do so.
- HAND IN for evaluation.
- HINT: use the template as well as the parse list, also, make use of the answer from your homework question 2.

Possible writing **template**:

- Give some context to the variables you are graphing based on what you know about the dataset (units and types of variables involved should be clear)

Either:

- Give the most striking features of the graphs (contrast or similarity).
-Synthesize these features and make a conclusion based on these features.

Or:

- Make a statement or conclusion based on your impression.
- Explain each of the features of the graphs (contrast or similarity) that support your statement or conclusion.

Histograms, boxplots, bar graphs:

1. Where it is centred (towards the left, right, middle)
2. How much spread?
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed & thin-tailed)
4. Modes (i.e. peak): where, how many? (unimodal, bimodal, multimodal, uniform)
5. Symmetric; Skewness (left-skewed, right-skewed)
6. Outliers; Extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

Scatterplots:

1. Strong / weak relationship
2. Linear (positive or negative) / non-linear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming.