Assignment 4: Capital Bikeshare

Linh Bui

2022-09-16

Exercise 1

i. Two continuous variables in this dataset: started_at & ended_at.

ii. Two categorical variables in this dataset: rideable_type & member_casual.

iii. Each row in the dataset represents an individual bicycle trip.

Exercise 2

```
bikeshare <- bikeshare %>%
  mutate(
    duration = ended_at - started_at
  )
```

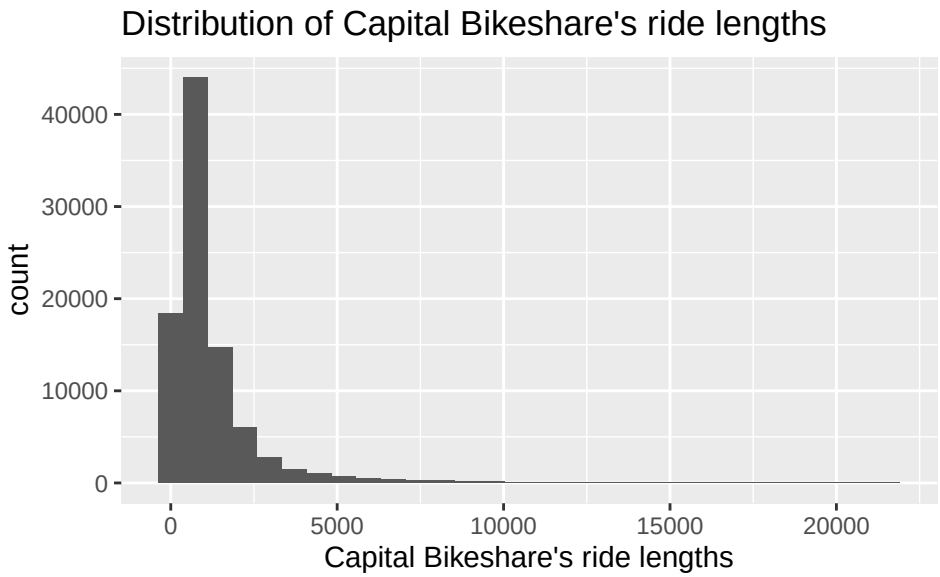
Exercise 3

i.

```
bikeshare %>%
  ggplot() +
  geom_histogram(mapping = aes(x = duration)) +
  labs(title = "Distribution of Capital Bikeshare's ride lengths",
       x = "Capital Bikeshare's ride lengths"
       )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

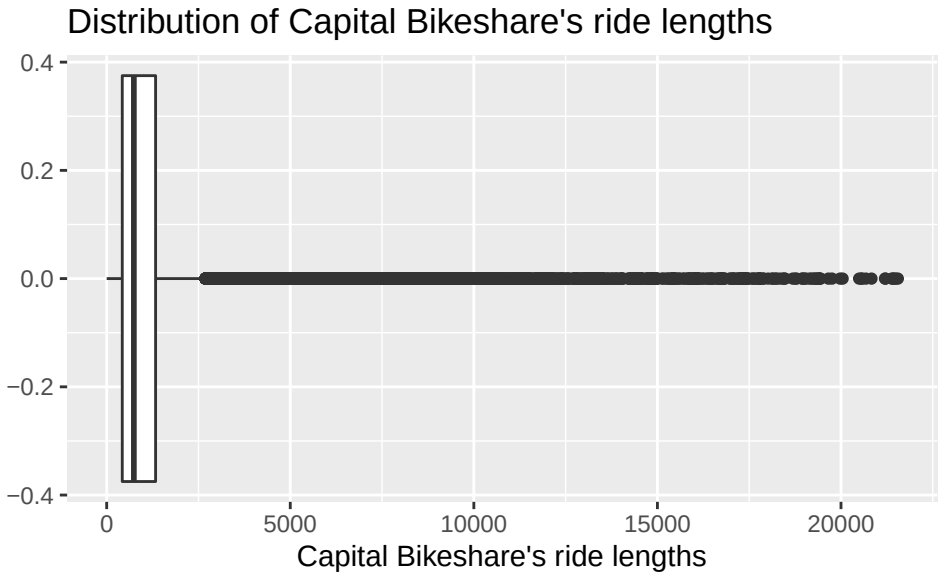


- This histogram is unimodal and right-skewed. It has only 1 mode.

ii.

```
bikeshare %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = duration)) +
  labs(title = "Distribution of Capital Bikeshare's ride lengths",
       x = "Capital Bikeshare's ride lengths"
       )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

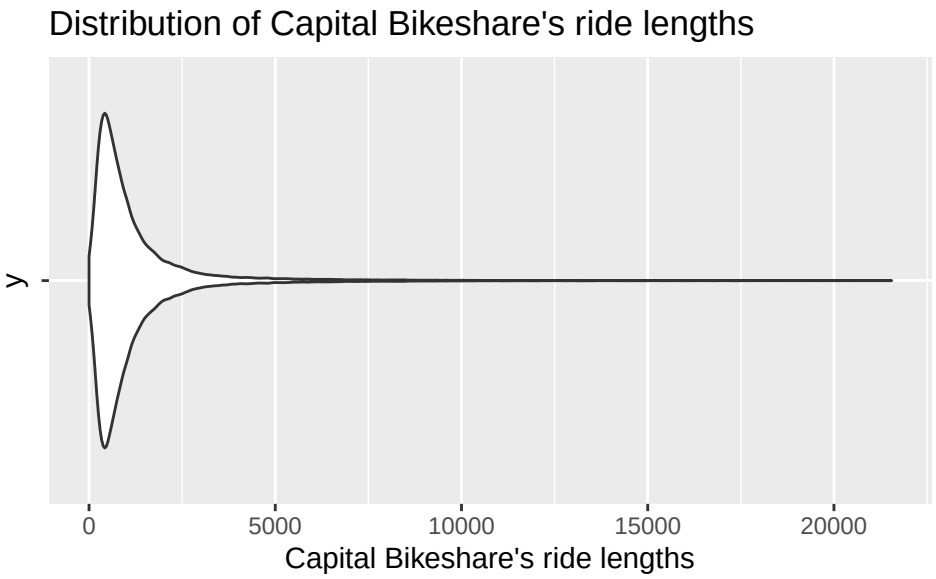


- In this graph, data is centered around 1000 and the distribution is right-skewed. Outliers can be seen here more clearly than in the histogram.

iii.

```
bikeshare %>%
  ggplot() +
  geom_violin(mapping = aes(x = duration, y = '')) +
  labs(title = "Distribution of Capital Bikeshare's ride lengths",
       x = "Capital Bikeshare's ride lengths"
       )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```



- This violin plot has the same shape as the histogram. It is unimodal and right-skewed. It is centered around 1000.

Exercise 4

```
bikeshare <- bikeshare %>%
  mutate(
    min_distance = distHaversine(
      cbind(start_lng, start_lat),
      cbind(end_lng, end_lat)
    )
  )
```

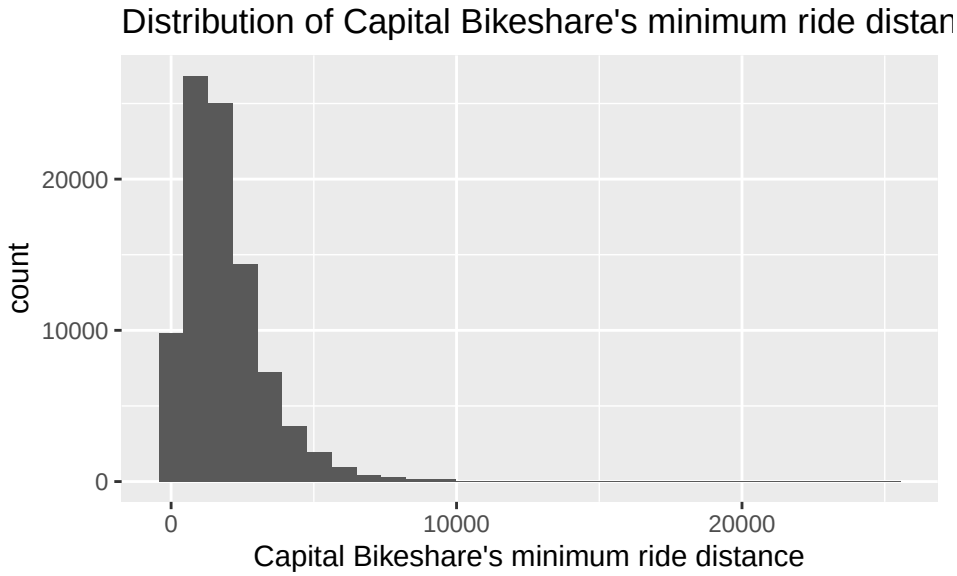
Exercise 5

i. Histogram

```
bikeshare %>%
  ggplot() +
  geom_histogram(mapping = aes(x = min_distance)) +
  labs(title = "Distribution of Capital Bikeshare's minimum ride distance",
       x = "Capital Bikeshare's minimum ride distance"
       )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 50 rows containing non-finite values (stat_bin).
```

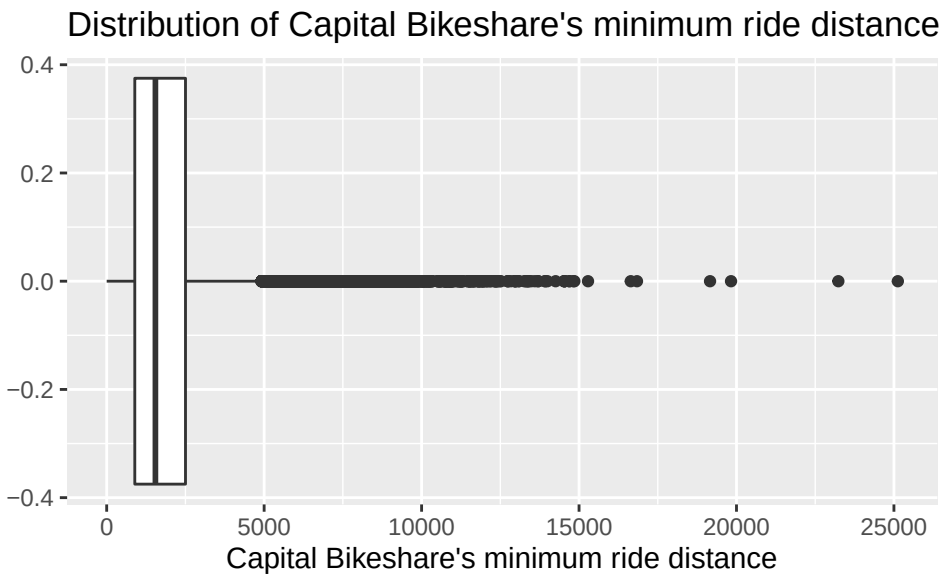


- This histogram is unimodal and right-skewed. It is centered around 2000.

ii. Box plot

```
bikeshare %>%
  ggplot() +
  geom_boxplot(mapping = aes(x = min_distance)) +
  labs(title = "Distribution of Capital Bikeshare's minimum ride distance",
       x = "Capital Bikeshare's minimum ride distance"
       )
```

```
## Warning: Removed 50 rows containing non-finite values (stat_boxplot).
```

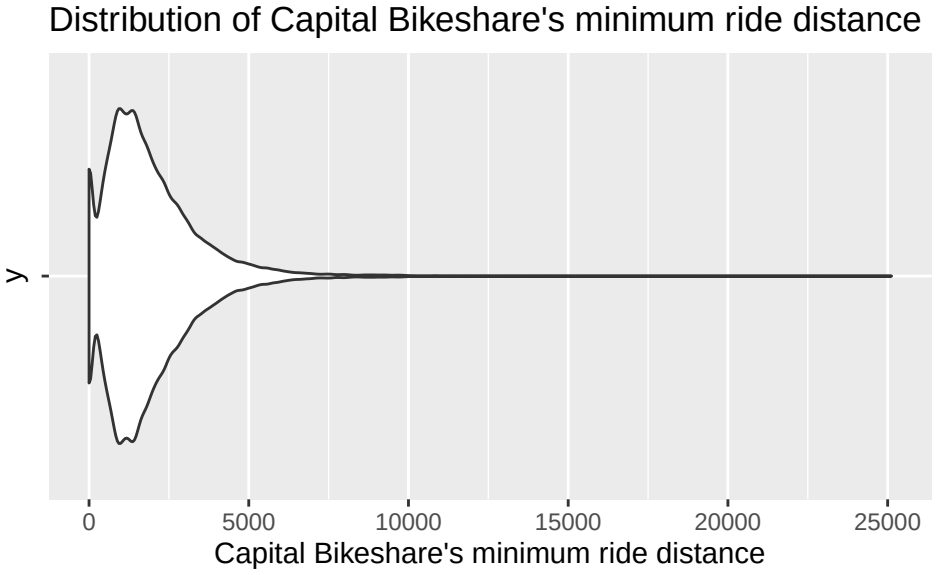


- In this graph, data is also centered around 2000. However, outliers can be seen more clearly here than in the other two plots.

iii. Violin plot

```
bikeshare %>%
  ggplot() +
  geom_violin(mapping = aes(x = min_distance, y = '')) +
  labs(title = "Distribution of Capital Bikeshare's minimum ride distance",
       x = "Capital Bikeshare's minimum ride distance"
       )
```

```
## Warning: Removed 50 rows containing non-finite values (stat_ydensity).
```



- This violin plot has quite the same shape as the histogram. It is also right-skewed. However, it seems like there are 2 modes in this graph. It is centered around 2000.

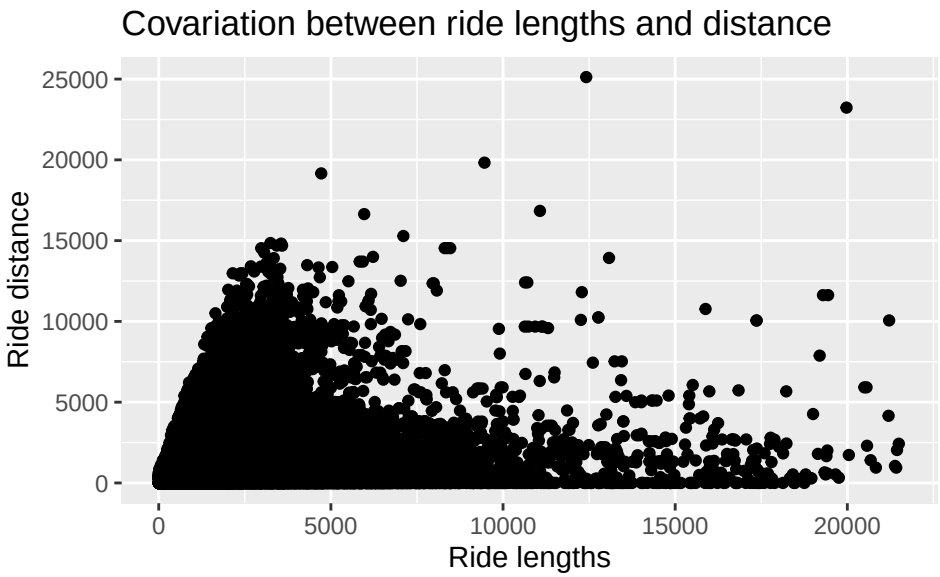
Exercise 6

i.

```
bikeshare %>%
  ggplot() +
  geom_point(mapping = aes(x = duration, y = min_distance)) +
  labs(
    title = "Covariation between ride lengths and distance",
    x = "Ride lengths",
    y = "Ride distance"
  )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```

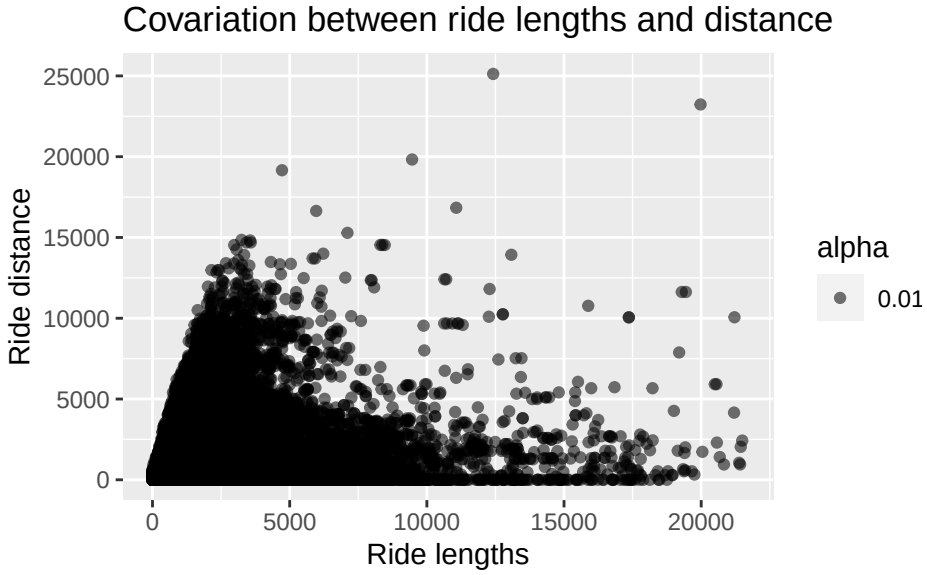


ii.

```
bikeshare %>%
  ggplot() +
  geom_point(mapping = aes(x = duration, y = min_distance, alpha = 0.01)) +
  labs(
    title = "Covariation between ride lengths and distance",
    x = "Ride lengths",
    y = "Ride distance"
  )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

```
## Warning: Removed 50 rows containing missing values (geom_point).
```

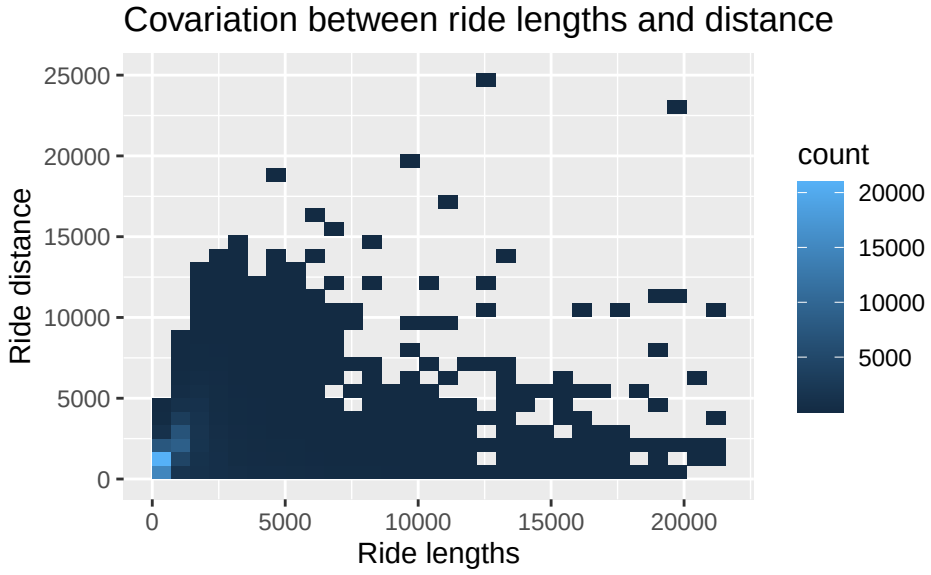


iii.

```
bikeshare %>%
  ggplot() +
  geom_bin2d(mapping = aes(x = duration, y = min_distance)) +
  labs(
    title = "Covariation between ride lengths and distance",
    x = "Ride lengths",
    y = "Ride distance"
  )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

```
## Warning: Removed 50 rows containing non-finite values (stat_bin2d).
```

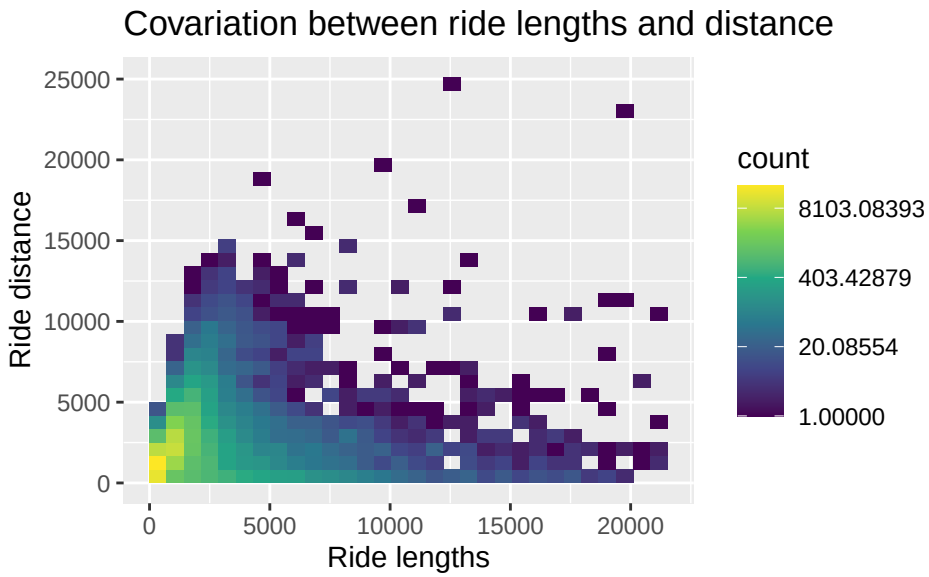


iv.

```
bikeshare %>%
  ggplot() +
  geom_bin2d(mapping = aes(x = duration, y = min_distance)) +
  labs(
    title = "Covariation between ride lengths and distance",
    x = "Ride lengths",
    y = "Ride distance") +
  scale_fill_viridis_c(trans = "log")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

```
## Warning: Removed 50 rows containing non-finite values (stat_bin2d).
```



Exercise 7

i.

```
bikeshare %>%
  summarize(
      mean = mean(duration),
      median = median(duration),
      standard_deviation = sd(duration),
      minimum = min(duration),
      maximum = max(duration)
  )
```

mean	median	standard_deviation	minimum	maximum
1190.265 secs	740 secs	1550.003	0 secs	21548 secs

- The mean bike ride longer than the median because there are a lot more number of people who have a short ride than number of people who ride a long ride. Since the median falls in the middle of the dataset, it falls at the low value (short ride). Meanwhile, mean is the sum of ride lengths divided by ride counts.

ii.

```
bikeshare %>%
  group_by(member_casual) %>%
  summarize(
      mean = mean(duration),
      median = median(duration),
      standard_deviation = sd(duration),
      minimum = min(duration),
      maximum = max(duration)
  )
```

member_casual	mean	median	standard_deviation	minimum	maximum
casual	1675.1181 secs	1014 secs	2023.4229	0 secs	21548 secs
member	772.3979 secs	586 secs	753.2179	0 secs	19339 secs

- Casual riders has the longest average ride length.
- This is not what I expected because I thought that members would be more likely to have long rides since members tend to have discounts and privileges.

iii.

```
bikeshare %>%
  group_by(member_casual) %>%
  summarize(
    n = n()
  )
```

member_casual	n
casual	42213
member	48980

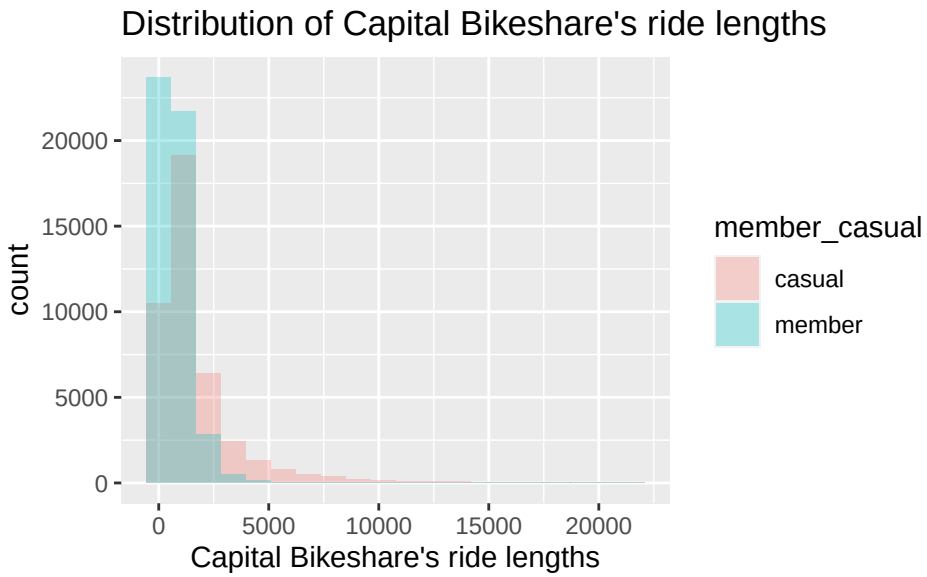
- Member riders have the most rides.

Exercise 8

i.

```
bikeshare %>%
  ggplot() +
  geom_histogram(mapping = aes(x = duration, fill = member_casual),
    bins = 20,
    alpha = 0.3,
    position = "identity") +
  labs(title = "Distribution of Capital Bikeshare's ride lengths",
       x = "Capital Bikeshare's ride lengths"
       )
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

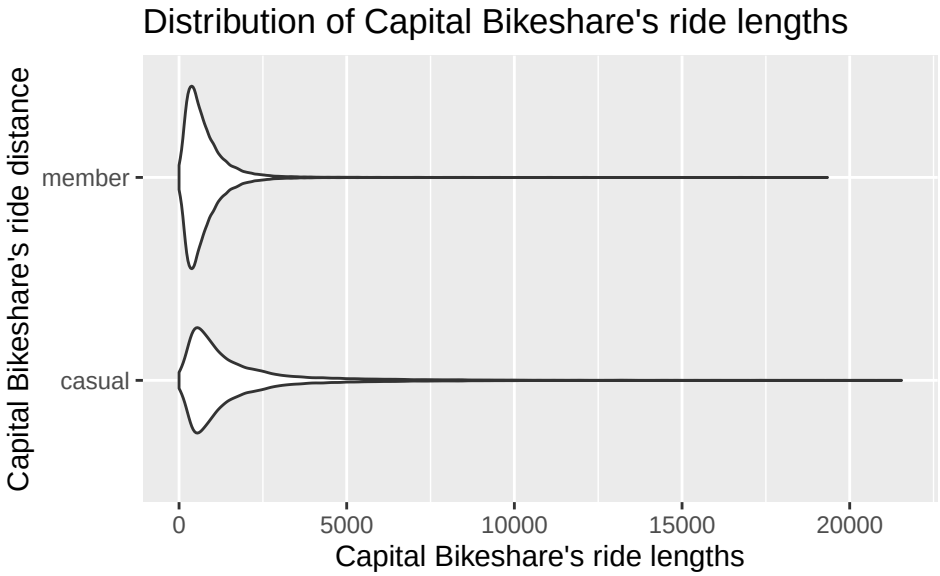


- It can be seen from the graph that both type of riders mostly have short rides. However, while the majority of Capital Bikeshare's members ride in short periods of time, casual riders' ride lengths are more diverse and they have much longer ride lengths than member riders do.

ii.

```
bikeshare %>%
  ggplot() +
  geom_violin(mapping = aes(x = duration, y = member_casual)) +
  labs(title = "Distribution of Capital Bikeshare's ride lengths",
       x = "Capital Bikeshare's ride lengths",
       y = "Capital Bikeshare's ride distance")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

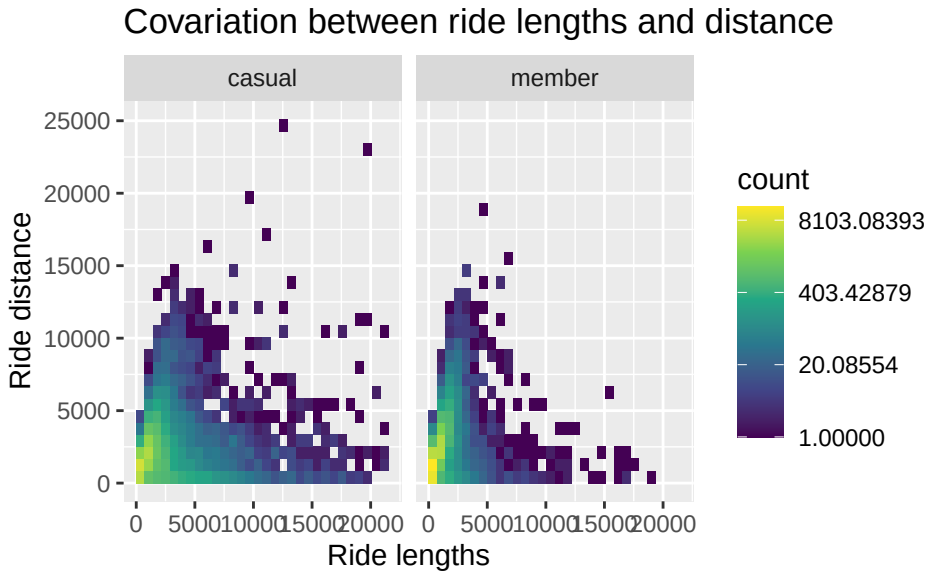


iii.

```r
bikeshare %>%
  ggplot() +
  geom_bin2d(mapping = aes(x = duration, y = min_distance)) +
  labs(
    title = "Covariation between ride lengths and distance",
    x = "Ride lengths",
    y = "Ride distance") +
  scale_fill_viridis_c(trans = "log")+
  facet_wrap( ~ member_casual)
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to contin
```

```
## Warning: Removed 50 rows containing non-finite values (stat_bin2d).
```



- Members tend to have more rides with longer duration but low minimum distance rides. Meanwhile, casual riders tend to ride further for longer periods of time. This is understandable because casual riders have less motivation to pay for short rides.

Exercise 9