Assignment 4: Mind the Gap

JIWON MOK

2022-02-24

Exercise 1

```
#View(gapminder)
```

i. Which variables in the dataset are categorical?

: country,continent, region, year

ii. Which variables in the dataset are continuous (i.e. numerical)?

: infant_mortality, life_expectancy, fertility, population, gdp

iii. What does each row in the dataset represent?

: each country's information

Exercise 2

```
gapminder %>%
  group_by(continent) %>%
  summarize(
    count = n()
  )
```

continent	count
Africa	2907
Americas	2052
Asia	2679
Europe	2223
Oceania	684

```
gapminder %>%
  summarize(
    mean = mean(infant_mortality, na.rm = TRUE),
    median = median(infant_mortality, na.rm = TRUE),
    standard_deviation = sd(infant_mortality, na.rm = TRUE),
    interquartile_range = IQR(infant_mortality, na.rm = TRUE)
  )
```

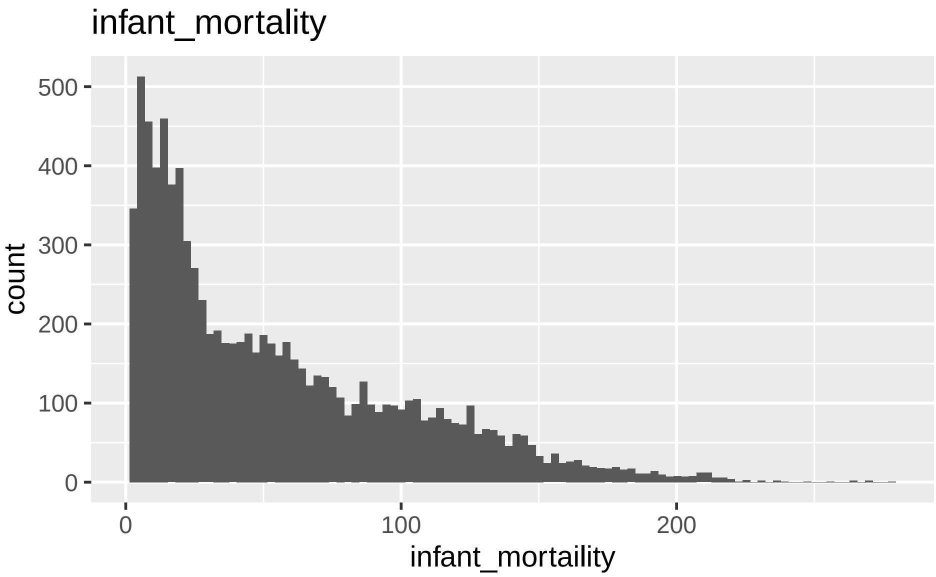
mean	median	standard_deviation	interquartile_range
55.30862	41.5	47.72805	69.1

Exercise 3

i.

```
gapminder %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = infant_mortality),
    bins = 100
  ) +
  labs(
    title = "infant_mortality",
    x = "infant_mortaility"
  )
```

```
## Warning: Removed 1453 rows containing non-finite values (stat_bin).
```

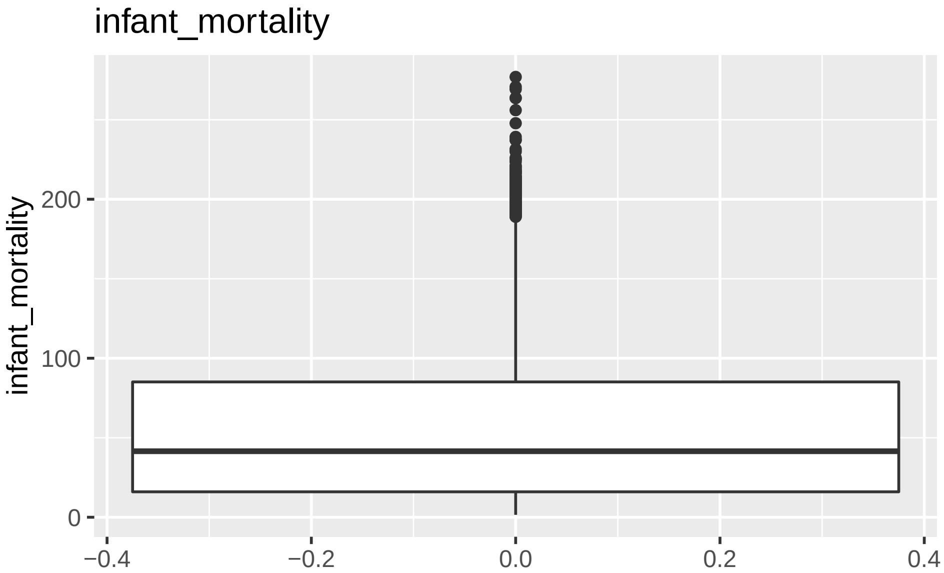


1) What is the shape of the distribution? : left skewed
2) Why do you think that most of the data points occur where they do (i.e. what is the real-world interpretation of this graph)? : infant_mortaility is barely appearing in the real-world

ii. box plot

```
gapminder %>%
  ggplot() +
  geom_boxplot(
    mapping = aes(y = infant_mortality)
  ) +
  labs(
    title = "infant_mortality",
    y = "infant_mortality"
  )
```

```
## Warning: Removed 1453 rows containing non-finite values (stat_boxplot).
```



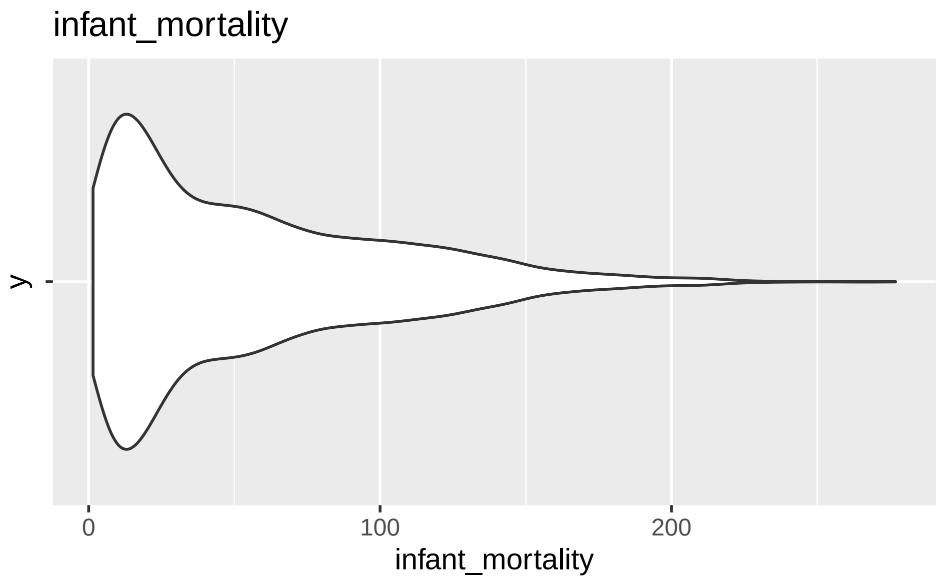
1) What is the shape and where is center of this distribution?

-> symmetric -> 0.0 is the center

iii. violin plot

```
gapminder %>%
  ggplot() +
  geom_violin(
    mapping = aes(x = infant_mortality, y="")
  ) +
  labs(
    title = "infant_mortality",
    x = "infant_mortality"
  )
```

```
## Warning: Removed 1453 rows containing non-finite values (stat_ydensity).
```



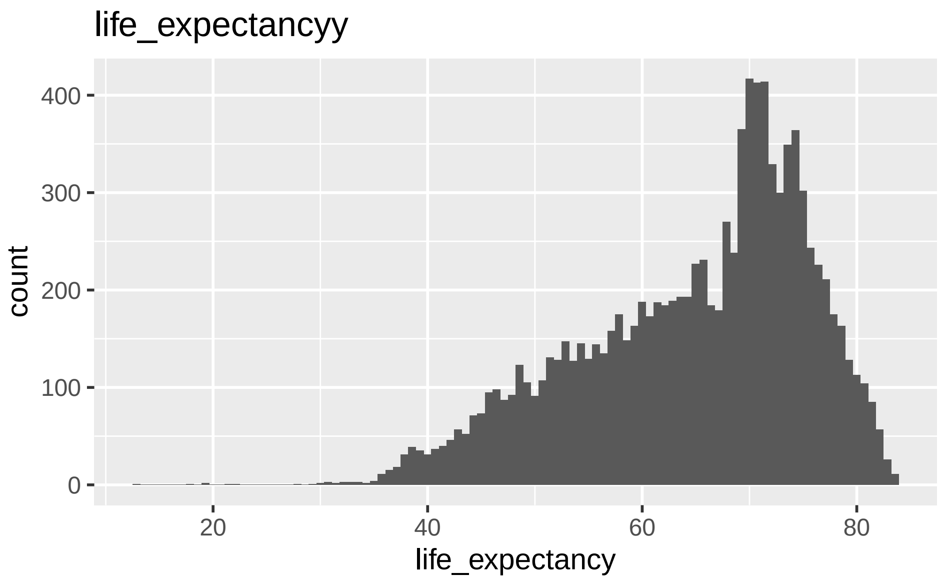
1) What is the shape and where is the center of this distribution? left skewed, center is 150

Exercise 4

i. histogram

1) **Describe each graph separately (shape and center).** Right skewed histogram/ center is at around 60

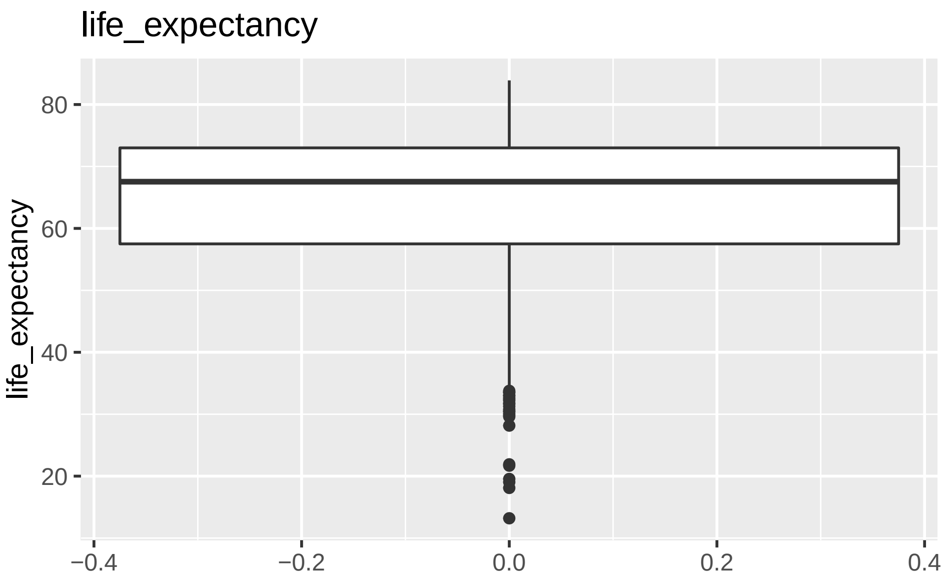
```
gapminder %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = life_expectancy),
    bins = 100
  ) +
  labs(
    title = "life_expectancyy",
    x = "life_expectancy"
  )
```



ii. boxplot

1) **Describe each graph separately (shape and center).** Symmetrical, center = 0.0

```
gapminder %>%
  ggplot() +
  geom_boxplot(
    mapping = aes(y = life_expectancy)
  ) +
  labs(
    title = "life_expectancy",
    y = "life_expectancy"
  )
```



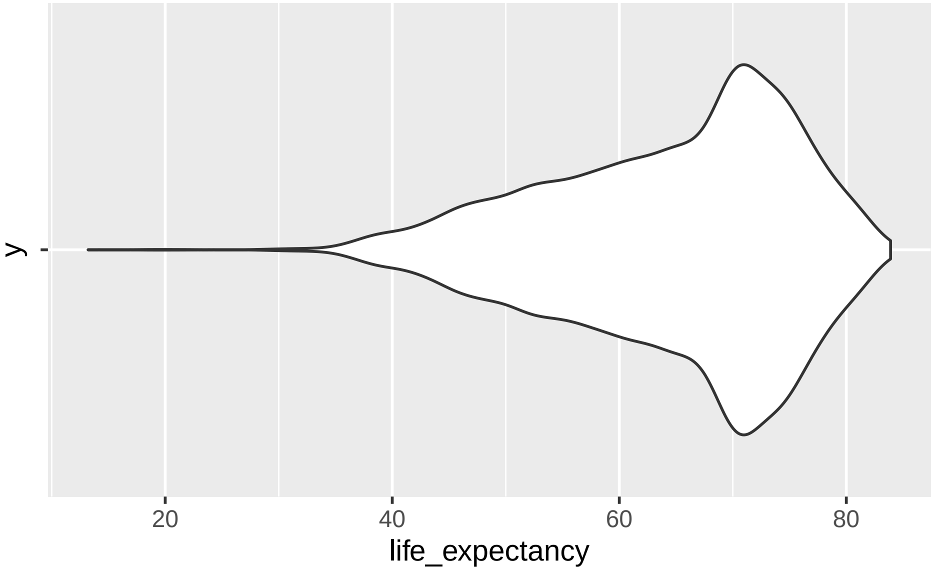
iii. violin plot

1) **Describe each graph separately (shape and center)** right skewed, center is at 60.

2) **Do all three graphs show the same pattern(s), or do any of the graphs display patterns in the**

The histogram and violin plot show the right skewed plot while box plot showed the symmetrial shape. Since the box plot is graphed under then condition that y is the parameter unlike other two plots are drawn upon x = life_expectancy.

```
gapminder %>%
  ggplot() +
  geom_violin(
    mapping = aes(x = life_expectancy, y="")
  )
```



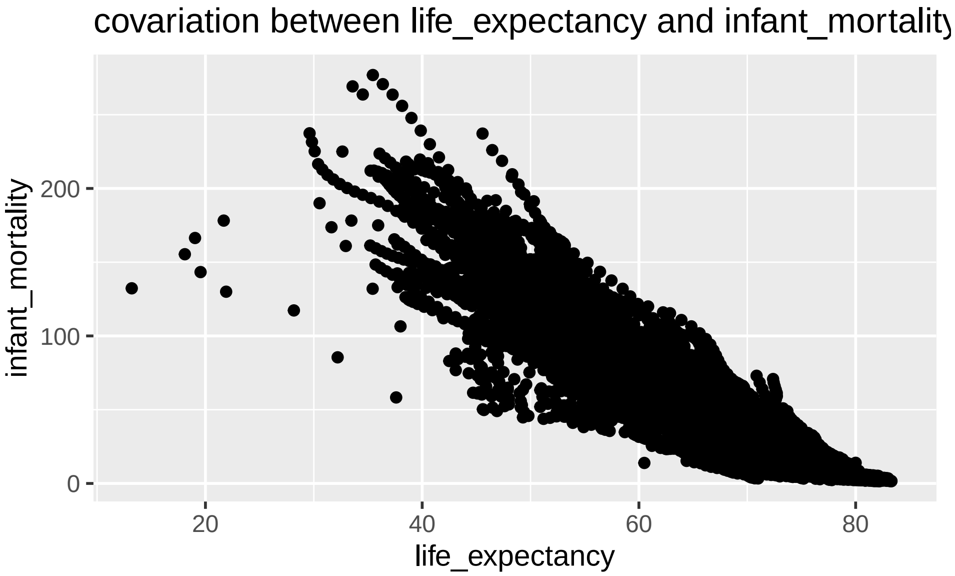
```
  labs(
    title = "life_expectancy",
    x = "life_expectancy"
  )
```

```
## $x
## [1] "life_expectancy"
##
## $title
## [1] "life_expectancy"
##
## attr(,"class")
## [1] "labels"
```

Exercise 5

```
gapminder %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = life_expectancy,
      y = infant_mortality
    )
  ) +
  labs(
    title = "covariation between life_expectancy and infant_mortality",
    x = "life_expectancy",
    y = "infant_mortality"
  )
```

```
## Warning: Removed 1453 rows containing missing values (geom_point).
```

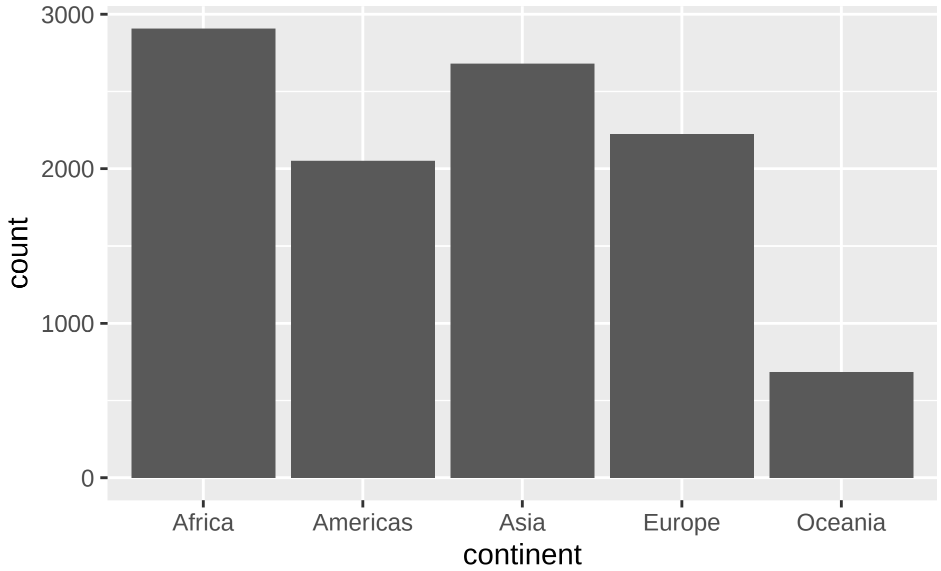


Describe any patterns you see in this graph.

infant_mortality and life_expectancy show the negative covariation.

Exercise 6

```
gapminder %>%
  ggplot() +
  geom_bar(mapping = aes(x = continent))
```



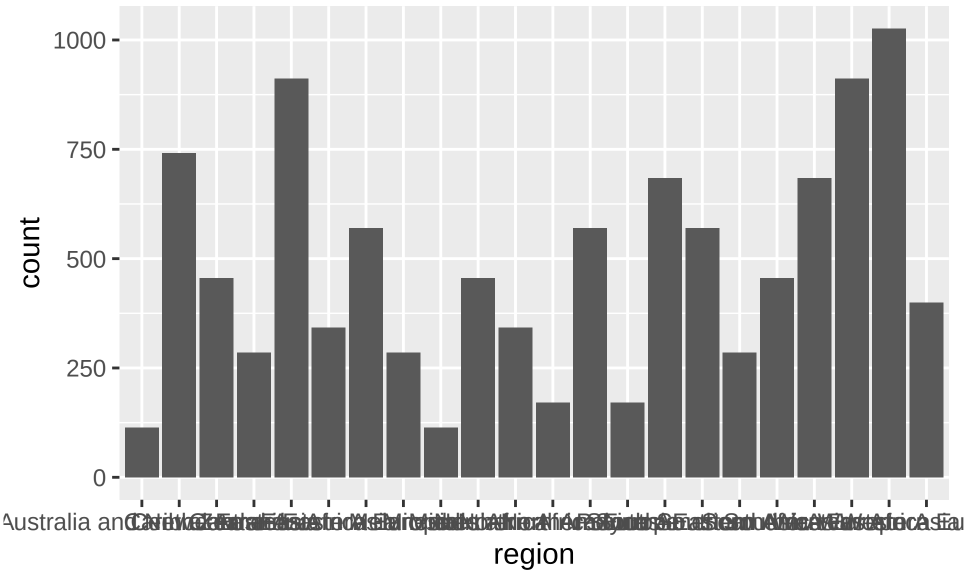
```
  theme(axis.text.x =element_text(angle = 45, hjust = 1)) +

  labs(
    title = "continent",
    x = "continent"
  )
```

```
## List of 3
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : num 45
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ x          : chr "continent"
##  $ title      : chr "continent"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
gapminder %>%
  ggplot() +
  geom_bar(mapping = aes(x = region), bins = 150)
```

```
## Warning: Ignoring unknown parameters: bins
```



```
theme(axis.text.x =element_text(angle = 45, hjust = 1)) +

labs(
  title = "region",
  x = "region"
)
```

```
## List of 3
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : NULL
##   ..$ angle        : num 45
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ x          : chr "region"
##  $ title      : chr "region"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

Exercise 7

- explore the covariation between two or more variables
- compare multiple countries at a single time or multiple times within a single country
- add time into analyses as a second variable (year, categorical variable -> create a new column

for this) as.factor(…)
- use the mutate() to create a new variable called year_cat, which contains the year variable as a categorical var(using as.factor(year))
- store the output dataframe in a new var, gapminder_cat

```
gapminder_cat <- gapminder%>%
  mutate(as.factor(year))
```

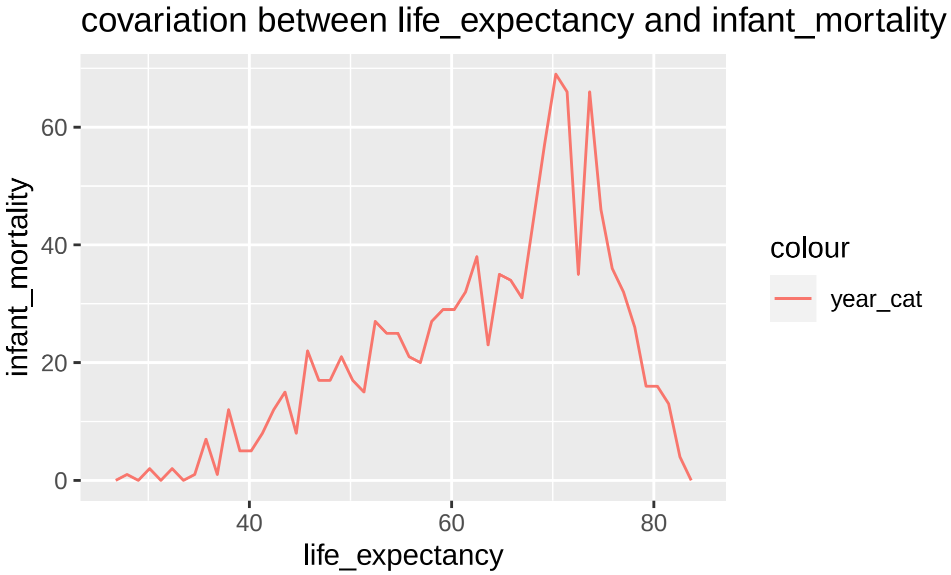
Exercise 8

```

gapminder_cat %>%
  filter(year%% 10==0) %>%
  ggplot() +
    geom_freqpoly(
      mapping = aes(
        x = life_expectancy,
        color = "year_cat",
        ),
      bins = 50,

    ) +

    labs(title = "covariation between life_expectancy and infant_mortality",
     x = "life_expectancy",
     y = "infant_mortality")
```



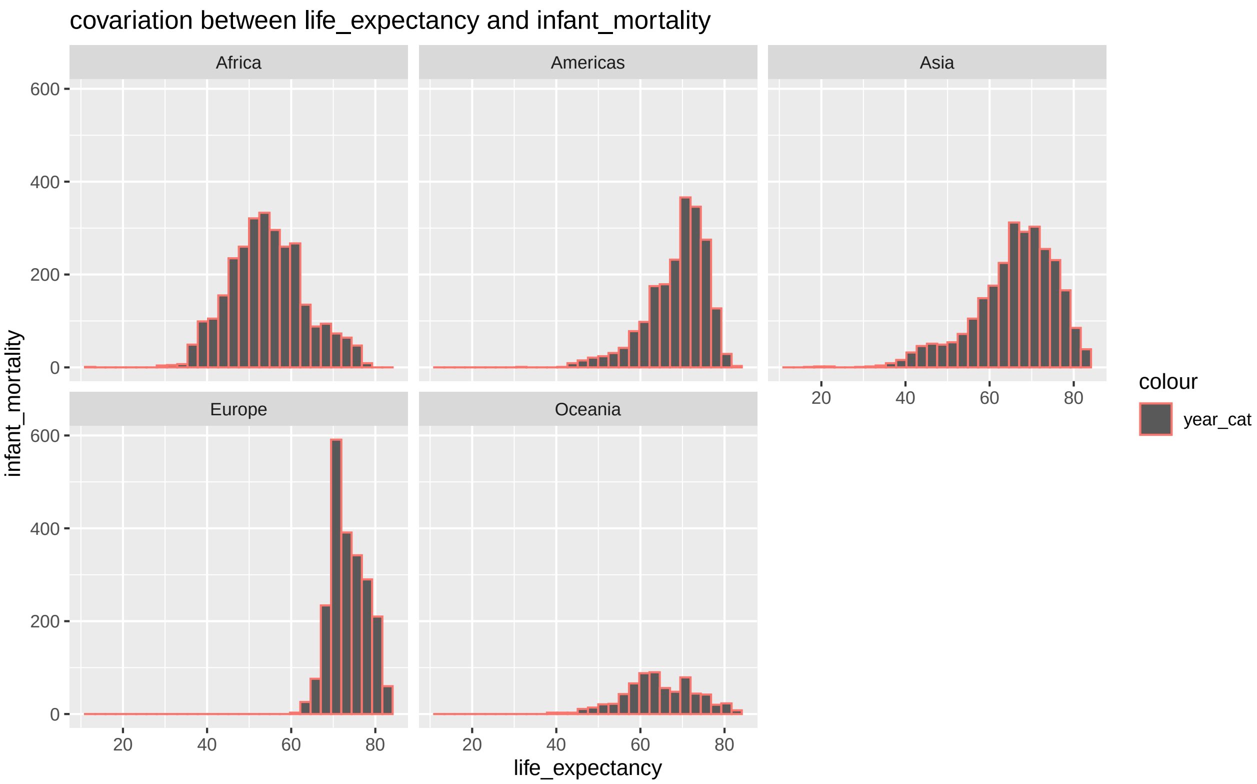
Exercise 9

```

  gapminder_cat %>%
    ggplot() +
      geom_histogram(
       mapping = aes(
        x = life_expectancy,
        color = "year_cat",
        ),
       bins=30,

    ) +
      facet_wrap(~ continent)+

  labs(
    title = "covariation between life_expectancy and infant_mortality",
     x = "life_expectancy",
     y = "infant_mortality")
```


covariation between life_expectancy and infant_mortality
Africa
Americas
Asia
Europe
Oceania
infant_mortality
life_expectancy
colour
■ year_cat