Assignment 4: Mind the Gap

James Jeong_Minsu Kang

2022-07-05

Exercise 1

i. Country, continent, and regions are variables in the dataset that are categorical.
ii. Year, infant_mortality, life_expectancy, fertility, population, and gdp are variables in the dataset that are continuous.
iii. Each row in the dataset represents countries of specific year.

Exercise 2

i.

```
gapminder %>%
  group_by(continent) %>%
  summarize( number = n(),
    mean = mean(infant_mortality, na.rm = TRUE),
    median = median(infant_mortality, na.rm = TRUE),
    stdev = sd(infant_mortality, na.rm = TRUE),
    interquartile_range = IQR(infant_mortality, na.rm = TRUE),
    minimum = min(infant_mortality, na.rm = TRUE),
    maximum = max(infant_mortality, na.rm = TRUE),
  )
```

continent	number	mean	median	stdev	interquartile_range	minimum	maximum
Africa	2907	95.12395	93.40	43.86734	62.500	11.4	237.4
Americas	2052	42.88145	30.80	34.60426	39.475	4.0	194.8
Asia	2679	55.26174	43.10	46.92980	58.950	2.0	276.9
Europe	2223	15.33022	11.25	14.19481	13.700	1.5	120.0
Oceania	684	39.10136	29.10	29.13022	35.875	3.0	134.6

ii.

```
gapminder %>%
  group_by(continent) %>%
  summarize(categorical_variable = n()) %>%
  mutate(categorical_variable)
```

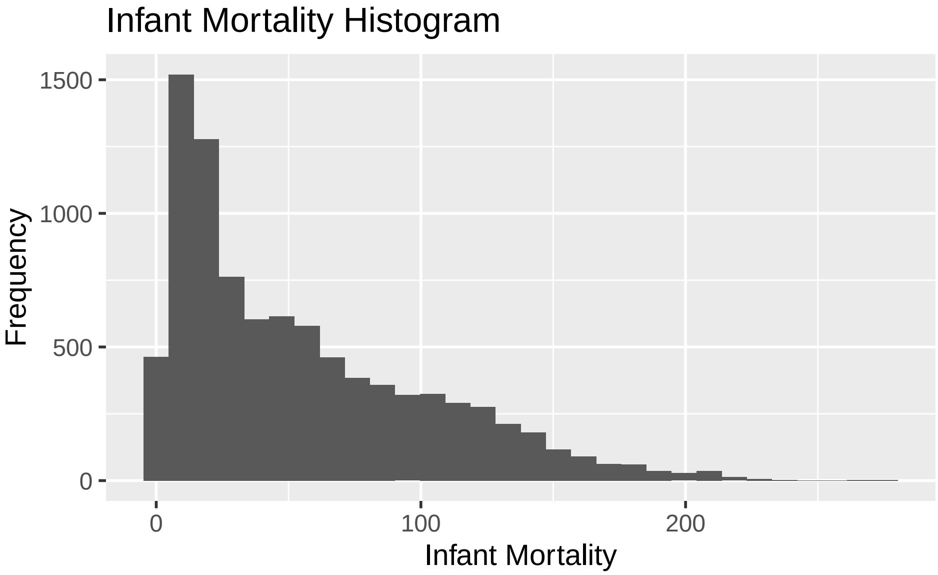
continent	categorical_variable
Africa	2907
Americas	2052
Asia	2679
Europe	2223
Oceania	684

Exercise 3

i.

```
gapminder %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = infant_mortality),
    bins = 30
  ) +
  labs(
    title = "Infant Mortality Histogram",
    x = "Infant Mortality",
    y = "Frequency"
  )
```

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

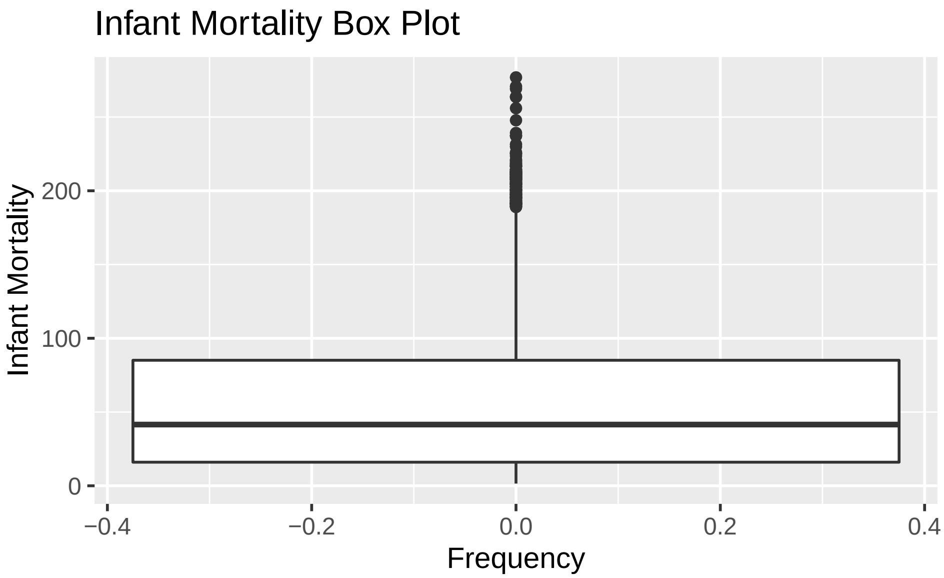


- The histogram shows right-skewed shape of distribution. The reason that the infant mortality gradually falls is because due to the development of the medical technology in the world within the time change.

ii.

```
gapminder %>%
  ggplot() +
  geom_boxplot(
    mapping = aes(y = infant_mortality)
  ) +
  labs(
    title = "Infant Mortality Box Plot",
    y = "Infant Mortality",
    x = "Frequency"
  )
```

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

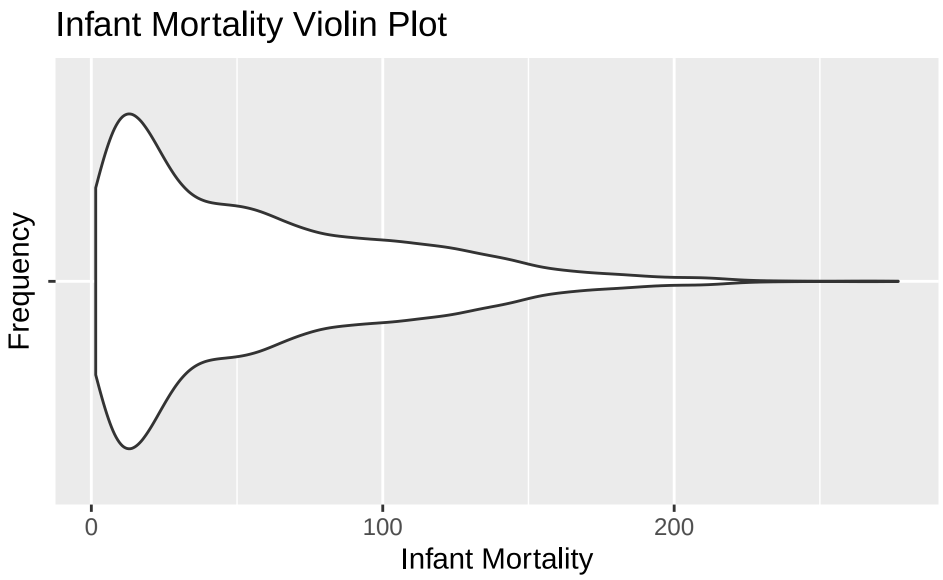


- The box plot is right-skewed as the median is located lower interquartile of the box. Center of the distribution is between around 40 to 50 based on the median line.

iii.

```
gapminder %>%
  ggplot() +
  geom_violin(
    mapping = aes(x = infant_mortality, y="")
  ) +
  labs(
    title = "Infant Mortality Violin Plot",
    x = "Infant Mortality",
    y = "Frequency"
  )
```

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

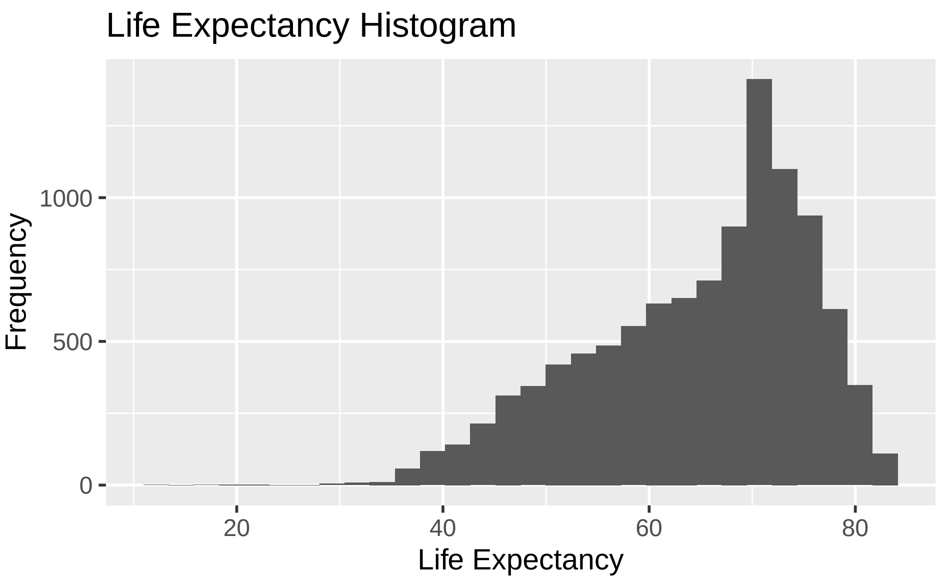


- Shape of the violin graph is right-skewed.

Exercise 4

i. Life Expectancy Histogram

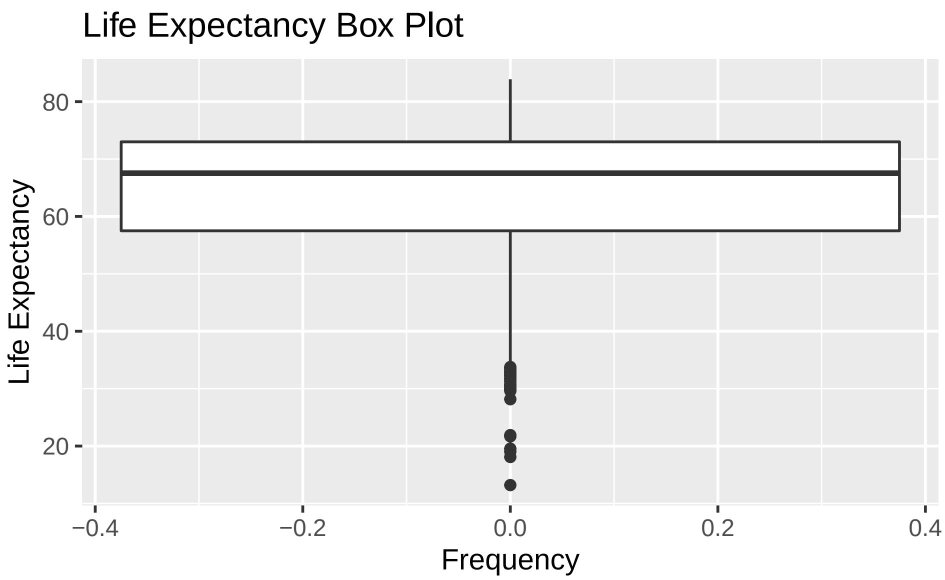
```
gapminder %>%
  ggplot() +
  geom_histogram(
    mapping = aes(x = life_expectancy),
    bins = 30
  ) +
  labs(
    title = "Life Expectancy Histogram",
    x = "Life Expectancy",
    y = "Frequency"
  )
```



- The graph is left-skewed shape and the center is at 67.54

ii. Life Expectancy Box Plot

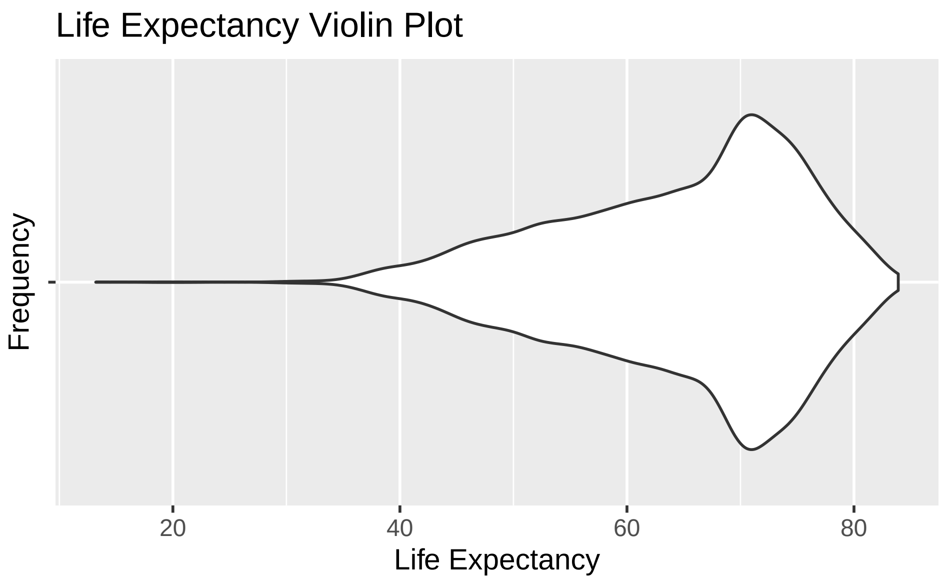
```
gapminder %>%
  ggplot() +
  geom_boxplot(
    mapping = aes(y = life_expectancy)
  ) +
  labs(
    title = "Life Expectancy Box Plot",
    y = "Life Expectancy",
    x = "Frequency"
  )
```



- The shape of the graph is left-skewed and the center is located at 67.54.

iii. Life Expectancy Violin Plot

```
gapminder %>%
  ggplot() +
  geom_violin(
    mapping = aes(x = life_expectancy, y="")
  ) +
  labs(
    title = "Life Expectancy Violin Plot",
    x = "Life Expectancy",
    y = "Frequency"
  )
```

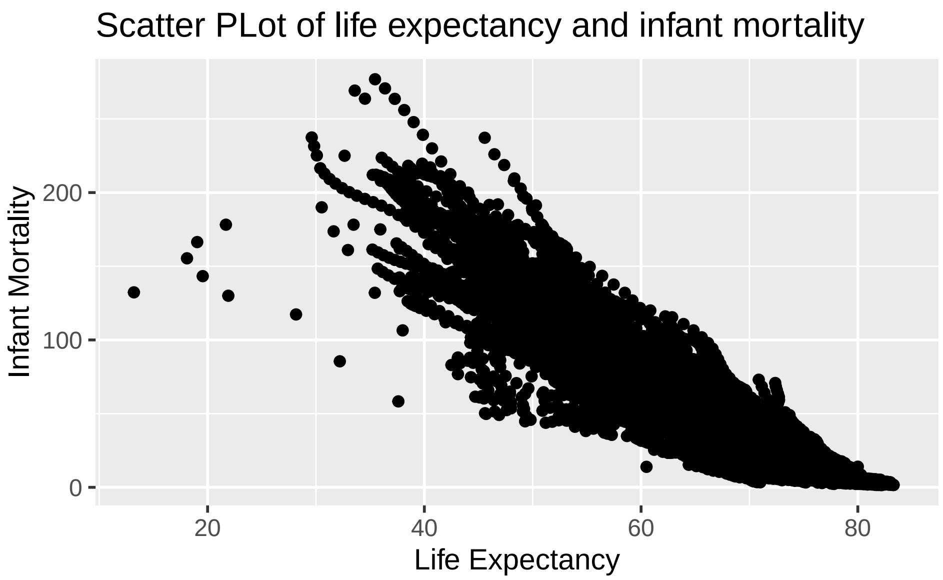


- The shape of the graph is left-skewed and the center is located at 67.54.
- All the three graphs show the similar patterns as they are different shape of graphs within the same dataset.

Exercise 5

```
gapminder %>%
  ggplot() +
  geom_point(
    mapping = aes(
      x = life_expectancy,
      y = infant_mortality
    )
  ) +
  labs(
    title = "Scatter PLot of life expectancy and infant mortality",
    x = "Life Expectancy",
    y = "Infant Mortality"
  )
```

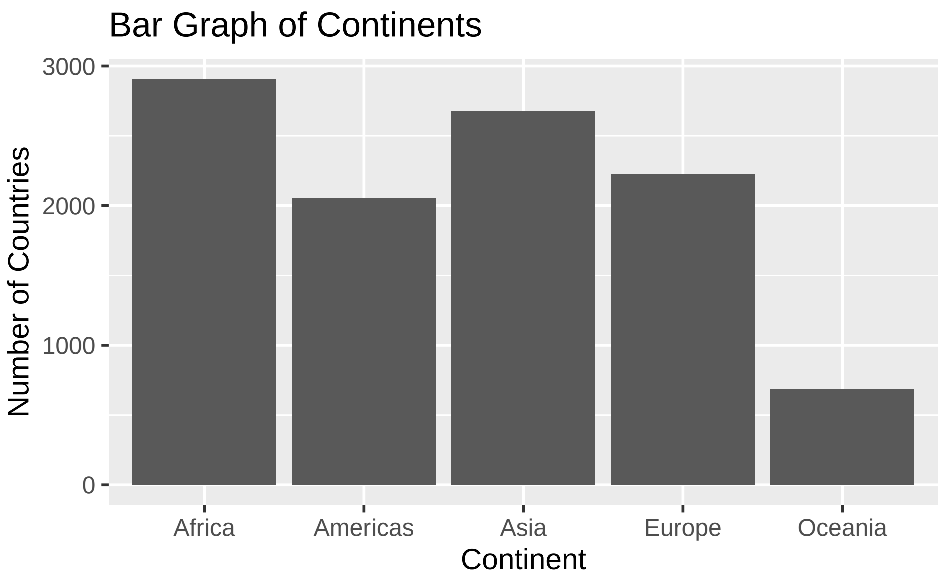
```
## Warning: Removed 1453 rows containing missing values (geom_point).
```



Exercise 6

i.

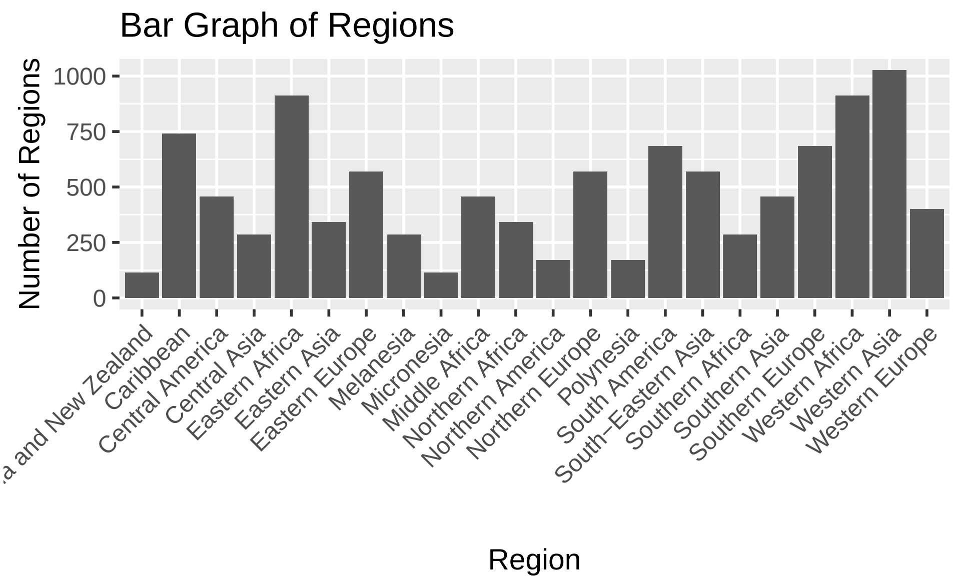
```
gapminder %>%
  ggplot() +
  geom_bar(
    mapping = aes(
      x = continent
    )
  ) +
  labs(
    title = "Bar Graph of Continents",
    x = "Continent",
    y = "Number of Countries"
  )
```



- The bars are high since the dataset includes variables with respect to each year.

ii.

```
gapminder %>%
  ggplot() +
  geom_bar(
    mapping = aes(
      x = region
    )
  ) +
  labs(
    title = "Bar Graph of Regions",
    x = "Region",
    y = "Number of Regions"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



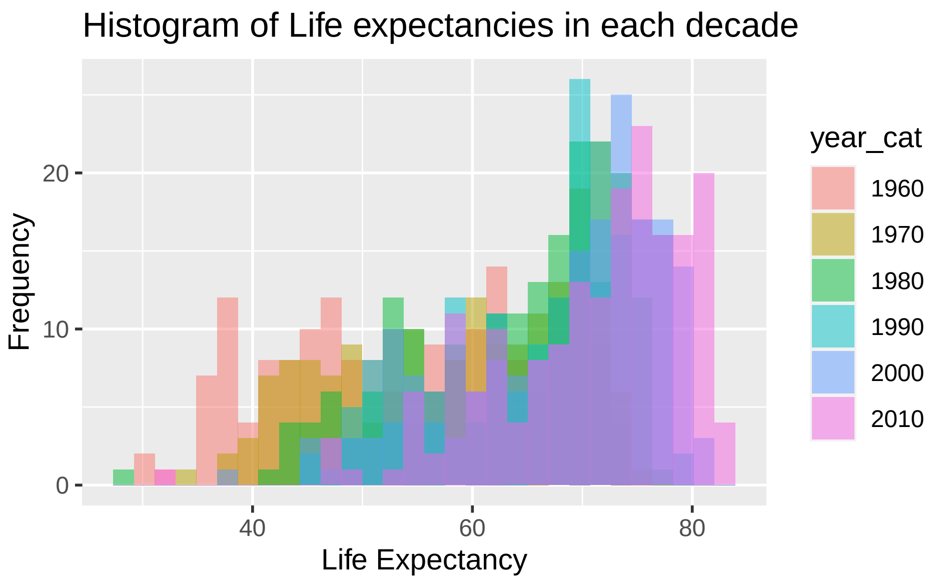
Exercise 7

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

Exercise 8

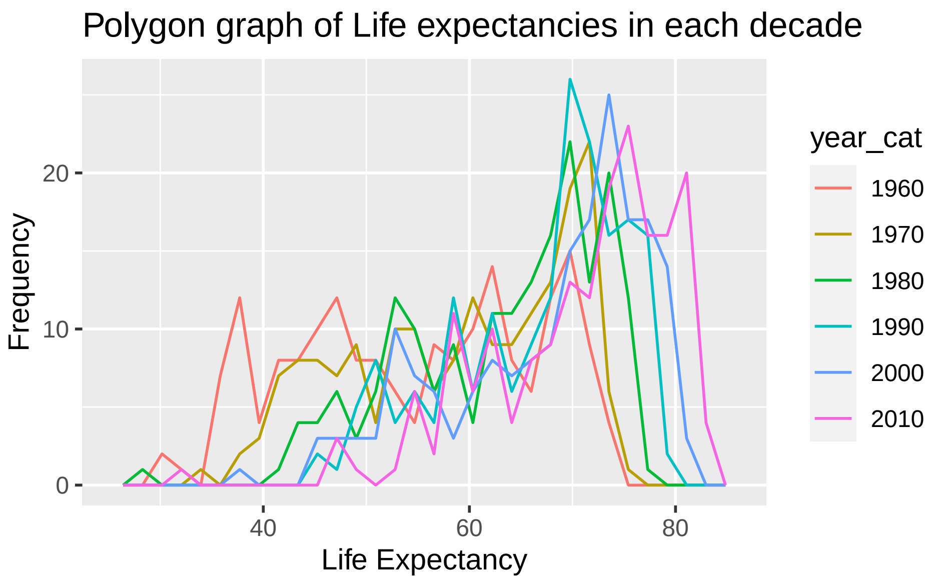
i.

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
    geom_histogram(
      mapping = aes(
        x = life_expectancy,
        fill = year_cat),
      bins = 30,
      position = "identity",
      alpha = 0.5
    ) +
  labs(title = "Histogram of Life expectancies in each decade",
    x = "Life Expectancy",
    y = "Frequency"
  )
```



ii.

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
    geom_freqpoly(
      mapping = aes(
        x = life_expectancy,
        color = year_cat),
      bins = 30
    ) +
  labs(title = "Polygon graph of Life expectancies in each decade",
    x = "Life Expectancy",
    y = "Frequency"
  )
```

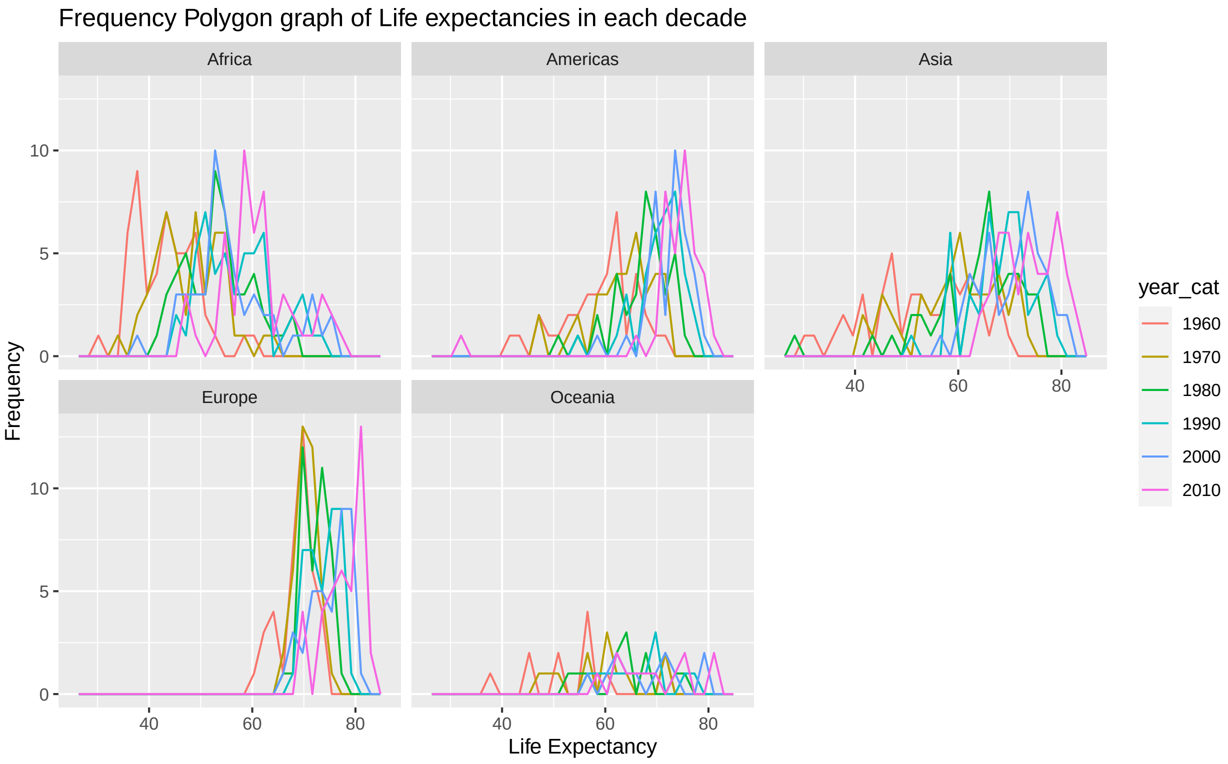


iii. With the change over time to the modern world, the life expectancy tends to rise based on the distribution of the both graphs.

Exercise 9

i.

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
    geom_freqpoly(
      mapping = aes(
        x = life_expectancy,
        color = year_cat),
      bins = 30
    ) +
  labs(title = "Frequency Polygon graph of Life expectancies in each decade",
    x = "Life Expectancy",
    y = "Frequency"
  ) +
  facet_wrap(~continent)
```

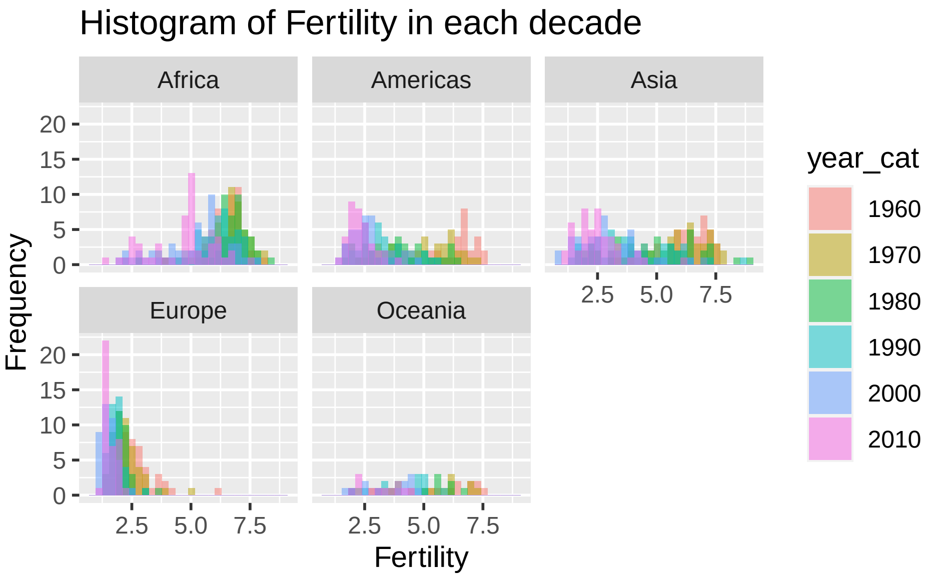


ii.

- Among the distribution of the life expectancies, America and Asia shows similar skewness within their distribution as it gradually goes up. However, Oceania comparably has stable frequency distribution over time. While Europe had a huge improvement in the level of life expectancy at certain period of time, Africa tended to have a decrease within the level of life expectancy.

##Exercise 10

```
gapminder_cat %>%
  filter(year %% 10 == 0) %>%
  ggplot() +
    geom_histogram(
      mapping = aes(
        x = fertility,
        fill = year_cat),
      bins = 30,
      position = "identity",
      alpha = 0.5
    ) +
  labs(title = "Histogram of Fertility in each decade",
    x = "Fertility",
    y = "Frequency"
  ) +
  facet_wrap(~continent)
```



- Each continent shows all different distribution of fertility rate with respect to time. It is definite fact that all the continents show less fertility rate over time, but specifically, Europe had high fertility rate since old days. On the other hand, Africa's fertility rate is higher and this could be linked with the social level of Africa which has the lowest level of GDP compared to other continents. For America and Asia, it shows typical bimodal shape as back in the past around 1960s, fertility rate was high but around 2010, it has low fertility rate. Compared to these continents, Oseania has stable freqency within the fertility which could be identified as a uniform. This could be derived from that fact that Oseania has low population compared to other continents.