346 Assignment 1 Haya Naviwala

Haya Naviwala

5/15/2022

R Markdown

Introduction: The source of this data is the Gapminder dataset which is an open source resource for which data is collected and global trends are identified.

As per the output below, the total number of variables and observations are 1,704 observations and 6 variables.

Overview of the variables- Country: the 142 countries from the world that Gapminder collects data on and reports on trends for, they are listed in alphabetical order from A-Z. Continent: there are 5 separate continents, they are also listed in alphabetical order. Year: These are the years that data has been collected from, it ranges from 1952 to 2007. lifeExp: these are the life expectancy values collected from each country per the years above there are 1626 unique life expectancy values. pop: this is the population per respective country per year. gdpPercap: this is the gdp per capita per country by year, it represents the economic output per it's population

This report will explore the intricacies of the gapminder data set and explore the trends, idiosyncracies, and nuances that come with it. There will be a range of analyses conducted and visualizations produced to aid in our analysis of the data and come to conclusions on it.

Description of the Variables: our variables fall into either category of Identifiers and Metrics. Identifiers are normally categorical, and in our case our variables of Country, Continent, and Year fall under the Identifiers category.

Metrics are usually measurable values, and tend to be numeric, in our case our variables of life expectancy, population, and gdp per capita fall under the Metrics classification.

Exploring the Identifiers: according to our analysis the number of unique values under the country category are 142 (meaning there is data on 142 different countries), for continent there are 5 (meaning data is collected on 5 different continents), for year there are 12 (meaning the data collected is over a range of 12 years), for life expectancy there are 1626 unique values, for population there are 1704, and for gdp per capita there are 1704 too. The list of unique values can also be observed next. We can also verify that there are no missing values, which there are none.

In general, we can say that some of these identifiers should be analyzed individually and some as a group. Identifiers such as a Country should be individually analyzed so that we can better make sense of what the data is telling us in proportion to the country itself. While something like population for example could be analyzed as a group as it's such a vast number that analyzing it in groups is more digestible and more comprehensible for us to glean information and trends from.

Exploring the Metrics

We have explored several descriptive statistics for each of the 6 variables

We have then calculated the statistics grouped by continent for metric variables and have used box plots to determine the existence of outliers. Our box plots show us that our population metrics and gdp per capita metrics have an existence of outliers.

We then showed the distribution of each metric variable through a histogram

We then showed the boxplots for each year by continent. For the Americas the trend is that the Life expectancy has steadily increased every single year. In Africa we notice a less steep increase in life expectancy over the years and even a dip in 2002. The increase is steep for Asia, Europe, and Oceania too however we can notice that the range for life expectancy over the years is largest for Asia (which makes sense since it is the largest continent), an the range is the smallest in Oceania meaning there is less variance (Oceania is also significantly smaller).

One metric I would love to explore more is this range and I would like to break it down by country within Asia to see what countries have higher and lower life expectancies.

Exploring the Relationships between variables: we first are showing the linear correlation between the two variables. At 0.58 we can say the correlation is moderate between gdp per capita and life expectancy.

We then explored the scatterplots of the metric variables of gdp per capita vs life expectancy by year and by continent. What's striking about this chart is how much life expectancy differs by continent and also it appears that there is a decent amount of correlation between life expectancy and gdp per capita, although not enough that it can be an explanatory factor. Comments on life expectancy: the range in Africa and Asia is massive compared to the other continents, this makes sense as there are more third world countries in these continents than in others plus they are larger population wise. Comments on gdp per capita: Europe and Asia tend to have the highest gdp per capita, but not necessarily a higher life expectancy- such as with asia.

Comments on relationships bw variables: There are also years such as 1977 where Asia has a high gdp per capita but still a lower life expectancy. This also explains our correlation coefficient since it only showed a moderate correlation between the two variables.

```
#explore the metrics-summary stats
#by lifeexp
min(gapminder$lifeExp)
```

```
## [1] 23.599
```

```
quantile(gapminder$lifeExp,0.25)
```

```
##    25%
## 48.198
```

```
median(gapminder$lifeExp)
```

```
## [1] 60.7125
```

```
mean(gapminder$lifeExp)
```

```
## [1] 59.47444
```

```
quantile(gapminder$lifeExp,0.75)
```

```
##     75%
## 70.8455
```

```
max(gapminder$lifeExp)
```

```
## [1] 82.603
```

```
sd(gapminder$lifeExp)
```

```
## [1] 12.91711
```

```
IQR(gapminder$lifeExp)
```

```
## [1] 22.6475
```

```r
#by year
min(gapminder$year)
```

```
## [1] 1952
```

```r
quantile(gapminder$year,0.25)
```

```
##     25%
## 1965.75
```

```r
median(gapminder$year)
```

```
## [1] 1979.5
```

```r
mean(gapminder$year)
```

```
## [1] 1979.5
```

```r
quantile(gapminder$year,0.75)
```

```
##     75%
## 1993.25
```

```r
max(gapminder$year)
```

```
## [1] 2007
```

```r
sd(gapminder$year)
```

```
## [1] 17.26533
```

```r
IQR(gapminder$year)
```

```
## [1] 27.5
```

```r
#by pop
min(gapminder$pop)
```

```
## [1] 60011
```

```r
quantile(gapminder$pop,0.25)
```

```
##     25%
## 2793664
```

```r
median(gapminder$pop)
```

```
## [1] 7023596
```

```r
mean(gapminder$pop)
```

```
## [1] 29601212
```

```r
quantile(gapminder$pop,0.75)
```

```
##      75%
## 19585222
```

```r
max(gapminder$pop)
```

```
## [1] 1318683096
```

```r
sd(gapminder$pop)
```

```
## [1] 106157897
```

```
IQR(gapminder$pop)
```

```
## [1] 16791558
```

```
#by gdpPerCap
min(gapminder$gdpPercap)
```

```
## [1] 241.1659
```

```
quantile(gapminder$gdpPercap,0.25)
```

```
##     25%
## 1202.06
```

```
median(gapminder$gdpPercap)
```

```
## [1] 3531.847
```

```
mean(gapminder$gdpPercap)
```

```
## [1] 7215.327
```

```
quantile(gapminder$gdpPercap,0.75)
```

```
##      75%
## 9325.462
```

```
max(gapminder$gdpPercap)
```

```
## [1] 113523.1
```

```
sd(gapminder$gdpPercap)
```

```
## [1] 9857.455
```

```
IQR(gapminder$gdpPercap)
```

```
## [1] 8123.402
```

```
#grouped by continent-lifeexp
summary_by_continent <-
  gapminder %>%
  group_by(continent) %>%
  summarize(min_lifeExp = min(gapminder$lifeExp),
            quart1_lifeExp = quantile(gapminder$lifeExp,0.25),
            med_lifeExp = median(gapminder$lifeExp),
            avg_lifeExp = mean(gapminder$lifeExp),
            quart3_lifeExp = quantile(gapminder$lifeExp,0.75),
            max_lifeExp = max(gapminder$lifeExp),
            sd_lifeExp = sd(gapminder$lifeExp),
            IQR_lifeExp = IQR(gapminder$lifeExp),
            .groups = 'drop_last')

summary_by_continent %>%
  kable(digits = 1)
```

continent	min_lifeExp	quart1_lifeExp	med_lifeExp	avg_lifeExp	quart3_lifeExp	max_lifeExp	sd_lifeExp	IQR_lifeExp
Africa	23.6	48.2	60.7	59.5	70.8	82.6	12.9	22.6

continent	min_lifeExp	quart1_lifeExp	med_lifeExp	avg_lifeExp	quart3_lifeExp	max_lifeExp	sd_lifeExp	IQR_lifeExp
Americas	23.6	48.2	60.7	59.5	70.8	82.6	12.9	22.6
Asia	23.6	48.2	60.7	59.5	70.8	82.6	12.9	22.6
Europe	23.6	48.2	60.7	59.5	70.8	82.6	12.9	22.6
Oceania	23.6	48.2	60.7	59.5	70.8	82.6	12.9	22.6

```
#grouped by continent-population
summary_by_continent <-
  gapminder %>%
  group_by(continent) %>%
  summarize(min_pop = min(gapminder$pop),
            quart1_pop= quantile(gapminder$pop,0.25),
            med_pop = median(gapminder$pop),
            avg_pop = mean(gapminder$pop),
            quart3_pop = quantile(gapminder$pop,0.75),
            max_pop = max(gapminder$pop),
            sd_pop = sd(gapminder$pop),
            IQR_pop = IQR(gapminder$pop),
            .groups = 'drop_last')

summary_by_continent %>%
  kable(digits = 1)
```

continent	min_pop	quart1_pop	med_pop	avg_pop	quart3_pop	max_pop	sd_pop	IQR_pop
Africa	60011	2793664	7023596	29601212	19585222	1318683096	106157897	16791558
Americas	60011	2793664	7023596	29601212	19585222	1318683096	106157897	16791558
Asia	60011	2793664	7023596	29601212	19585222	1318683096	106157897	16791558
Europe	60011	2793664	7023596	29601212	19585222	1318683096	106157897	16791558
Oceania	60011	2793664	7023596	29601212	19585222	1318683096	106157897	16791558

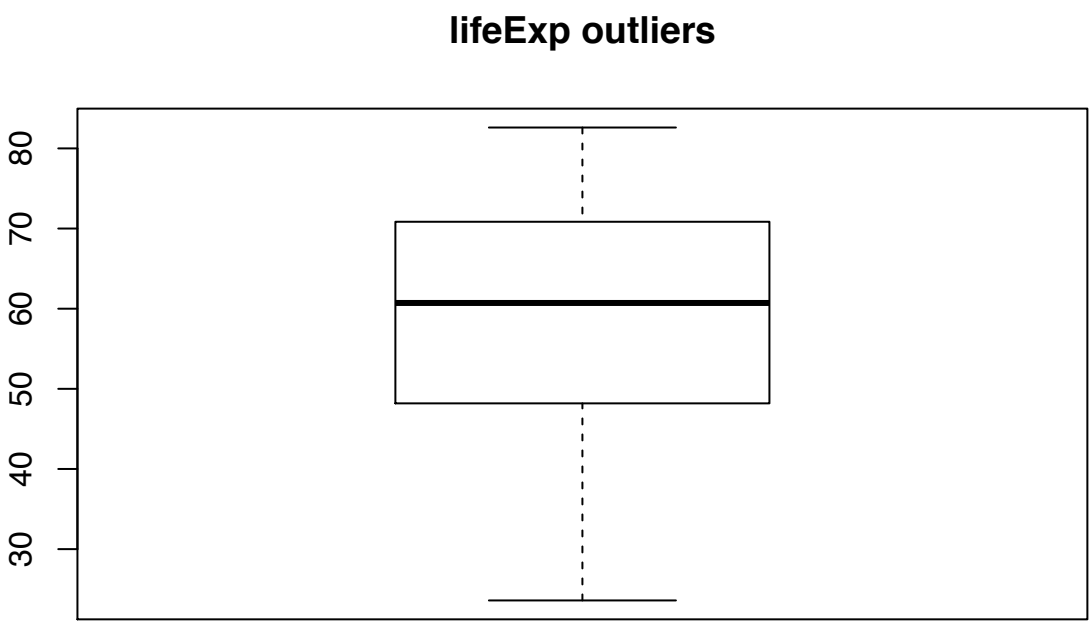
```
#grouped by continent-gdpPercap
summary_by_continent <-
  gapminder %>%
  group_by(continent) %>%
  summarize(min_gdpPercap = min(gapminder$gdpPercap),
            quart1_gdpPercap= quantile(gapminder$gdpPercap,0.25),
            med_gdpPercap = median(gapminder$gdpPercap),
            avg_gdpPercap = mean(gapminder$gdpPercap),
            quart3_gdpPercap = quantile(gapminder$gdpPercap,0.75),
            max_gdpPercap = max(gapminder$gdpPercap),
            sd_gdpPercap = sd(gapminder$gdpPercap),
            IQR_gdpPercap = IQR(gapminder$gdpPercap),
            .groups = 'drop_last')

summary_by_continent %>%
  kable(digits = 1)
```

continent	min_gdpPercap	quart1_gdpPercap	med_gdpPercap	avg_gdpPercap	quart3_gdpPercap	max_gdpPercap	sd_gdpPercap	IQR_gdpPercap
Africa	241.2	1202.1	3531.8	7215.3	9325.5	113523.1	9857.5	8123.4
Americas	241.2	1202.1	3531.8	7215.3	9325.5	113523.1	9857.5	8123.4

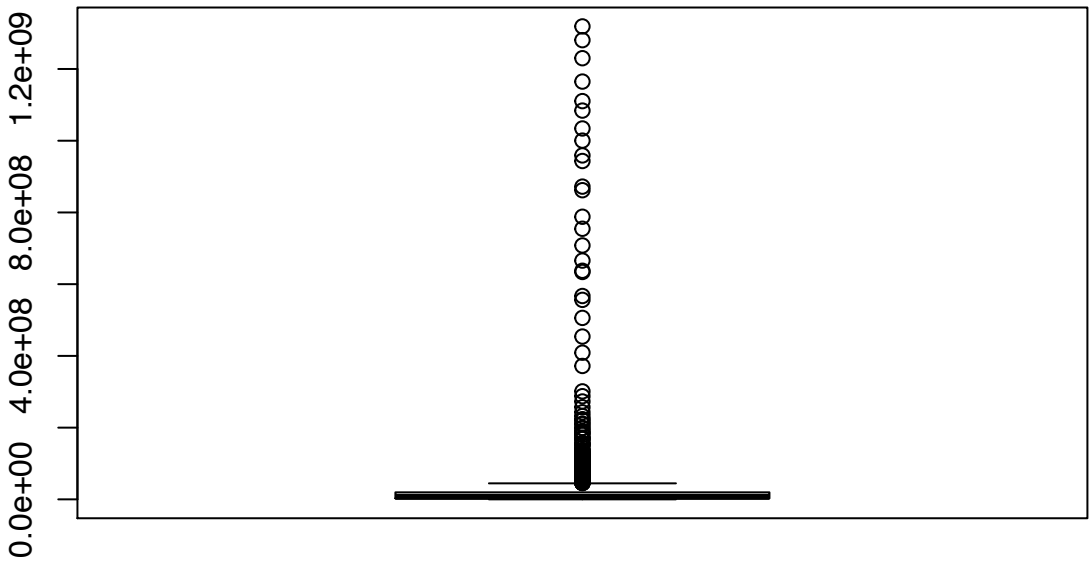
continent	min_gdpPercap	qrtr1_gdpPercap	med_gdpPercap	avg_gdpPercap	qrtr3_gdpPercap	max_gdpPercap	sd_gdpPercap	IQR_gdpPercap
Asia	241.2	1202.1	3531.8	7215.3	9325.5	113523.1	9857.5	8123.4
Europe	241.2	1202.1	3531.8	7215.3	9325.5	113523.1	9857.5	8123.4
Oceania	241.2	1202.1	3531.8	7215.3	9325.5	113523.1	9857.5	8123.4

```
#finding outliers
boxplot(gapminder$lifeExp,main = "lifeExp outliers")
```

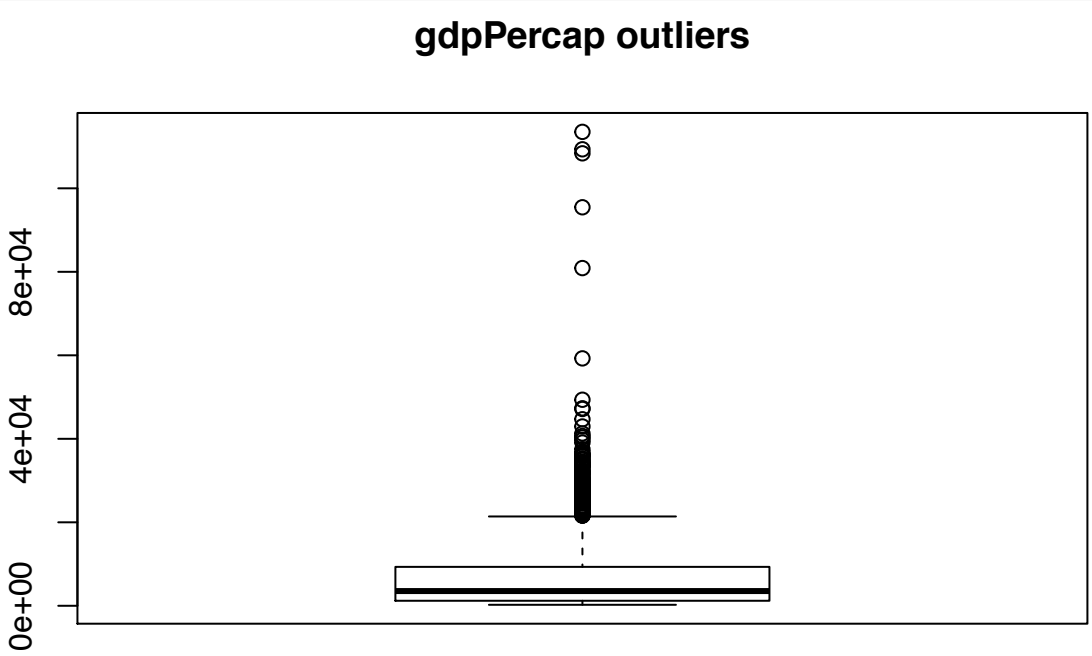



```
boxplot(gapminder$pop,main = "population outliers")
```

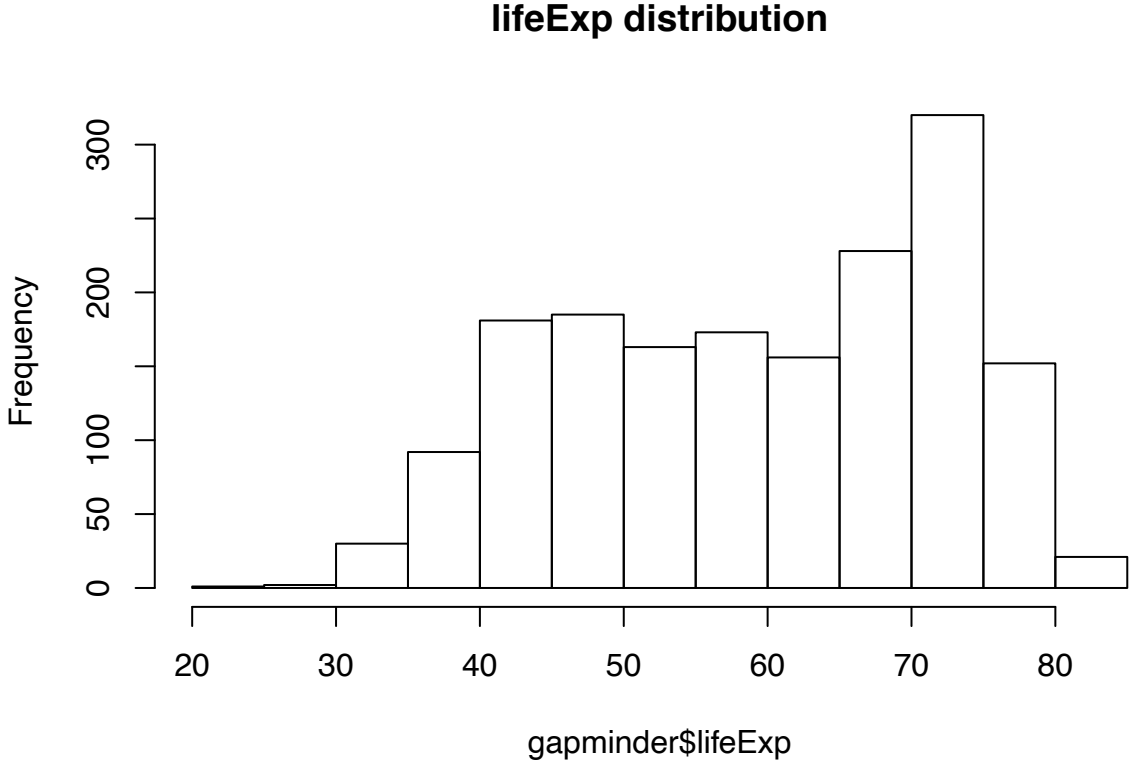
```
## Warning in x[floor(d)] + x[ceiling(d)]: NAs produced by integer overflow
```



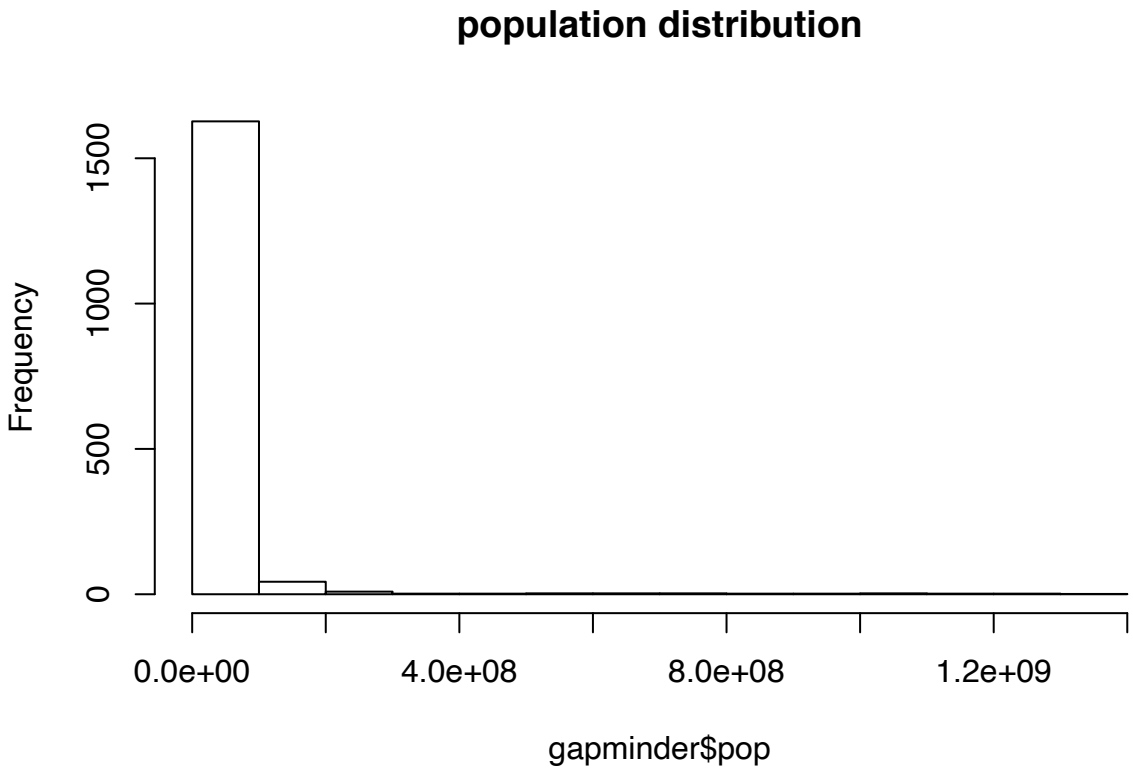
```r
boxplot(gapminder$gdpPercap,main = "gdpPercap outliers")
```



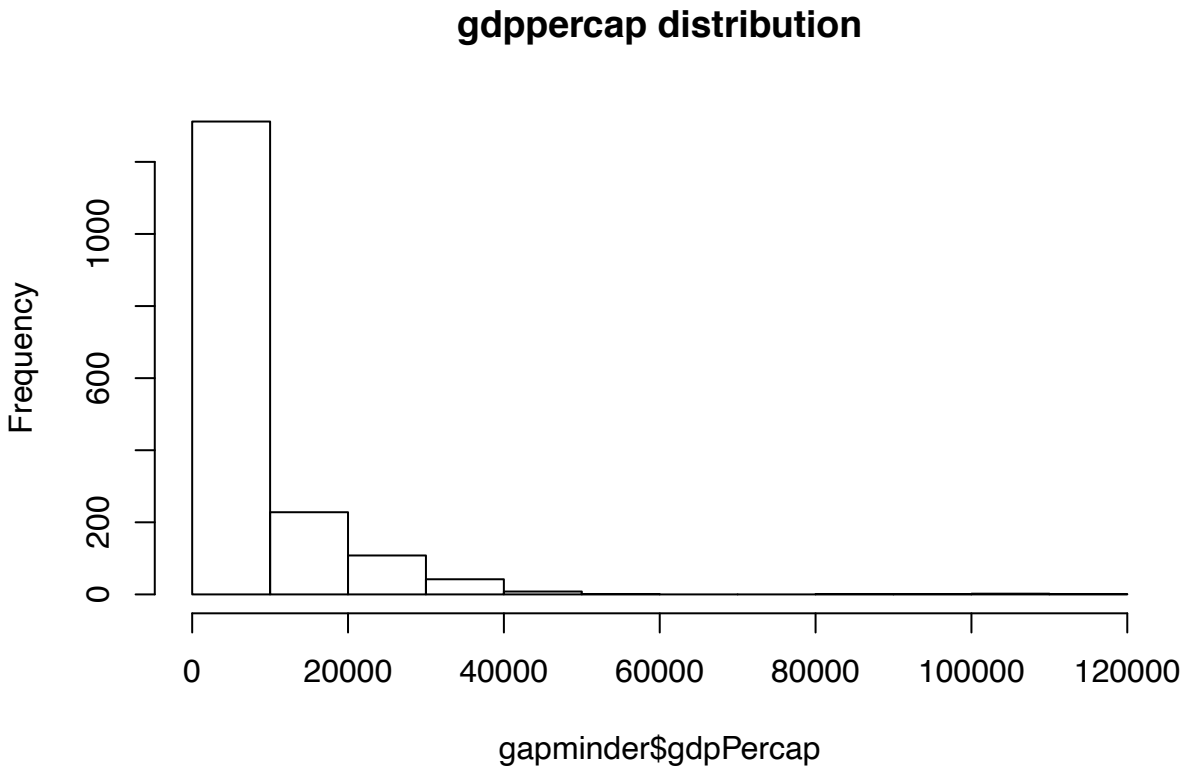

```r
#histogram
hist(gapminder$lifeExp, main = "lifeExp distribution")
```




```r
hist(gapminder$pop, main = "population distribution")
```

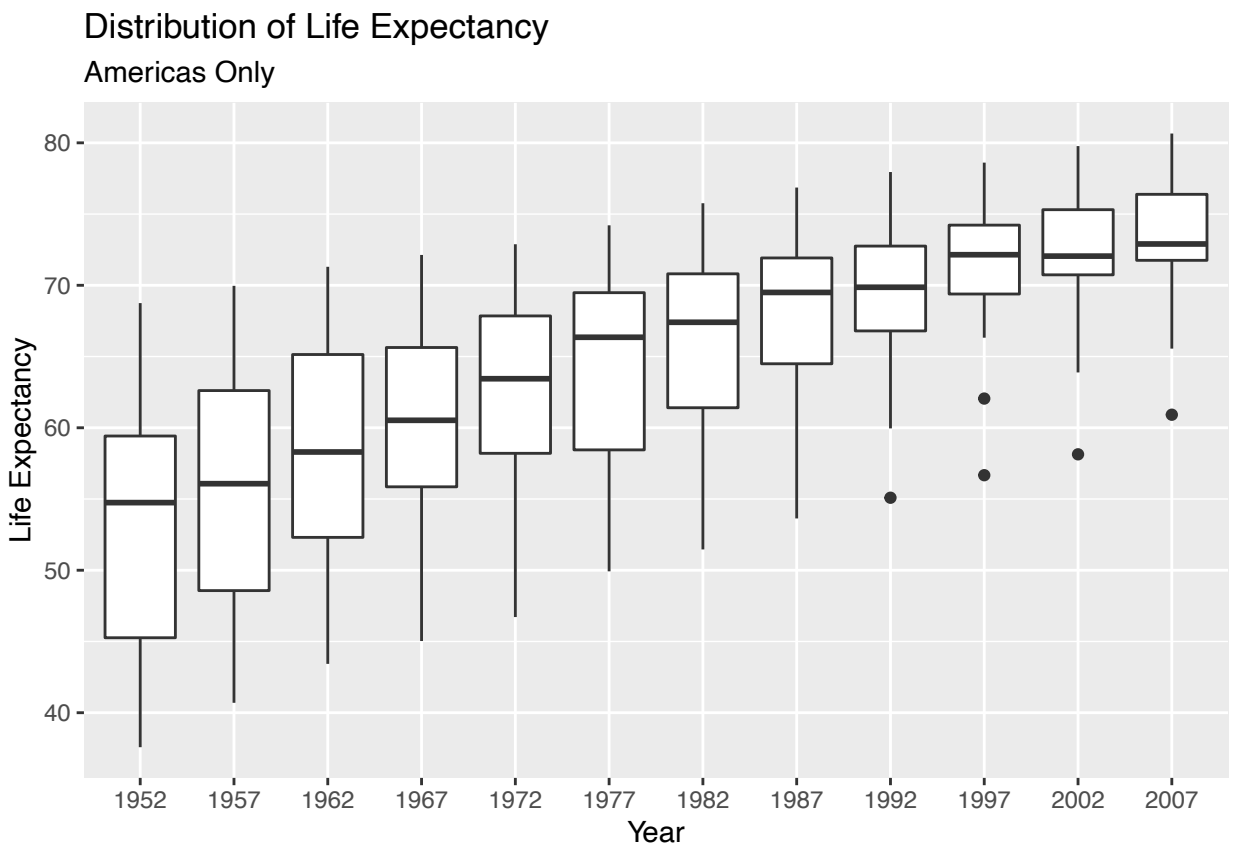



```
hist(gapminder$gdpPercap, main = "gdppercap distribution")
```

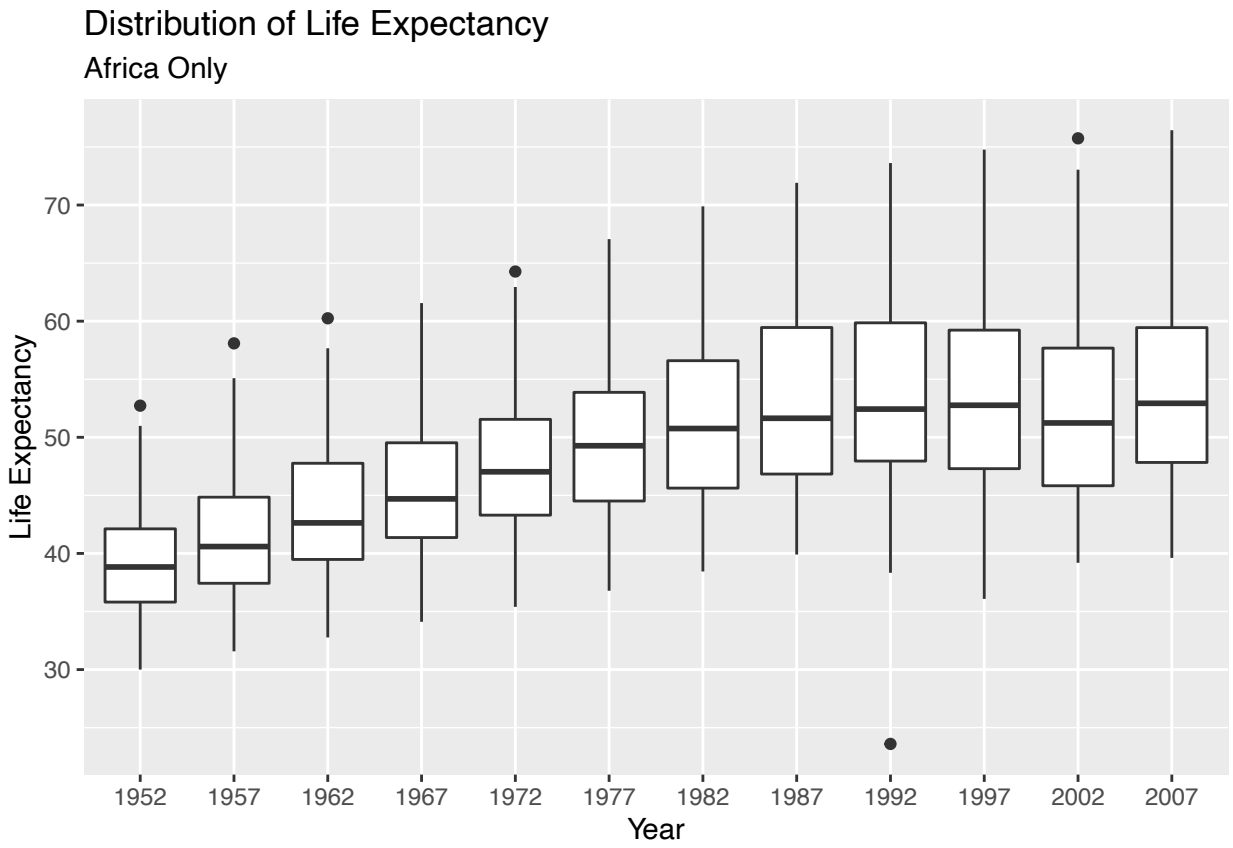



```
#boxplots by continent
#americas
library(ggplot2)
```

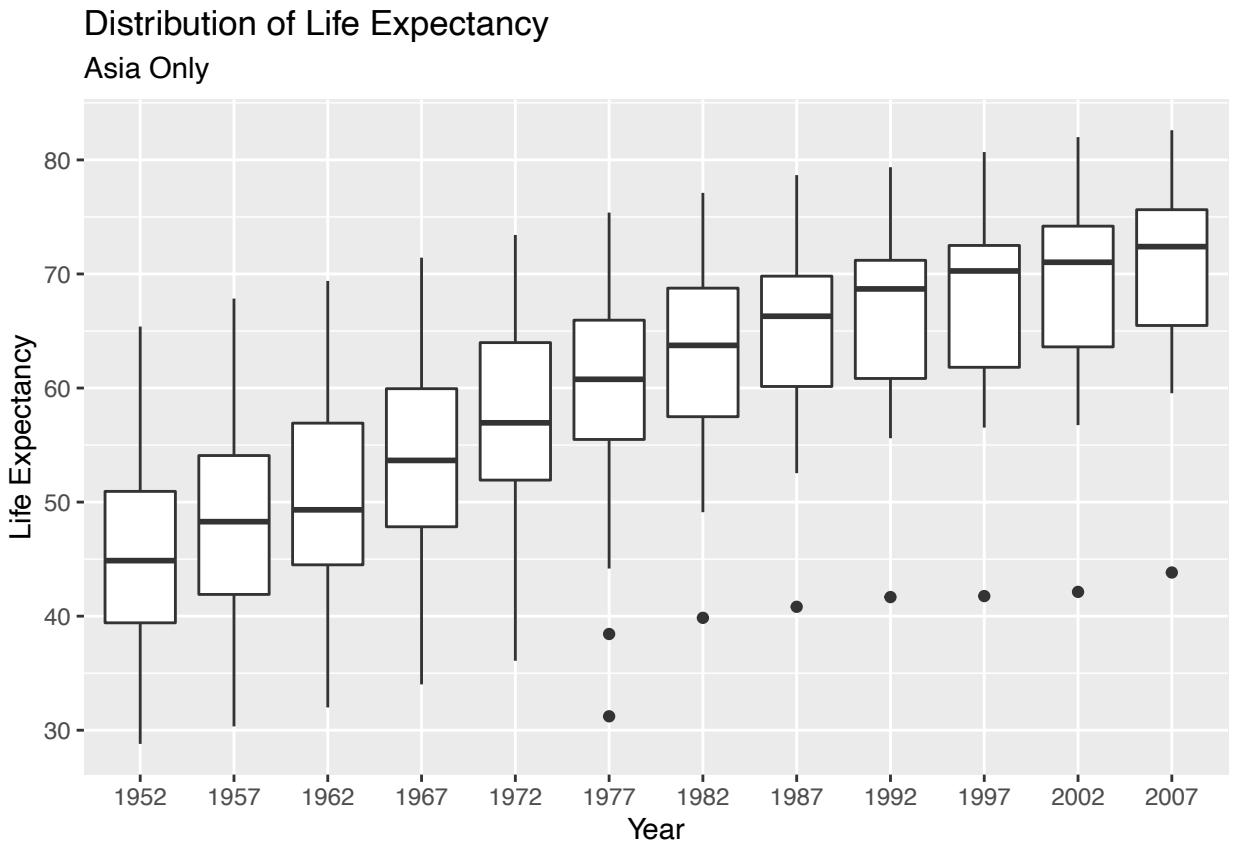
```r
gapminder_americas <- gapminder %>%
  filter(continent == 'Americas')
gapminder_americas %>%
  ggplot(aes(x = as_factor(year), y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy",
       subtitle = "Americas Only",
       x = 'Year',
       y = 'Life Expectancy')
```



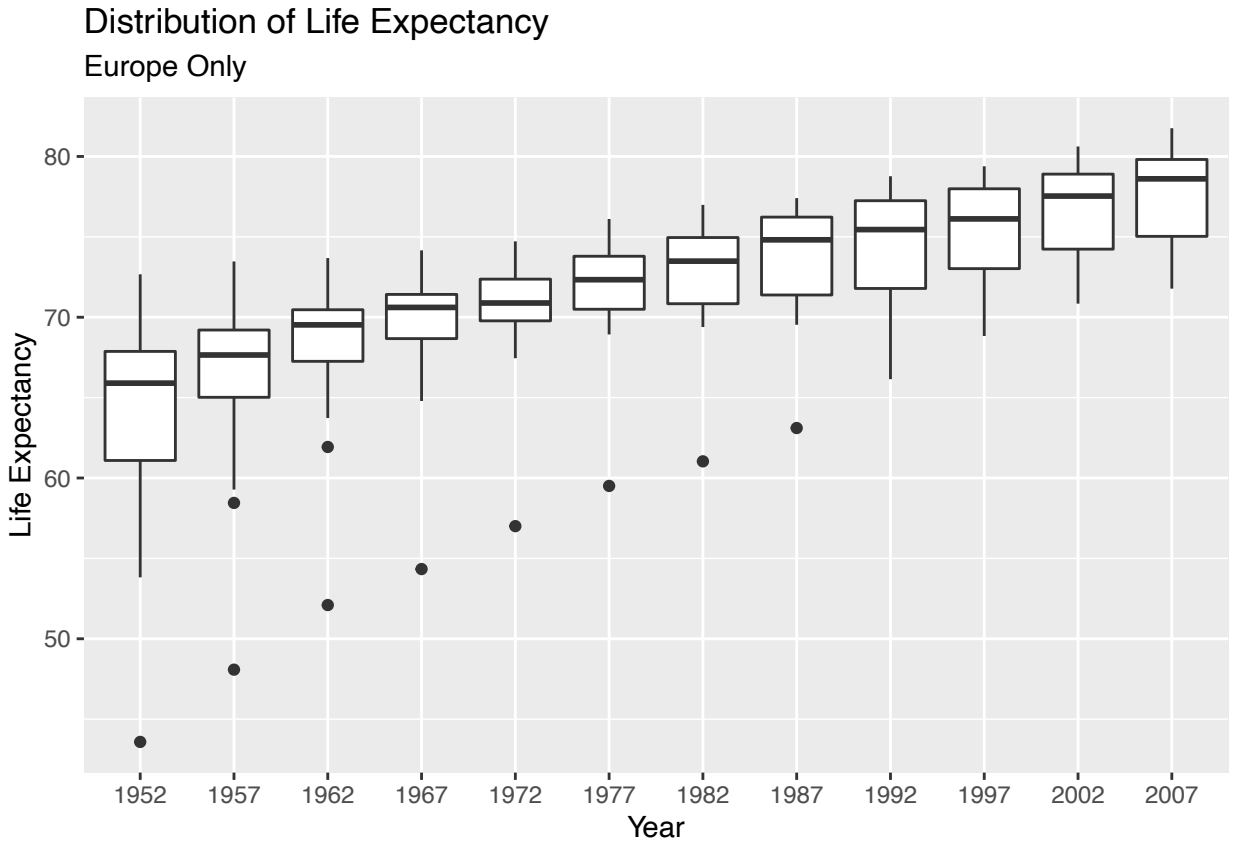

```r
#africa
gapminder_africa <- gapminder %>%
  filter(continent == 'Africa')
gapminder_africa %>%
  ggplot(aes(x = as_factor(year), y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy",
       subtitle = "Africa Only",
       x = 'Year',
       y = 'Life Expectancy')
```



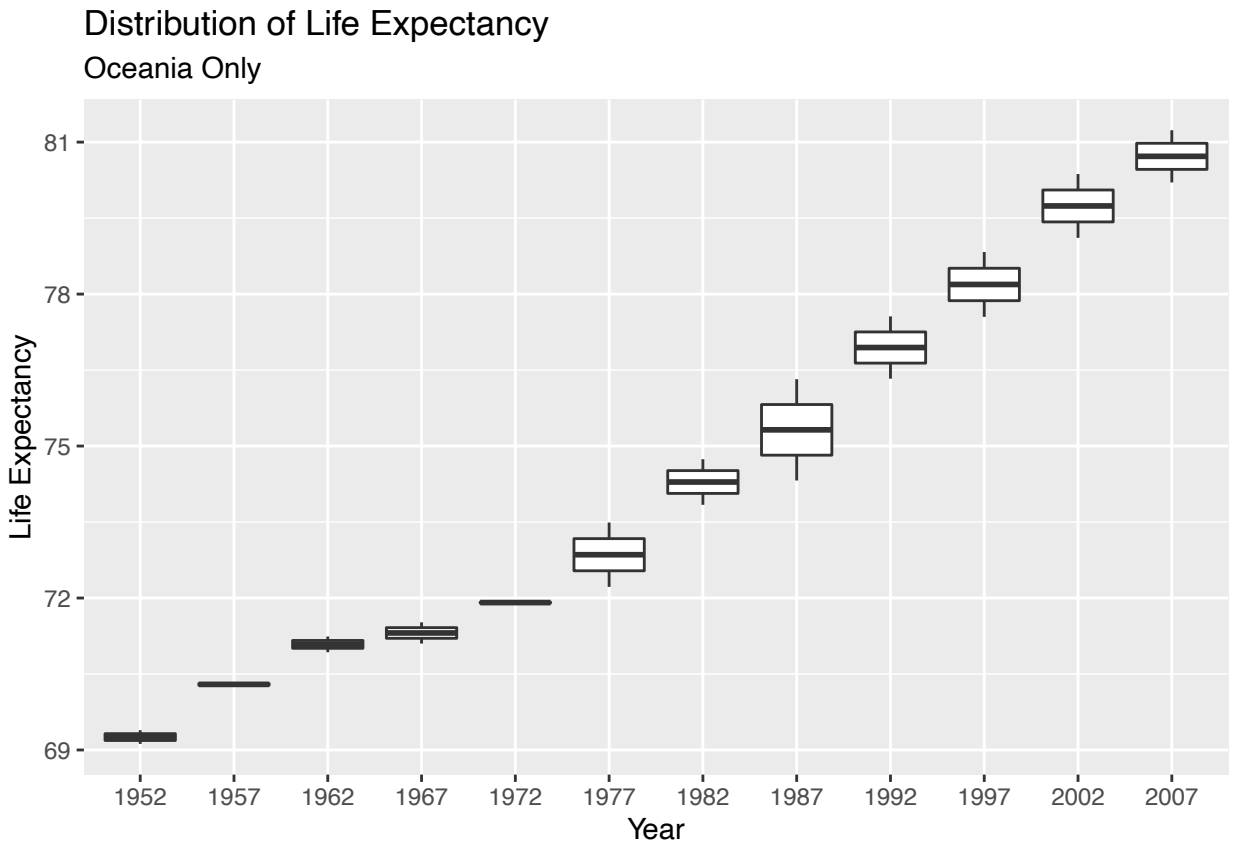

```
#asia
gapminder_asia <- gapminder %>%
  filter(continent == 'Asia')
gapminder_asia %>%
  ggplot(aes(x = as_factor(year), y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy",
       subtitle = "Asia Only",
       x = 'Year',
       y = 'Life Expectancy')
```




```
#europe
gapminder_europe <- gapminder %>%
  filter(continent == 'Europe')
gapminder_europe %>%
  ggplot(aes(x = as_factor(year), y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy",
       subtitle = "Europe Only",
       x = 'Year',
       y = 'Life Expectancy')
```



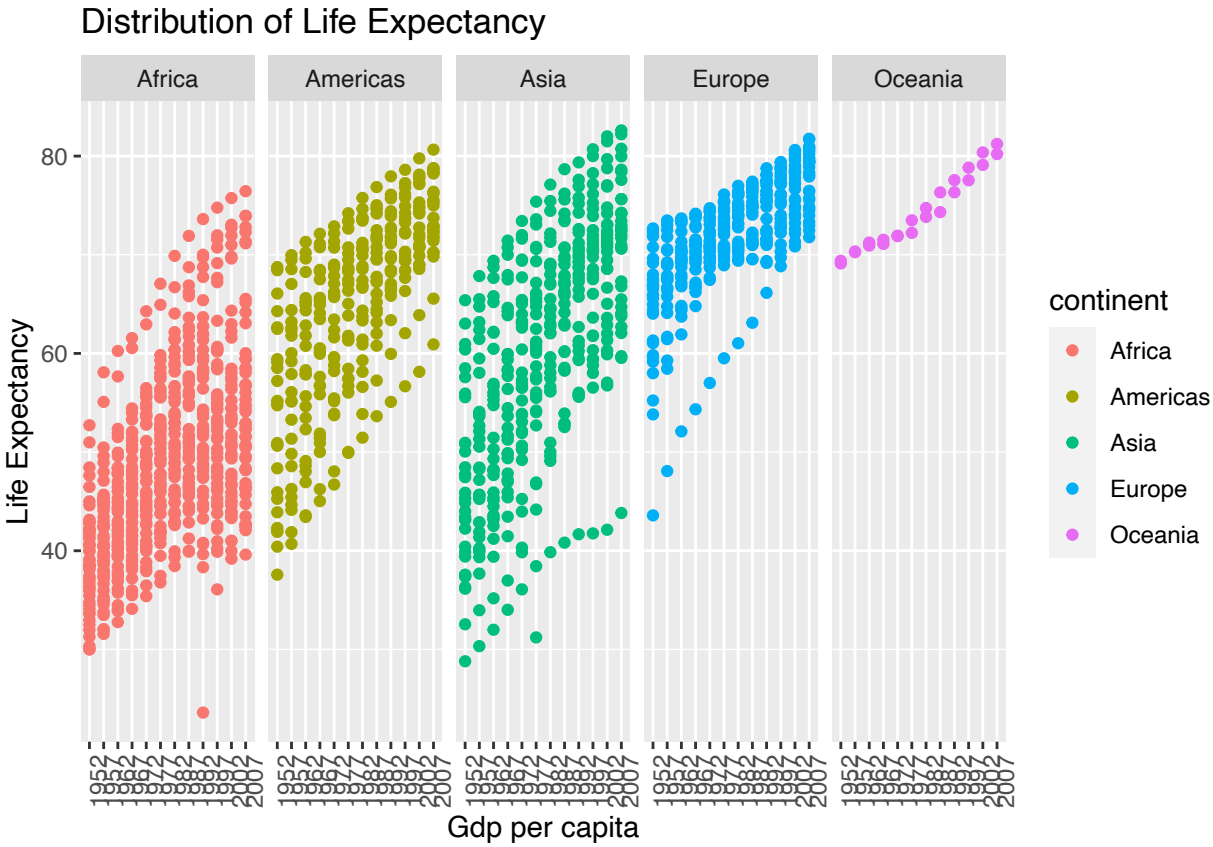

```
#oceania
gapminder_oceania <- gapminder %>%
  filter(continent == 'Oceania')
gapminder_oceania %>%
  ggplot(aes(x = as_factor(year), y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy",
       subtitle = "Oceania Only",
       x = 'Year',
       y = 'Life Expectancy')
```




```
#exploring relationships bw vars
#correlation
cor(gapminder$gdpPercap, gapminder$lifeExp)
```

```
## [1] 0.5837062
```

```
#scatterplot
gapminder %>%
  ggplot(aes(x = as_factor(year), y = lifeExp, color= continent)) +
  geom_point() +
  facet_wrap(vars(continent), nrow = 1) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Distribution of Life Expectancy",
       x = 'Gdp per capita',
       y = 'Life Expectancy')
```



In conclusion, some of the questions I think are worth further studying through my analysis would be:

1) Why does Asia have such a large range in life expectancy? and what is the distribution amongst countries specifically.

2) Has the correlation between life expectancy and gdp per capita differred or stayed the same since the emergence of COVID-19?

3) What other key factors are responsible for the lack of variance in the life expectancy when it comes to certain continents?

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.