hw2

Ning Shen

2019/9/21

Question 1.1

```
gapminder %>%
  filter(country %in% c("China","Canada","Poland") & year %in% 1970:1979)
```

```
## # A tibble: 6 x 6
##   country continent  year lifeExp       pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>     <int>     <dbl>
## 1 Canada  Americas   1972    72.9  22284500    18971.
## 2 Canada  Americas   1977    74.2  23796400    22091.
## 3 China   Asia       1972    63.1 862030000      677.
## 4 China   Asia       1977    64.0 943455000      741.
## 5 Poland  Europe     1972    70.8  33039545     8007.
## 6 Poland  Europe     1977    70.7  34621254     9508.
```

Question 1.2

```
gapminder %>%
  filter(country %in% c("China","Canada","Poland") & year %in% 1970:1979) %>%
  select(country, gdpPercap)
```

```
## # A tibble: 6 x 2
##   country gdpPercap
##   <fct>       <dbl>
## 1 Canada     18971.
## 2 Canada     22091.
## 3 China        677.
## 4 China        741.
## 5 Poland      8007.
## 6 Poland      9508.
```

Question 1.3

```
gapminder %>%
  mutate(increase = c(NA, diff(lifeExp))) %>%
  filter(increase < 0) %>%
  head(6)
```

```
## # A tibble: 6 x 7
##   country   continent  year lifeExp     pop gdpPercap increase
##   <fct>     <fct>     <int>   <dbl>   <int>     <dbl>    <dbl>
## 1 Albania   Europe     1992    71.6 3326498     2497.   -0.419
## 2 Algeria   Africa     1952    43.1 9279525     2449.  -33.3
## 3 Angola    Africa     1952    30.0 4232095     3521.  -42.3
## 4 Angola    Africa     1987    39.9 7874230     2430.   -0.036
## 5 Australia Oceania    1952    69.1 8691212    10040.   -6.20
## 6 Austria   Europe     1952    66.8 6927772     6137.  -14.4
```

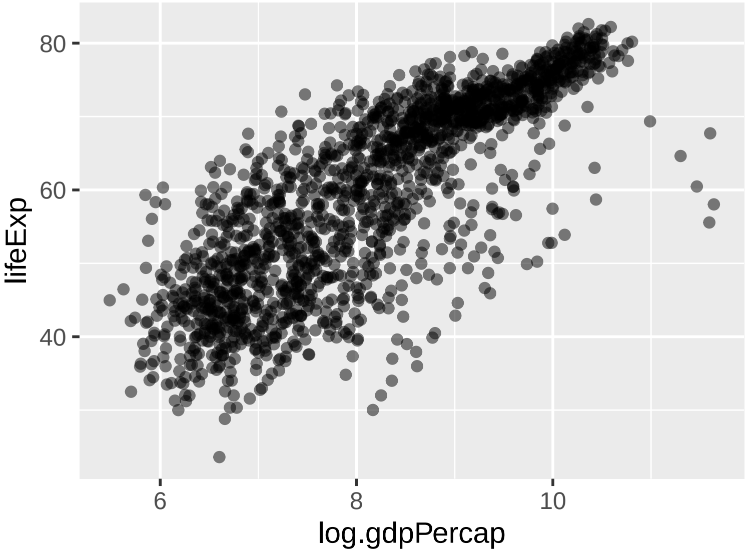
Question 1.4

```
gapminder %>%
  group_by(country) %>%
  filter(gdpPercap == max(gdpPercap)) %>%
  head(6)
```

```
## # A tibble: 6 x 6
## # Groups:   country [6]
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1982    39.9 12881816      978.
## 2 Albania     Europe     2007    76.4  3600523     5937.
## 3 Algeria     Africa     2007    72.3 33333216     6223.
## 4 Angola      Africa     1967    36.0  5247469     5523.
## 5 Argentina   Americas   2007    75.3 40301927    12779.
## 6 Australia   Oceania    2007    81.2 20434176    34435.
```

Question 1.5

```
gapminder %>%
  transmute(log.gdpPercap = log(gdpPercap), lifeExp) %>%
  ggplot(aes(log.gdpPercap,lifeExp)) + geom_point(alpha = 0.5)
```



Question 2

`continent` is a categorical variable and `lifeExp` is a quantitative variable, which are both from dataset `gapminder`.

What are possible values (or range, whichever is appropriate) of each variable?

```
gapminder %>%
  select(continent, lifeExp) %>%
  summary()
```

```
##     continent      lifeExp
##  Africa  :624   Min.   :23.60
##  Americas:300   1st Qu.:48.20
##  Asia    :396   Median :60.71
##  Europe  :360   Mean   :59.47
##  Oceania : 24   3rd Qu.:70.85
##                 Max.   :82.60
```

The summary of these two variable indicates that `continent` can only take values in “Africa”, “Americas”, “Asia”, “Europe”, “Oceania”; and `lifeExp` only takes value from 23.60 to 82.60.

What values are typical? What’s the spread? What’s the distribution? Etc., tailored to the variable at hand.

- For `continent`:

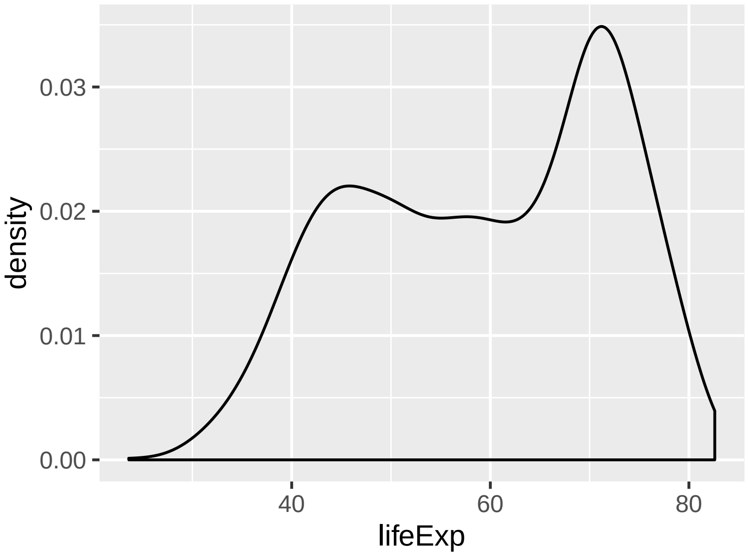
```
gapminder %>%
  select(continent) %>%
  table() %>%
  kable(col.names = c("continent","Freq"))
```

continent	Freq
Africa	624
Americas	300
Asia	396
Europe	360
Oceania	24

From the contingency table of `continent` above, “Africa” appears to be the most frequent (624 times), followed by “Asia”, " Europe“,”Americas“. And”Oceania" has the lowest frequency of only 24 times.

- For `lifeExp`:

```
gapminder %>%
  ggplot(aes(lifeExp)) + geom_density()
```

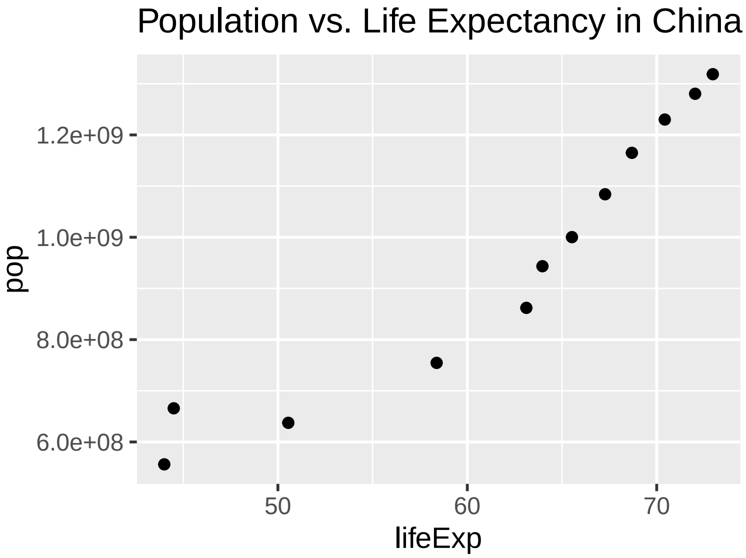


There are two peaks in the density plot of `lifeExp` above, i.e. it follows a bimodal distribution. The two peaks are arround 45 and 72 where the right one is higher.

Question 3

Population vs. Life Expectancy in China

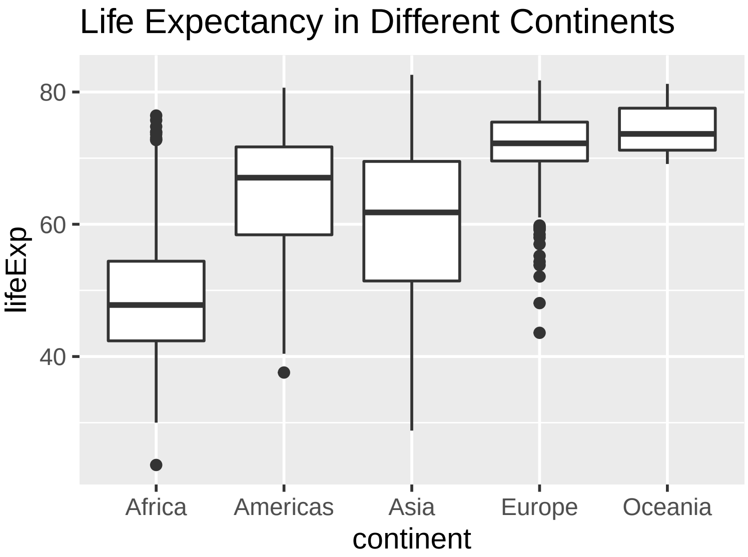
```
gapminder %>%
  filter(country == "China") %>%
  ggplot(aes(lifeExp, pop)) +
  labs(title = "Population vs. Life Expectancy in China") +
  geom_point()
```



From the scatter plot above, it seems that population and life expectancy have some sort of positive linear relationship, espectially when life expectancy is larger than 60.

Life Expectancy in Different Continents

```
gapminder %>%
  ggplot(aes(x = continent, y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Life Expectancy in Different Continents")
```



From the boxplot above, it seems that Oceania and Europe have relatively high life expectancy, followed by Americas, Asia and Africa. In addition, Asia has the largest variance based on the length of the box.