Exploring Datasets (HW4 - STAT 545)

Phuong (Sam) Can

08/10/2019

Contents

Univariate Data Reshaping

Let's first load the `gapminder` dataset.

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(tsibble)
library(gapminder)
attach(gapminder)
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

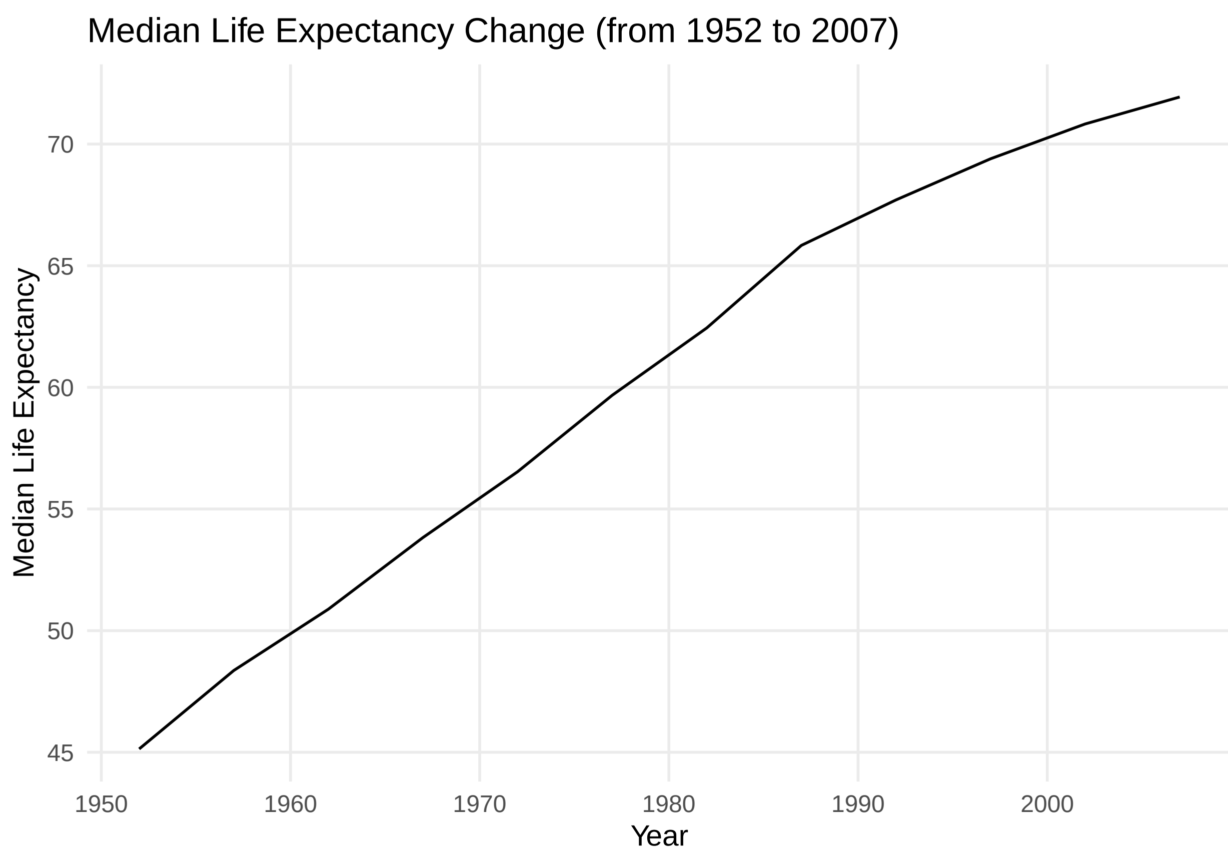
Simply by examining three variables in this dataset, `continent`, `year`, and `lifeExp`, we can answer many questions, including:

1. What is the minimum, maximum, mean, median life expectancy in the world within each year from 1952 to 2007?

To answer this question, we will create a new tibble called `gap_lifeExp_world` that displays these numbers and plot them in a line graph as following:

```
gap_lifeExp_world <- gapminder %>%
  group_by(year) %>%
  mutate(min_lifeExp = min(lifeExp)) %>%
  mutate(max_lifeExp = max(lifeExp)) %>%
  mutate(mean_lifeExp = mean(lifeExp)) %>%
  mutate(med_lifeExp = median(lifeExp)) %>%
  select(-c(continent, country, pop, gdpPercap))

ggplot(gap_lifeExp_world, aes(year, med_lifeExp)) +
  geom_line() +
  labs(title="Median Life Expectancy Change (from 1952 to 2007)", x="Year", y="Median Life Expectancy")
  theme_minimal() +
  theme(panel.grid.minor = element_blank())
```

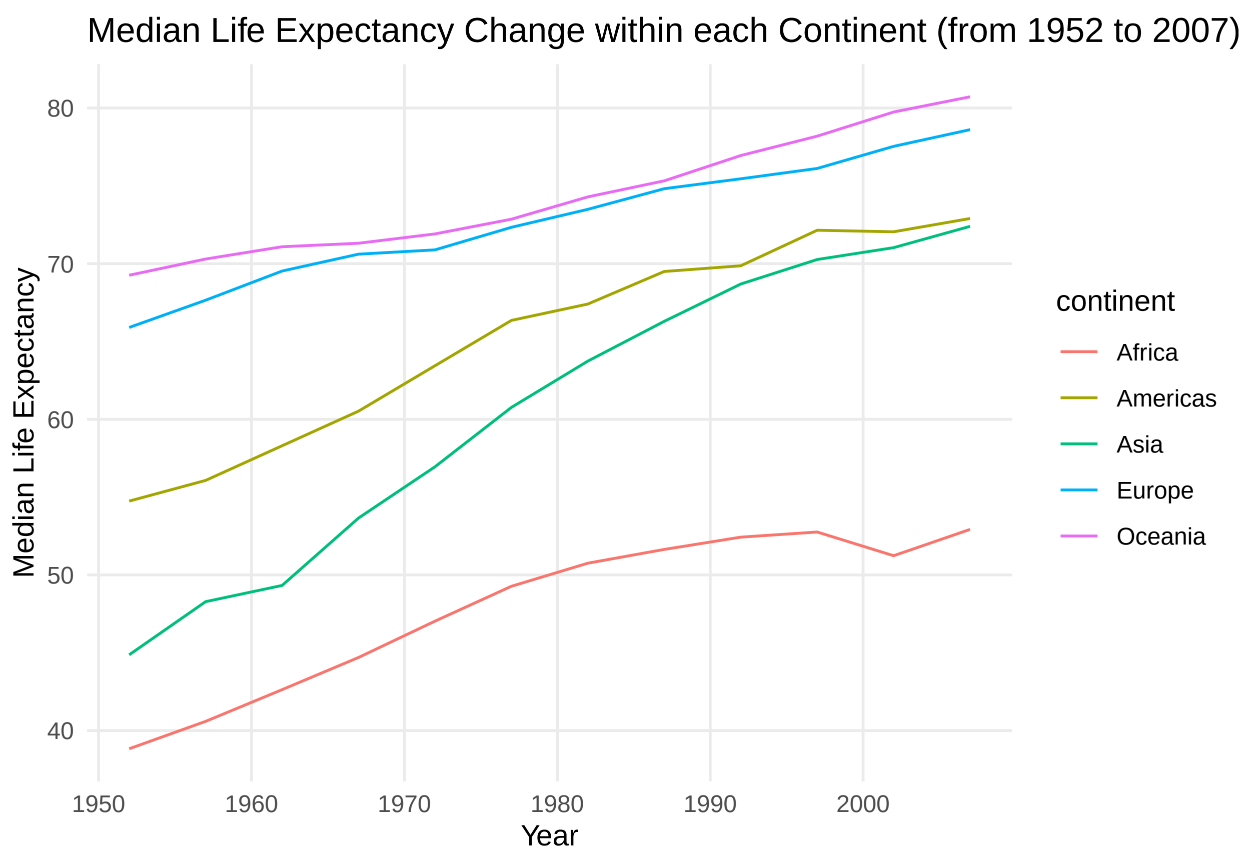


2. What is the minimum, maximum, mean, median life expectancy within each year within each continent from 1952 to 2007?

To answer this question, we will create a new table called `gap_lifeExp_continent_year` that showcases these statistics and plot them in a line graph as following:

```
gap_lifeExp_continent_year <- gapminder %>%
  group_by(year, continent) %>%
  mutate(min_lifeExp = min(lifeExp[continent == continent])) %>%
  mutate(max_lifeExp = max(lifeExp[continent == continent])) %>%
  mutate(mean_lifeExp = mean(lifeExp[continent == continent])) %>%
  mutate(med_lifeExp = median(lifeExp[continent == continent])) %>%
  select(-c(country, pop, gdpPercap))

ggplot(gap_lifeExp_continent_year, aes(year, med_lifeExp, color = continent)) +
  geom_line() +
  labs(title="Median Life Expectancy Change within each Continent (from 1952 to 2007)", x="Year", y="Med
  theme_minimal() +
  theme(panel.grid.minor = element_blank())
```

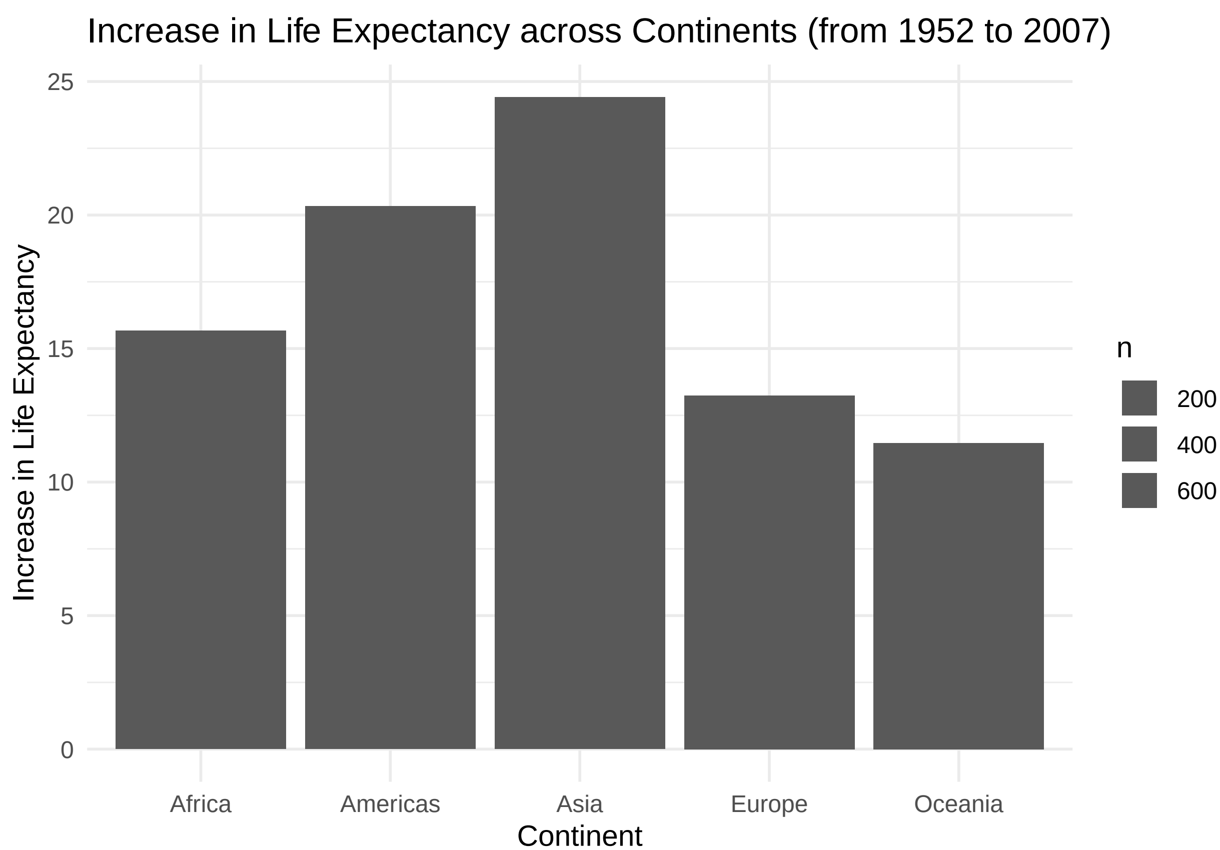


3. What is the increase in life expectancy (maximum minus minimum) within each continent throughout the years from 1952 to 2007?

We will again create a new tibble called `gap_lifeExp_continent_through` with all relevant statistics, then plot them in a bar graph to show the magnitude of the change that each continent has made in life expectancy throughout the years.

```
gap_lifeExp_continent_through <- gapminder %>%
  group_by(country) %>%
  mutate(lifeExp_inc = diff(lifeExp, lag = 11)) %>%
  group_by(continent) %>%
  mutate(n_countries = n_distinct(country)) %>%
  mutate(lifeExp_inc = mean(lifeExp_inc))

ggplot(gap_lifeExp_continent_through, aes(continent, lifeExp_inc)) +
  geom_bar(stat = "sum") +
  scale_fill_grey() +
  theme(legend.position = "none", panel.grid = element_blank()) +
  theme_minimal() +
  labs(title="Increase in Life Expectancy across Continents (from 1952 to 2007)", x="Continent", y="Inc
```



We can also try to make a sub-dataset of gapminder wider to see if we can make plots easier that way. Let's take the `gap_lifeExp_continent_year` tibble that we created before and select only the median life expectancy within each continent.

```
gap_lifeExp_continent_year
```

```
## # A tibble: 1,704 x 7
## # Groups:   year, continent [60]
##    continent  year lifeExp min_lifeExp max_lifeExp mean_lifeExp med_lifeExp
##    <fct>     <int>   <dbl>       <dbl>       <dbl>        <dbl>       <dbl>
##  1 Asia       1952    28.8        28.8        65.4         46.3        44.9
##  2 Asia       1957    30.3        30.3        67.8         49.3        48.3
##  3 Asia       1962    32.0        32.0        69.4         51.6        49.3
##  4 Asia       1967    34.0        34.0        71.4         54.7        53.7
##  5 Asia       1972    36.1        36.1        73.4         57.3        57.0
##  6 Asia       1977    38.4        31.2        75.4         59.6        60.8
##  7 Asia       1982    39.9        39.9        77.1         62.6        63.7
##  8 Asia       1987    40.8        40.8        78.7         64.9        66.3
##  9 Asia       1992    41.7        41.7        79.4         66.5        68.7
## 10 Asia       1997    41.8        41.8        80.7         68.0        70.3
## # ... with 1,694 more rows
```

```
gap_lifeExp_continent_year <-
  gap_lifeExp_continent_year %>%
  distinct(continent, year, med_lifeExp)
(gap_lifeExp_continent_year <-
  gap_lifeExp_continent_year %>%
  pivot_wider(id_cols = year,
            names_from = continent,
            values_from = med_lifeExp))
```

```
## # A tibble: 12 x 6
## # Groups:   year [12]
##     year  Asia Europe Africa Americas Oceania
##    <int> <dbl>  <dbl>  <dbl>    <dbl>   <dbl>
##  1  1952  44.9   65.9   38.8     54.7    69.3
##  2  1957  48.3   67.6   40.6     56.1    70.3
##  3  1962  49.3   69.5   42.6     58.3    71.1
##  4  1967  53.7   70.6   44.7     60.5    71.3
##  5  1972  57.0   70.9   47.0     63.4    71.9
##  6  1977  60.8   72.3   49.3     66.4    72.9
##  7  1982  63.7   73.5   50.8     67.4    74.3
##  8  1987  66.3   74.8   51.6     69.5    75.3
##  9  1992  68.7   75.5   52.4     69.9    76.9
## 10  1997  70.3   76.1   52.8     72.1    78.2
## 11  2002  71.0   77.5   51.2     72.0    79.7
## 12  2007  72.4   78.6   52.9     72.9    80.7
```

Let's try plotting this tibble. We will graph life expectancy in 2007 across continents.

```
## # A tibble: 1 x 6
## # Groups:   year [1]
##    year  Asia Europe Africa Americas Oceania
##   <int> <dbl>  <dbl>  <dbl>    <dbl>   <dbl>
## 1  2007  72.4   78.6   52.9     72.9    80.7
```

Running the commented codes above would give errors when we try to put multiple columns that represent `continent` in the `x` argument, and their values in the `y` argument for ggplot aesthetics. In fact, there is no way to knit such a plot if we keep using a wide tibble.

Solution: re-lengthening the tibble as following.

```
(gap_lifeExp_continent_year %>%
  pivot_longer(cols = -year,
               names_to = "continent",
               values_to = "med_lifeExp"))
```

```
## # A tibble: 5 x 3
## # Groups:   year [1]
##    year continent med_lifeExp
##   <int> <chr>           <dbl>
## 1  2007 Asia             72.4
## 2  2007 Europe           78.6
## 3  2007 Africa           52.9
## 4  2007 Americas         72.9
## 5  2007 Oceania          80.7
```

Multivariate Data Reshaping

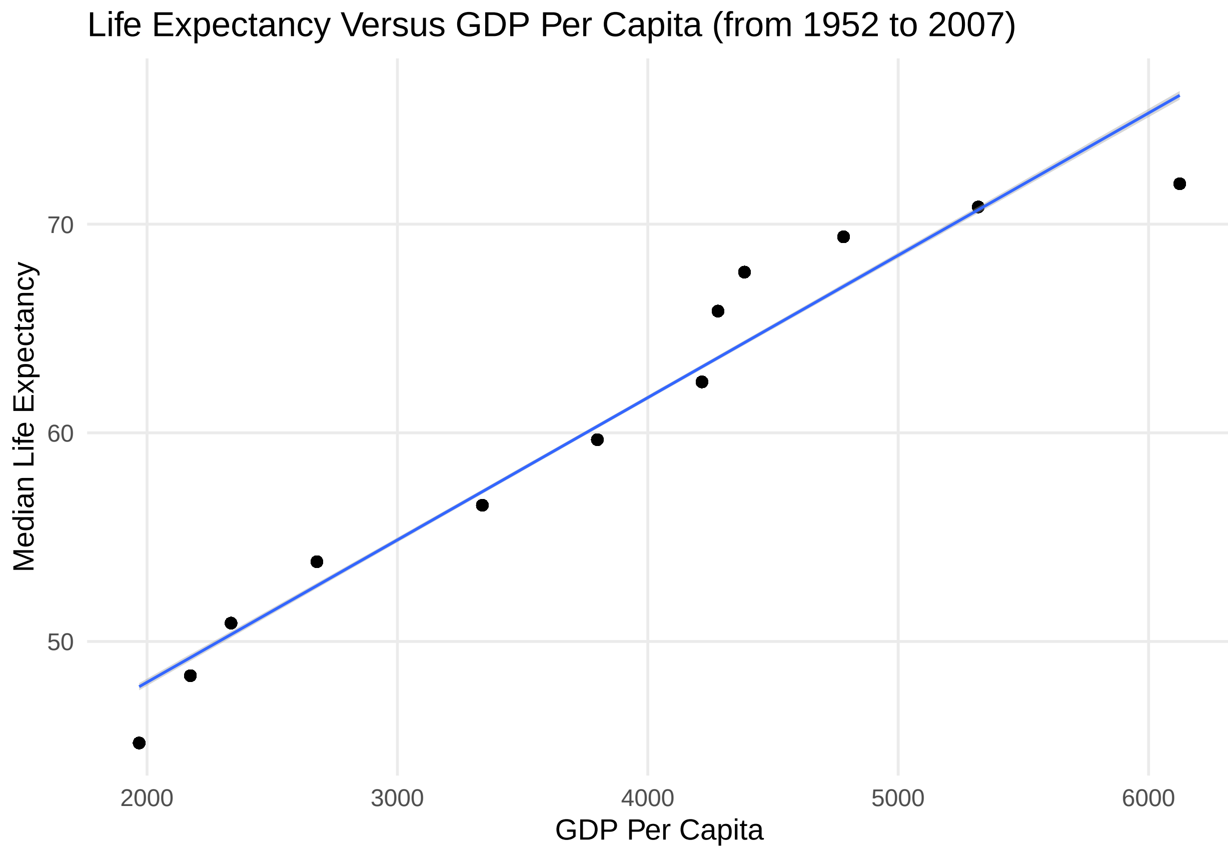
Now, let's look at four variables: `continent`, `year`, `lifeExp`, and `gdpPercap` and tackle some questions related to them.

1. What is the minimum, maximum, mean, median life expectancy and GDP per capita in the world within each year from 1952 to 2007?

To answer this question, we will create a new tibble called `gap_lifeExp_gdp_world` that displays these numbers and plot the relationship between the median life expectancy and median GDP per capita in a scatterplot (with a regression line) as following:

```
gap_lifeExp_gdp_world <- gapminder %>%
  group_by(year) %>%
  mutate(min_lifeExp = min(lifeExp)) %>%
  mutate(min_gdpPercap = min(gdpPercap)) %>%
  mutate(max_lifeExp = max(lifeExp)) %>%
  mutate(max_gdpPercap = max(gdpPercap)) %>%
  mutate(mean_lifeExp = mean(lifeExp)) %>%
  mutate(mean_gdpPercap = mean(gdpPercap)) %>%
  mutate(med_lifeExp = median(lifeExp)) %>%
  mutate(med_gdpPercap = median(gdpPercap)) %>%
  select(-c(continent, country, pop))

ggplot(gap_lifeExp_gdp_world, aes(med_gdpPercap, med_lifeExp)) +
  geom_point() +
  geom_smooth(method='lm',formula=y~x, size = 0.5) +
  labs(title="Life Expectancy Versus GDP Per Capita (from 1952 to 2007)", x="GDP Per Capita", y="Median
  theme_minimal() +
  theme(panel.grid.minor = element_blank())
```



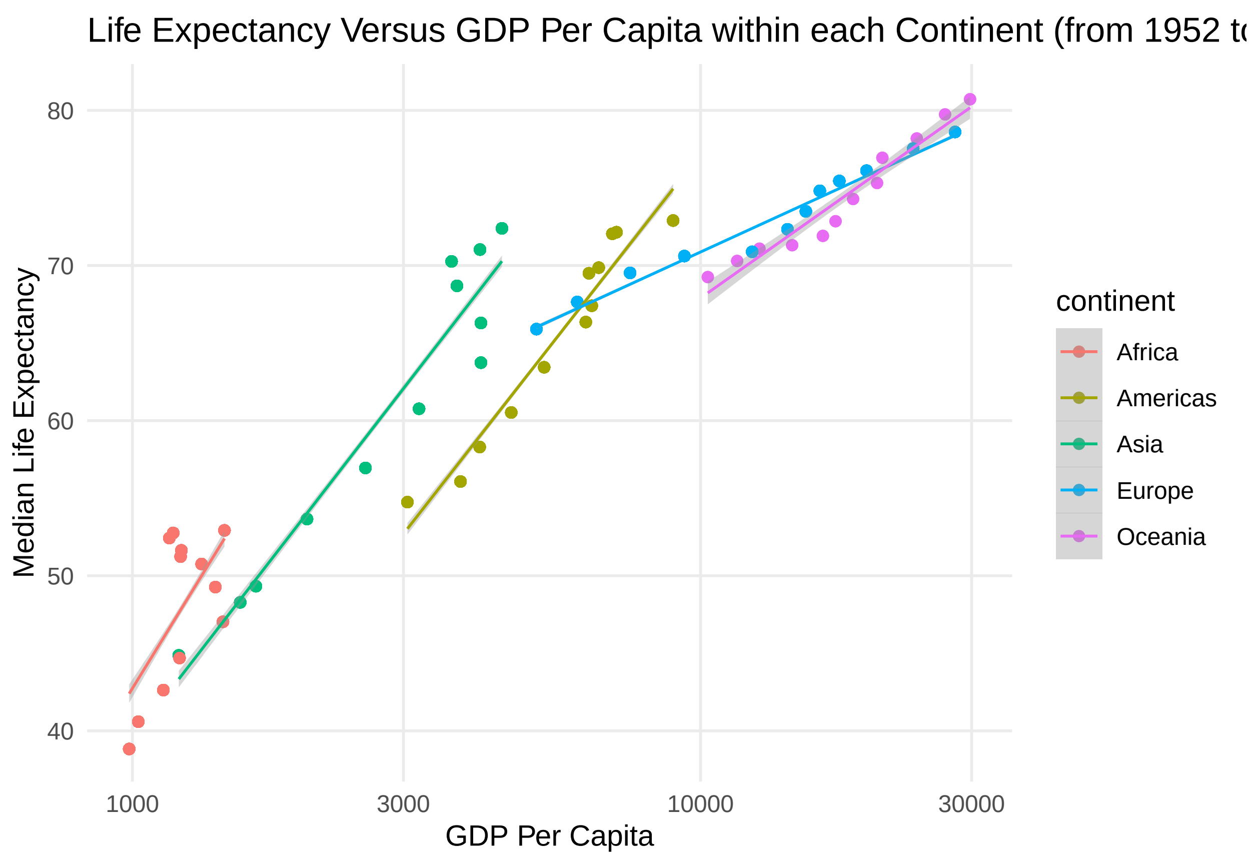
Life Expectancy Versus GDP Per Capita (from 1952 to 2007)
Median Life Expectancy
70
60
50
2000
3000
4000
5000
6000
GDP Per Capita

2. What is the minimum, maximum, mean, median life expectancy and GDP per capita within each year within each continent from 1952 to 2007?

To answer this question, we will create a new table called `gap_lifeExp_gdp_continent_year` that showcases these statistics and plot them in a line graph as following:

```
gap_lifeExp_gdp_continent_year <- gapminder %>%
  group_by(year, continent) %>%
  mutate(min_lifeExp = min(lifeExp[continent == continent])) %>%
  mutate(min_gdpPercap = min(gdpPercap[continent == continent])) %>%
  mutate(max_lifeExp = max(lifeExp[continent == continent])) %>%
  mutate(max_gdpPercap = max(gdpPercap[continent == continent])) %>%
  mutate(mean_lifeExp = mean(lifeExp[continent == continent])) %>%
  mutate(mean_gdpPercap = mean(gdpPercap[continent == continent])) %>%
  mutate(med_lifeExp = median(lifeExp[continent == continent])) %>%
  mutate(med_gdpPercap = median(gdpPercap[continent == continent])) %>%
  select(-c(country, pop))

ggplot(gap_lifeExp_gdp_continent_year, aes(med_gdpPercap, med_lifeExp, color = continent)) +
  geom_point() +
  geom_smooth(method='lm',formula=y~x, size = 0.5) +
  labs(title="Life Expectancy Versus GDP Per Capita within each Continent (from 1952 to 2007)", x="GDP 
  theme_minimal() +
  scale_x_log10() +
  theme(panel.grid.minor = element_blank())
```

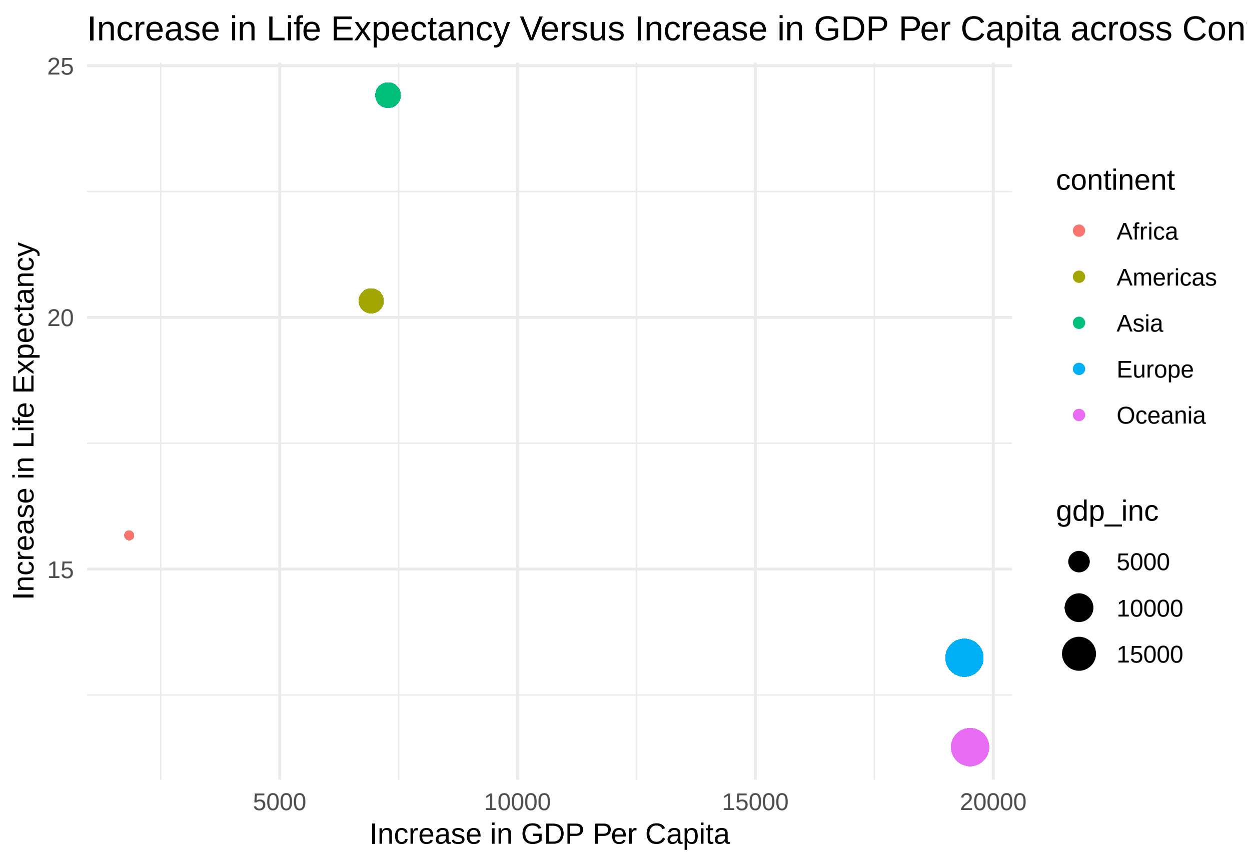


3. What is the increase in life expectancy (maximum minus minimum) within each continent throughout the years from 1952 to 2007?

We will again create a new tibble called `gap_lifeExp_gdp_continent_through` with all relevant statistics, then plot the increase in life expectancy versus increase in GDP per capita across continents, with size scales based on the increase in GDP per capita each continent has.

```
gap_lifeExp_gdp_continent_through <- gapminder %>%
  group_by(country) %>%
  mutate(lifeExp_inc = diff(lifeExp, lag = 11)) %>%
  mutate(gdp_inc = diff(gdpPercap, lag = 11)) %>%
  group_by(continent) %>%
  mutate(n_countries = n_distinct(country)) %>%
  mutate(lifeExp_inc = mean(lifeExp_inc)) %>%
  mutate(gdp_inc = mean(gdp_inc))

ggplot(gap_lifeExp_gdp_continent_through, aes(gdp_inc, lifeExp_inc, color = continent)) +
  geom_point(aes(size = gdp_inc)) +
  theme_minimal() +
  labs(title="Increase in Life Expectancy Versus Increase in GDP Per Capita across Continents", x="Incre
```



Again, we can make another subset of gapminder wider to see if we can make plots easier that way. Let's take the `gap_lifeExp_gdp_continent_year` tibble that we created before and select only the median life expectancy and median GDP per capita within each continent.

```r
gap_lifeExp_gdp_continent_year
```

```
## # A tibble: 1,704 x 12
## # Groups:   year, continent [60]
##    continent  year lifeExp gdpPercap min_lifeExp min_gdpPercap max_lifeExp
##    <fct>     <int>   <dbl>     <dbl>       <dbl>         <dbl>       <dbl>
##  1 Asia       1952    28.8      779.        28.8           331        65.4
##  2 Asia       1957    30.3      821.        30.3           350        67.8
##  3 Asia       1962    32.0      853.        32.0           388        69.4
##  4 Asia       1967    34.0      836.        34.0           349        71.4
##  5 Asia       1972    36.1      740.        36.1           357        73.4
##  6 Asia       1977    38.4      786.        31.2           371        75.4
##  7 Asia       1982    39.9      978.        39.9           424        77.1
##  8 Asia       1987    40.8      852.        40.8           385        78.7
##  9 Asia       1992    41.7      649.        41.7           347        79.4
## 10 Asia       1997    41.8      635.        41.8           415        80.7
## # ... with 1,694 more rows, and 5 more variables: max_gdpPercap <dbl>,
## #   mean_lifeExp <dbl>, mean_gdpPercap <dbl>, med_lifeExp <dbl>,
## #   med_gdpPercap <dbl>
```

```r
gap_lifeExp_gdp_continent_year <-
  gap_lifeExp_gdp_continent_year %>%
  distinct(continent, year, med_lifeExp, med_gdpPercap)
gap_lifeExp_gdp_continent_year
```

```
## # A tibble: 60 x 4
## # Groups:   year, continent [60]
##    continent  year med_lifeExp med_gdpPercap
##    <fct>     <int>       <dbl>         <dbl>
##  1 Asia       1952        44.9         1207.
##  2 Asia       1957        48.3         1548.
##  3 Asia       1962        49.3         1650.
##  4 Asia       1967        53.7         2029.
##  5 Asia       1972        57.0         2571.
##  6 Asia       1977        60.8         3195.
##  7 Asia       1982        63.7         4107.
##  8 Asia       1987        66.3         4106.
##  9 Asia       1992        68.7         3726.
## 10 Asia       1997        70.3         3645.
## # ... with 50 more rows
```

```r
(gap_lifeExp_gdp_continent_year <-
  gap_lifeExp_gdp_continent_year %>%
  pivot_wider(id_cols = year,
              names_from = continent,
              names_sep = "-",
              values_from = c(med_lifeExp, med_gdpPercap)))
```

```
## # A tibble: 12 x 11
## # Groups:   year [12]
##     year `med_lifeExp-As~ `med_lifeExp-Eu~ `med_lifeExp-Af~
##    <int>            <dbl>            <dbl>            <dbl>
```

```
## 1  1952              44.9             65.9             38.8
## 2  1957              48.3             67.6             40.6
## 3  1962              49.3             69.5             42.6
## 4  1967              53.7             70.6             44.7
## 5  1972              57.0             70.9             47.0
## 6  1977              60.8             72.3             49.3
## 7  1982              63.7             73.5             50.8
## 8  1987              66.3             74.8             51.6
## 9  1992              68.7             75.5             52.4
## 10 1997              70.3             76.1             52.8
## 11 2002              71.0             77.5             51.2
## 12 2007              72.4             78.6             52.9
## # ... with 7 more variables: `med_lifeExp-Americas` <dbl>,
## #   `med_lifeExp-Oceania` <dbl>, `med_gdpPercap-Asia` <dbl>,
## #   `med_gdpPercap-Europe` <dbl>, `med_gdpPercap-Africa` <dbl>,
## #   `med_gdpPercap-Americas` <dbl>, `med_gdpPercap-Oceania` <dbl>
```

Let's try plotting this tibble. We will graph life expectancy versus GDP per capita in 2007 across continents.

```
knitr::opts_chunk$set(error = TRUE)
gap_lifeExp_gdp_continent_year07 <-
  gap_lifeExp_gdp_continent_year %>%
  filter(year == 2007)
# ggplot(gap_lifeExp_continent_year07,  aes(med_lifeExp-Asia:med_lifeExp-Oceania, med_gdpPercap-Asia:me
#  geom_point()
```

Running the commented codes above would give errors when we try to put multiple cells in each of the `x` and `y` arguments for ggplot aesthetics. In fact, there is no way to knit such a plot if we keep using a wide tibble.

Solution: re-lengthening the tibble as following.

```
gap_lifeExp_gdp_continent_year <-
  gap_lifeExp_gdp_continent_year %>%
  pivot_longer(cols = -year,
               names_to = c("med", "continent"),
               names_sep = "-",
               values_to = c("value"))
(gap_lifeExp_gdp_continent_year %>%
  pivot_wider(id_cols = c(year, continent),
               names_from = c("med"),
               values_from = c("value")))
```

```
## # A tibble: 60 x 4
## # Groups:   year [12]
##     year continent med_lifeExp med_gdpPercap
##    <int> <chr>           <dbl>         <dbl>
## 1   1952 Asia             44.9         1207.
## 2   1952 Europe           65.9         5142.
## 3   1952 Africa           38.8          987.
## 4   1952 Americas         54.7         3048.
## 5   1952 Oceania          69.3        10298.
## 6   1957 Asia             48.3         1548.
## 7   1957 Europe           67.6         6067.
## 8   1957 Africa           40.6         1024.
## 9   1957 Americas         56.1         3781.
## 10  1957 Oceania          70.3        11599.
```

```
## # ... with 50 more rows
```

Table Joins

```
guest <- read_csv("https://raw.githubusercontent.com/STAT545-UBC/Classroom/master/data/wedding/attend.c
email <- read_csv("https://raw.githubusercontent.com/STAT545-UBC/Classroom/master/data/wedding/emails.c
```

1. To add emails for participants in `guest`, we first have to separate the names in `email`, then left join with `guest` by `name` variable.

```
email <- email %>%
  separate_rows(guest, sep = ",")
(guest <- guest %>%
  rename(guest = name) %>%
  left_join(email, by = "guest"))
```

```
## # A tibble: 30 x 8
##    party guest meal_wedding meal_brunch attendance_wedd~ attendance_brun~
##    <dbl> <chr> <chr>        <chr>       <chr>            <chr>
##  1     1 Somm~ PENDING      PENDING     PENDING          PENDING
##  2     1 Phil~ vegetarian   Menu C      CONFIRMED        CONFIRMED
##  3     1 Blan~ chicken      Menu A      CONFIRMED        CONFIRMED
##  4     1 Emaa~ PENDING      PENDING     PENDING          PENDING
##  5     2 Blai~ chicken      Menu C      CONFIRMED        CONFIRMED
##  6     2 Nige~ <NA>         <NA>        CANCELLED        CANCELLED
##  7     3 Sine~ PENDING      PENDING     PENDING          PENDING
##  8     4 Ayra~ vegetarian   Menu B      PENDING          PENDING
##  9     5 Atla~ PENDING      PENDING     PENDING          PENDING
## 10     5 Denz~ fish         Menu B      CONFIRMED        CONFIRMED
## # ... with 20 more rows, and 2 more variables: attendance_golf <chr>,
## #   email <chr>
```

2. To find out the ones we have emails for but are not on the guestlist, we can have `guest` anti join with the previously left-joined `email` dataset by `email` variable.

```
(email %>%
  anti_join(guest, by = "email"))
```

```
## # A tibble: 3 x 2
##   guest              email
##   <chr>              <chr>
## 1 Turner Jones       tjjones12@hotmail.ca
## 2 Albert Marshall    themarshallfamily1234@gmail.com
## 3 " Vivian Marshall" themarshallfamily1234@gmail.com
```

3. To include everyone we have emails for, we full join the left-joined guestlist with the original `email` dataset by `email` variable.

```
(guest %>%
  full_join(email, by = "email"))
```

```
## # A tibble: 46 x 9
##    party guest.x meal_wedding meal_brunch attendance_wedd~ attendance_brun~
##    <dbl> <chr>   <chr>        <chr>       <chr>            <chr>
##  1     1 Sommer~ PENDING      PENDING     PENDING          PENDING
##  2     1 Sommer~ PENDING      PENDING     PENDING          PENDING
##  3     1 Sommer~ PENDING      PENDING     PENDING          PENDING
##  4     1 Sommer~ PENDING      PENDING     PENDING          PENDING
##  5     1 Philli~ vegetarian   Menu C      CONFIRMED        CONFIRMED
```

```
## 6      1 Blanka~ chicken      Menu A      CONFIRMED        CONFIRMED
## 7      1 Emaan ~ PENDING      PENDING     PENDING          PENDING
## 8      2 Blair ~ chicken      Menu C      CONFIRMED        CONFIRMED
## 9      2 Blair ~ chicken      Menu C      CONFIRMED        CONFIRMED
## 10     2 Nigel ~ <NA>         <NA>        CANCELLED        CANCELLED
## # ... with 36 more rows, and 3 more variables: attendance_golf <chr>,
## #   email <chr>, guest.y <chr>
```