Homework 02

Kimberly Sharpe

Exercise 1: Basic dplyr

1.1 subsetting gapminder to three countries in the 1970s

```
gapminder %>%
  filter(year > 1969 & year < 1980,
        country == "Afghanistan" |
        country == "Canada" |
        country == "Denmark")
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1972    36.1 13079460      740.
## 2 Afghanistan Asia       1977    38.4 14880372      786.
## 3 Canada      Americas   1972    72.9 22284500    18971.
## 4 Canada      Americas   1977    74.2 23796400    22091.
## 5 Denmark     Europe     1972    73.5  4991596    18866.
## 6 Denmark     Europe     1977    74.7  5088419    20423.
```

1.2 using a pipe operator to select country & gdpPercap from filtered dataset

GDP per capita in the 1970s

```
gapminder %>%
  filter(year > 1969 & year < 1980,
        country == "Afghanistan" |
        country == "Canada" |
        country == "Denmark") %>%
  select(country, gdpPercap, year)
```

```
## # A tibble: 6 x 3
##   country     gdpPercap  year
##   <fct>           <dbl> <int>
## 1 Afghanistan      740.  1972
## 2 Afghanistan      786.  1977
## 3 Canada         18971.  1972
## 4 Canada         22091.  1977
## 5 Denmark        18866.  1972
## 6 Denmark        20423.  1977
```

1.3 filtering gapminder to all entries that have experienced a drop in life expectancy

```
gapminder %>%
  arrange(year) %>%
  group_by(country) %>%
  mutate(diff_LifeExp = lifeExp - lag(lifeExp)) %>%
  filter(diff_LifeExp < 0) %>%
  arrange(diff_LifeExp)
```

```
## # A tibble: 102 x 7
## # Groups:   country [52]
##    country      continent  year lifeExp       pop gdpPercap diff_LifeExp
##    <fct>        <fct>     <int>   <dbl>     <int>     <dbl>        <dbl>
##  1 Rwanda       Africa     1992    23.6   7290203      737.       -20.4 
##  2 Zimbabwe     Africa     1997    46.8  11404948      792.       -13.6 
##  3 Lesotho      Africa     2002    44.6   2046772     1275.       -11.0 
##  4 Swaziland    Africa     2002    43.9   1130269     4128.       -10.4 
##  5 Botswana     Africa     1997    52.6   1536536     8647.       -10.2 
##  6 Cambodia     Asia       1977    31.2   6978607      525.        -9.10
##  7 Namibia      Africa     2002    51.5   1972153     4072.        -7.43
##  8 South Africa Africa     2002    53.4  44433622     7711.        -6.87
##  9 Zimbabwe     Africa     2002    40.0  11926563      672.        -6.82
## 10 China        Asia       1962    44.5 665770000      488.        -6.05
## # ... with 92 more rows
```

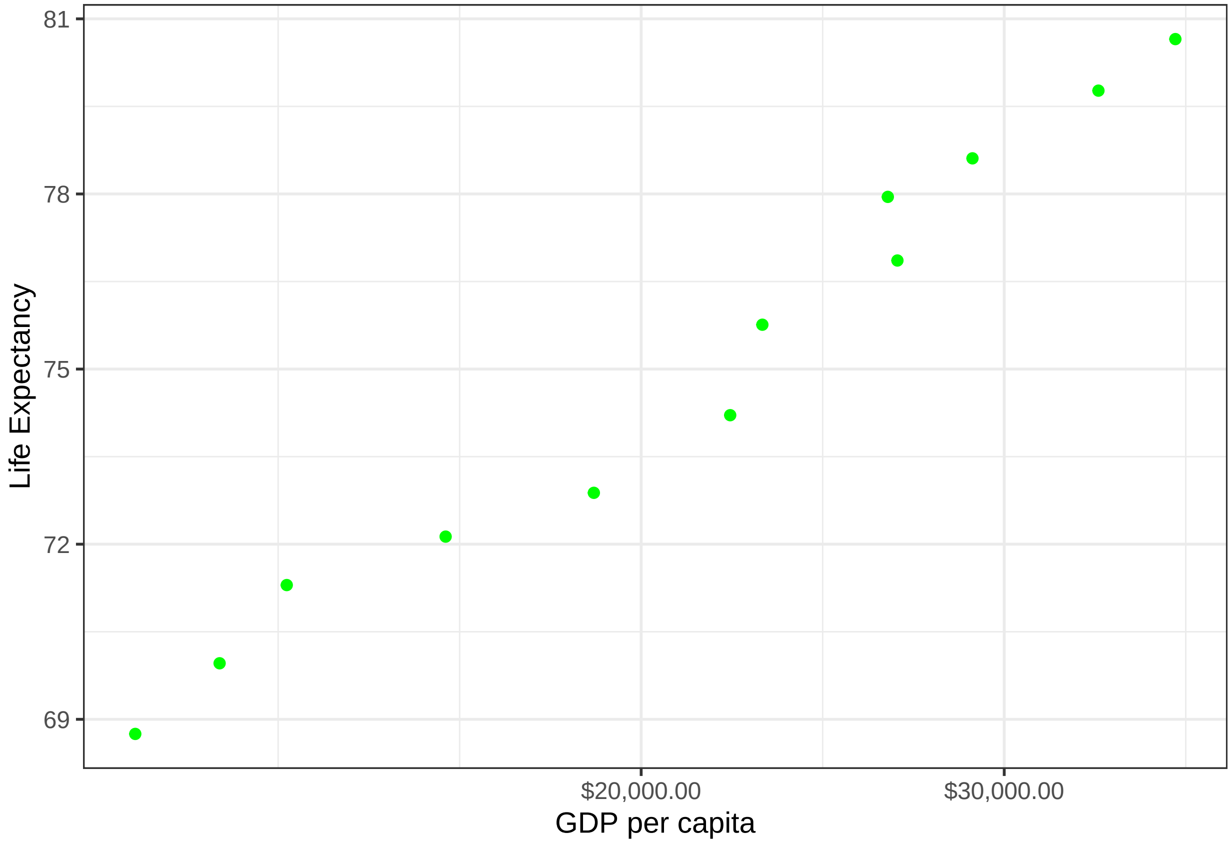
1.4 showing max GDP per capita experienced by each country

```
gapminder %>%
  select(country, year, gdpPercap) %>%
  group_by(country) %>%
  filter(gdpPercap == max(gdpPercap))
```

```
## # A tibble: 142 x 3
## # Groups:   country [142]
##    country      year gdpPercap
##    <fct>       <int>     <dbl>
##  1 Afghanistan  1982      978.
##  2 Albania      2007     5937.
##  3 Algeria      2007     6223.
##  4 Angola       1967     5523.
##  5 Argentina    2007    12779.
##  6 Australia    2007    34435.
##  7 Austria      2007    36126.
##  8 Bahrain      2007    29796.
##  9 Bangladesh   2007     1391.
## 10 Belgium      2007    33693.
## # ... with 132 more rows
```

1.5 producing a scatterplot of Canada's life expectancy vs GDP

```
gapminder %>%
  filter(country == "Canada") %>%
  ggplot(aes(gdpPercap, lifeExp)) +
  geom_point(color = "green") +
  scale_x_log10("GDP per capita", labels = scales::dollar_format()) +
  theme_bw() +
  ylab("Life Expectancy")
```



Exercise 2: Explore individual variables with dplyr

Choose one categorial and one quantitative variable: Categorical variable: continent Quantative variable: gdpPercap

What are the possible values of each variable?

Continent:

How many continents are in the dataset?

```
gapminder %>%
  select(continent) %>%
  summarize(n_unique = n_distinct(continent))
```

```
## # A tibble: 1 x 1
##   n_unique
##      <int>
## 1        5
```

What continents are included in the dataset?

```
gapminder %>%
  group_by(continent) %>%
  summarize(n_unique = n_distinct(continent))
```

```
## # A tibble: 5 x 2
##   continent n_unique
##   <fct>        <int>
## 1 Africa           1
## 2 Americas         1
## 3 Asia             1
## 4 Europe           1
## 5 Oceania          1
```

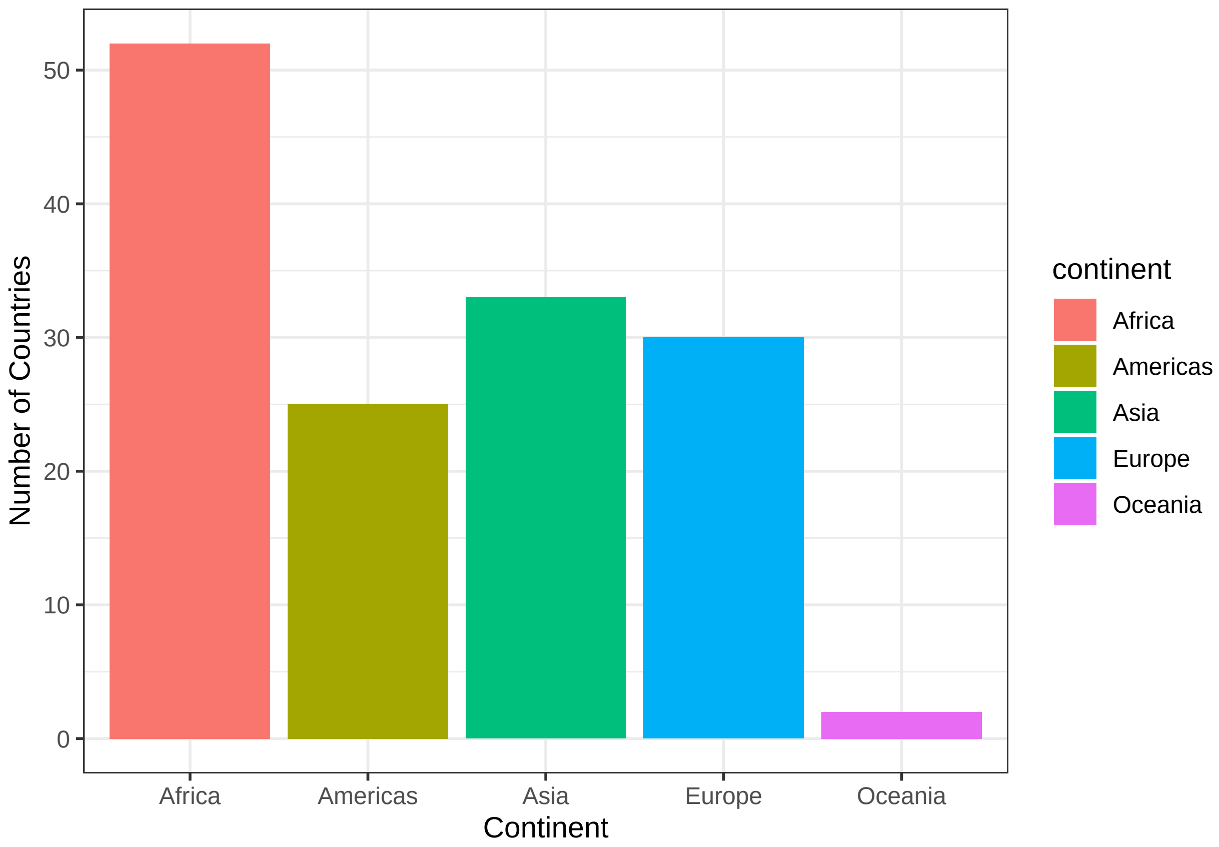
How many countries are in each continent?

```
gapminder %>%
  group_by(continent) %>%
  count(n_distinct(country))
```

```
## # A tibble: 5 x 3
## # Groups:   continent [5]
##   continent `n_distinct(country)`     n
##   <fct>                       <int> <int>
## 1 Africa                         52   624
## 2 Americas                       25   300
## 3 Asia                           33   396
## 4 Europe                         30   360
## 5 Oceania                         2    24
```

Can I visualize this in a graph? Note: divide by 12 because there are 12 time points per country

```
ggplot(gapminder) +
  geom_bar(aes(continent, ..count../12, fill=continent)) +
  xlab("Continent") +
  ylab("Number of Countries") +
  theme_bw()
```



GDP per capita

Let's get run summary to find the range, median and mean for GDP per capita

```
gapminder %>%
  select(gdpPercap) %>%
  summary()
```

```
##    gdpPercap
##  Min.   :   241.2
##  1st Qu.:  1202.1
##  Median :  3531.8
##  Mean   :  7215.3
##  3rd Qu.:  9325.5
##  Max.   :113523.1
```

Which country had the lowest GDP per capita and which country had the highest GDP per capita in this dataset?

```
gapminder %>%
  select(country, year, gdpPercap) %>%
  filter(gdpPercap == min(gdpPercap) | gdpPercap == max(gdpPercap)) %>%
    group_by(country)
```

```
## # A tibble: 2 x 3
## # Groups:   country [2]
##   country           year gdpPercap
##   <fct>            <int>     <dbl>
## 1 Congo, Dem. Rep.  2002      241.
## 2 Kuwait            1957   113523.
```

The Democratic Republic of Congo (in 2002) had the lowest GDP per capita. Kuwait (in 1957) had the highest GDP per capita.

What other countries had the lowest GDP per capita?

```r
gapminder %>%
  group_by(continent, year) %>%
  summarize(min_GDP = min(gdpPercap),
            country = country[gdpPercap == min_GDP]) %>%
  arrange(min_GDP)
```

```
## # A tibble: 60 x 4
## # Groups:   continent [5]
##    continent  year min_GDP country
##    <fct>     <int>   <dbl> <fct>
##  1 Africa     2002    241. Congo, Dem. Rep.
##  2 Africa     2007    278. Congo, Dem. Rep.
##  3 Africa     1952    299. Lesotho
##  4 Africa     1997    312. Congo, Dem. Rep.
##  5 Asia       1952    331  Myanmar
##  6 Africa     1957    336. Lesotho
##  7 Asia       1992    347  Myanmar
##  8 Asia       1967    349  Myanmar
##  9 Asia       1957    350  Myanmar
## 10 Africa     1962    355. Burundi
## # ... with 50 more rows
```

We can also see when each country had their lowest verses their highest GDP per capita

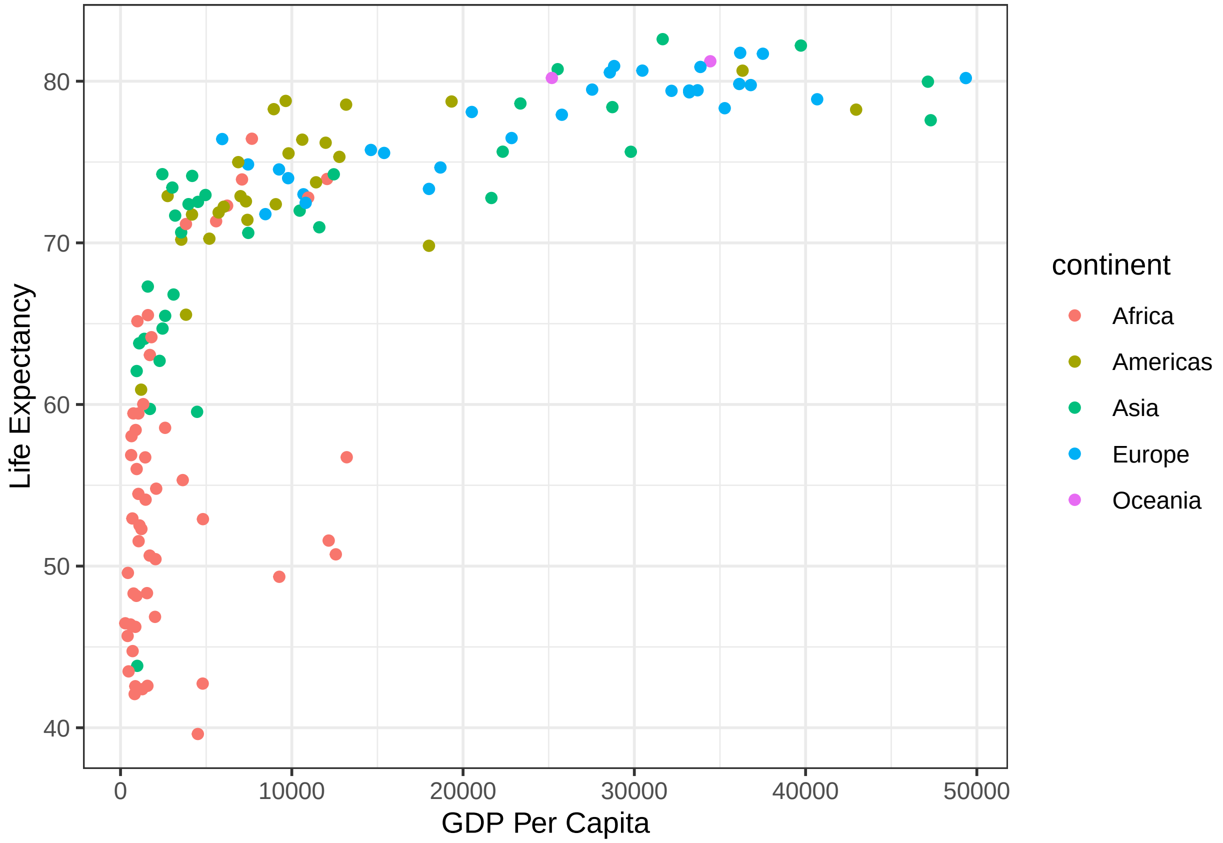
```r
gapminder %>%
  select(country, year, gdpPercap) %>%
  group_by(country) %>%
  filter(gdpPercap == min(gdpPercap) | gdpPercap == max(gdpPercap))
```

```
## # A tibble: 284 x 3
## # Groups:   country [142]
##    country      year gdpPercap
##    <fct>       <int>     <dbl>
##  1 Afghanistan  1982      978.
##  2 Afghanistan  1997      635.
##  3 Albania      1952     1601.
##  4 Albania      2007     5937.
##  5 Algeria      1952     2449.
##  6 Algeria      2007     6223.
##  7 Angola       1967     5523.
##  8 Angola       1997     2277.
##  9 Argentina    1952     5911.
## 10 Argentina    2007    12779.
## # ... with 274 more rows
```

Exercise #3: Explore various plot types

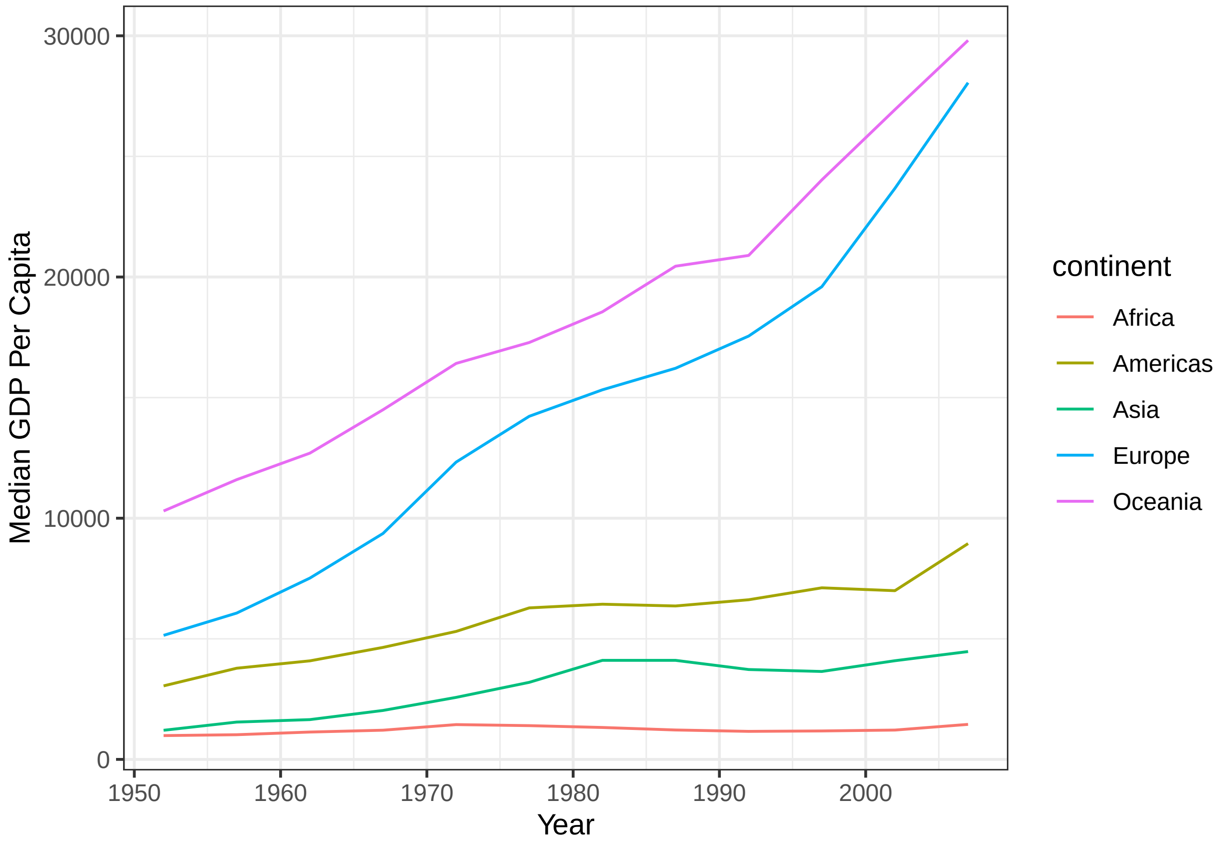
GDP by life expectancy in 2007

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(gdpPercap,lifeExp)) +
  geom_point(aes(color=continent)) +
  theme_bw() +
  labs(x="GDP Per Capita", y="Life Expectancy")
```



Median GDP per capita by continent over time

```
gapminder %>%
  group_by(year, continent) %>%
  summarize(median_GDP = median(gdpPercap)) %>%
  ggplot(aes(x=year, y=median_GDP, colour=continent)) +
  geom_line() +
  theme_bw() +
  labs(x="Year", y ="Median GDP Per Capita")
```


Median GDP Per Capita
30000
20000
10000
0
1950
1960
1970
1980
1990
2000
Year
continent
Africa
Americas
Asia
Europe
Oceania