Assignment 01

Sorif Hossain Course:AST-532

January 21, 2018

Contents

```
## Warning: package 'tidyverse' was built under R version 3.4.3

## -- Attaching packages ------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1      v purrr   0.2.4
## v tibble  1.4.1      v dplyr   0.7.4
## v tidyr   0.7.2      v stringr 1.2.0
## v readr   1.1.1      v forcats 0.2.0

## Warning: package 'tibble' was built under R version 3.4.3

## Warning: package 'tidyr' was built under R version 3.4.3

## Warning: package 'readr' was built under R version 3.4.3

## Warning: package 'purrr' was built under R version 3.4.3

## Warning: package 'dplyr' was built under R version 3.4.3

## Warning: package 'forcats' was built under R version 3.4.3

## -- Conflicts ---------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Warning: package 'gapminder' was built under R version 3.4.3
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed          dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Exercise-01: The gapminder data for just 2007.

```
dat <-gapminder

hdat <- subset(dat , subset = year==2007)
hdat
```

```
## # A tibble: 142 x 6
##    country     continent  year lifeExp       pop gdpPercap
##    <fct>       <fct>     <int>  <dbl>     <int>     <dbl>
##  1 Afghanistan Asia       2007   43.8  31889923       975
##  2 Albania     Europe     2007   76.4   3600523      5937
##  3 Algeria     Africa     2007   72.3  33333216      6223
##  4 Angola      Africa     2007   42.7  12420476      4797
##  5 Argentina   Americas   2007   75.3  40301927     12779
##  6 Australia   Oceania    2007   81.2  20434176     34435
##  7 Austria     Europe     2007   79.8   8199783     36126
##  8 Bahrain     Asia       2007   75.6    708573     29796
##  9 Bangladesh  Asia       2007   64.1 150448339      1391
## 10 Belgium     Europe     2007   79.4  10392226     33693
## # ... with 132 more rows
```
```
#number of rows :

nrow(dat) #number of rows
```

```
## [1] 1704
```
```
table(dat$continent) #number of observations per continent
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##      624      300      396      360       24
```
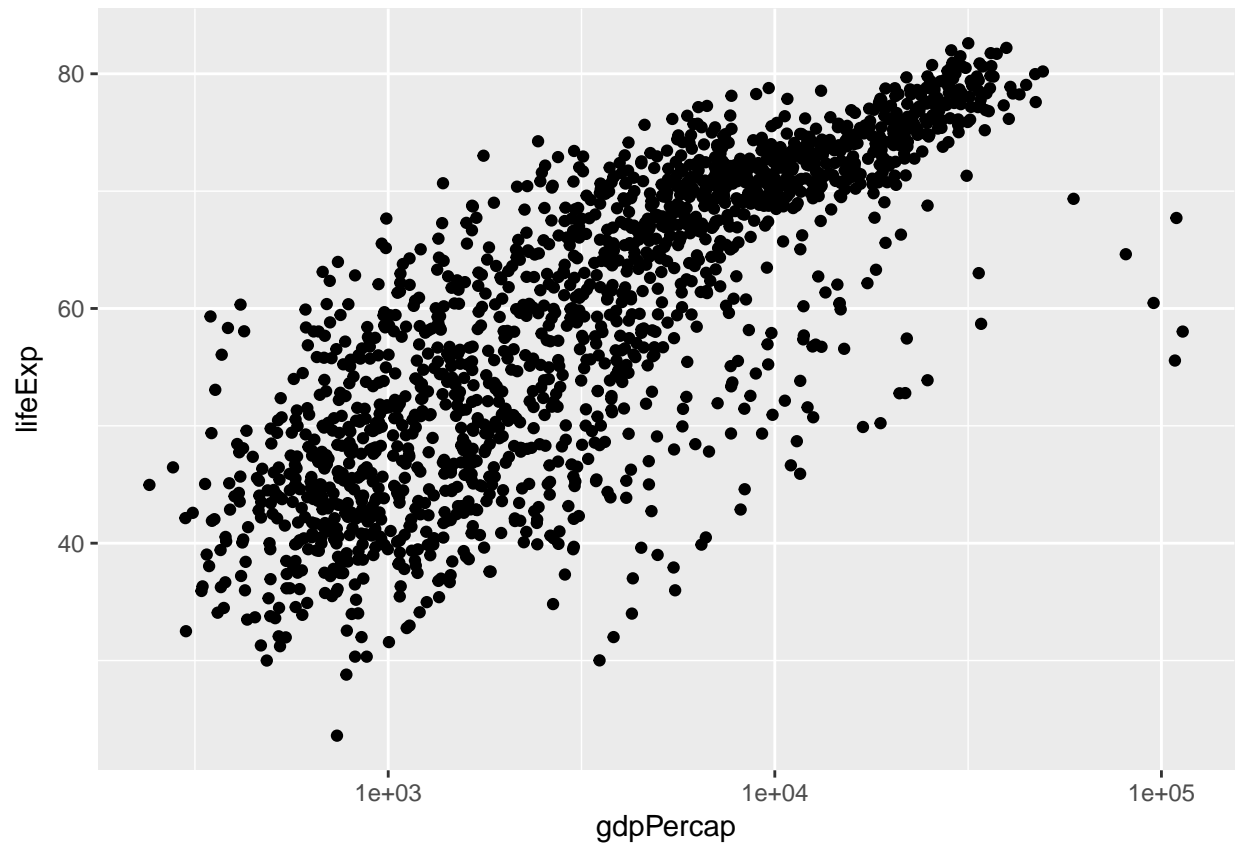```
p <- ggplot(filter(gapminder, continent!= "Oceania"),
                    aes(x = gdpPercap, y = lifeExp)) # just initializes
```
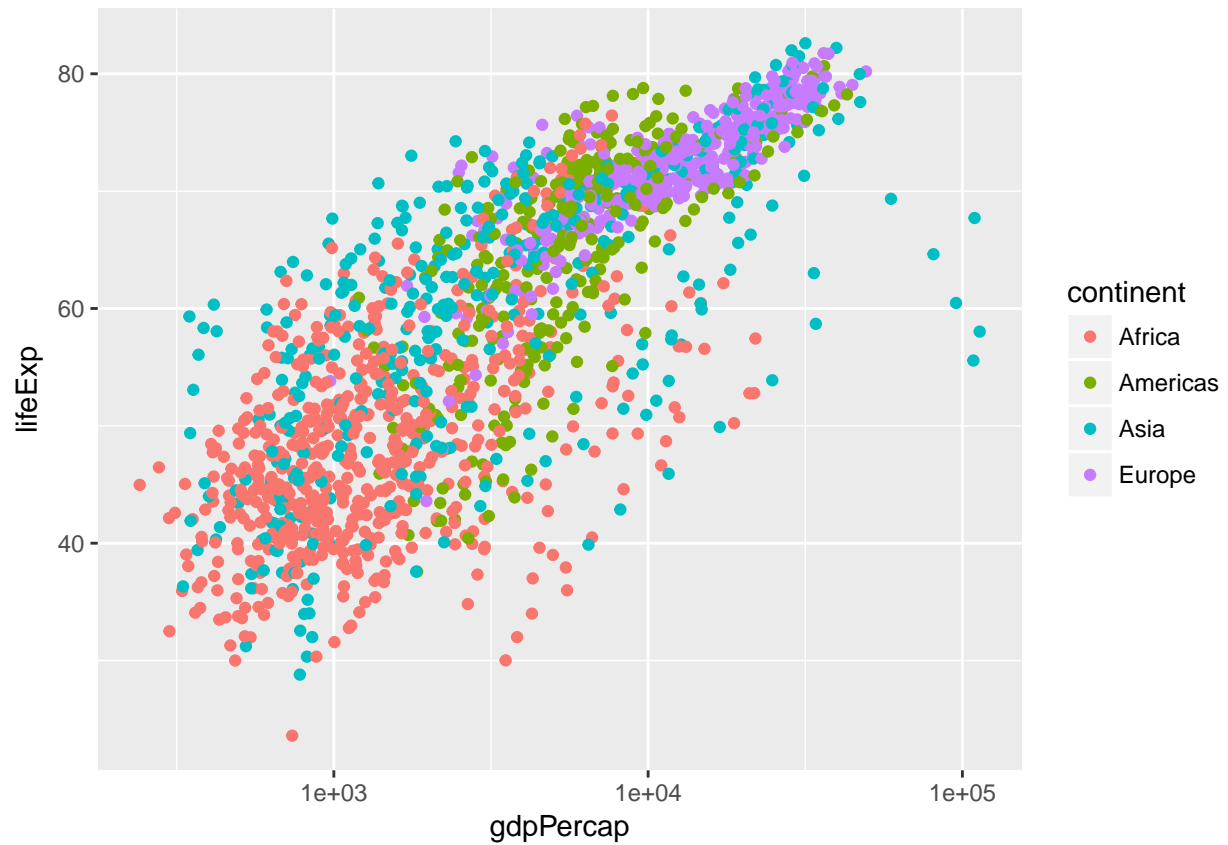
```
Warning: package 'bindrcpp' was built under R version 3.4.3
```
```
            p <- p + scale_x_log10() # log the x axis the right way
            p + geom_point() # scatterplot
```
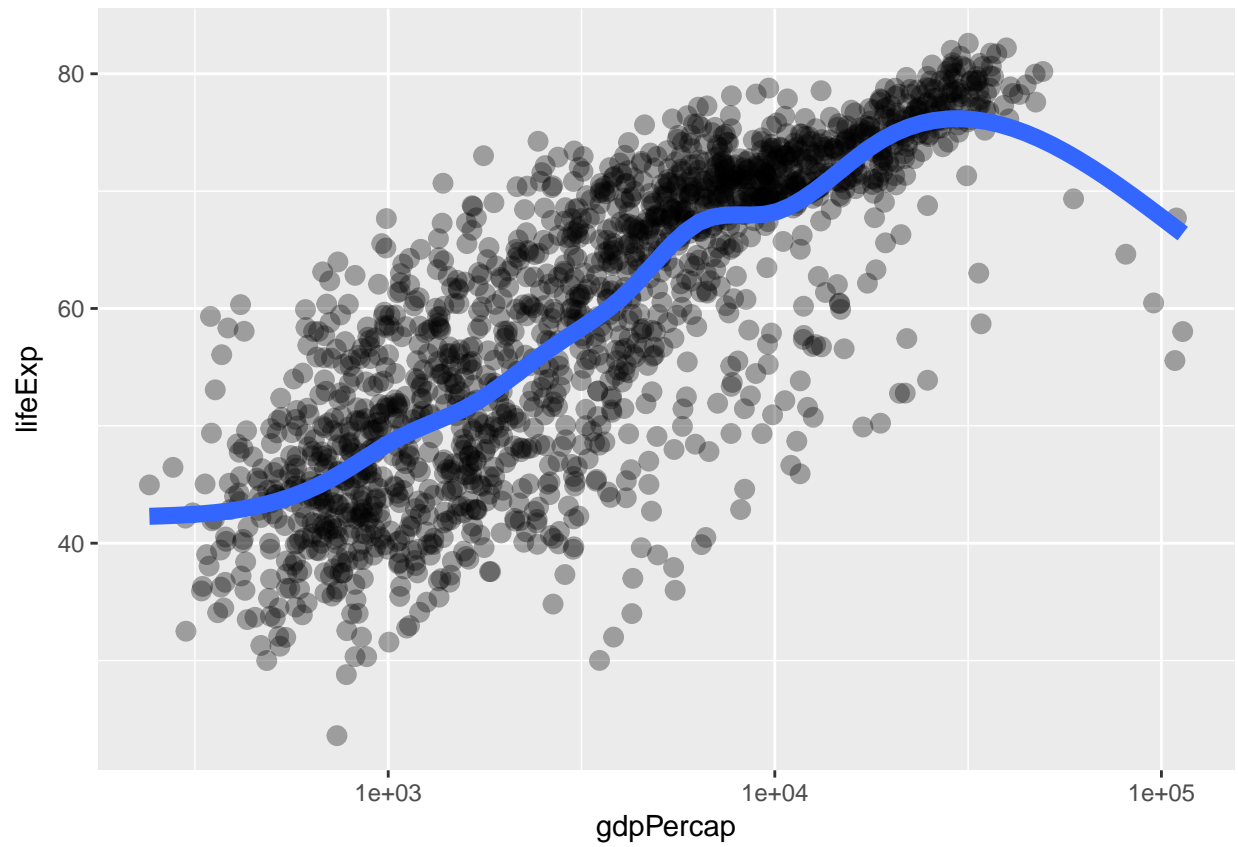
```
p + geom_point(aes(color = continent)) # map continent to color
```
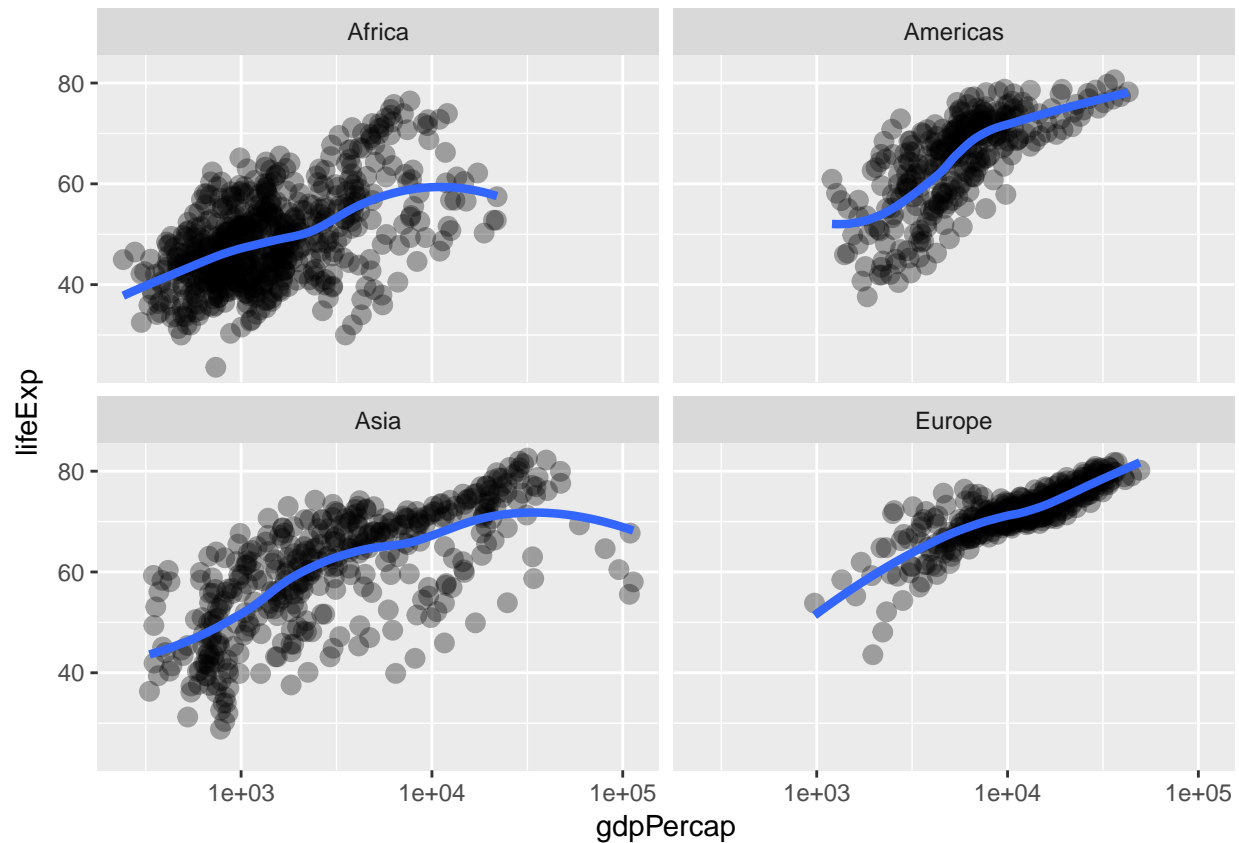
```
p + geom_point(alpha = (1/3), size = 3) + geom_smooth(lwd = 3, se = FALSE)
```

```
## `geom_smooth()` using method = 'gam'
```

```
p + geom_point(alpha = (1/3), size = 3) + facet_wrap(~ continent)+geom_smooth(lwd =1.5,se
```

```
## `geom_smooth()` using method = 'loess'
```

Exercise -02:

```r
dat_less <- subset(dat , subset = dat$lifeExp <32) #The life expectancy less than 32 years.

nrow(dat_less) #number of rows
```

```
## [1] 12
```

```r
table(dat_less$continent) #number of observations per continent.
```

```
##
##    Africa Americas     Asia   Europe  Oceania
##         8        0        4        0        0
```

Exercise -03:

```r
length(iris) #number of variables in the iris data set
```

```
## [1] 5
```

```r
names(iris) # their names
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"
```

```r
nrow(iris) # number of rows
```

```
## [1] 150
```

```r
dat1 <-iris[,-5]
```

```r
apply(as.matrix(dat1) , 2 , min) #smallest observations for each numeric variables.
```
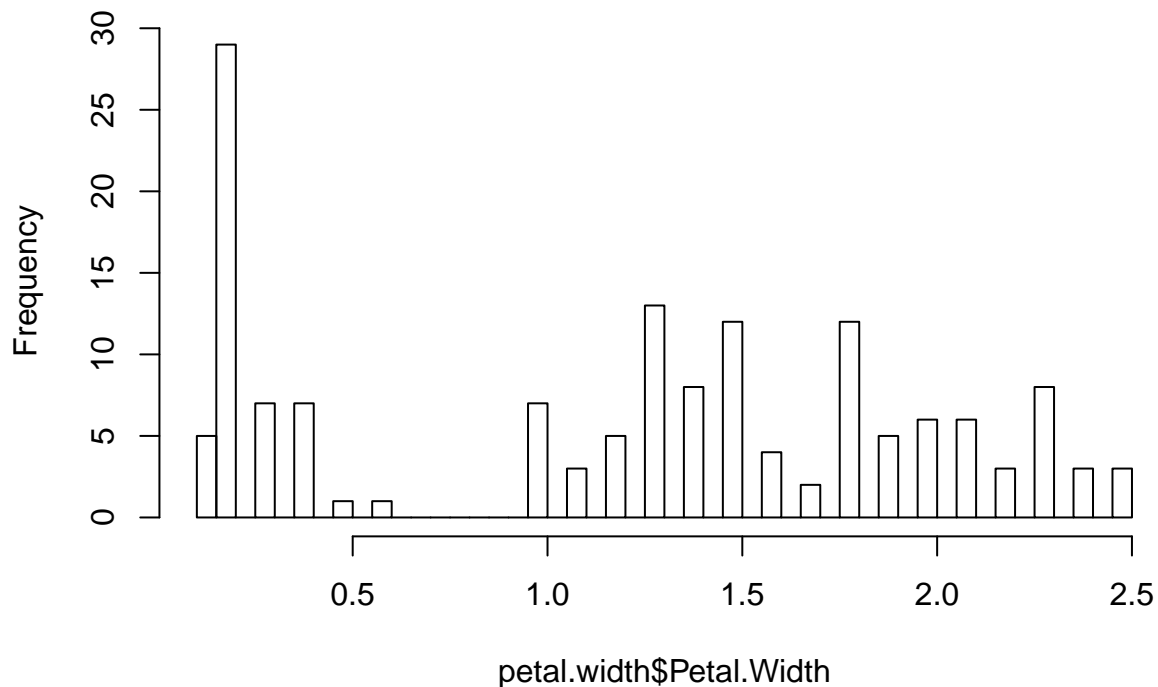
```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##          4.3          2.0          1.0          0.1
```

```r
petal.width <- subset(iris , select = Petal.Width) #Extract the Petal.width variables
```

```r
hist(petal.width$Petal.Width, breaks = 50) #make a histogram
```

Histogram of petal.width$Petal.Width


```r
table(cut(petal.width$Petal.Width , breaks = 10)) #make a table of frequencies
```

```
##
## (0.0976,0.34]    (0.34,0.58]    (0.58,0.82]    (0.82,1.06]     (1.06,1.3]
##            41              8              1              7             21
##     (1.3,1.54]    (1.54,1.78]    (1.78,2.02]    (2.02,2.26]     (2.26,2.5]
##            20              6             23              9             14
```

Exercise -04:

```r
filter(gapminder , country==c("Canada" , "Algeria")) #all entries of Canada and Algeria.
```

```
## # A tibble: 12 x 6
##     country continent  year lifeExp      pop gdpPercap
##     <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Algeria Africa      1957    45.7 10270856      3014
##  2 Algeria Africa      1967    51.4 12760499      3247
##  3 Algeria Africa      1977    58.0 17152804      4910
##  4 Algeria Africa      1987    65.8 23254956      5681
```

```
##  5 Algeria Africa      1997     69.2 29072015      4797
##  6 Algeria Africa      2007     72.3 33333216      6223
##  7 Canada  Americas    1952     68.8 14785584     11367
##  8 Canada  Americas    1962     71.3 18985849     13462
##  9 Canada  Americas    1972     72.9 22284500     18971
## 10 Canada  Americas    1982     75.8 25201900     22899
## 11 Canada  Americas    1992     78.0 28523502     26343
## 12 Canada  Americas    2002     79.8 31902268     33329
```

```r
filter(gapminder,
       country %in% c("Canada", "Algeria"), year < 1970, year >= 1960) #all entries of Canada and Alger
```

```
## # A tibble: 4 x 6
##    country continent  year lifeExp      pop gdpPercap
##    <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Algeria Africa     1962    48.3 11000948      2551
## 2 Algeria Africa     1967    51.4 12760499      3247
## 3 Canada  Americas   1962    71.3 18985849     13462
## 4 Canada  Americas   1967    72.1 20819767     16077
```

```r
filter(gapminder,
       (country == "Canada") |
         (country == "Algeria" &
          year %in% 1960:1969))#all entries of Canada, and entries of Algeria occuring in the '60s.
```

```
## # A tibble: 14 x 6
##     country continent  year lifeExp      pop gdpPercap
##     <fct>   <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Algeria Africa     1962    48.3 11000948      2551
##  2 Algeria Africa     1967    51.4 12760499      3247
##  3 Canada  Americas   1952    68.8 14785584     11367
##  4 Canada  Americas   1957    70.0 17010154     12490
##  5 Canada  Americas   1962    71.3 18985849     13462
##  6 Canada  Americas   1967    72.1 20819767     16077
##  7 Canada  Americas   1972    72.9 22284500     18971
##  8 Canada  Americas   1977    74.2 23796400     22091
##  9 Canada  Americas   1982    75.8 25201900     22899
## 10 Canada  Americas   1987    76.9 26549700     26627
## 11 Canada  Americas   1992    78.0 28523502     26343
## 12 Canada  Americas   1997    78.6 30305843     28955
## 13 Canada  Americas   2002    79.8 31902268     33329
## 14 Canada  Americas   2007    80.7 33390141     36319
```

```r
filter(gapminder,
       continent != "Europe") #all entries _not_ including European countries.
```

```
## # A tibble: 1,344 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 1,334 more rows
```

Exercise -04: Take all countries in Europe that have a GPD per capita greater than 10000, and select all variables except `gdpPercap`. (Hint: use `-`).

```
gapminder %>%
  filter(country %in%country, gdpPercap >10000) %>%
  select(country, year,continent,lifeExp,pop)
```

```
## # A tibble: 392 x 5
##    country     year continent lifeExp      pop
##    <fct>      <int> <fct>        <dbl>    <int>
##  1 Argentina  1977 Americas      68.5 26983828
##  2 Argentina  1997 Americas      73.3 36203463
##  3 Argentina  2007 Americas      75.3 40301927
##  4 Australia  1952 Oceania       69.1  8691212
##  5 Australia  1957 Oceania       70.3  9712569
##  6 Australia  1962 Oceania       70.9 10794968
##  7 Australia  1967 Oceania       71.1 11872264
##  8 Australia  1972 Oceania       71.9 13177000
##  9 Australia  1977 Oceania       73.5 14074100
## 10 Australia  1982 Oceania       74.7 15184200
## # ... with 382 more rows
```

Exercise -05:

```
x <- arrange(gapminder , year ,desc(lifeExp)) # Order the data frame by year, then descending by life e

df <-x[,c(3,4,1,2,5,6)] #In addition to the above exercise, rearrange the variables so that `year` come
```

Exercise -06: Make a new column called `cc` that pastes the country name followed by the continent, separated by a comma. (Hint: use the `paste` function with the `sep=", "` argument).

```
mutate(gapminder , cc = paste(country ,continent , sep = ","))
```

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

Exercise -07:

```
#Find the minimum GDP per capita experienced by each country
gapminder%>%
  group_by(country) %>%
```

```r
  summarise(minimum=min(gdpPercap))
```

```
## # A tibble: 142 x 2
##    country     minimum
##    <fct>         <dbl>
##  1 Afghanistan     635
##  2 Albania        1601
##  3 Algeria        2449
##  4 Angola         2277
##  5 Argentina      5911
##  6 Australia     10040
##  7 Austria        6137
##  8 Bahrain        9867
##  9 Bangladesh      630
## 10 Belgium        8343
## # ... with 132 more rows
```

```r
#How many years of record does each country have?
gapminder%>%
  group_by(country) %>%
  summarise(total=length(year))
```

```
## # A tibble: 142 x 2
##    country     total
##    <fct>       <int>
##  1 Afghanistan    12
##  2 Albania        12
##  3 Algeria        12
##  4 Angola         12
##  5 Argentina      12
##  6 Australia      12
##  7 Austria        12
##  8 Bahrain        12
##  9 Bangladesh     12
## 10 Belgium        12
## # ... with 132 more rows
```

```r
#Within Asia, what are the min and max life expectancies experienced in each year?
gapminder%>%
  group_by(country=="Asia" , year) %>%
  summarise(minimum=min(lifeExp) , maximum=max(lifeExp))
```

```
## # A tibble: 12 x 4
## # Groups: country == "Asia" [?]
##    `country == "Asia"`  year minimum maximum
##    <lgl>               <int>   <dbl>   <dbl>
##  1 F                    1952    28.8    72.7
##  2 F                    1957    30.3    73.5
##  3 F                    1962    32.0    73.7
##  4 F                    1967    34.0    74.2
##  5 F                    1972    35.4    74.7
##  6 F                    1977    31.2    76.1
##  7 F                    1982    38.4    77.1
##  8 F                    1987    39.9    78.7
##  9 F                    1992    23.6    79.4
```

```
## 10 F                      1997     36.1     80.7
## 11 F                      2002     39.2     82.0
## 12 F                      2007     39.6     82.6
```