# Assignment 1_Matei_Penca_s4039696

Name: Matei Penca
Studentnumber: s4039696

## Assignment 1

### Packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

There will be some warning because I do not have the latest version of R... If I try to install R 4.1.1 I cannot use Rmarkdown because it cannot install package "mime" so I will be using my old version which works fine.

### Install and load new package

Install the `gapminder`-package which allows you to make use of the gapminder dataset:

```
install.packages("gapminder")
```

(you only have to do this once!)

Load the gapminder data:

```
library(gapminder)
```

```
## Warning: package 'gapminder' was built under R version 4.1.1
```

Get a quick overview of the dataset

```
head(gapminder) # gapminder is name of dataset
```

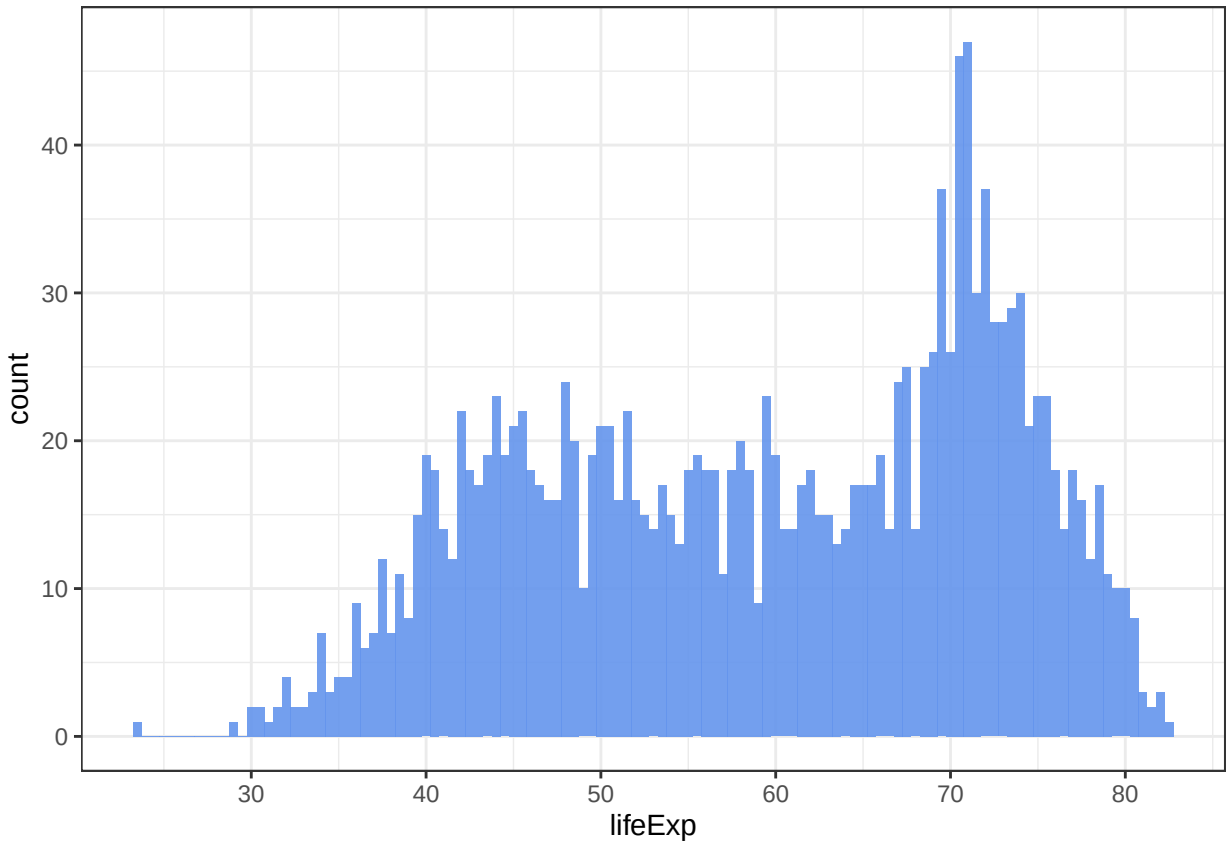
```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    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.
```

For explanation of variables, click here.

## Assignment

1. Create a histogram of `lifeExp`, the life-expectancy in years for the different countries in the dataset across the different years. Select an appropriate 'binwidth'. Change colours to your liking.

```
ggplot(gapminder, aes(lifeExp)) +
  geom_histogram(binwidth = 0.5, alpha=0.9, fill="cornflowerblue") +
  scale_x_continuous(breaks = seq(20, max(gapminder$lifeExp), by = 10)) +
  theme_bw()
```



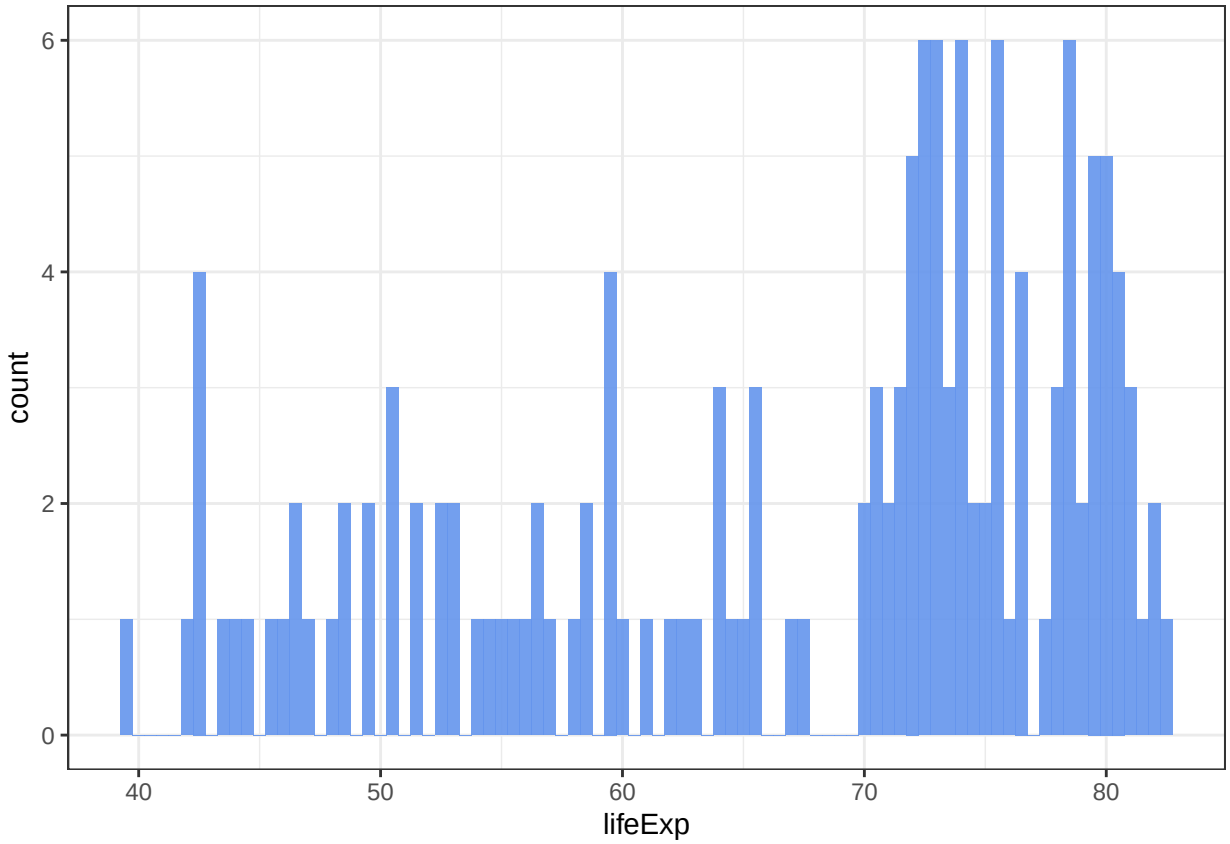

2. Describe in a few sentences what you see in the graph.

Looking at the graph we can see that the highest count is for the life expectancy of around 73 years old with more than 40 entries. While the life expectancy for 70+ years is high we can also see a lot of entries also in the intervals [40,50] which might be from third world countries where the expectancy is lower. Using this graph we could guess that the differences between a developed country and a developing one.

3. Create the same histogram, but only for the last year of measurement, which is 2007. You can create a new dataset for 2007 only by running the following code:

```
gapminder_2007 <- gapminder[gapminder$year == 2007, ]
```

```
ggplot(gapminder_2007, aes(lifeExp)) +
  geom_histogram(binwidth = 0.5, alpha=0.9, fill="cornflowerblue") +
  scale_x_continuous(breaks = seq(20, max(gapminder$lifeExp), by = 10)) +
  theme_bw()
```

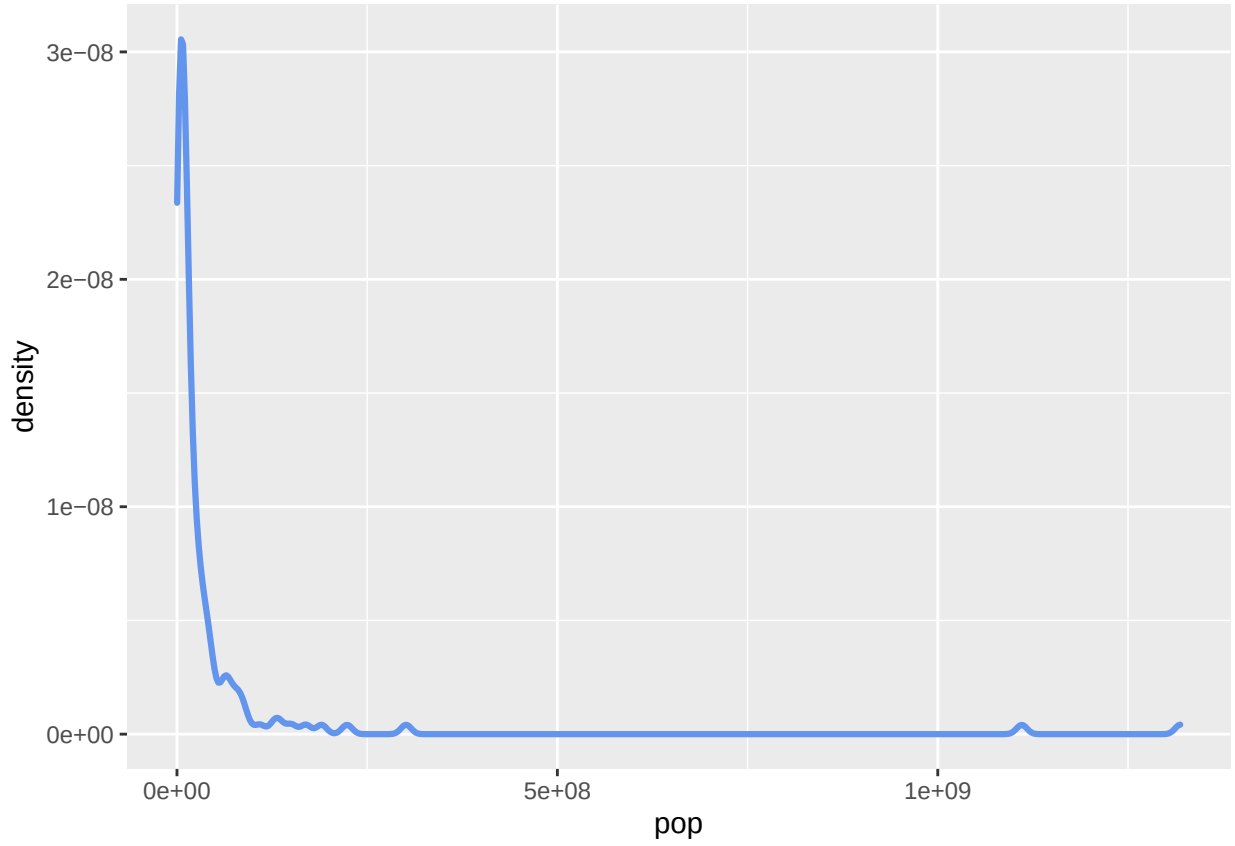



4. Explain in a few sentences how the graph is different from the one you created in 1.

First of all because there is less data, the graph looks a bit weird as there are a lot of columns(ages) with no data points. Looking at this graph we can see that in 2007 the life expectancy seems higher than in the graph with all the years. A difference between the graph would be the count of people with a lower life expectancy [40,60].

5. For the dataset `gapminder_2007`, create density-plot for the variable `pop` (referring to population size of a country). Choose colours to your liking.

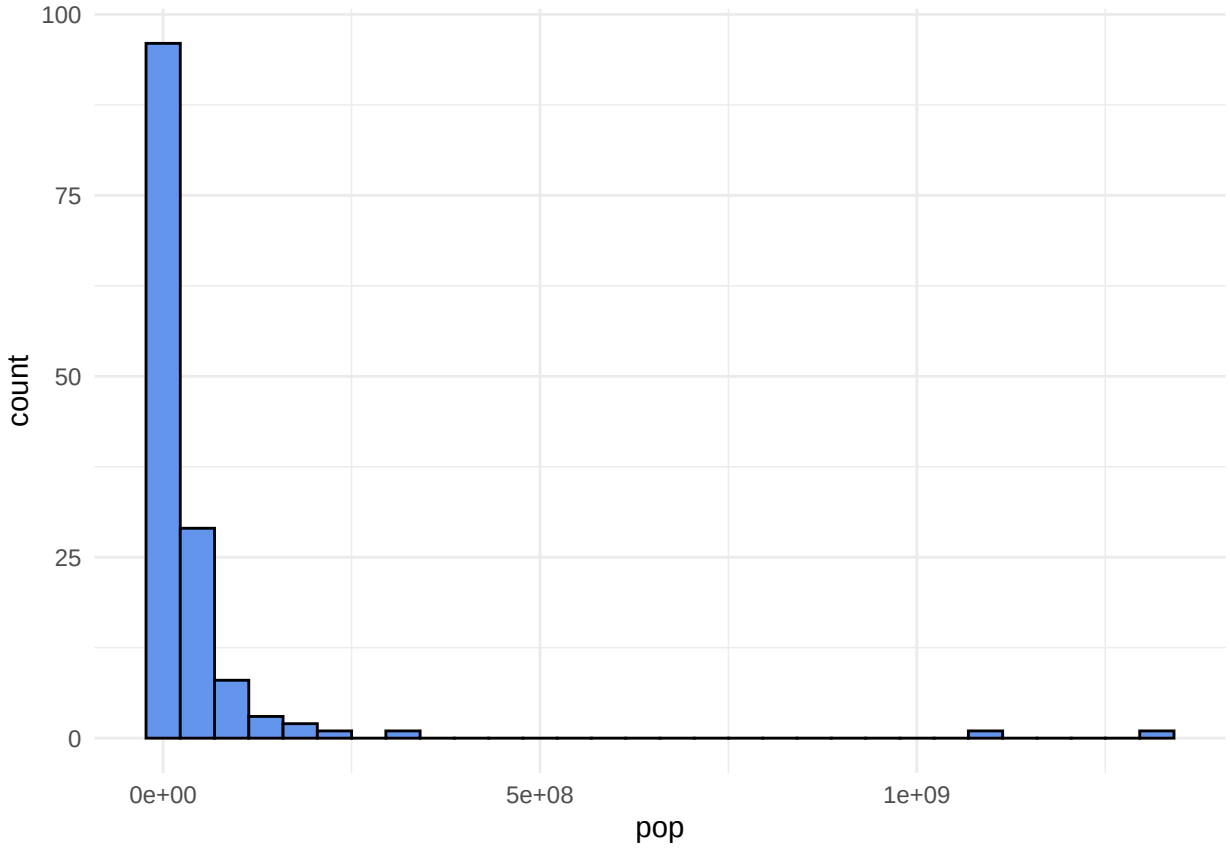
```
ggplot(gapminder_2007, aes(pop)) +
 geom_density(color="cornflowerblue", size=1.1)
```




6. Also create a histogram. Add a different theme!

```
ggplot(gapminder_2007, aes(pop)) +
 geom_histogram(fill="cornflowerblue", color="black") +
 theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




7. Why are both of the graphs not very informative?

Both the density plot and histogram have the common problem that the interval for the population is very big and there is little data points for it(most of the data points are at the beginning, but there are also a few at the far right). As we were showed in the first lecture, both the density and the histogram show more or less the same thing, but for our example I would say there is not that much insight to gather from these 2 graphs other than that most countries have smaller populations.

8. Do you have a suggestion how to make this graph more informative?

First of all we could zoom in on a smaller part of the interval and we could also force R to not show the scientific notation which can confuse people reading the graph.

9. Make a density-plot of the variable 'countries'. Why doesn't this work?

```
graph <- ggplot(gapminder_2007, aes(country)) +
  geom_density()
suppressWarnings(print(graph))
```


The graph won't work because it is supposed to be used on a continuous variable while country is categorical. This graph throws 50 warnings which I have suppressed so that they do not show up in the PDF.

The error is :Groups with fewer than two data points have been dropped.

10. What would be a more appropriate type of graph for this variable?

Country variable is categorical variable so it would make more sense to use a graph geom for categorical variables.

## Submit your pdf-file online

When you are done with your assignment, try clicking the "Knit"-button in Rstudio so that your code will be knitted into a **pdf**-file. Submit this file on Nestor.

If you see an error appear, try to fix the error and try again. If you can't get the "knitting" to work, then you can copy your code, text, and graphs into a word-file and submit it on Nestor.

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
```

```
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gapminder_0.3.0 ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9        pillar_1.4.6     compiler_4.0.5   tools_4.0.5
##  [5] digest_0.6.28    evaluate_0.14    lifecycle_1.0.1  tibble_3.0.3
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.11     cli_3.0.1
## [13] rstudioapi_0.13  yaml_2.2.1       xfun_0.26        fastmap_1.1.0
## [17] withr_2.4.2      stringr_1.4.0    dplyr_1.0.2      knitr_1.36
## [21] generics_0.1.0   vctrs_0.3.2      grid_4.0.5       tidyselect_1.1.1
## [25] glue_1.4.1       R6_2.5.1         fansi_0.5.0      rmarkdown_2.11
## [29] farver_2.1.0     purrr_0.3.4      magrittr_2.0.1   scales_1.1.1
## [33] ellipsis_0.3.1   htmltools_0.5.2  colorspace_1.4-1 labeling_0.4.2
## [37] utf8_1.2.2       stringi_1.4.6    munsell_0.5.0    crayon_1.4.1
```