Workbook #1 - Introduction of R

Aldrich Wang

Sept 15, 2022

Contents

Introduction

R programming and exploratory data analysis have been very significant in the process of economic research. We cannot testify a hypothesis or test a model without collecting data or performing quantitative analysis on our collected databases. Similarly, mastering the R language is an extremely important step in producing a research project by ourselves. In our cohort, we will learn how to use R step by step in order to succeed in economic research.

Normally, data analysis includes data import, tidying, transformation, visualization, modeling, and conclusion. In this first workbook, we are gonna focus on data visualization.

Note: This workbook is prepared for members who have little to no previous experience with R. If you have been using this language for quite a while, feel free to skip this workbook exercise.

Intro to Dataframes

The following code chunk showcases the packages that we need for our work - we can use different functions from these packages. This could be a head start on our work.

```
# load required packages
library(tibble)
library(ggplot2)
library(gapminder)
```

In the following chunk, we are able to view this displayed data frame, named "gapminder". The size of this data frame is 1704 * 6. The first row of this data frame tells us the variable names - country, continent, year, lifeExp (life expectancy at birth), pop (total population), gdpPercap (GDP per capita). As we can see in the variables, the data stored in a data frame can be of **numeric**, **factor**, or **character** type. In a **tidy** data frame, each variable forms a column; each observation forms a row; each cell is a single measurement.

```
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
```

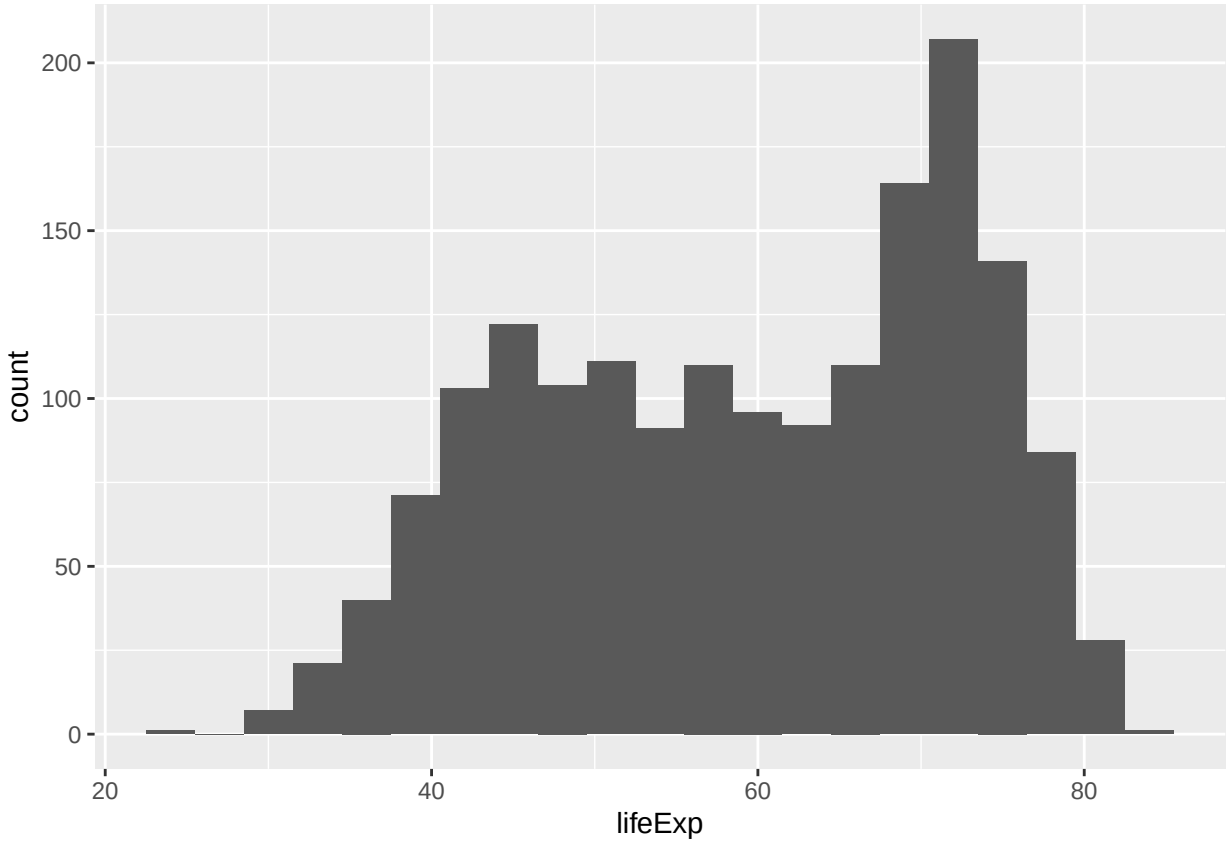
Data Visualization - different types of plot in R

R is a powerful tool when it comes to data visualization. With the ggplot2 package, we are able to plot histograms, box plots, violin plots, scatter plots, etc.

Histograms

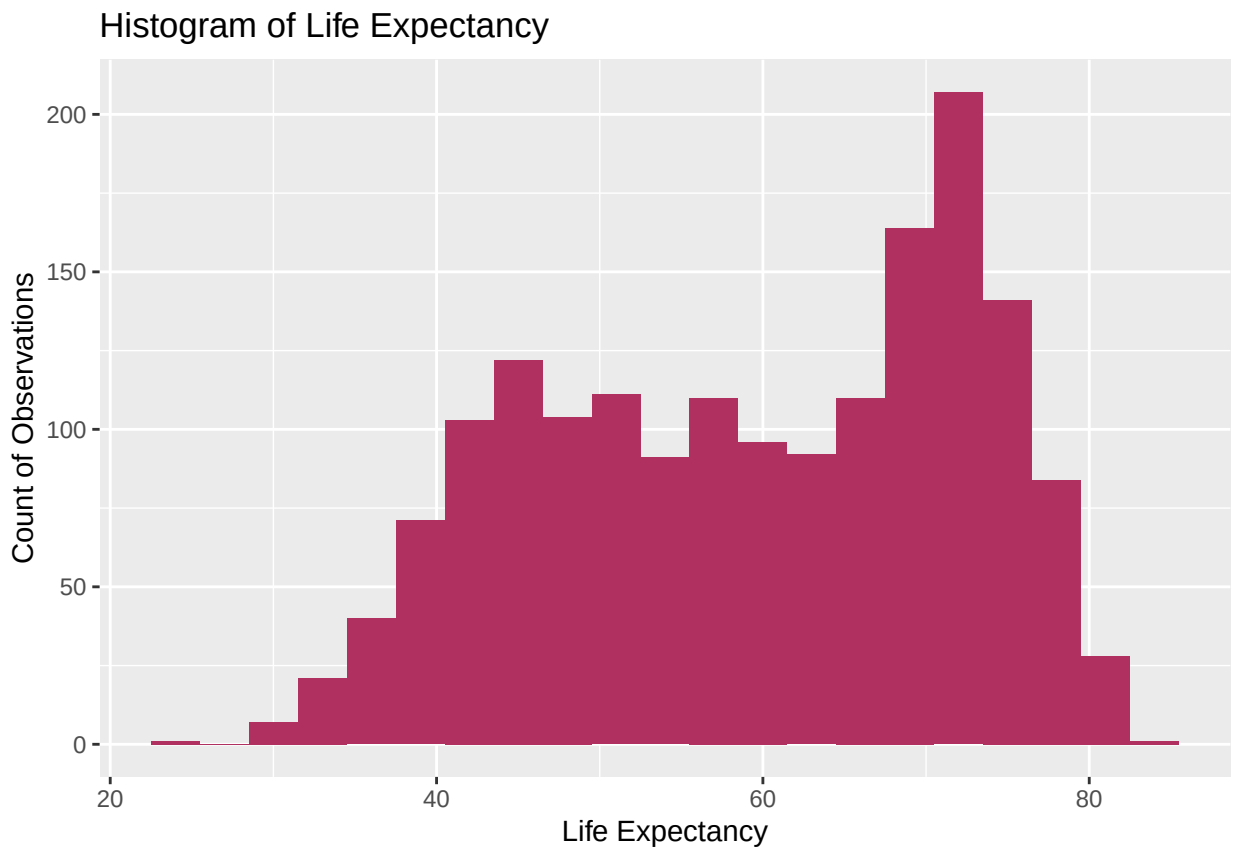
Let's generate a histogram of life expectancy from the gapminder data frame. We'll build on this plain histogram and make it look fancier.

```
ggplot(data = gapminder, mapping = aes(x = lifeExp)) +
  geom_histogram(binwidth = 3)
```



Then, we add labels to this graph to make it more elaborate for our readers and assign a color to this graph for maroon pride.

```
ggplot(data = gapminder,
       mapping = aes(x = lifeExp)) +
  geom_histogram(binwidth = 3,
                 fill = "maroon") +
  labs(title = "Histogram of Life Expectancy",
       x = "Life Expectancy",
       y = "Count of Observations")
```

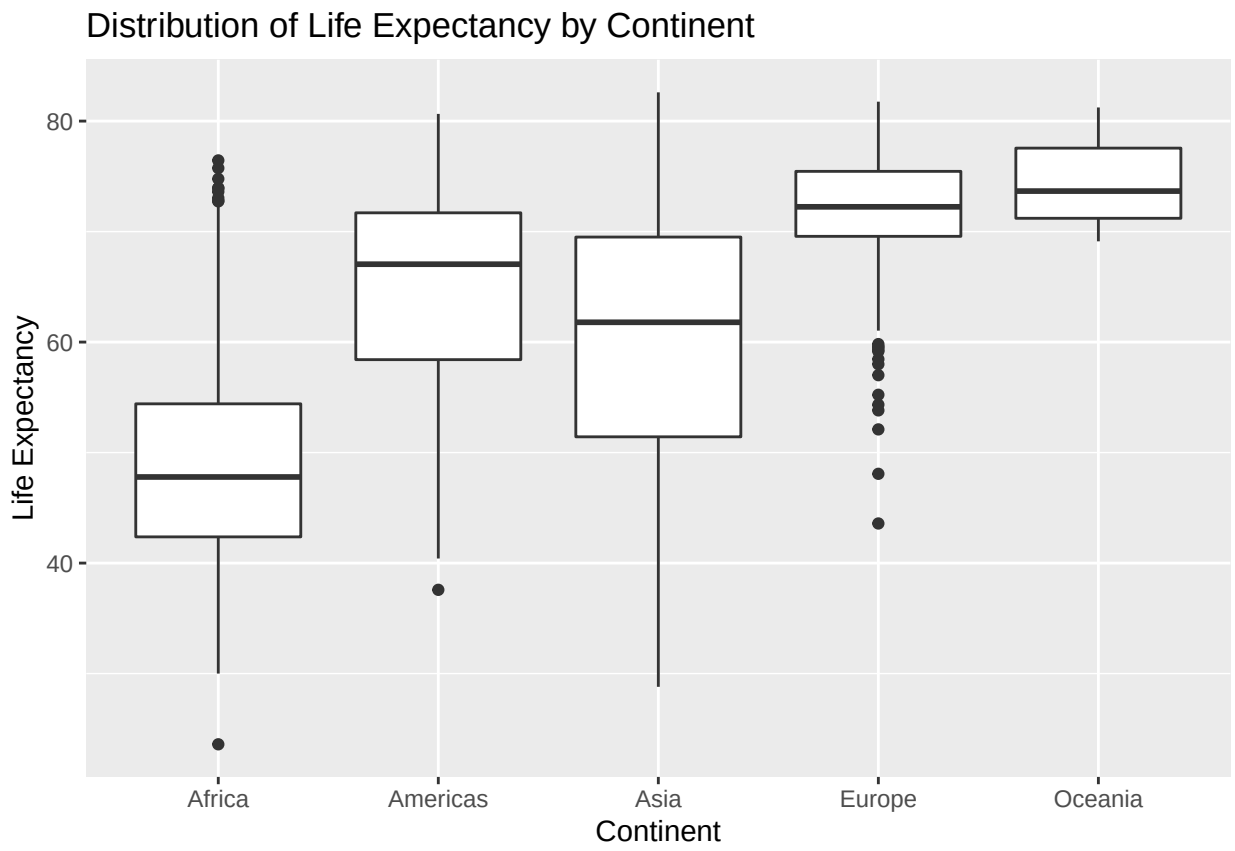


Exercise: Can you try to generate separate histograms of life expectancy for each continent? Feel free to google and talk with others!

Box plots

After the introduction of histograms, box plots are usually good at giving us information on distributions. Let's compare the distribution of life expectancy by continent from a box plot. (Feel free to play around with the code to change the color you like!)

```
ggplot(data = gapminder, mapping = aes(x = continent, y = lifeExp)) +
  geom_boxplot() +
  labs(title = "Distribution of Life Expectancy by Continent",
       x = "Continent",
       y = "Life Expectancy")
```



Violin Plots

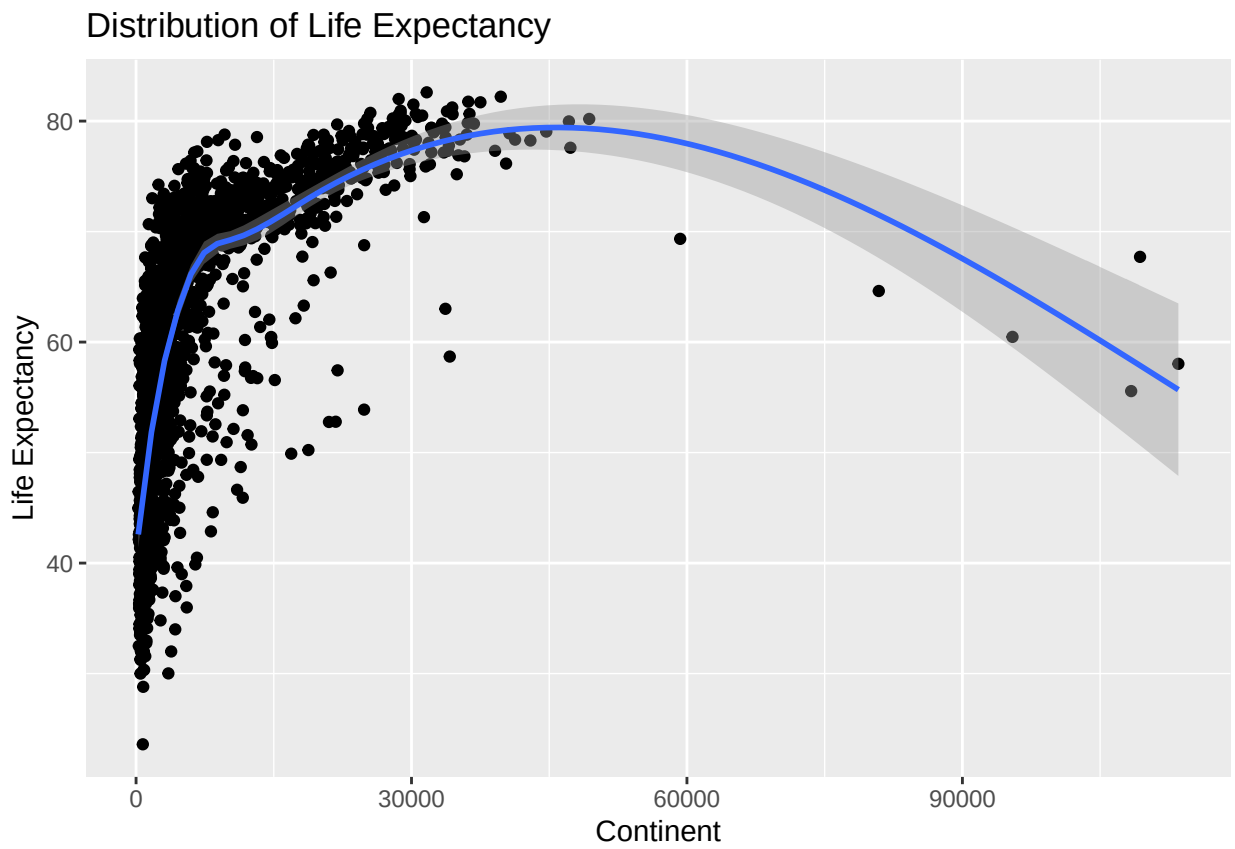
Exercise: By a similar logic to creating a box plot, please try to create a violin plot detailing the distribution of life expectancy by continent.

Scatter Plots

Now let's do scatter plots. Use the scatter plot to showcase the relationship between per capita GDP and life expectancy. We also need a smoothing line for these scatter dots on the graph, as this line itself can roughly tell us the relationship between two variables.

```
ggplot(data = gapminder,
       mapping = aes(x = gdpPercap,
                     y = lifeExp)) +
  geom_point() +
  geom_smooth(method = 'gam') + # the smoothing line
  labs(title = "Distribution of Life Expectancy",
       x = "Continent",
       y = "Life Expectancy")
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



Exercise: Use color to identify how this relationship differs by continent.

Session Info

```
sessioninfo::session_info()
```

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.2.0 (2022-04-22)
##  os       Red Hat Enterprise Linux 8.6 (Ootpa)
##  system   x86_64, linux-gnu
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       America/Chicago
##  date     2022-09-15
##  pandoc   2.17.1.1 @ /usr/lib/rstudio-server/bin/quarto/bin/ (via rmarkdown)
##
## - Packages -------------------------------------------------------------------
##  package     * version date (UTC) lib source
##  assertthat    0.2.1   2019-03-21 [2] CRAN (R 4.2.0)
##  cli           3.3.0   2022-04-25 [2] CRAN (R 4.2.0)
##  colorspace    2.0-3   2022-02-21 [2] CRAN (R 4.2.0)
```

```
##  crayon        1.5.1   2022-03-26 [2] CRAN (R 4.2.0)
##  DBI           1.1.2   2021-12-20 [2] CRAN (R 4.2.0)
##  digest        0.6.29  2021-12-01 [2] CRAN (R 4.2.0)
##  dplyr         1.0.9   2022-04-28 [2] CRAN (R 4.2.0)
##  ellipsis      0.3.2   2021-04-29 [2] CRAN (R 4.2.0)
##  evaluate      0.15    2022-02-18 [2] CRAN (R 4.2.0)
##  fansi         1.0.3   2022-03-24 [2] CRAN (R 4.2.0)
##  farver        2.1.0   2021-02-28 [2] CRAN (R 4.2.0)
##  fastmap       1.1.0   2021-01-25 [2] CRAN (R 4.2.0)
##  gapminder   * 0.3.0   2017-10-31 [2] CRAN (R 4.2.0)
##  generics      0.1.2   2022-01-31 [2] CRAN (R 4.2.0)
##  ggplot2     * 3.3.6   2022-05-03 [2] CRAN (R 4.2.0)
##  glue          1.6.2   2022-02-24 [2] CRAN (R 4.2.0)
##  gtable        0.3.0   2019-03-25 [2] CRAN (R 4.2.0)
##  highr         0.9     2021-04-16 [2] CRAN (R 4.2.0)
##  htmltools     0.5.2   2021-08-25 [2] CRAN (R 4.2.0)
##  knitr         1.39    2022-04-26 [2] CRAN (R 4.2.0)
##  labeling      0.4.2   2020-10-20 [2] CRAN (R 4.2.0)
##  lattice       0.20-45 2021-09-22 [2] CRAN (R 4.2.0)
##  lifecycle     1.0.1   2021-09-24 [2] CRAN (R 4.2.0)
##  magrittr      2.0.3   2022-03-30 [2] CRAN (R 4.2.0)
##  Matrix        1.4-1   2022-03-23 [2] CRAN (R 4.2.0)
##  mgcv          1.8-40  2022-03-29 [2] CRAN (R 4.2.0)
##  munsell       0.5.0   2018-06-12 [2] CRAN (R 4.2.0)
##  nlme          3.1-157 2022-03-25 [2] CRAN (R 4.2.0)
##  pillar        1.7.0   2022-02-01 [2] CRAN (R 4.2.0)
##  pkgconfig     2.0.3   2019-09-22 [2] CRAN (R 4.2.0)
##  purrr         0.3.4   2020-04-17 [2] CRAN (R 4.2.0)
##  R6            2.5.1   2021-08-19 [2] CRAN (R 4.2.0)
##  rlang         1.0.2   2022-03-04 [2] CRAN (R 4.2.0)
##  rmarkdown     2.15    2022-08-16 [1] CRAN (R 4.2.0)
##  rstudioapi    0.13    2020-11-12 [2] CRAN (R 4.2.0)
##  scales        1.2.0   2022-04-13 [2] CRAN (R 4.2.0)
##  sessioninfo   1.2.2   2021-12-06 [2] CRAN (R 4.2.0)
##  stringi       1.7.6   2021-11-29 [2] CRAN (R 4.2.0)
##  stringr       1.4.0   2019-02-10 [2] CRAN (R 4.2.0)
##  tibble      * 3.1.7   2022-05-03 [2] CRAN (R 4.2.0)
##  tidyselect    1.1.2   2022-02-21 [2] CRAN (R 4.2.0)
##  utf8          1.2.2   2021-07-24 [2] CRAN (R 4.2.0)
##  vctrs         0.4.1   2022-04-13 [2] CRAN (R 4.2.0)
##  withr         2.5.0   2022-03-03 [2] CRAN (R 4.2.0)
##  xfun          0.31    2022-05-10 [2] CRAN (R 4.2.0)
##  yaml          2.3.5   2022-02-21 [2] CRAN (R 4.2.0)
##
##  [1] /home/chenghaow/R/x86_64-pc-linux-gnu-library/4.2
##  [2] /opt/R/4.2.0/lib/R/library
##
## ------------------------------------------------------------------------------
```