OLAH DATA KUANTITATIF MENGGUNAKAN BAHASA R (R-LANGUAGE)

Lala Septem Riza

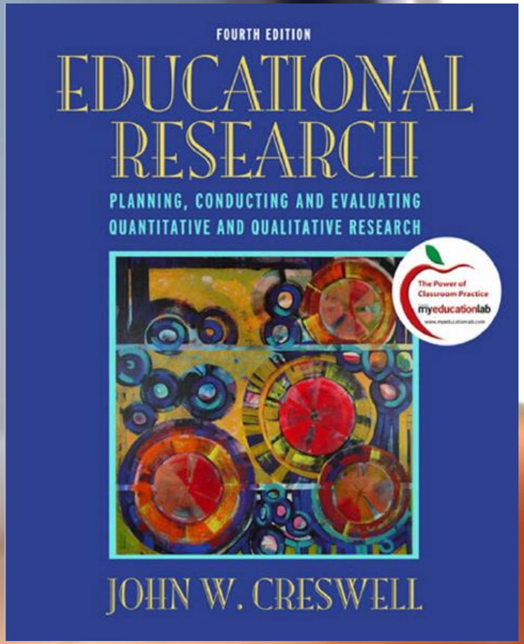
Department of Computer Science Education,
Universitas Pendidikan Indonesia
lala.s.riza@upi.edu

Outline

- Pengenalan Penelitian Kuantitatif dan Applied Statistics
- Pengenalan Bahasa Pemrograman R
- Applied Statistics di R
  - Visualisasi Data di R
  - Descriptive Statistics di R
  - Inference Statistics di R
- Wawasan: Penelitian dengan R untuk implementasi dan model komputasi.

Penelitian Kuantitatif dan Applied Statistics

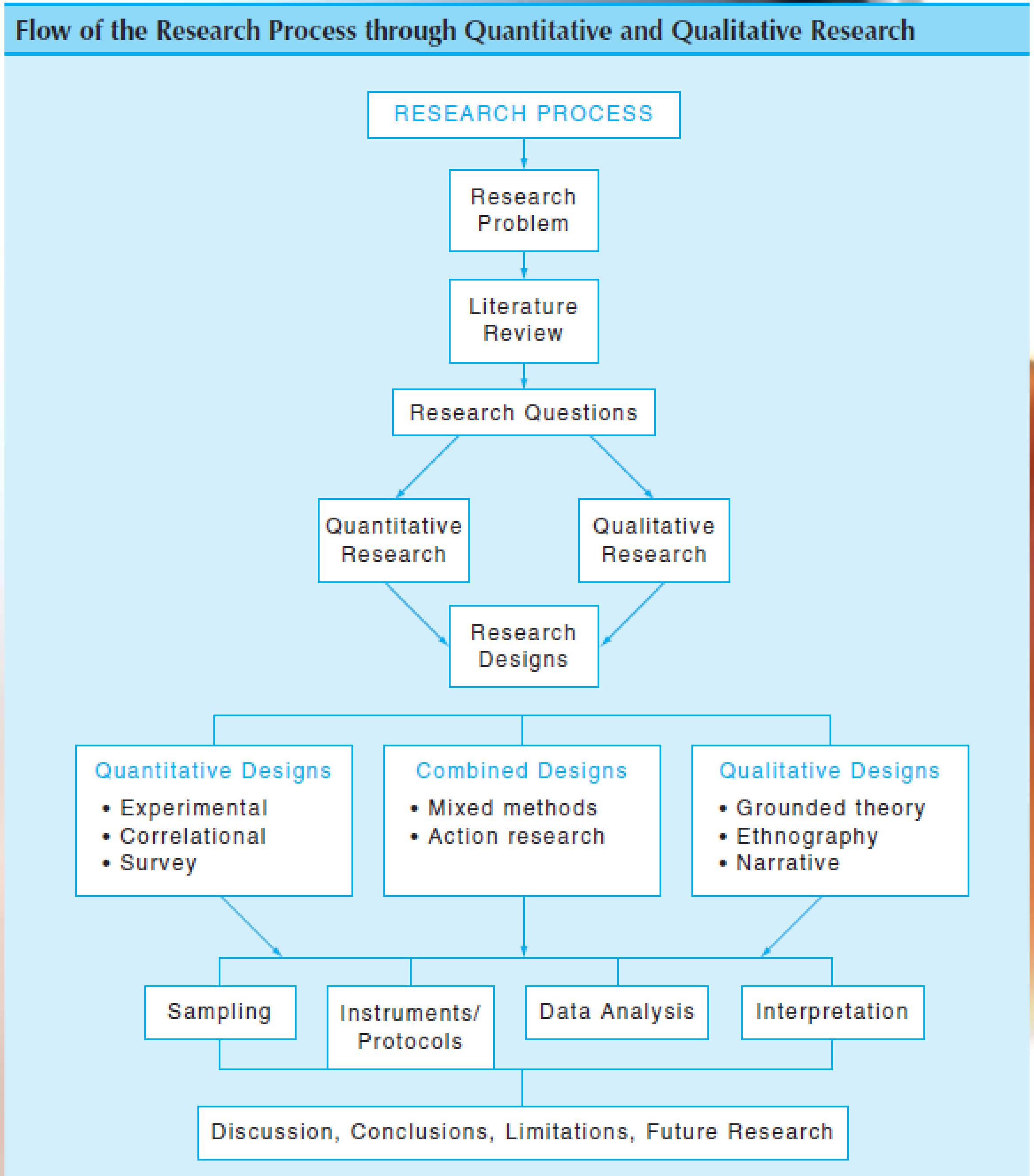
Metode Kuantitatif



- Quantitative research is defined as a **systematic investigation** of phenomena by gathering **quantifiable data** and performing **statistical, mathematical, or computational techniques**.

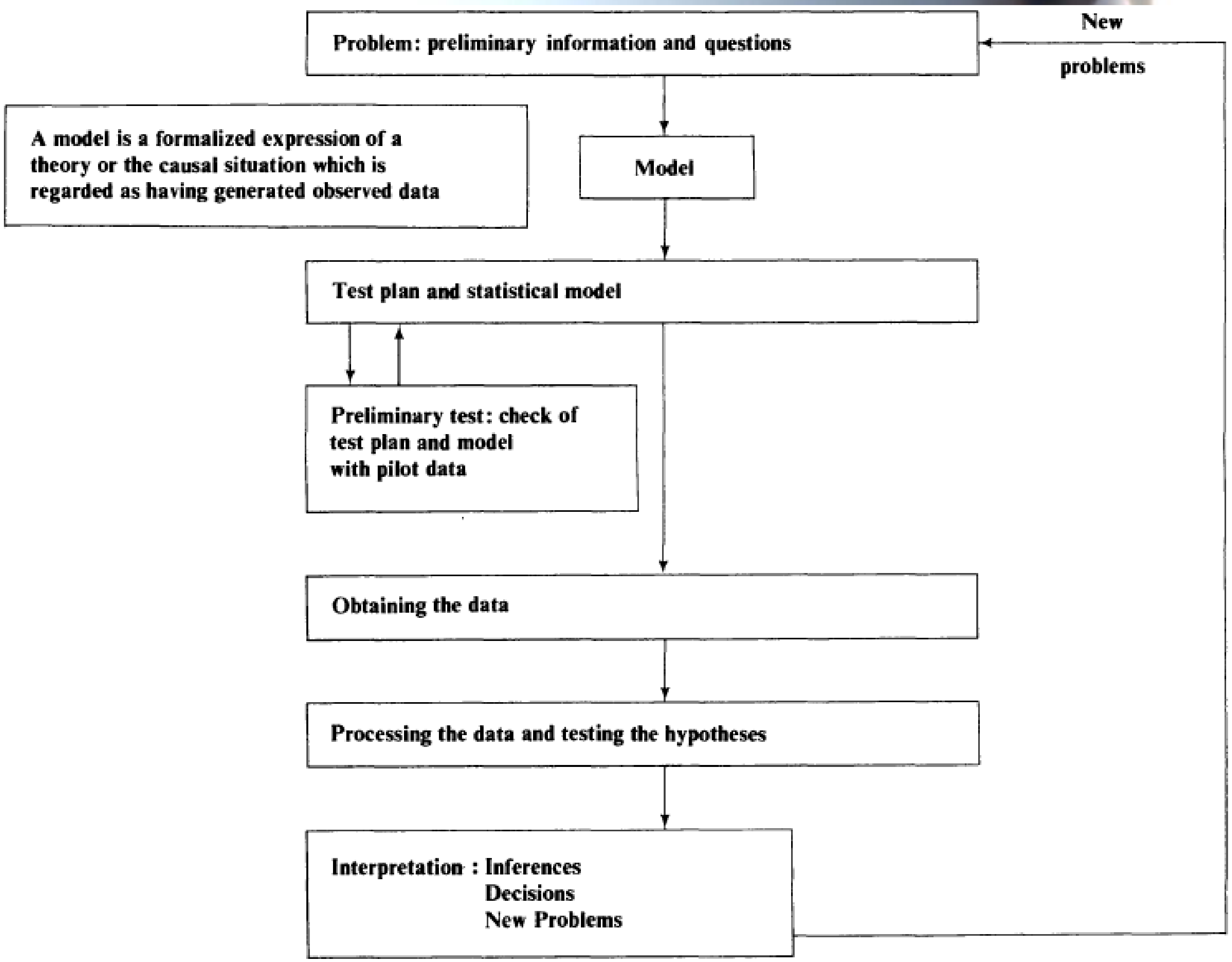
Quantitative Designs:

- **Experimental**: To determine whether an activity or materials **make a difference** in results for participants.
- **Correlational**: To examine/measure the **association or relation** of one or more variables than in testing the impact of activities or materials.
- **Survey**: to describe trends in a large population of individuals by giving **questionnaire** to a small group of people



Statistics in Research

- Research is an **endeavour** to discover answers to intellectual and practical problems through the application of **scientific method**.
- Research is a systematized effort to gain **new knowledge**.



Tipe Statistics

1. **Applied statistics**: **Descriptive statistics** and the application of **inferential statistics.**
2. **Mathematical statistics**: The manipulation of probability distributions necessary for deriving results related to **methods of estimation and inference**, various aspects of **computational statistics** and the **design of experiments**.
3. **Theoretical statistics**: The **logical arguments** underlying justification of approaches to statistical inference, as well as encompassing *mathematical statistics*.

Applied Statistics

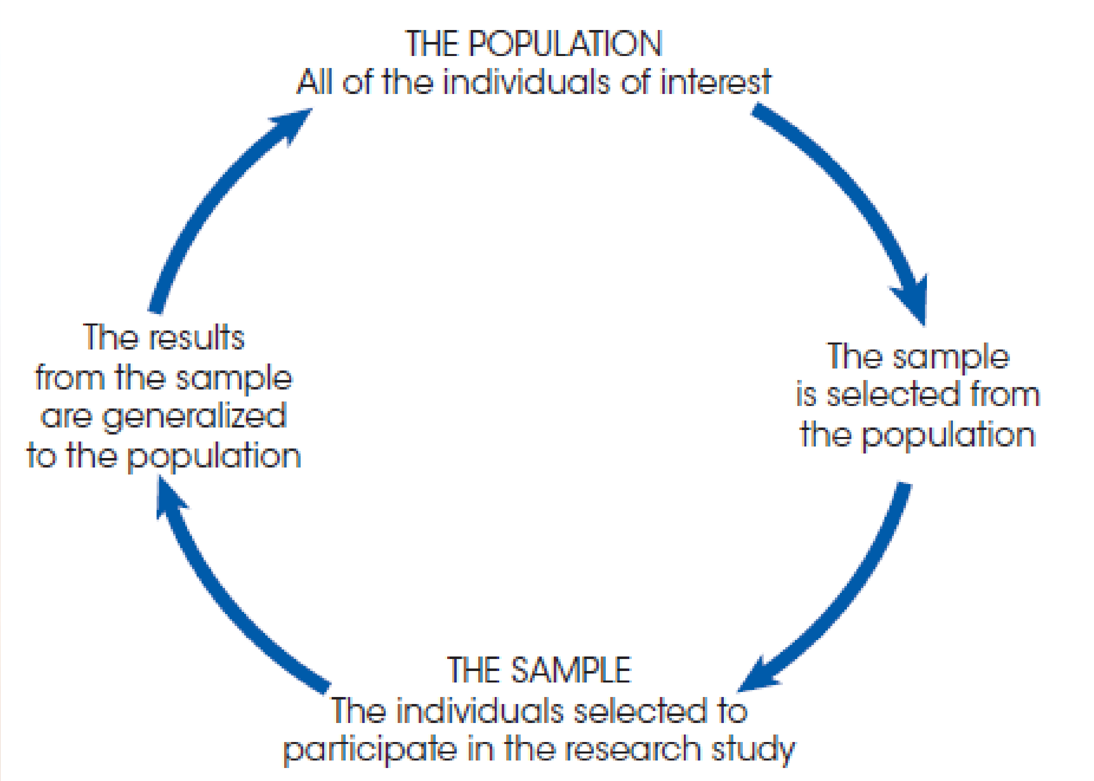
- Descriptive Statistics: to **describe** the characteristics of the sample in an accurate and unambiguous fashion in such a way that the information will be easily **communicated** to others ➔ Distribution (frequency distribution), Central tendency (mean, median, mode), Dispersion (range and std deviation), and their visualizations.

- Inferential Statistics: **working with a sample** we introduce some unknown amount of error due to the effects of chance. Inferential statistics allow **conclusions about a population based on data from a sample**.

Inferential statistics consist of techniques that allow us to **study samples** and then **make generalizations about the populations** from which they were selected.

Relationship between population and sample

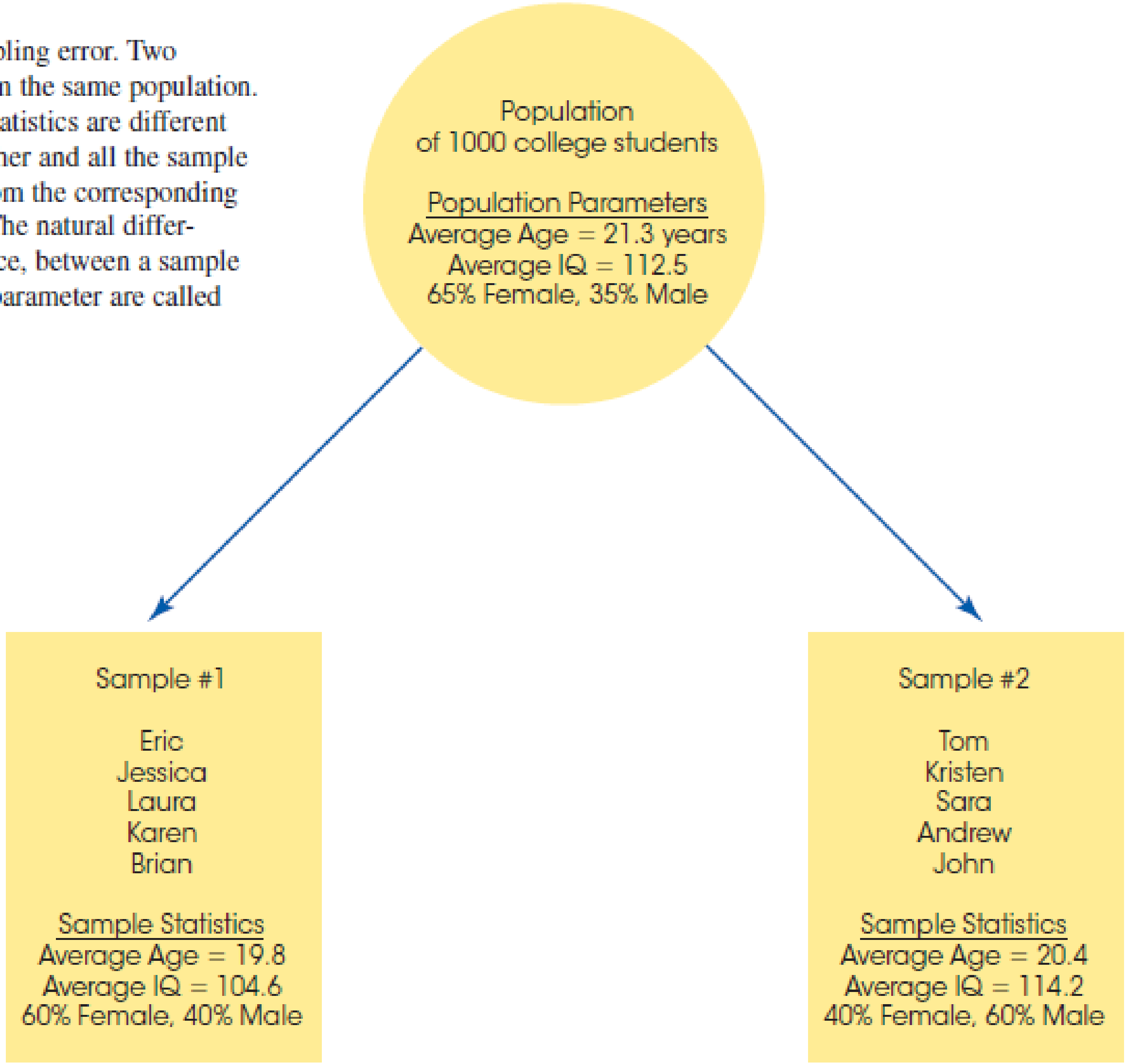
- Specifically, when a researcher finishes **examining the sample**, the goal is **to generalize** the results back to the entire population.
- Remember that the research started with **a general question about the population**.
- To answer the question, a researcher **studies a sample** and then **generalizes the results** from the sample to the population.



Sampling Error

Sampling error is the naturally occurring discrepancy, or error, that exists between a sample statistic and the corresponding population parameter.

FIGURE 1.2
A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another and all the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and population parameter are called sampling error.



Step 1
Experiment:
Compare two studying methods

Population of College Students

Data
Test scores for the students in each sample

Sample A: Read from printed pages			Sample B: Read from computer screen		
25	26	28	20	22	27
27	21	27	23	17	23
30	28	24	25	28	21
19	23	26	22	19	22
29	26	22	18	24	19

Step 2
Descriptive statistics:
Organize and simplify

Average Score = 26

Average Score = 22

Step 3
Inferential statistics:
Interpret results

The sample data show a 4-point difference between the two methods of studying. However, there are two ways to interpret the results.

1. There actually is no difference between the two studying methods, and the sample difference is due to chance (sampling error).
2. There really is a difference between the two methods, and the sample data accurately reflect this difference.

The goal of inferential statistics is to help researchers decide between the two interpretations.

FIGURE 1.3
The role of statistics in experimental research.

Reproducibility and Replicability

- **Reproducibility** is the ability to get the same research results using the raw data and computer programs provided by the researchers.
- **Replicability** is the ability to independently achieve similar conclusions when differences in sampling, research procedures and data analysis methods may exist.
- **Reproducibility and replicability** together are among the main principles of the scientific method.

Introduction to R Programming Language

What is R?

- R is a programming language and software environment for **statistical computing and graphics**.
- R is an implementation of the **S programming** language combined with lexical scoping semantics inspired by Scheme.
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.
- R is a GNU project/Open Source.
- R is an **interpreted language**; users typically access it through a command-line interpreter.

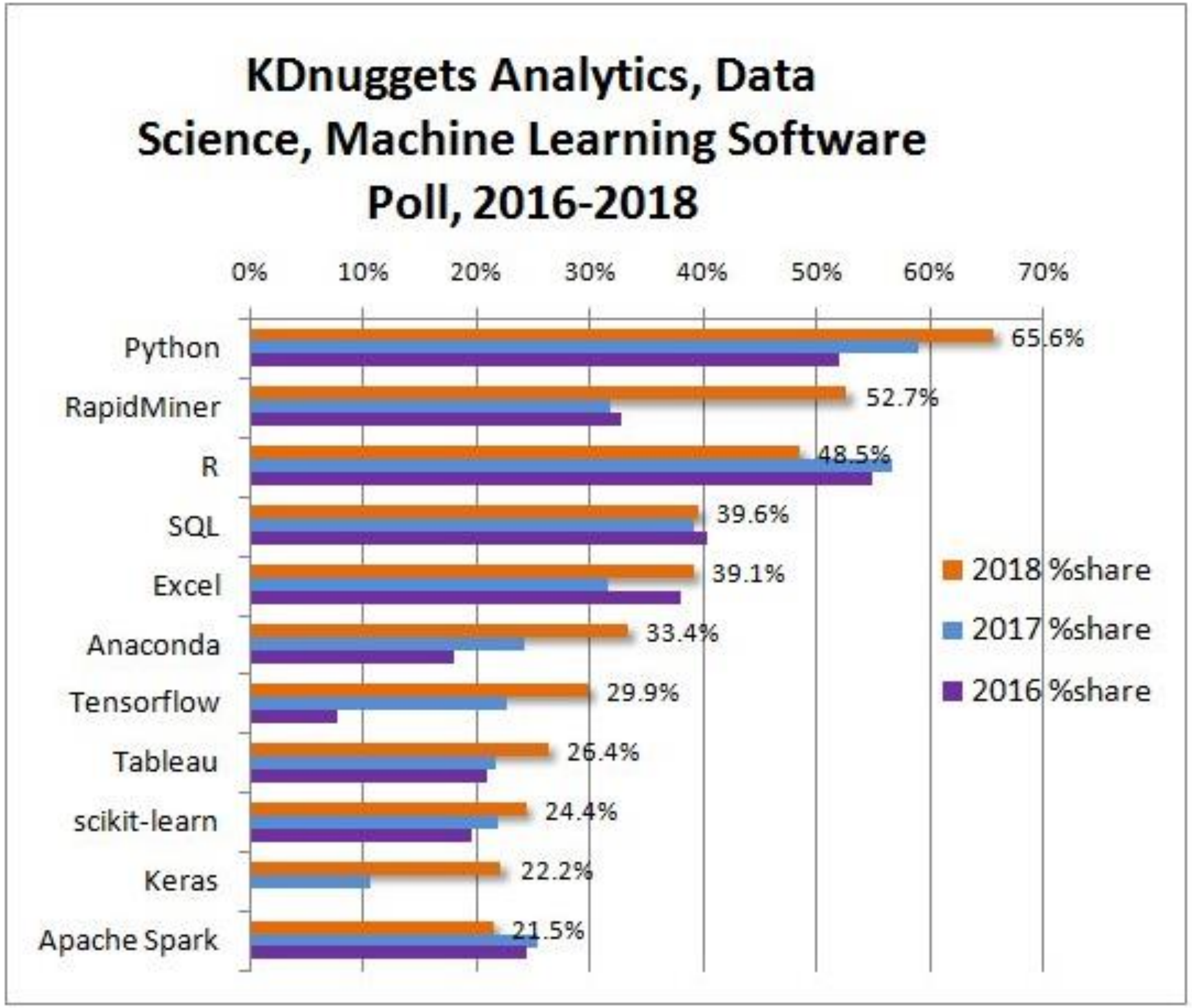
Why do we use R ?

- It has two repositories (>16.000 R Packages):
  1. Comprehensive R Archive Network (CRAN, https://cran.r-project.org/)
  2. The Bioconductor project (https://www.bioconductor.org/).
- The quality of many packages is backed through the following highly reputed academic journals: **Journal of Statistical Software, the R Journal, and Bioinformatics**.

Survey 2018

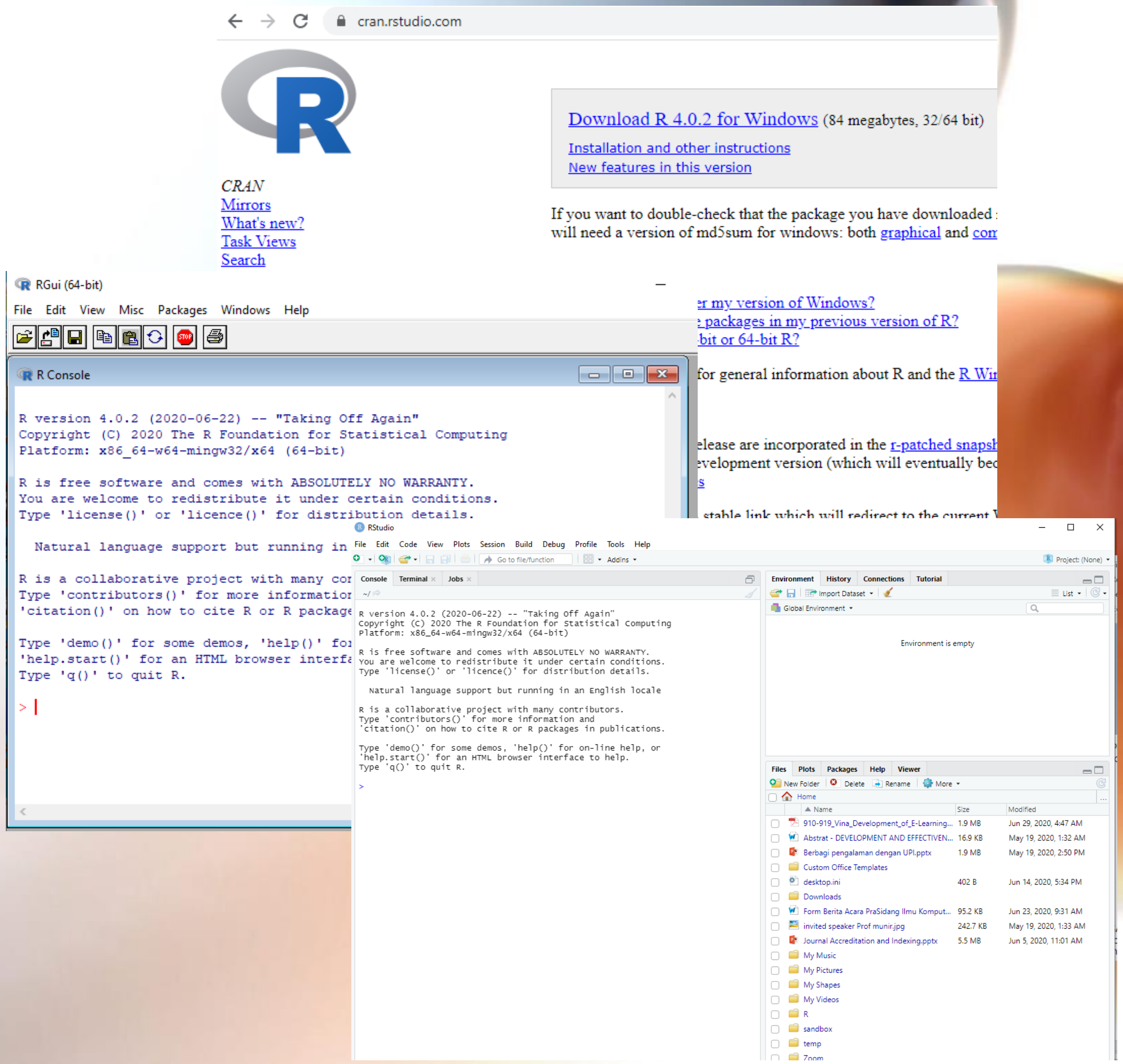
- https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

Top Analytics, Data Science, Machine Learning Tools



R Installation

- R installer: Go to https://cran.rstudio.com/
- rStudio: Go to https://www.rstudio.com/products/rstudio/download/
- Follow the instructions.
- Open R or rStudio.



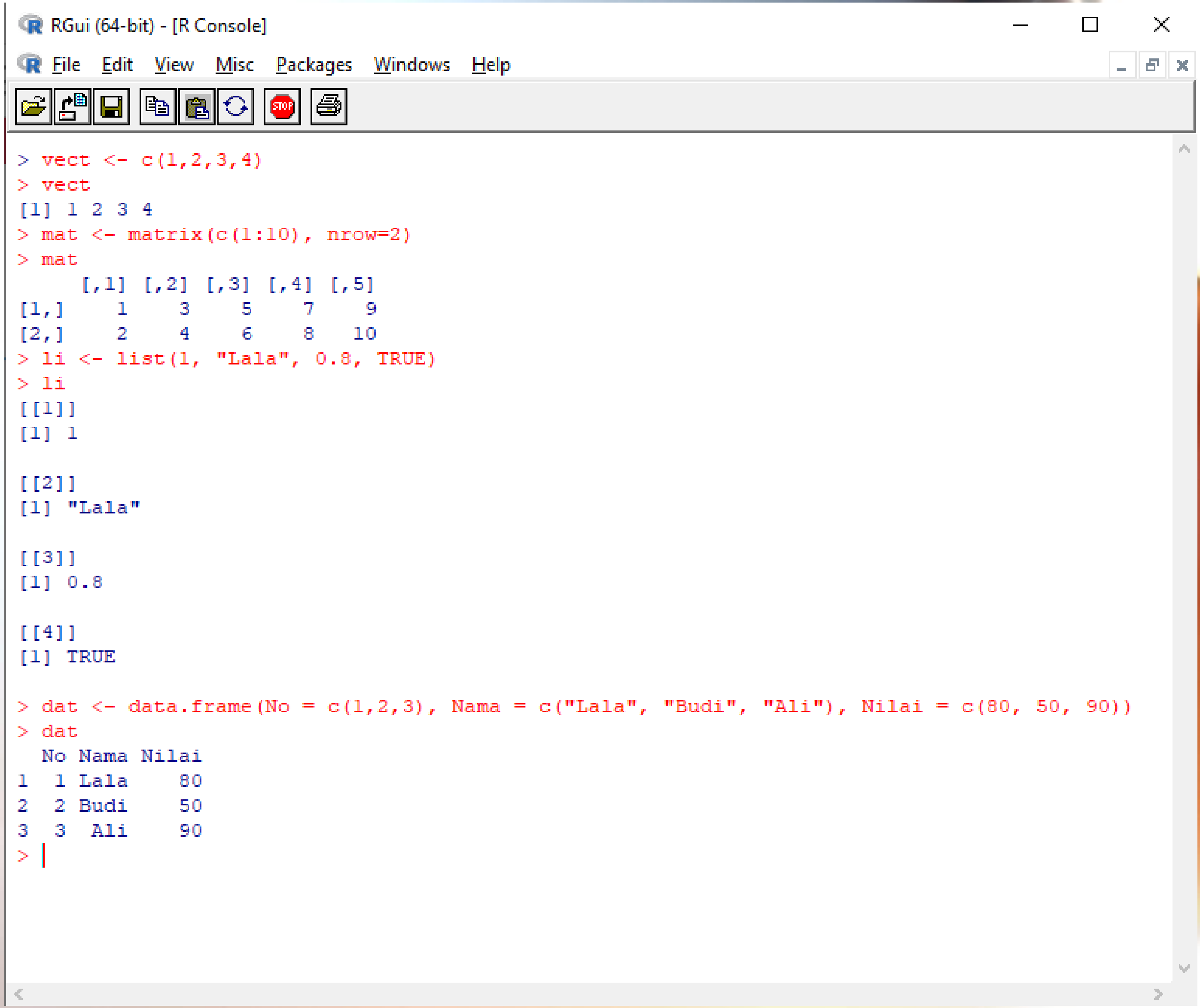
Simple Commands in R

- At the R prompt we type expressions. The <- symbol is the assignment operator.

```
> x <- 1
> print(x)
[1] 1
> x
[1] 1
> msg <- "hello"
> msg
[1] "hello“
> x <- c(0, 2, 5, 8, 9)
> sum(x)/length(x)
[1] 4.8
> mean(x)
[1] 4.8
```

R Objects

- R has 5 basic or atomic classes of objects:
  - character,
  - numeric (real numbers),
  - integer,
  - complex,
  - logical (True/False).
- Vector, Matrix, List, Data Frame



Functions in R

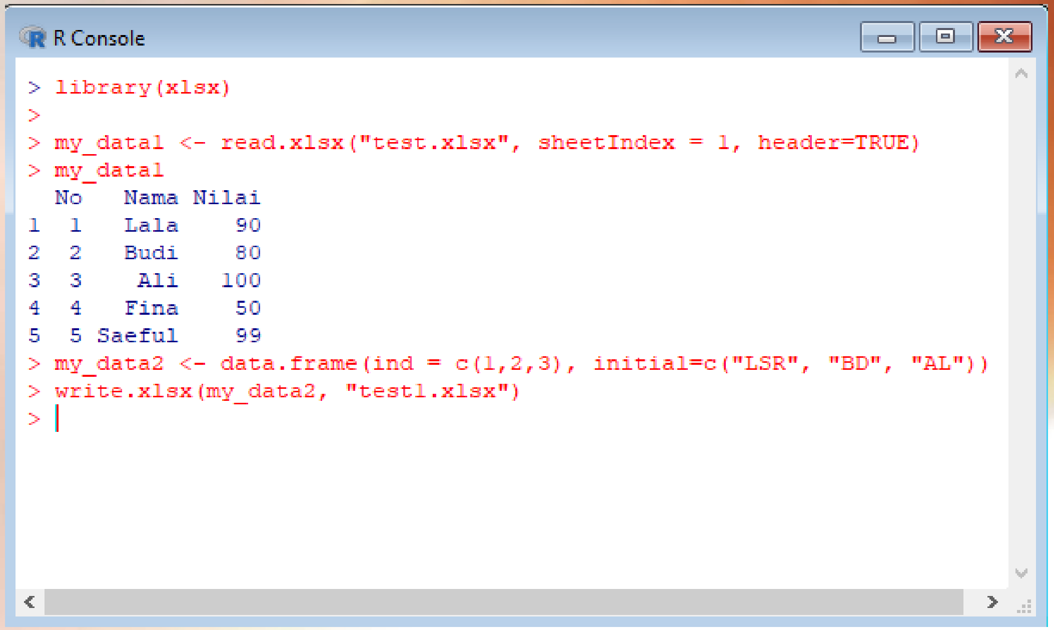
RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

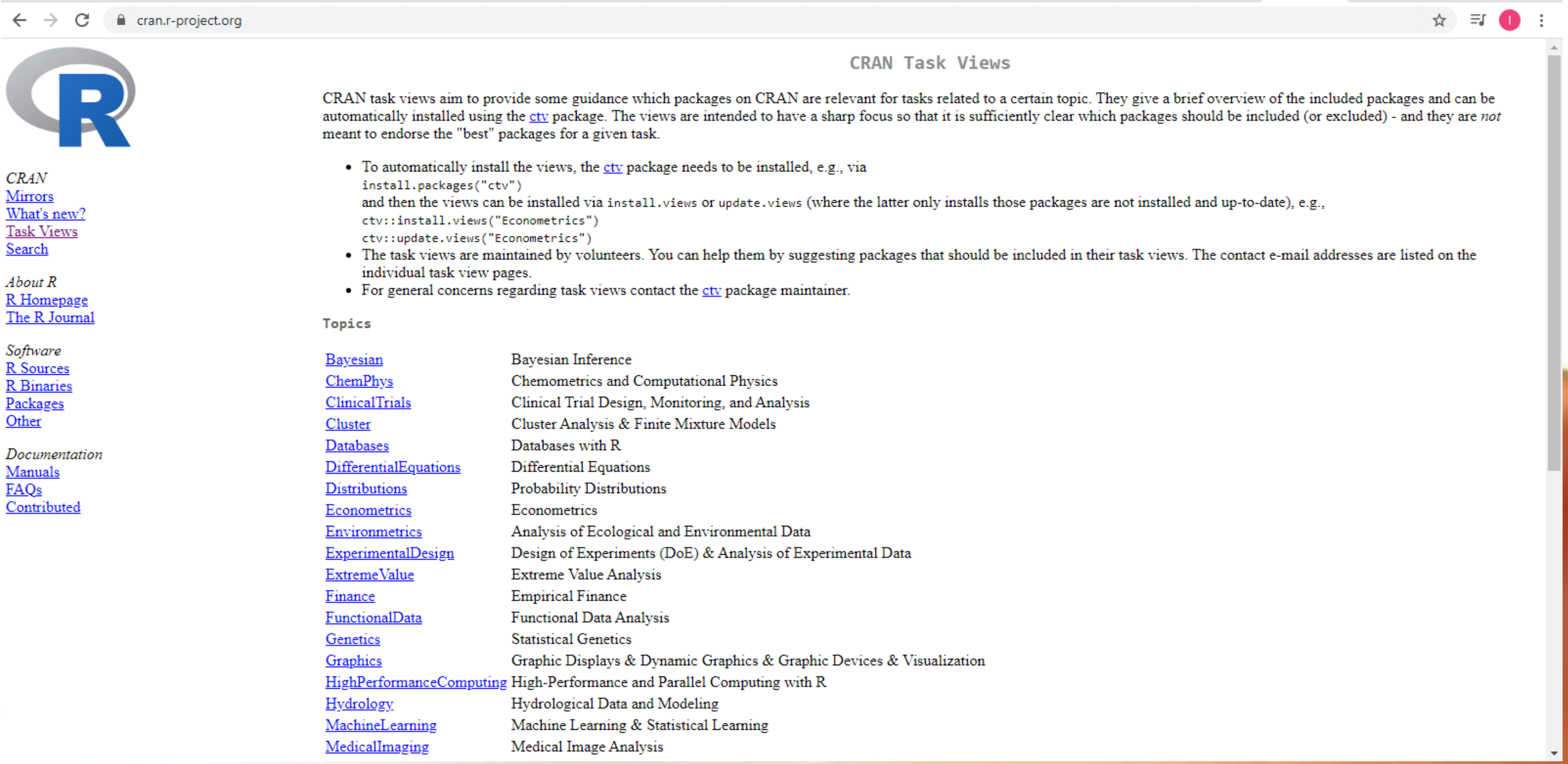
```
> area <- function(panjang, lebar){
+ return(panjang*lebar)}
> area
function(panjang, lebar){
return(panjang*lebar)}
> area(10,5)
[1] 50
> f <- function(num){
+ for (i in seq_len(num)){
+ cat("Hello, world!\n")
+ }
+ }
> f(3)
Hello, world!
Hello, world!
Hello, world!
> 
```

Data from/to a file

```
> install.packages("xlsx")
> library(xlsx)

> my_data1 <- read.xlsx("test.xlsx", sheetIndex =
1, header=TRUE)
> my_data2 <- data.frame(ind = c(1,2,3),
initial=c("LSR", "BD", "AL"))
> write.xlsx(my_data2, "test1.xlsx")
```



Install R Package



- R Package == Software Library

```
> install.packages("car")
> install.packages(c("car", "MASS"))
> library(car)
```

Applied Statistics di R

- Data Visualization di R
- Descriptive Statistics di R
- Inference Statistics di R

Data Visualization in R

- Packages for visualization: “**ggplot2**”, “**tidyverse**”.

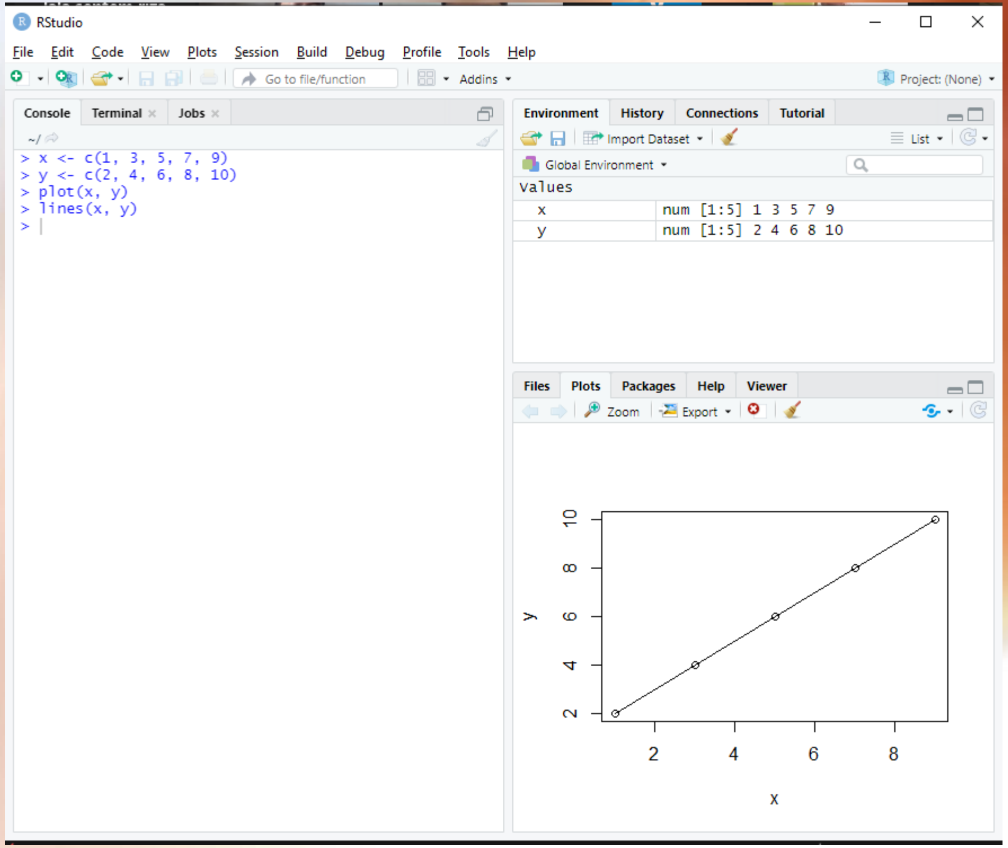
```
> install.packages(“tidyverse”)
```

- To make it available for use, we code:

```
> library(tidyverse)
```

- R Package: plot()

```
> x <- c(1, 3, 5, 7, 9)
> y <- c(2, 4, 6, 8, 10)
> plot(x, y)
> lines(x, y)
```

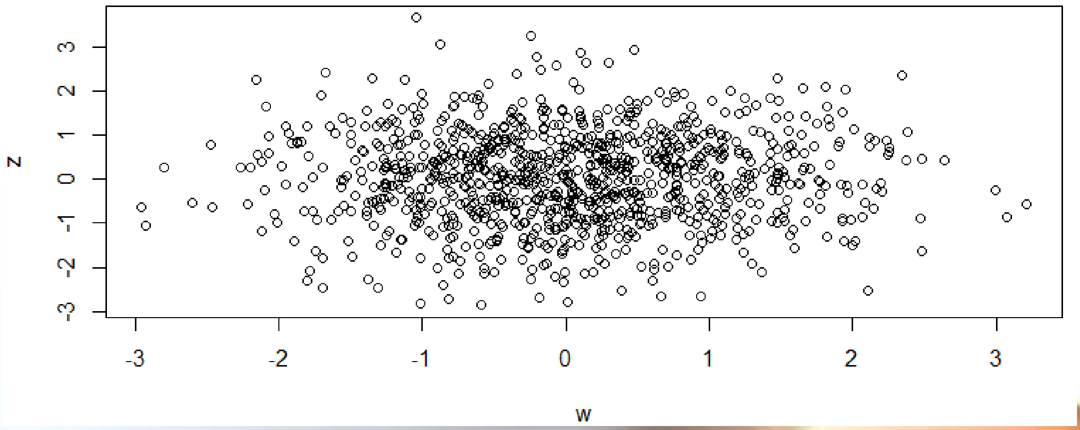


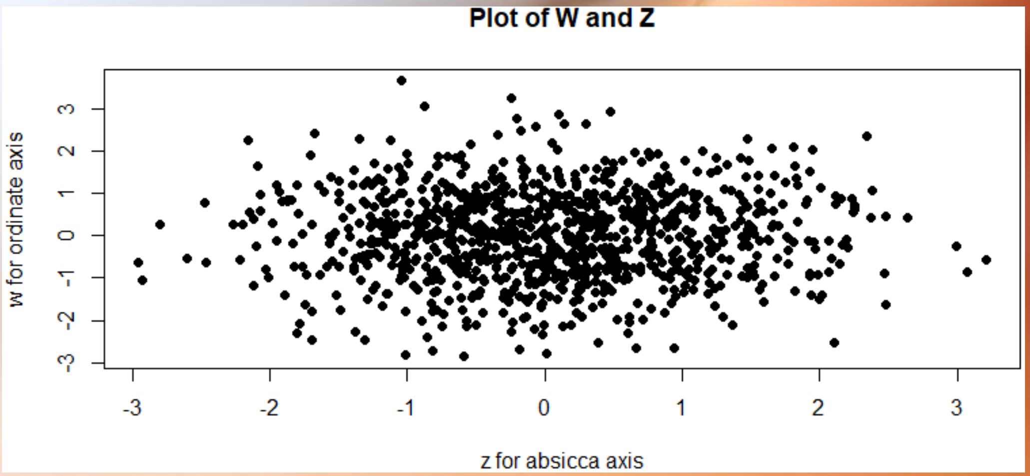
Plotting

```
> z = rnorm(1000)
> w = rnorm(1000)
> plot(w, z)
> plot(w, z, main = "Plot of W and Z",
xlab="z for absicca axis", ylab="w for
ordinate axis", pch = 19)
```

```
> pdf(file = "plot1.pdf", width = 12, height
= 17, family = "Helvetica")
> plot(w, z)
> dev.off()
```

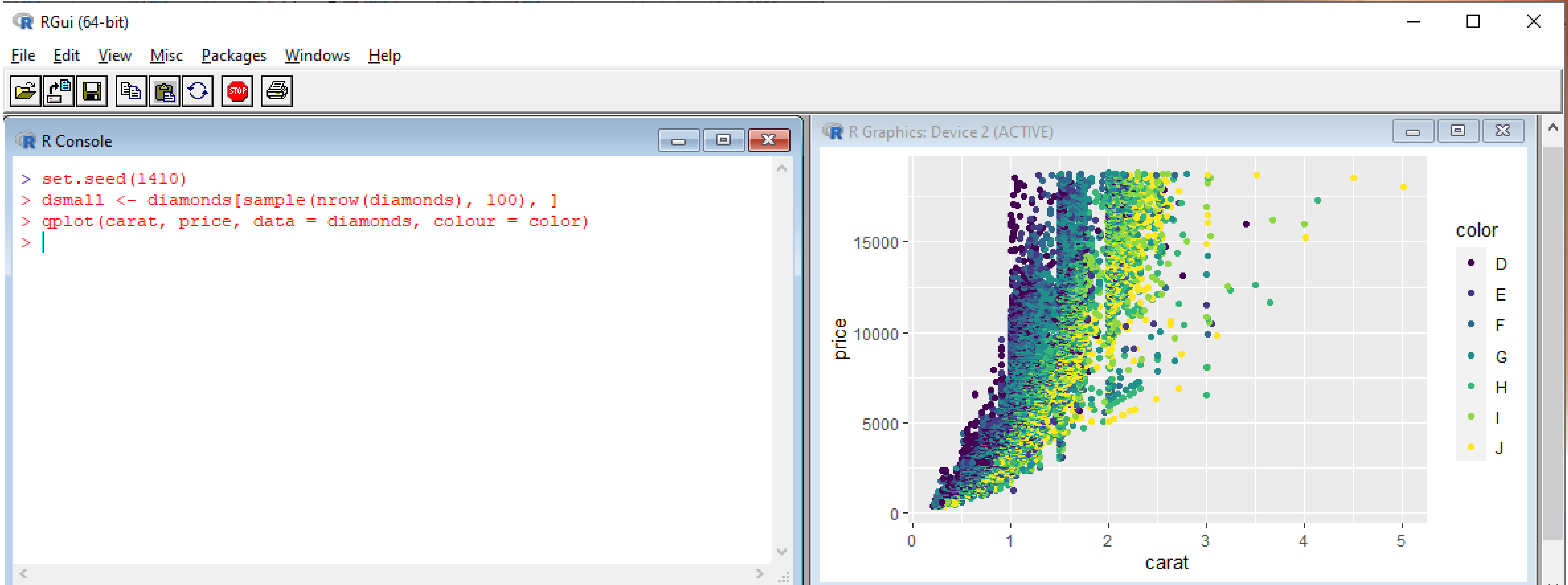
```
> tiff("Plot3.tiff", width = 4, height = 4,
units = 'in', res = 300)
> plot(w, z)
> dev.off()
```





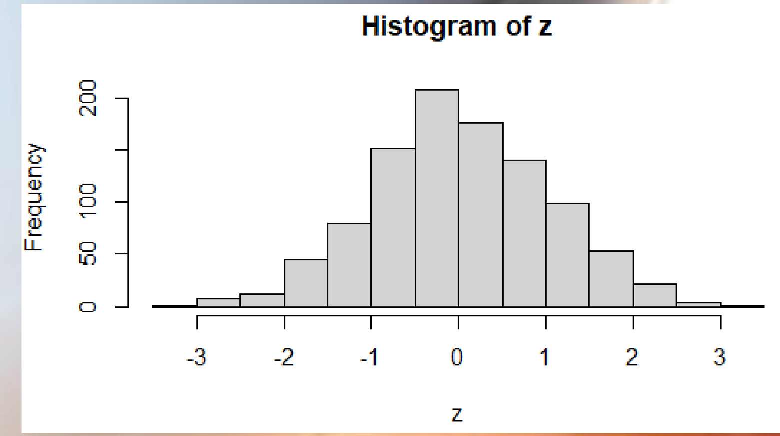
Advance Plot: qplot

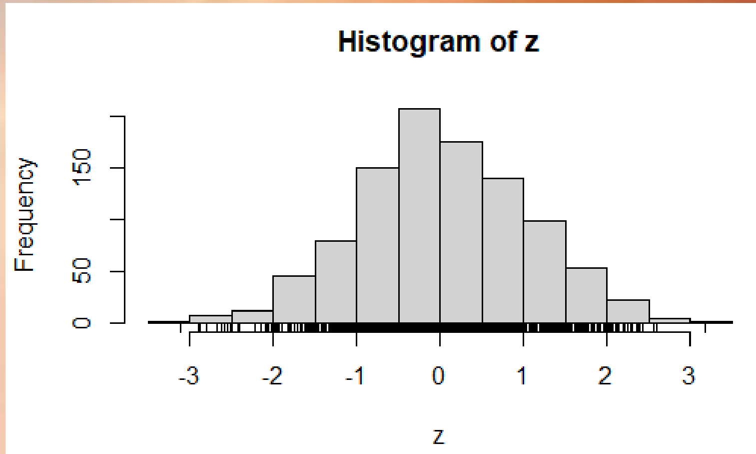
```
> library(ggplot2)
> set.seed(1410)
> dsmall <- diamonds[sample(nrow(diamonds), 100), ]
> qplot(carat, price, data = diamonds, colour = color)
```



Histogram

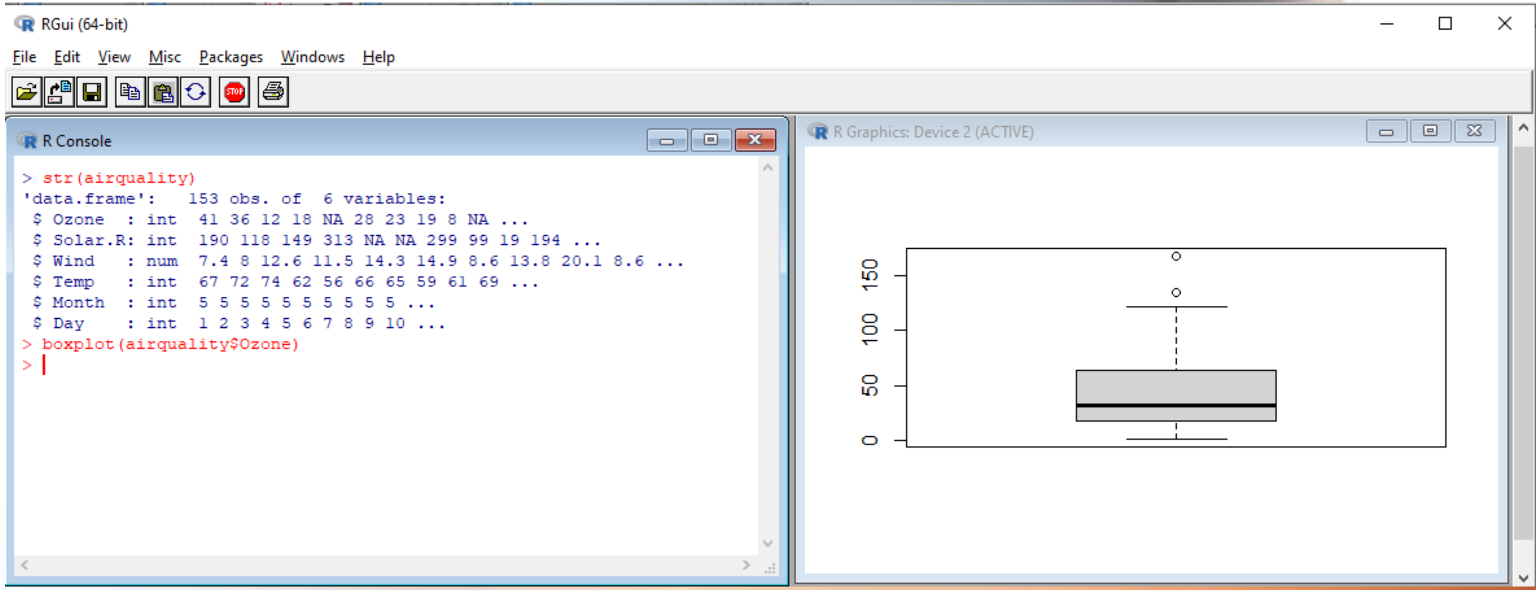
```
> z = rnorm(1000)
> w = rnorm(1000)

> hist(z)
> rug(z)
> qplot(z, geom = "histogram")
```

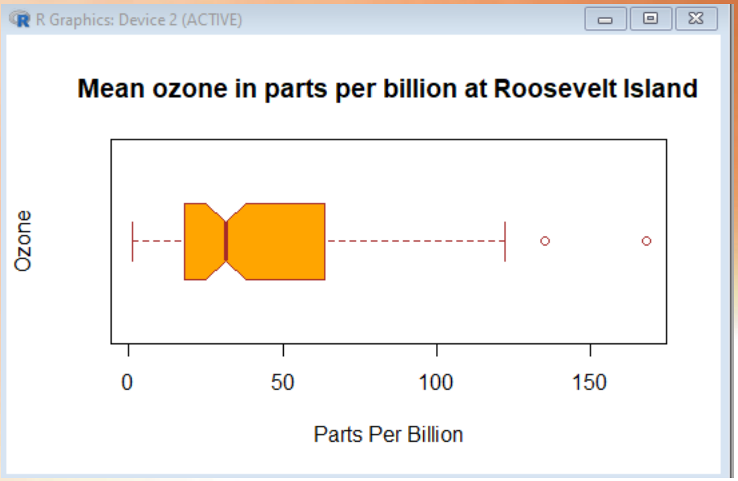




Box Plot

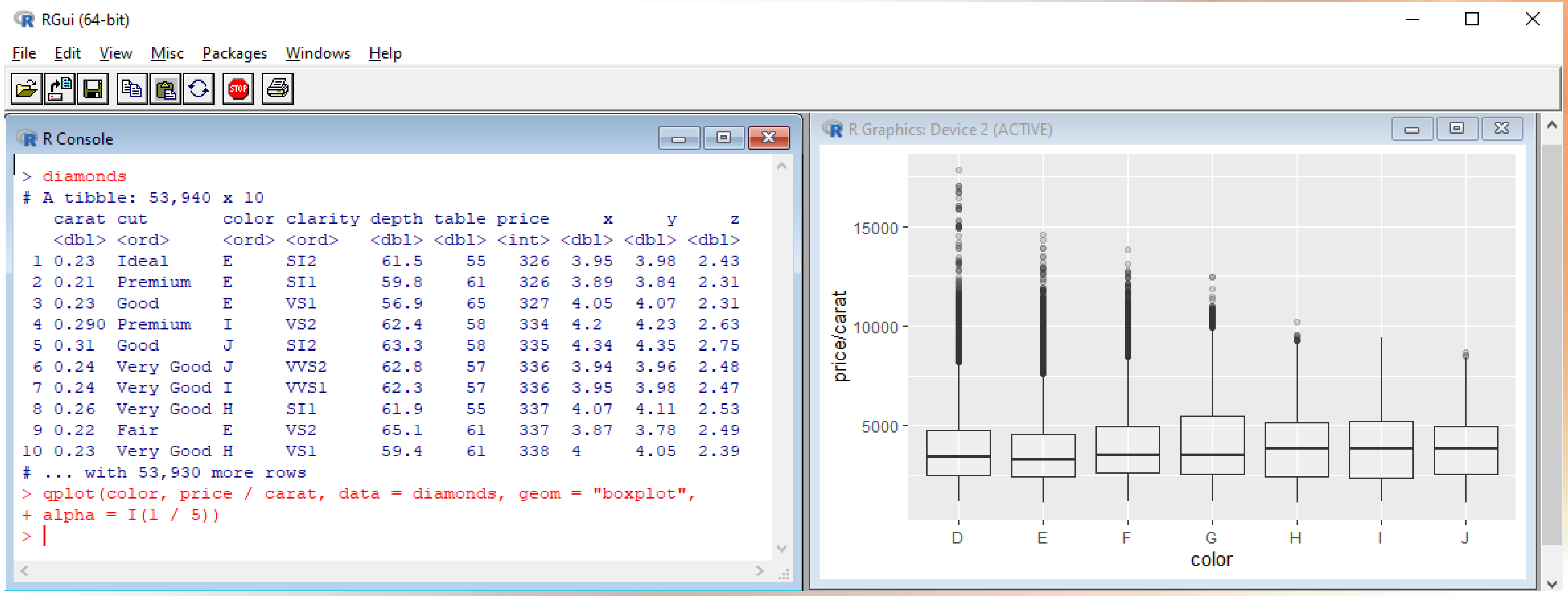
```
> str(airquality)
> boxplot(airquality$Ozone)
> boxplot(airquality$Ozone,
main = "Mean ozone in parts
per billion at Roosevelt
Island",
xlab = "Parts Per Billion",
ylab = "Ozone",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```





Advance Boxplot

```
> library(ggplot2)
> diamonds
> qplot(color, price / carat, data = diamonds, geom =
"boxplot",alpha=I(1/5))
```



Descriptive Statistics

- It is aimed at summarizing, describing and presenting a series of values or a dataset.
- Two types:
  1. Location measures (mean, median, mode)
  2. Dispersion measures (variance, std deviation, quartile)

R Console

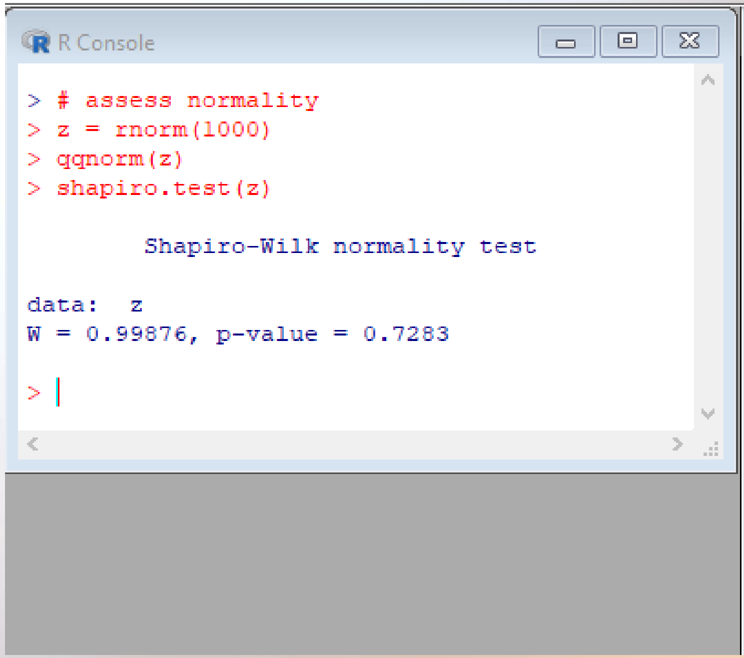
```
> dat <- iris
> head(dat)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> min(dat$Sepal.Length)
[1] 4.3
> rng <- range(dat$Sepal.Length)
> mean(dat$Sepal.Length)
[1] 5.843333
> median(dat$Sepal.Length)
[1] 5.8
> quantile(dat$Sepal.Length, 0.25) # first quartile
25%
5.1
> sd(dat$Sepal.Length)
[1] 0.8280661
> var(dat$Sepal.Length)
[1] 0.6856935
> summary(dat)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

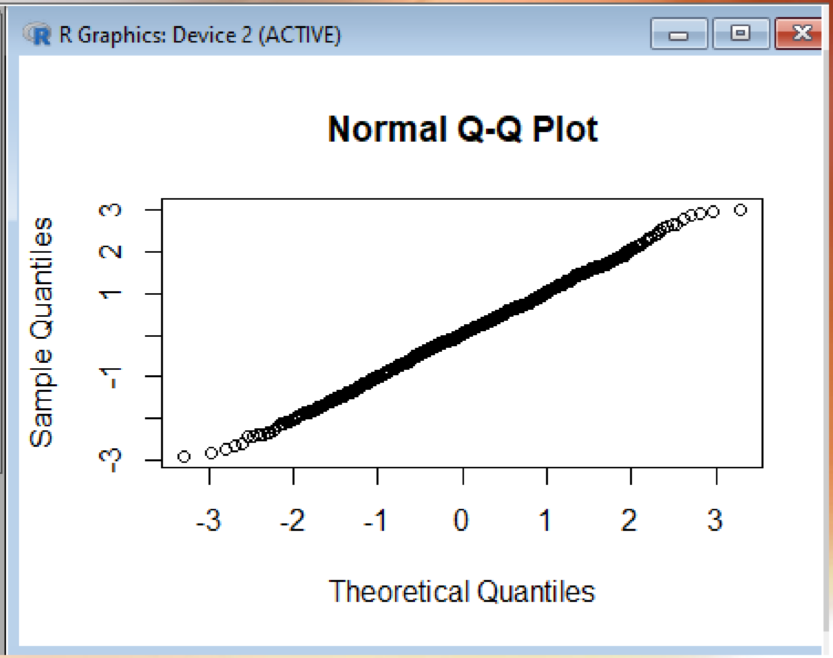
Statistical Inference

- Statistical inference is the process of using data analysis to deduce properties of an underlying distribution of probability.

Assess Normality

```
> z = rnorm(1000)
> qqnorm(z)
> shapiro.test(z)
```



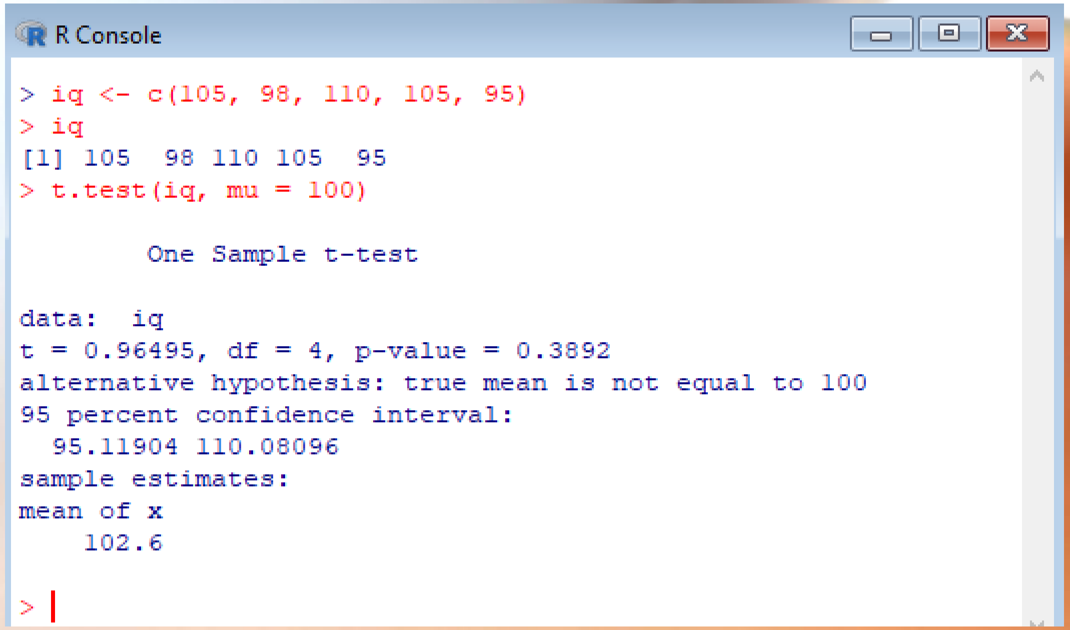


- The null hypothesis is that the data are normally distributed.
- Since p is quite high (> 0.05), we **fail to reject** the null hypothesis

Test of Mean Differences

- t-Tests for One Sample

```
> iq <- c(105, 98, 110, 105, 95)
> t.test(iq, mu = 100)
```



Note:

- The **null hypothesis** was that the sample was drawn from a population with mean equal to 100.
- Since p = 0.3892 is relatively large (certainly not smaller than some conventional level such as 0.05 or 0.01), we **fail to reject** the null hypothesis.

- **Two-Sample *t*-Test:** A two-sample test is used to evaluate the null hypothesis that two population means are **equal, or equivalently**, that both samples were selected from the same population.

```
> grade.0 <- c(30, 25, 59, 42, 31)
> grade.1 <- c(140, 90, 95, 170, 120)
> t.test(grade.0, grade.1)
```

R Console

```
> grade.0 <- c(30, 25, 59, 42, 31)
> grade.1 <- c(140, 90, 95, 170, 120)
> t.test(grade.0, grade.1)

        Welch Two Sample t-test

data:  grade.0 and grade.1
t = -5.3515, df = 5.3094, p-value = 0.002549
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -126.00773  -45.19227
sample estimates:
mean of x mean of y 
     37.4     123.0 
```

Since p-value is equal to 0.002549, we have evidence to **reject the null hypothesis**

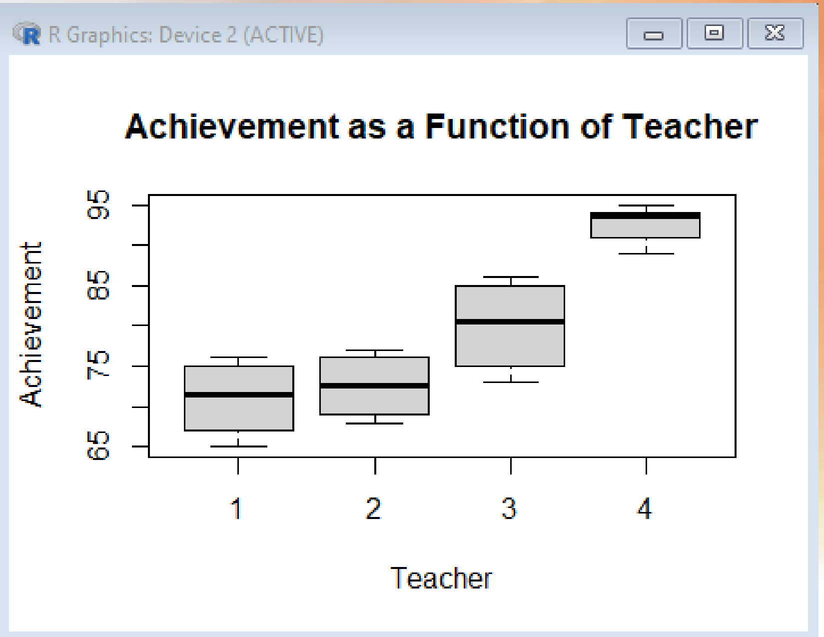
Analysis of Variance (ANOVA)

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: $\mu_1 \neq \mu_2$, $\mu_3 \neq \mu_4$, etc

Table 6.1 Achievement as a function of teacher.

Teacher			
1	2	3	4
70	69	85	95
67	68	86	94
65	70	85	89
75	76	76	94
76	77	75	93
73	75	73	91
$M = 71.00$	$M = 72.5$	$M = 80.0$	$M = 92.67$



```r
> achiev <- read.table("achiev.txt", header = T)
> attach(achiev)
> boxplot(ac ~ teach, data = achiev,
main="Achievement as a Function of Teacher",
xlab = "Teacher", ylab = "Achievement")
> shapiro.test(ac)
> library(FSA)
> f.teach <- factor(teach)
> hist(ac~f.teach, data = achiev)
> fligner.test(ac~f.teach, data = achiev)
> aggregate(ac ~ f.teach, FUN = var)
> aggregate(ac ~ f.teach, FUN = mean)

> anova.fit <- aov(ac ~ f.teach, data = achiev)
> summary(anova.fit)
```

```
> boxplot(ac ~ teach, data = achiev, main="Achievement as a Function o$
> shapiro.test(ac)

        Shapiro-Wilk normality test

data:  ac
W = 0.90565, p-value = 0.02842

>
> library(FSA)
> f.teach <- factor(teach)
> hist(ac~f.teach, data = achiev)
> fligner.test(ac~f.teach, data = achiev)

        Fligner-Killeen test of homogeneity of variances

data:  ac by f.teach
Fligner-Killeen:med chi-squared = 10.813, df = 3, p-value =
0.01278

> aggregate(ac ~ f.teach, FUN = var)
  f.teach        ac
1       1 19.600000
2       2 15.500000
3       3 35.200000
4       4  5.066667
> aggregate(ac ~ f.teach, FUN = mean)
  f.teach       ac
1       1 71.00000
2       2 72.50000
3       3 80.00000
4       4 92.66667
>
> anova.fit <- aov(ac ~ f.teach, data = achiev)
> summary(anova.fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
f.teach      3 1764.1   588.0   31.21 9.68e-08 ***
Residuals   20  376.8    18.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value equal to 9.68e-08, So We **reject** the null hypothesis that population achievement means are equal across teachers

Wawasan: Penelitian dengan R

Research using R language: R Packages

1. **Riza, L. S.**, Handian, D., Megasari, R., Abdullah, A. G., Nandiyanto, A. B. D., & Nazir, S. (2018). Development of **R package** and experimental analysis on prediction of the CO2 compressibility factor using **gradient descent**. *Journal of Engineering Science and Technology*, *13*(8), 2342-2351.
2. **Riza, L. S.**, & Nugroho, E. P. (2018). **MetaheuristicOpt**: An R Package for Optimisation Based on Meta-Heuristics Algorithms. *Pertanika Journal of Science & Technology*, *26*(3).
3. **Riza, L. S.**, Bergmeir, C. N., Herrera Triguero, F., & Benítez Sánchez, J. M. (2015). **frbs**: Fuzzy rule-based systems for classification and regression in R. American Statistical Association.
4. **Riza, L. S.**, Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Śle, D., & Benítez, J. M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "**RoughSets**". *Information Sciences*, *287*, 68-89.

Research using R language: Implementations/Computational Models

1. Mediayani, M., Wibisono, Y., **Riza, L. S.,** & Pérez, A. R. (2019). Determining **trending topics in twitter** with a data-streaming method in R. *Indonesian Journal of Science and Technology*, *4*(1), 148-157.
2. **Riza, L. S.,** Putra, B., Wihardi, Y., & Paramita, B. (2019). **Data to text** for generating information of weather and air quality in the R programming language. *Journal of Engineering Science and Technology*, *14*(1), 498-508.
3. **Riza, L. S.,** Rachmat, A. B., Munir, T. H., & Nazir, S. (2019). **Genomic Repeat Detection** Using the Knuth-Morris-Pratt Algorithm on R High-Performance-Computing Package. *Int. J. Advance Soft Compu. Appl*, *11*(1).
4. **Riza, L. S.,** Dhiba, T. F., Setiawan, W., Hidayat, T., & Fahsi, M. (2019). Parallel random projection using R high performance computing for **planted motif search**. *TELKOMNIKA*, *17*(3), 1352-1359.
5. **Riza, L. S.,** Utama, J. A., Putra, S. M., Simatupang, F. M., & Nugroho, E. P. (2018). Parallel Exponential Smoothing Using the Bootstrap Method in R for Forecasting **Asteroid's Orbital Elements**. *Pertanika Journal of Science & Technology*, *26*(1).
6. Nazir, S., Shahzad, S., & **Riza, L. S.** (2017). **Birthmark-based software** classification using rough sets. *Arabian Journal for Science and Engineering*, *42*(2), 859-871.

References

Cresswell, J. W. (2005). Educational research: Planning, conducting and evaluating qualitative and quantitative research. *Upper Saddle River, NJ: Merrill & Prentice Hall*.

Denis, D. J. (2020). *Univariate, Bivariate, and Multivariate Statistics Using R: Quantitative Tools for Data Analysis and Data Science*. John Wiley & Sons.

Tattar, P. N., Ramaiah, S., & Manjunath, B. G. (2016). *A Course in Statistics with R*. John Wiley & Sons.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. springer.