Output development using R & R markdown

Ziv Shkedy and Thi Huyen Nguyen
Hasselt University

Foundations for inference
Ha Noi
03/03/25-07/03/25

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

twitter @erbiostat

UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

The `airquality` data

Introduction

What do we do in this session ?

- We conduct a simple analysis for the variable Wind speed in the `airquality` data:
  - Summary statistics.
  - Graphical display: histogram.
  - Confidence interval.
  - Test of hypothesis.
- Focus:
  - How to conduct the analysis in R ?
  - How to produce an output ?
  - How to combine text and software output in the same document ?

UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

The `airquality` data

R Program: Case studies_Prog1_V1.R

Part 1

The dataset

The `airquality` data in R

```
> dim(airquality)
[1] 153   6
> names(airquality)
 [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

The R object for the data: 153 observations and 6 variables.

Variables names:

Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island.

Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport.

The `airquality` data in R

```
> head(airquality)

   zone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

The `airquality` data

Part 2

Analysis plan + analysis in R

Analysis plan

- Response: wind speed.
- Analysis:
  - Calculate mean and SD.
  - Construct 95% C.I. for the population mean.
  - Test the hypothesis for the population mean.

Case 1: using a $N(0,1)$

If $X \sim N(\mu, \sigma^2)$

then: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

And $Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

1. X has a normal distribution with unknown μ and known σ^2.

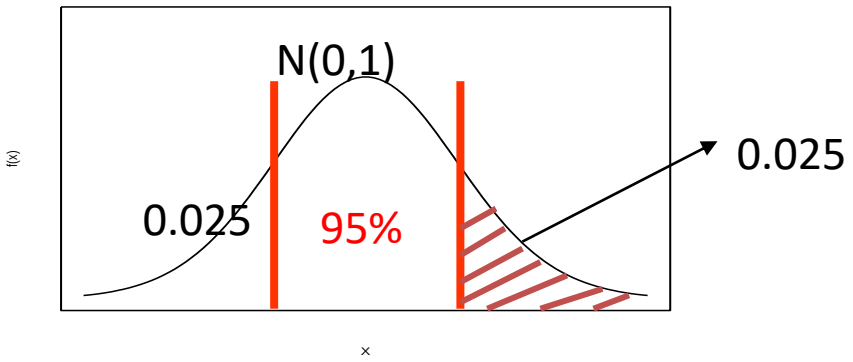
CI for case 1

Stap 1: example, choose 1-α = 0.95

Stap 2: Case 1, thus:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

Stap 3: critical point:



CI for case 1

Step 4: calculate the point estimator : $\bar{\chi}$

Step 5: calculate the CI

For this, we know:

$$P\left(-1.96 \le \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \le 1.96\right) = 0.95$$

or, after the conversion of the formula:

$$P\left(\overline{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}} \le \mu \le \overline{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

CI for case 1

$$P\left(\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}} \le \mu \le \bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

L R

$$P(L \le \mu \le R) = 1 - \alpha$$

So, a (1-α) CI for μ is :

$$\left[\bar{x} - z\sqrt{\frac{\sigma^2}{n}}, \bar{x} + z\sqrt{\frac{\sigma^2}{n}}\right]$$

Two sided hypothesis testing

The mean under H_0 is not equal to the mean under H_1 :

$$H_0 : \mu = \mu_{H_0}$$

null hypothesis

$$H_1 : \mu \neq \mu_{H_0}$$

alternative hypothesis

two sided test problem

Test statistic

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

R functions for the analysis

- `mean()`
- `var()`
- `z.test()`

R program for the analysis: `Airquality.R`

R program

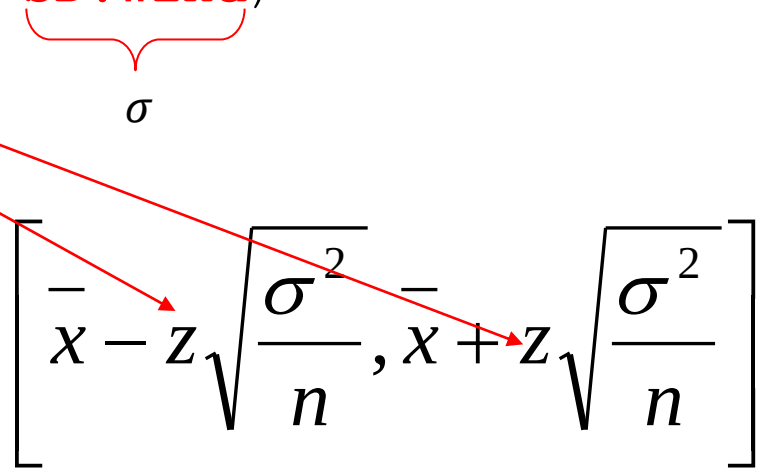
```
wind<-airquality$Wind
M.wind<-mean(wind)
SD.wind<-sqrt(var(wind))
```

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

R program

```
wind=na.omit(airquality$Wind)
library(TeachingDemos)
z.test(wind,sd=SD.wind)
```



R program

```
z.test(wind, SD.wind, mu=9)
```

σ

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

$H_0: \mu = 9$
$H_1: \mu \neq 9$

Output

```
> wind<-airquality$Wind
> M.wind<-mean(wind)
> SD.wind<-sqrt(var(wind))
> M.wind
[1] 9.957516
> SD.wind
[1] 3.523001
```

Output (C.I)

```
> wind=na.omit(airquality$Wind)
> library(TeachingDemos)
> z.test(wind,sd=SD.wind)

        One Sample z-test

data:  wind
z = 34.961, n = 153.00000, Std. Dev. = 3.52300, Std. Dev. of the sample
mean = 0.28482, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.399284 10.515749
sample estimates:
mean of wind
    9.957516
```

$$\left[\overline{x} - z\sqrt{\frac{\sigma^2}{n}}, \overline{x} + z\sqrt{\frac{\sigma^2}{n}}\right] = [9.399, 10.515]$$

Output (test of hypothesis)

```
> z.test(wind, SD.wind, mu=9)

        One Sample z-test

data:  wind
z = 3.3619, n = 153.00000, Std. Dev. = 3.52300, Std. Dev. of the sample
mean = 0.28482, p-value = 0.0007742
alternative hypothesis: true mean is not equal to 9
95 percent confidence interval:
  9.399284 10.515749
sample estimates:
mean of wind
    9.957516
```

$H_0: \mu = 9$

$H_1: \mu \neq 9$

A two sided test

Case 3: using a $t_{(n-1)}$ distribution

An estimate for the standard error of the sample mean.

$$SE = \frac{s}{\sqrt{n}}$$

s → The standard deviation.

n → The sample size.

If the sample comes from a population whose distribution is normal and the variance is unknown, and the sample size is small , then :

$$T_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$$

A Student t-distribution with (n-1) degrees of freedom is denoted by t (n-1).

R functions for the analysis

- In our case, n=153 (n>30 !!!) but, for illustration, we will use t distribution for the test.

- `mean()`
- `var()`
- `t.test()`

R program

```
t.test(wind,mu=9)
```

$$t_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}} \sim t_{(n-1)}$$

$H_0: \mu = 9$
$H_1: \mu \neq 9$

- Two sided t test with unknown variance.
- n>30.

Output (C.I)

```
> t.test(wind,mu=0)

        One Sample t-test

data:  wind
t = 34.961, df = 152, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.394804 10.520229
sample estimates:
mean of x
 9.957516
```

Not relevant

$$\left[\bar{x} - t\sqrt{\frac{s^2}{n}}, \bar{x} + t\sqrt{\frac{s^2}{n}}\right] = [9.394, 10.520]$$

Output (test of hypothesis)

```
 t.test(wind,mu=9,var.equal=TRUE)

        One Sample t-test

data:  wind
t = 3.3619, df = 152, p-value = 0.0009794
alternative hypothesis: true mean is not equal to 9
95 percent confidence interval:
  9.394804 10.520229
sample estimates:
mean of x
 9.957516
```

$H_0: \mu = 9$

$H_1: \mu \neq 9$

A two sided test

UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

The `airquality` data

Part 3

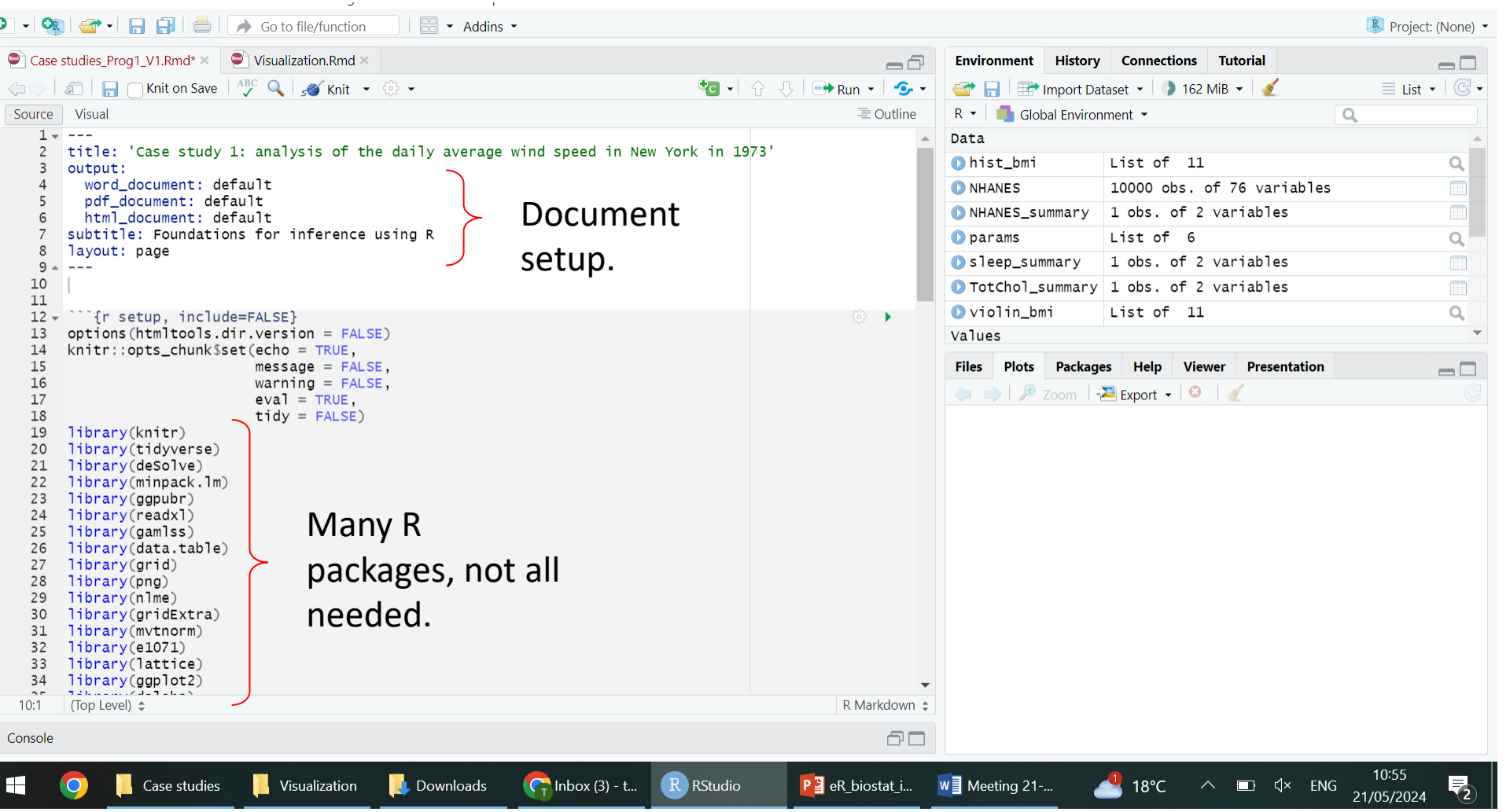
The Rmd program

R program for the analysis: Case Study_ Prog1_V1.Rmd

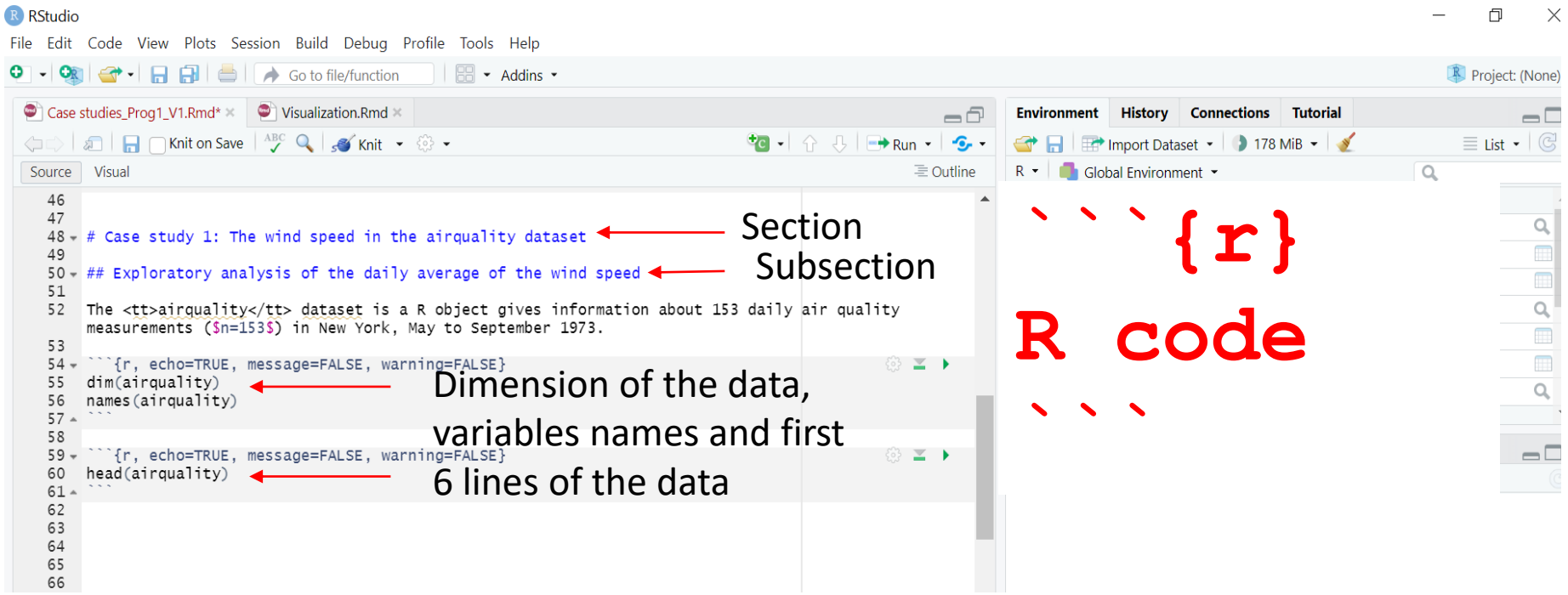
The output

- We run the R markdown file.
- Produce output format: Word document.

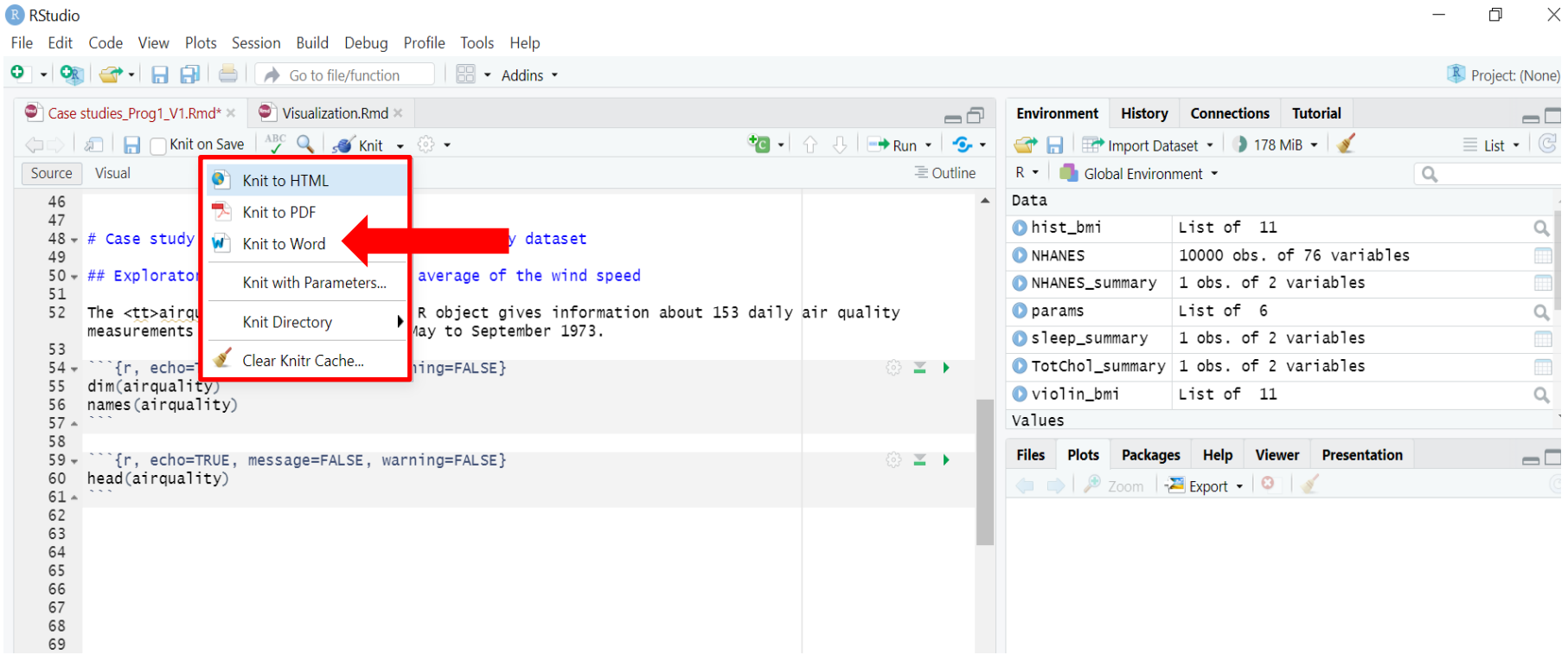
The Rmd program



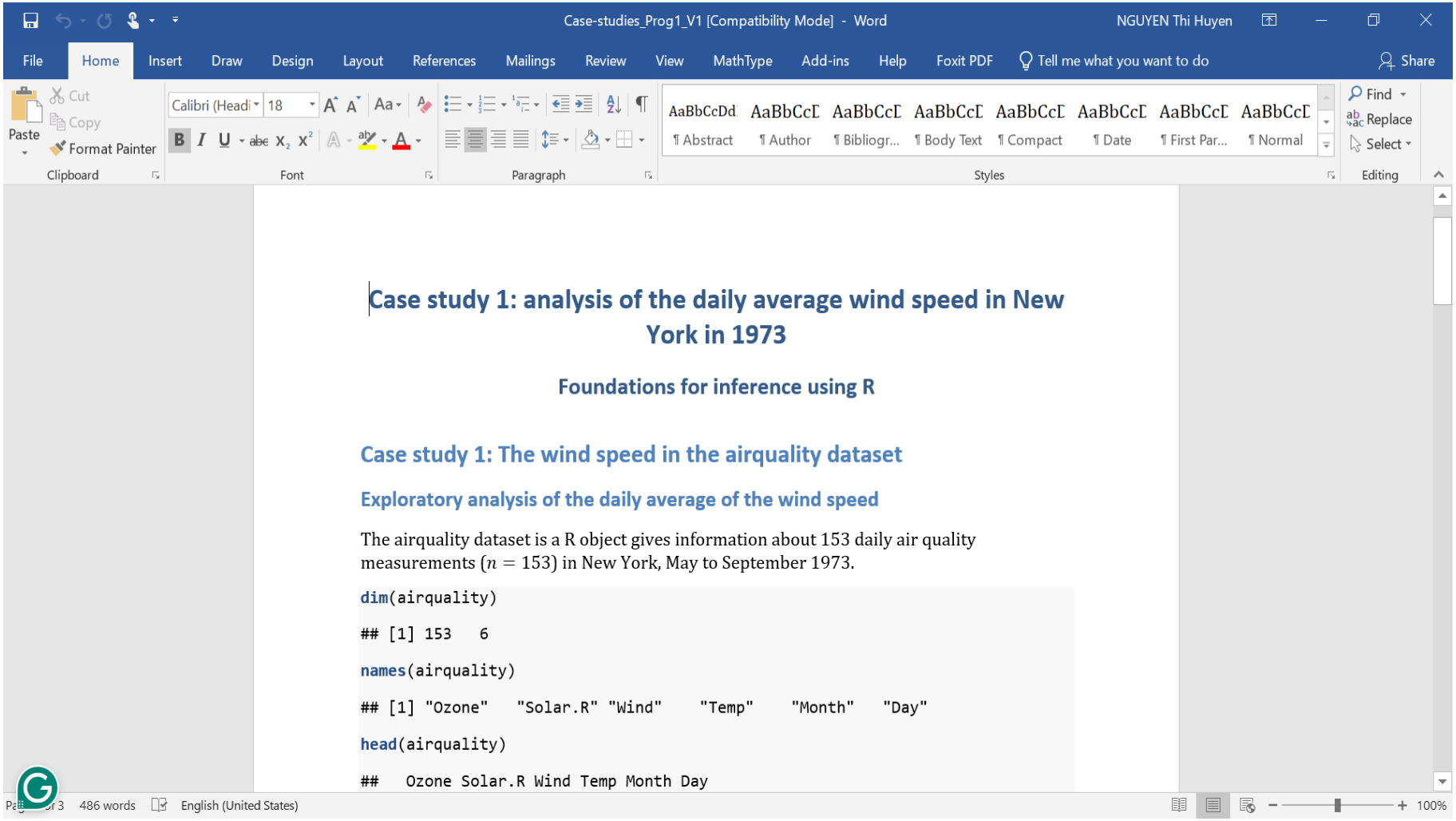
The Rmd program



Choose the output



The Word document output



UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

The `airquality` data

Part 4

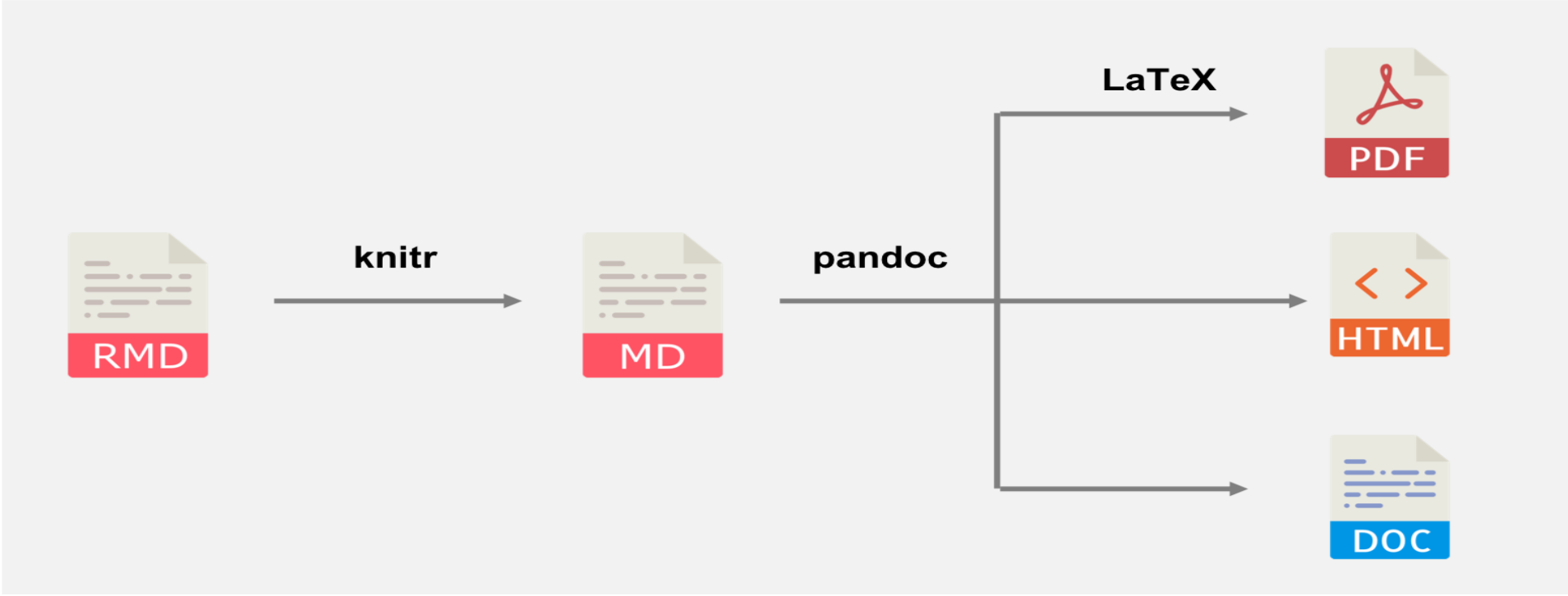
How to produce the Word output ?

Reproducible Research

- Aim: create an output in a Word document.
- Can be used to communicate the analysis' results with other people in the organization.
- Not all potential readers are interested on "how to do the analysis".
- We DO NOT aim to develop a report for the analysis but to provide a document from which the results can be seen and discuss by different people in the organization.

The Rmd file

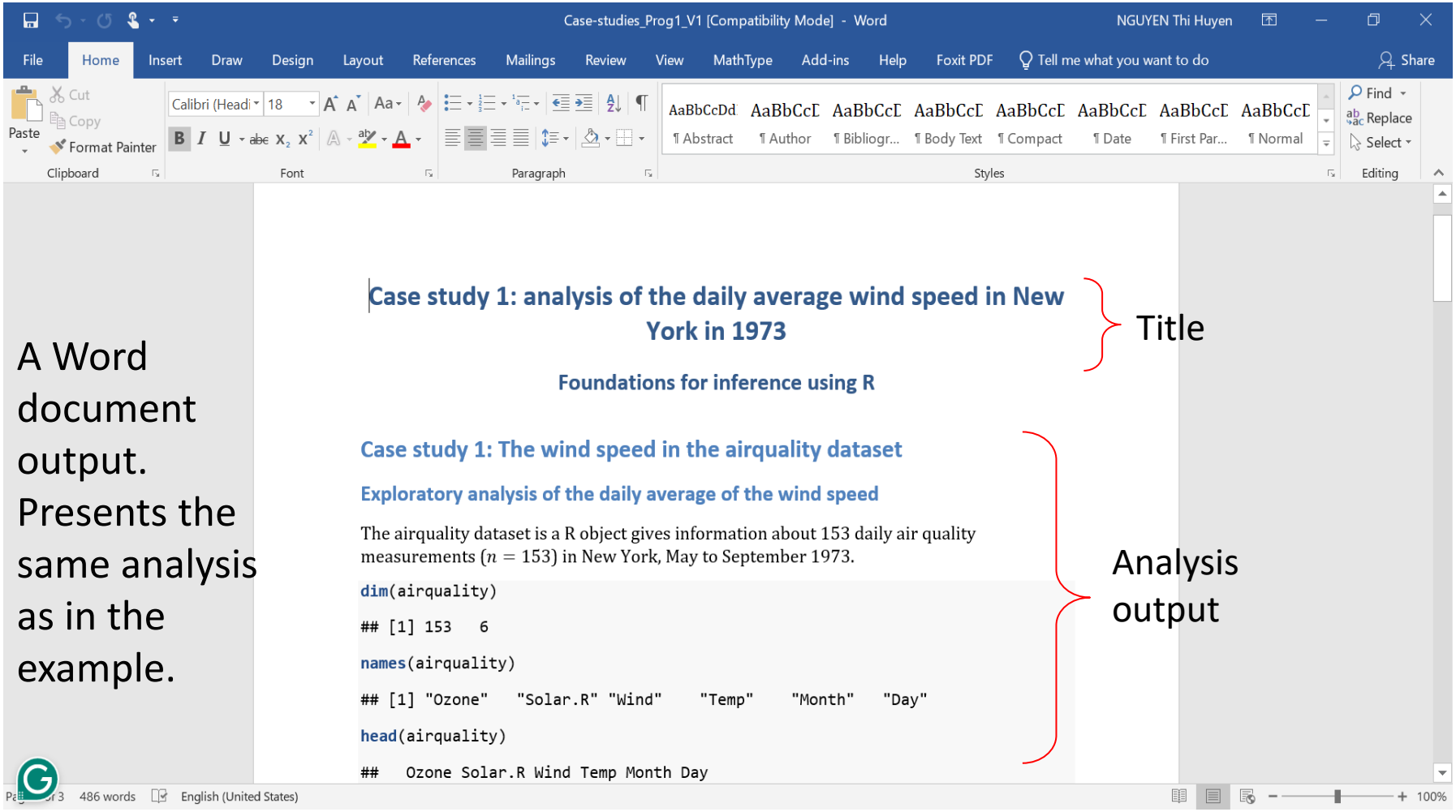
- Analyses ⇨ high quality report.
- Rmarkdown – Different dynamic and statistic formats (html, pdf, word, books, dashboard, e.t.c).



Knitr: A General-Purpose Package for Dynamic Report Generation in r. https://yihui.org/knitr/.

The Word output

- A Word document output.
- Presents the same analysis as in the example.



UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

Part 4.1:
How to set up the Word file?

The Rmd file

- We use Rmd file to
  - Conduct the analysis.
  - Set up the document.
- We use a Word file in order to
  - Present & communicate the result.

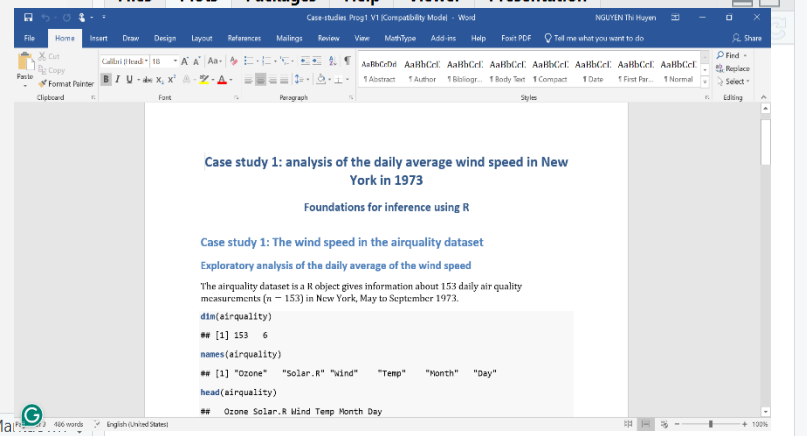
Set up the document

```
---
title: 'Case study 1: analysis of the daily average wind speed in New York in 1973'
output:
  word_document: default
  pdf_document: default
  html_document: default
subtitle: Foundations for inference using R
layout: page
---

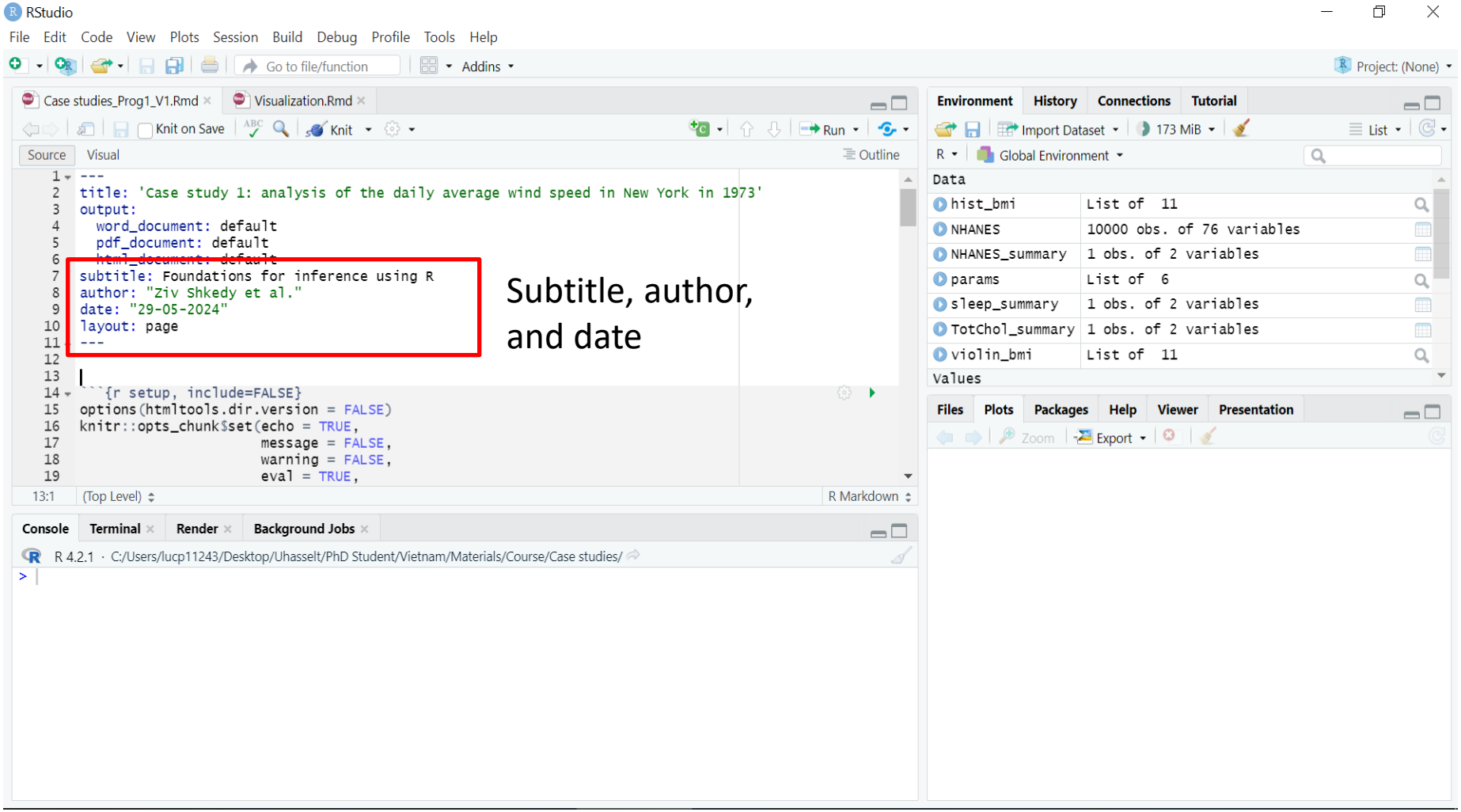
```{r setup, include=FALSE}
options(htmltools.dir.version = FALSE)
knitr::opts_chunk$set(echo = TRUE,
                      message = FALSE,
                      warning = FALSE,
                      eval = TRUE,
                      tidy = FALSE)
library(knitr)
library(tidyverse)
library(deSolve)
library(minpack.lm)
library(ggpubr)
library(readxl)
library(gamlss)
library(data.table)
library(grid)
library(png)
library(nlme)
library(gridExtra)
library(mvtnorm)
library(e1071)
library(lattice)
library(ggplot2)
```

Set up the Word document:

Word_document: default

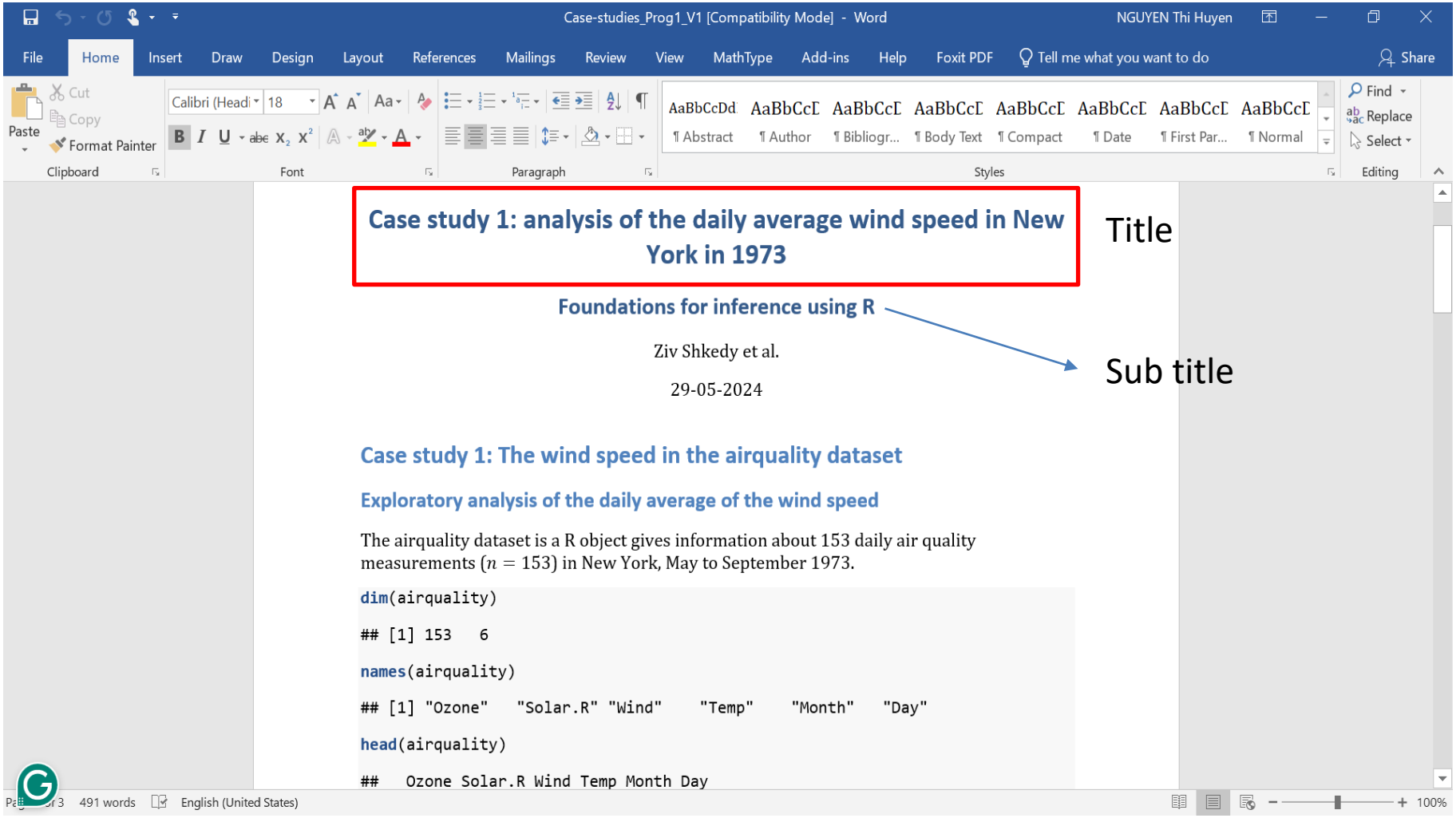

```

Titles, authors and dates



Subtitle, author, and date

Titles, authors and dates



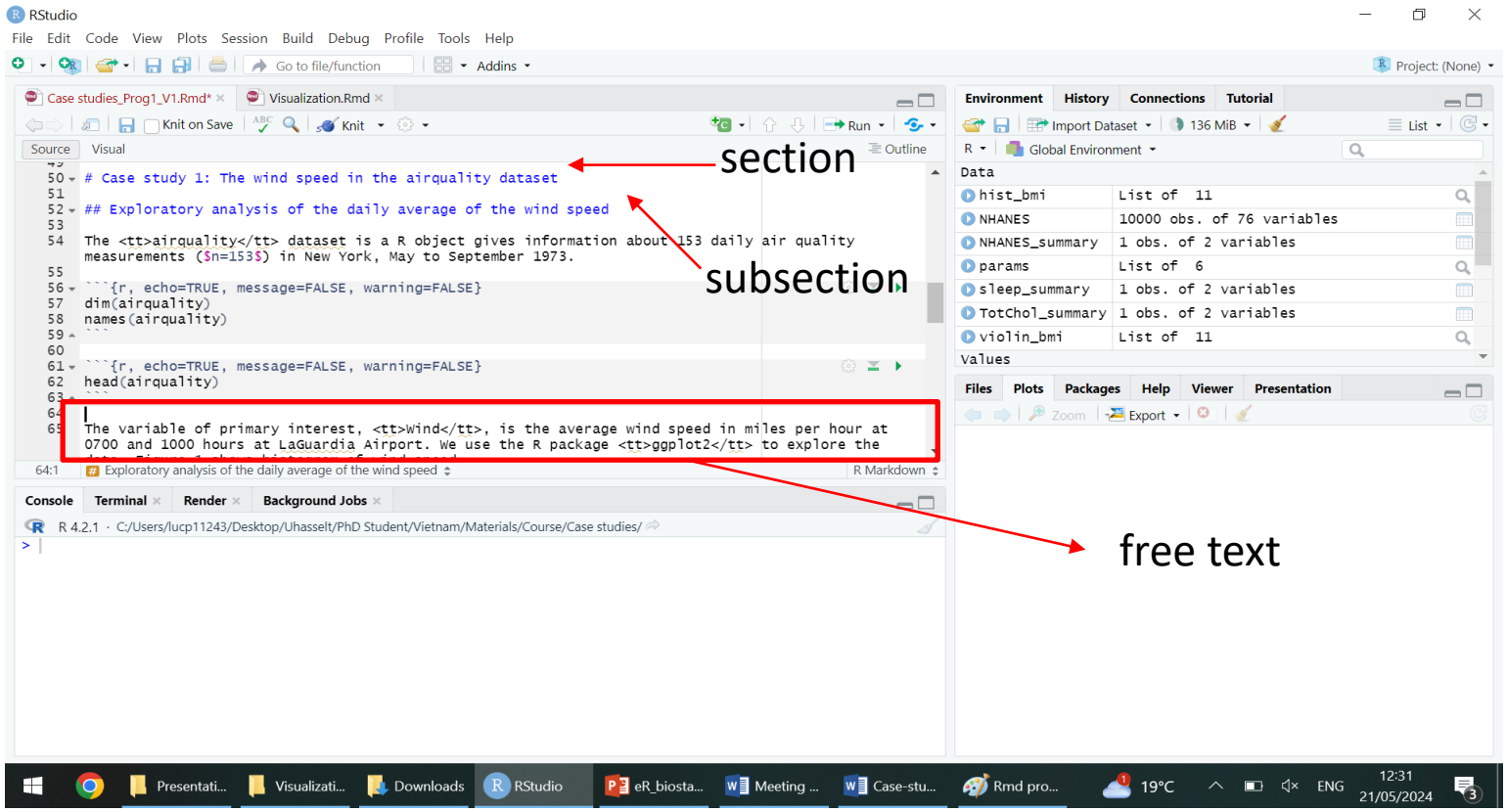
UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

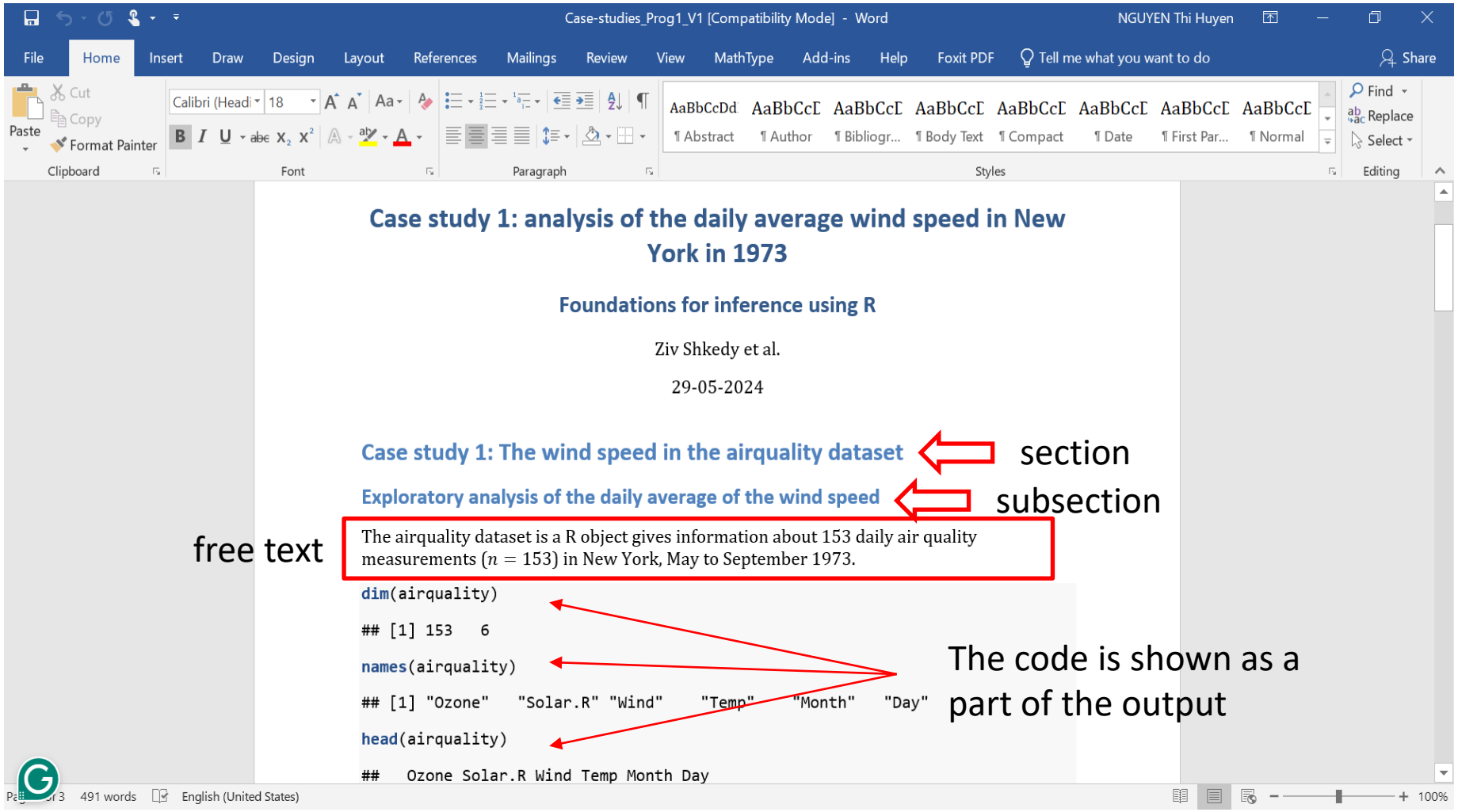
Part 4.2:
The Word document and the Rmd program in details.

Section, subsection, subsubsection

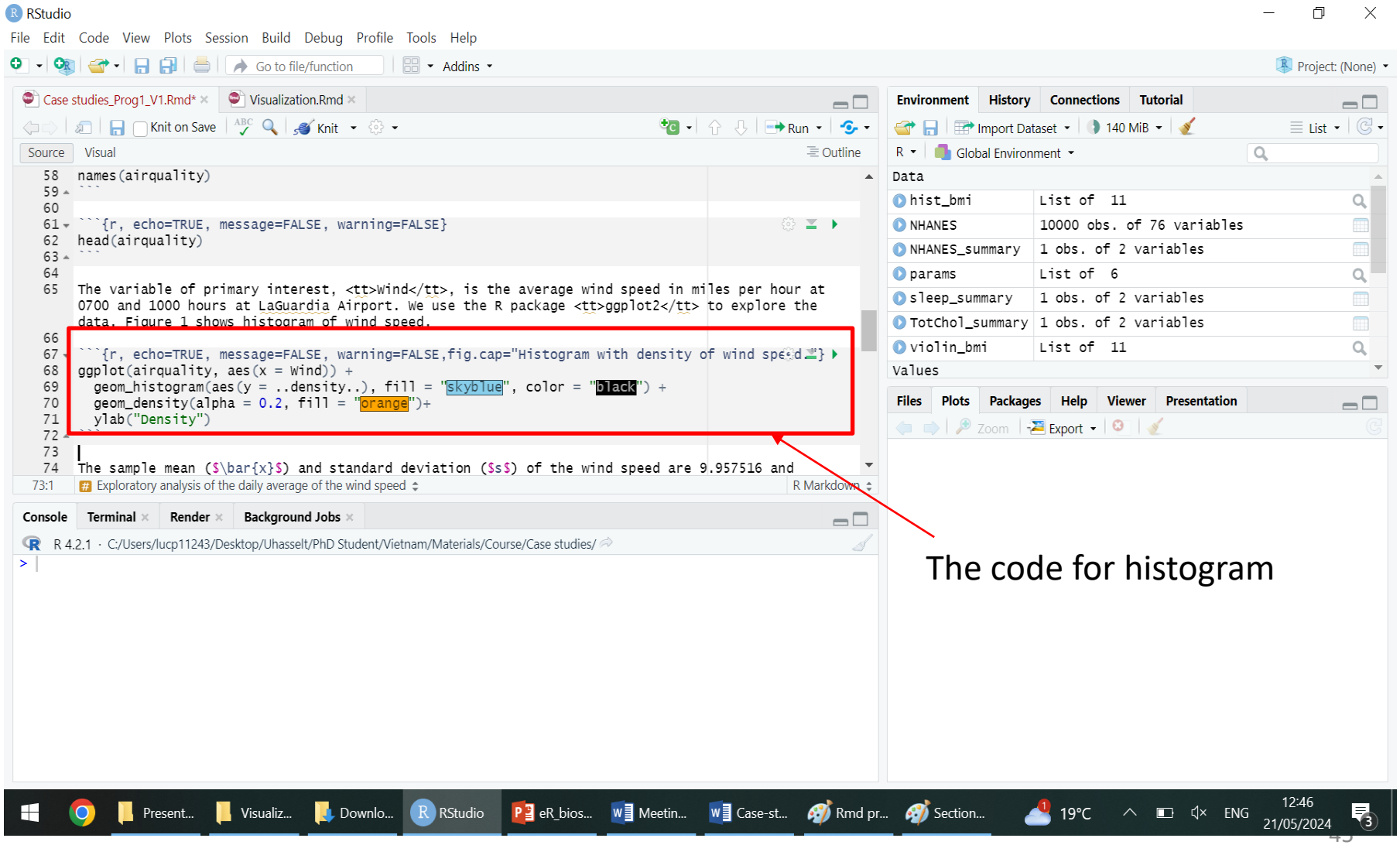


- In addition to the code, we can add free text in the Rmd file.

Section, subsection, subsubsection



Code in the Rmd file



The output in the Word file

Case-studies_Prog1_V1 [Compatibility Mode] - Word — NGUYEN Thi Huyen

File Home Insert Draw Design Layout References Mailings Review View MathType Add-ins Help Foxit PDF Tell me what you want to do Share

Paste Cut Copy Format Painter — Clipboard

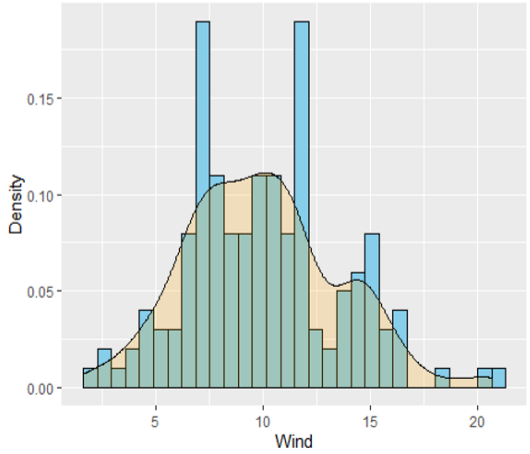
Cambria (Boc 12 — Font

Paragraph

AaBbCcDd ¶ Abstract | AaBbCcE ¶ Author | AaBbCcE ¶ Bibliogr... | AaBbCcE ¶ Body Text | AaBbCcE ¶ Compact | AaBbCcE ¶ Date | AaBbCcE ¶ First Par... | AaBbCcE ¶ Normal — Styles

Find Replace Select — Editing

```
ggplot(airquality, aes(x = Wind)) +
  geom_histogram(aes(y = ..density..), fill = "skyblue", color = "black") +
  geom_density(alpha = 0.2, fill = "orange")+
  ylab("Density")
```

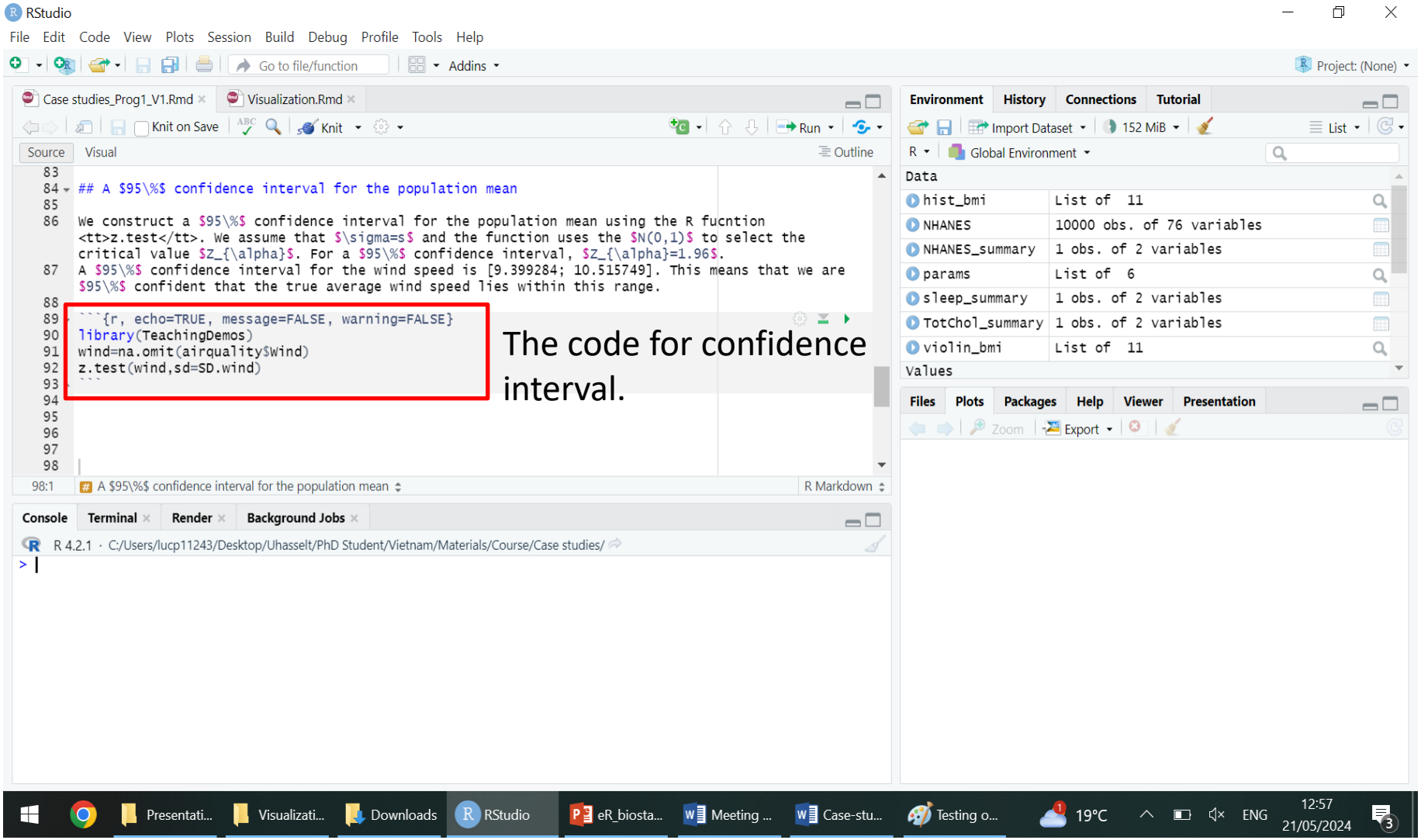


Histogram with density of wind speed.

Page 3 of 3 501 words English (United States) 80%

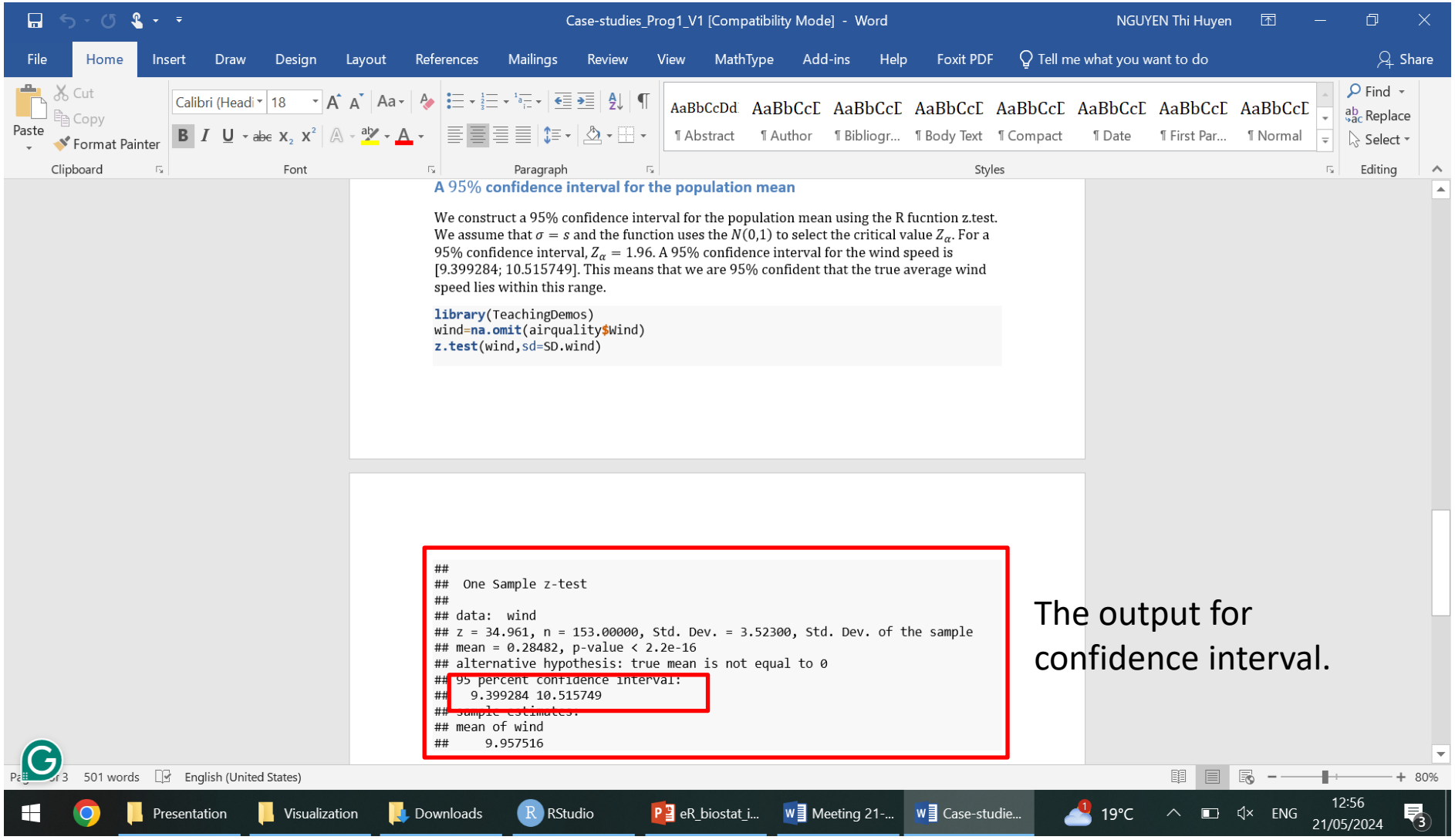
Presentation Visualization Downloads RStudio eR_biostat_i... Meeting 21-... Case-studie... 19°C ENG 12:50 21/05/2024

Code in the Rmd file



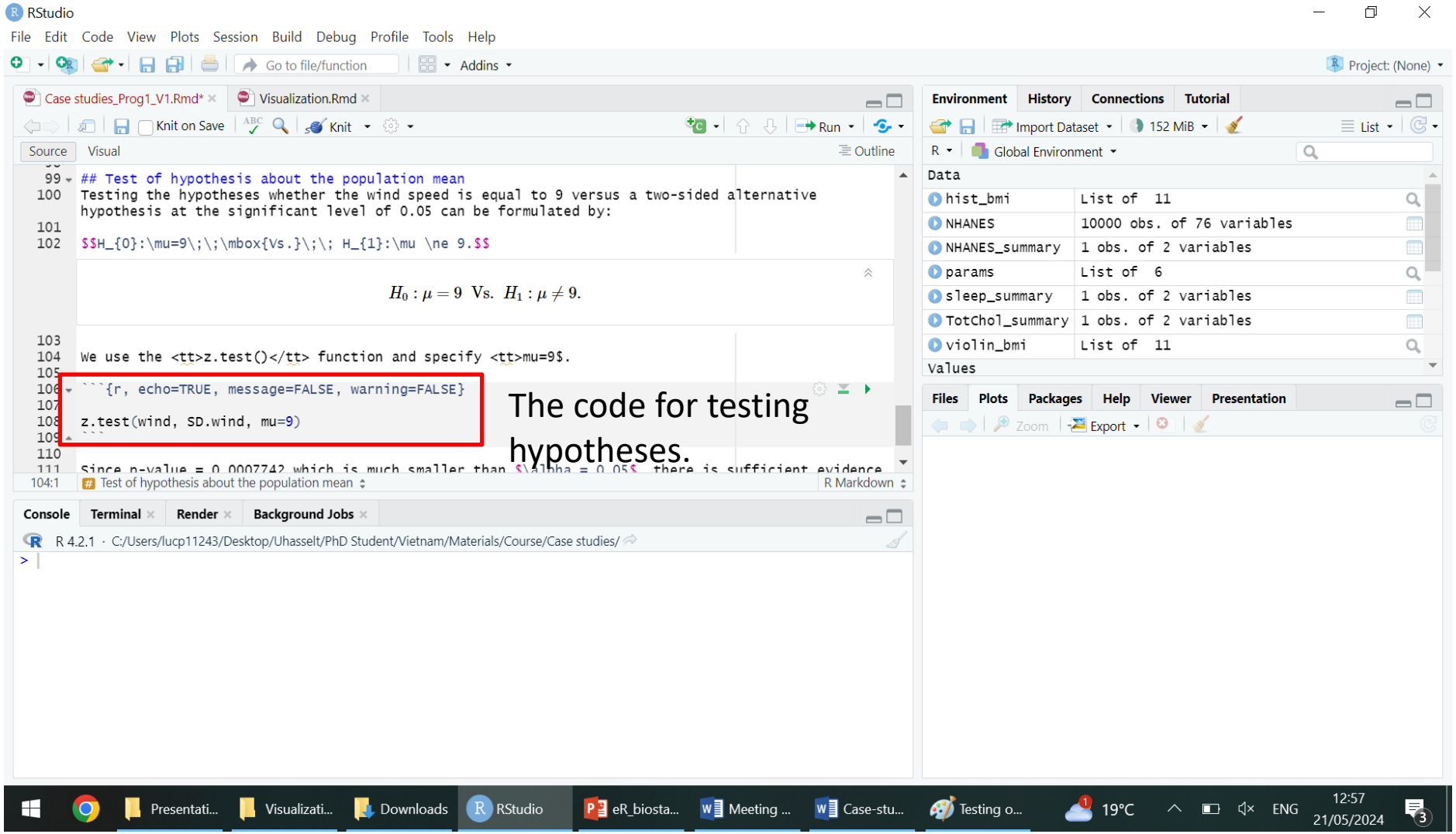
The code for confidence interval.

The output in the Word file



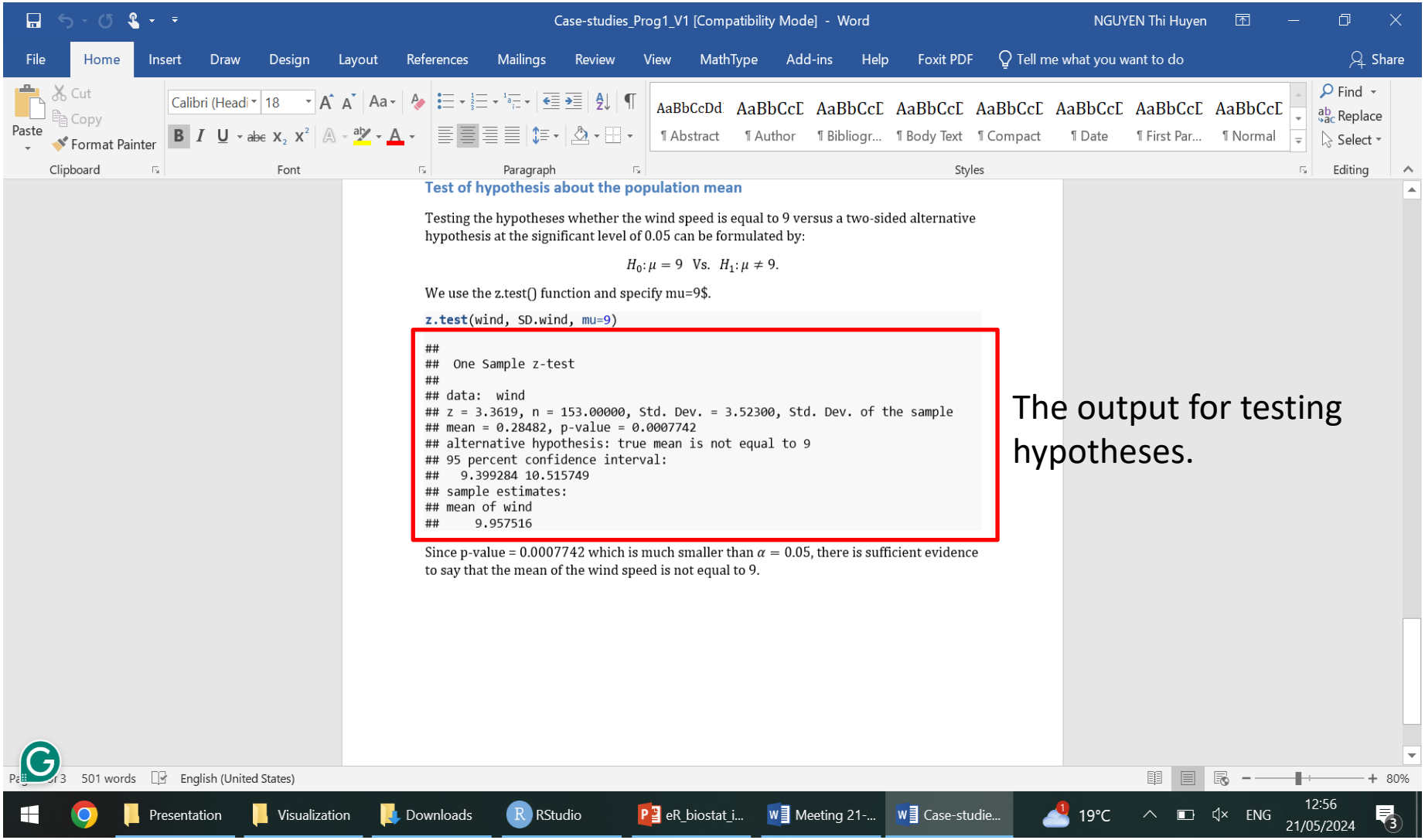
The output for confidence interval.

Code in the Rmd file



The code for testing hypotheses.

The output in the Word file



The output for testing hypotheses.

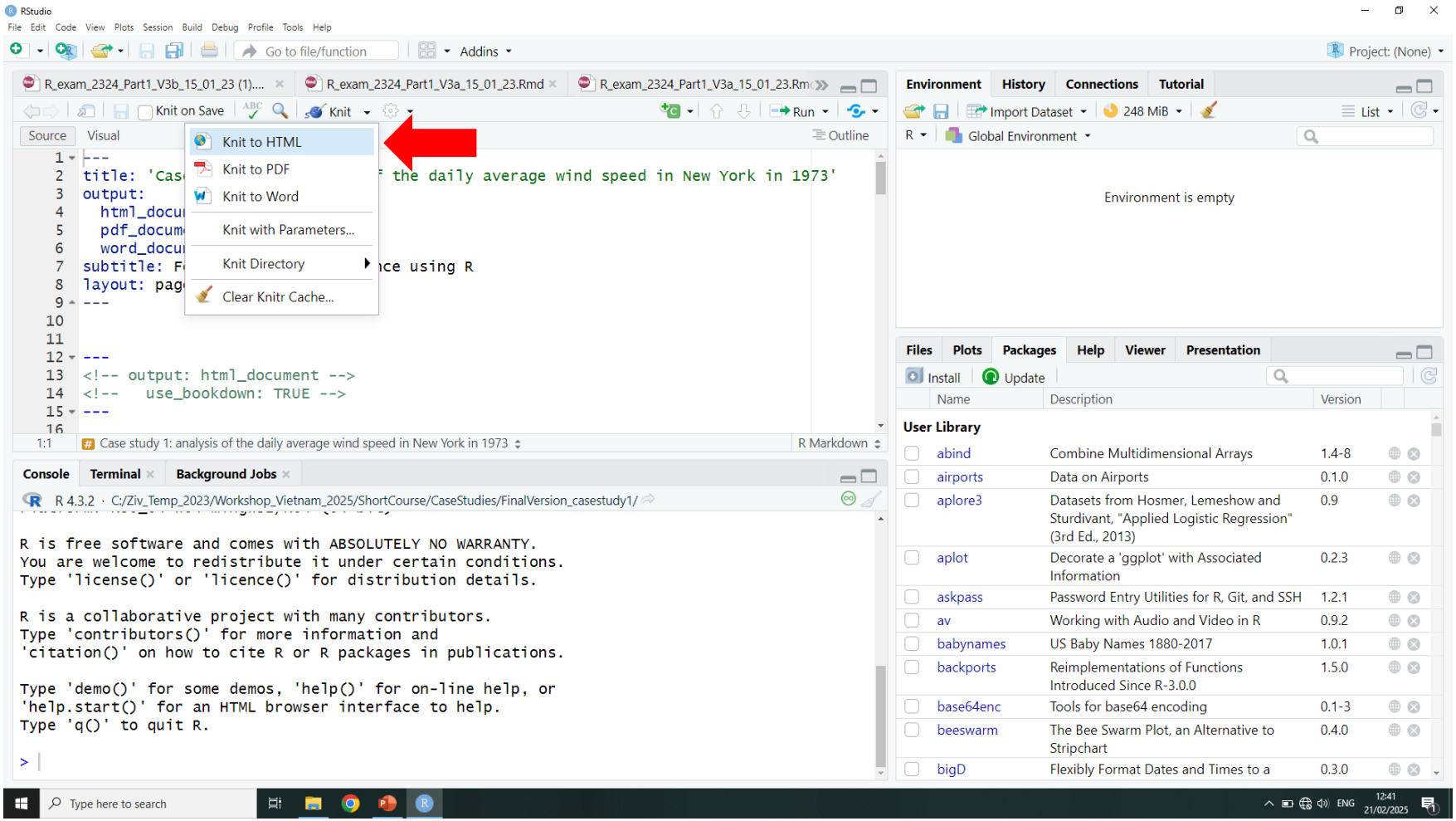
UHASSELT I-BioStat KU LEUVEN

Interuniversity Institute for Biostatistics and statistical Bioinformatics

vliruos SHARING MINDS, CHANGING LIVES

Part 5:
How to set up the HTML file?

Setup the HTML document



The output

Case study 1: analysis of the da

File C:/Ziv_Temp_2023/Workshop_Vietnam_2025/ShortCourse/CaseStudies/FinalVersion_casestudy1/Case-studies_Prog1_V1.html

uhasselt.be bookmarks

All Bookmarks

Case study 1: analysis of the daily average wind speed in New York in 1973

Foundations for inference using R

Case study 1: The wind speed in the airquality dataset

Exploratory analysis of the daily average of the wind speed

The `airquality` dataset is a R object gives information about 153 daily air quality measurements (\(n=153\)) in New York, May to September 1973.

```
dim(airquality)
```

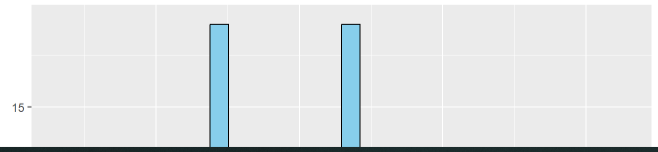
```
## [1] 153   6
```

```
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

The variable of primary interest, `Wind`, is the average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport. We use the R package `ggplot2` to explore the data. Figure 1 shows histogram of wind speed.

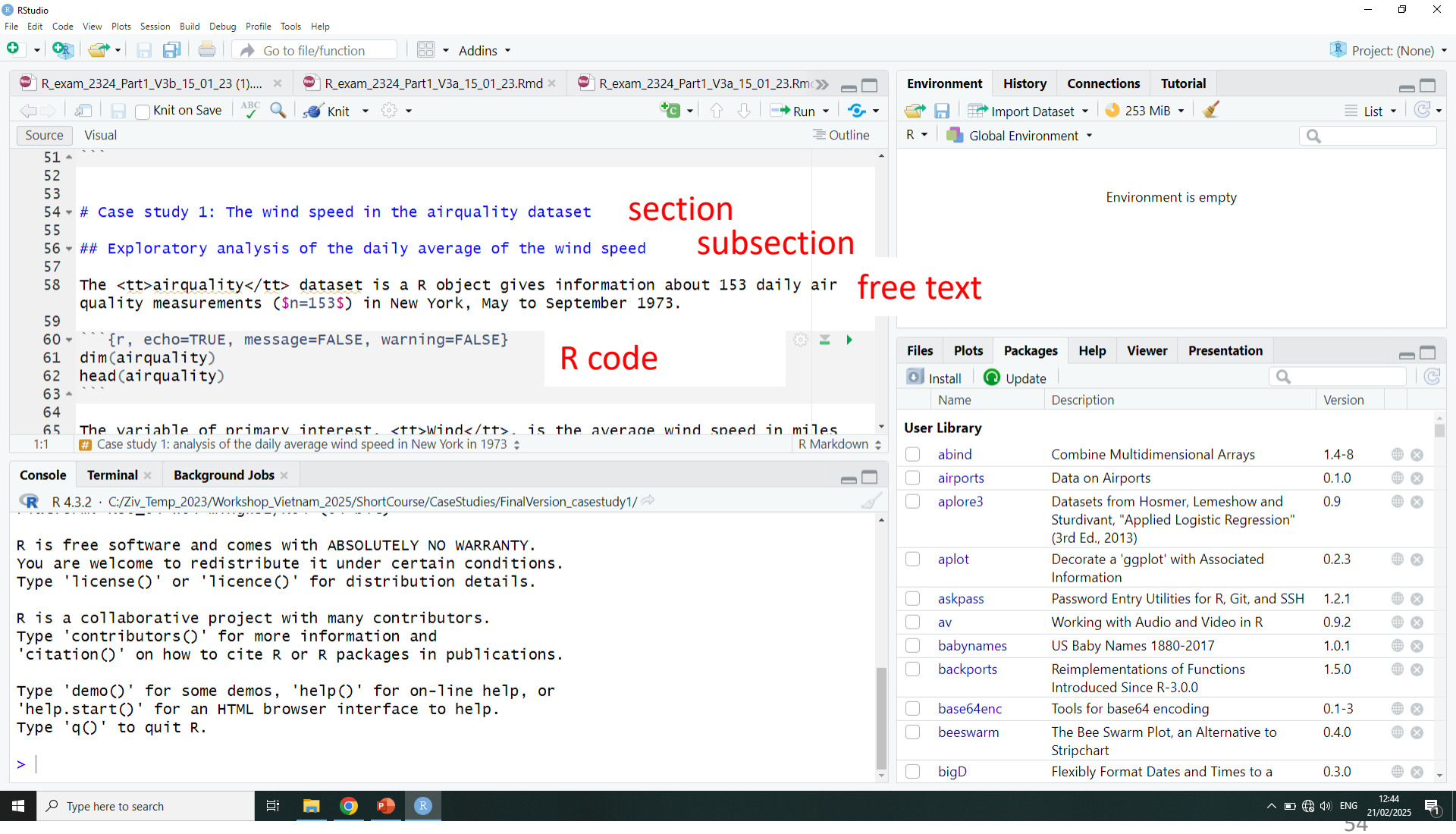
```
ggplot(airquality, aes(x = Wind)) +
  geom_histogram(fill = "skyblue", color = "black")+
  ylab("Frequency")
```



Type here to search

ENG 12:44 21/02/2025

The Rmd file


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
R_exam_2324_Part1_V3b_15_01_23 (1)....
R_exam_2324_Part1_V3a_15_01_23.Rmd
R_exam_2324_Part1_V3a_15_01_23.Rm
Knit on Save
Knit
Run
Source
Visual
Outline
51 ```
52
53
54 # Case study 1: The wind speed in the airquality dataset
55
56 ## Exploratory analysis of the daily average of the wind speed
57
58 The <tt>airquality</tt> dataset is a R object gives information about 153 daily air quality measurements ($n=153$) in New York, May to September 1973.
59
60 ```{r, echo=TRUE, message=FALSE, warning=FALSE}
61 dim(airquality)
62 head(airquality)
63 ```
64
65 The variable of primary interest, <tt>Wind</tt>, is the average wind speed in miles
1:1
Case study 1: analysis of the daily average wind speed in New York in 1973
R Markdown
section
subsection
free text
R code
Environment
History
Connections
Tutorial
Import Dataset
253 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Presentation
Install
Update
Name
Description
Version
User Library
abind
Combine Multidimensional Arrays
1.4-8
airports
Data on Airports
0.1.0
aplore3
Datasets from Hosmer, Lemeshow and Sturdivant, "Applied Logistic Regression" (3rd Ed., 2013)
0.9
aplot
Decorate a 'ggplot' with Associated Information
0.2.3
askpass
Password Entry Utilities for R, Git, and SSH
1.2.1
av
Working with Audio and Video in R
0.9.2
babynames
US Baby Names 1880-2017
1.0.1
backports
Reimplementations of Functions Introduced Since R-3.0.0
1.5.0
base64enc
Tools for base64 encoding
0.1-3
beeswarm
The Bee Swarm Plot, an Alternative to Stripchart
0.4.0
bigD
Flexibly Format Dates and Times to a
0.3.0
Console
Terminal
Background Jobs
R 4.3.2 · C:/Ziv_Temp_2023/Workshop_Vietnam_2025/ShortCourse/CaseStudies/FinalVersion_casestudy1/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Type here to search
ENG
12:44
21/02/2025

The output

Case study 1: analysis of the da

File C:/Ziv_Temp_2023/Workshop_Vietnam_2025/ShortCourse/CaseStudies/FinalVersion_casestudy1/Case-studies_Prog1_V1.html

uhasselt.be bookmarks

All Bookmarks

Case study 1: analysis of the daily average wind speed in New York in 1973

section

Foundations for inference using R

Case study 1: The wind speed in the airquality dataset

subsection

Exploratory analysis of the daily average of the wind speed

free text

The `airquality` dataset is a R object gives information about 153 daily air quality measurements (\(n=153\)) in New York, May to September 1973.

```
dim(airquality)
```

```
## [1] 153   6
```

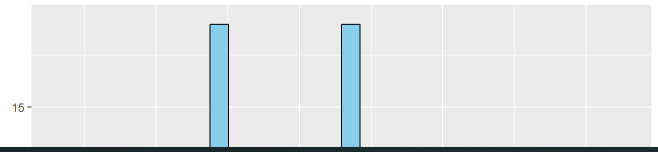
```
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

R code and

The variable of primary interest, `Wind`, is the average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport. We use the R package `ggplot2` to explore the data. Figure 1 shows histogram of wind speed.

```
ggplot(airquality, aes(x = Wind)) +
  geom_histogram(fill = "skyblue", color = "black")+
  ylab("Frequency")
```



Type here to search

ENG 12:44 21/02/2025

Discussion

- R Studio + R markdown:
- Easy to use.
- Text + code.
- Output:
  - Standard: HTML, PDF, DOC.
  - Example: Word.doc.