Output development using R & R markdown

Ziv Shkedy and Thi Huyen Nguyen

Hasselt University

Foundations for inference

Ha Noi, 29/05/2024

Visit us on Facebook ER-BioStat

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

twitter @erbiostat

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 produce an output ?
  - How to combine text and software output in the same document ?

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
```

The `airquality` data

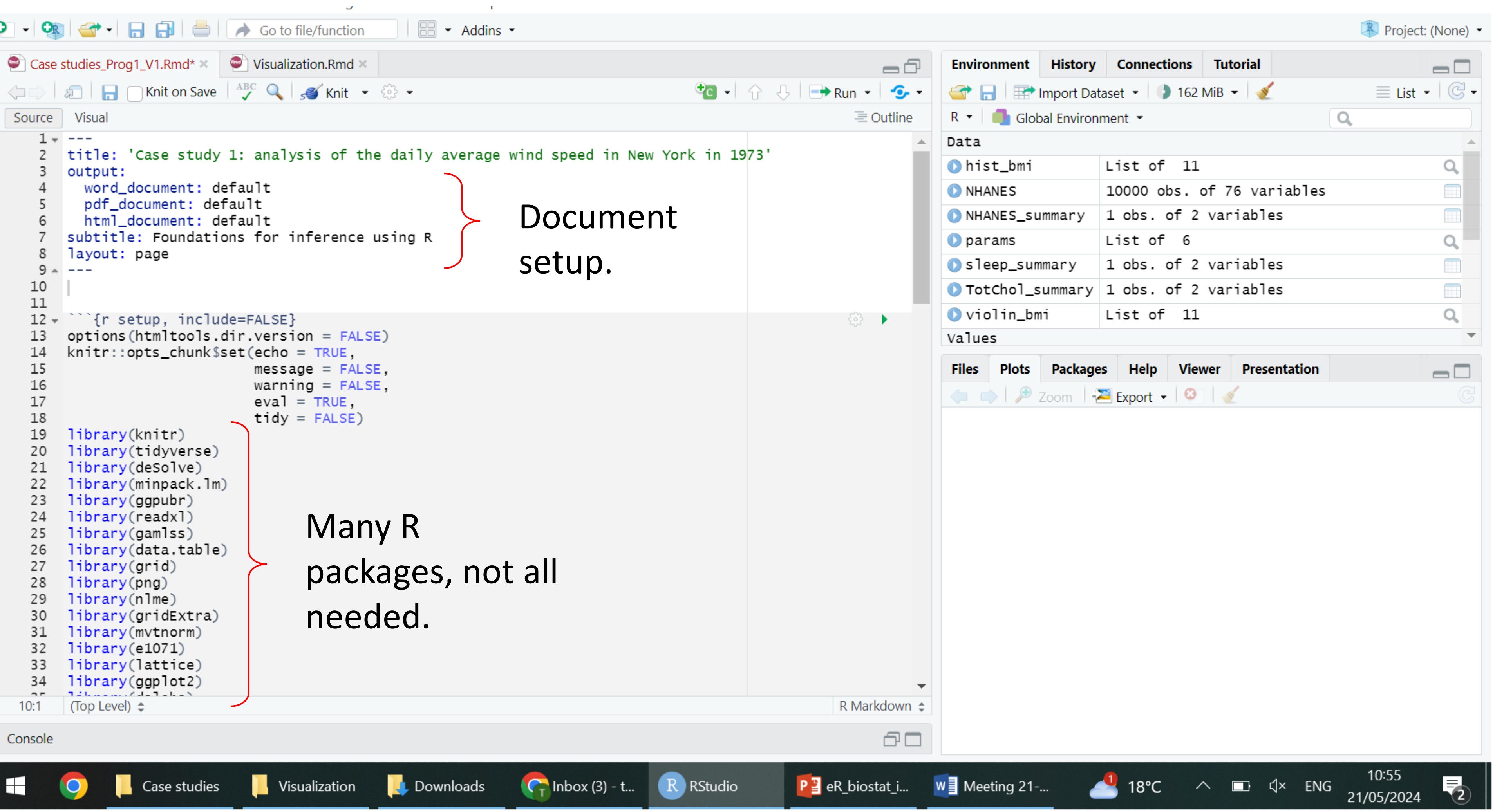
Part 2

The Rmd program

The output

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

The Rmd program



~~~
---
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)
~~~

Data

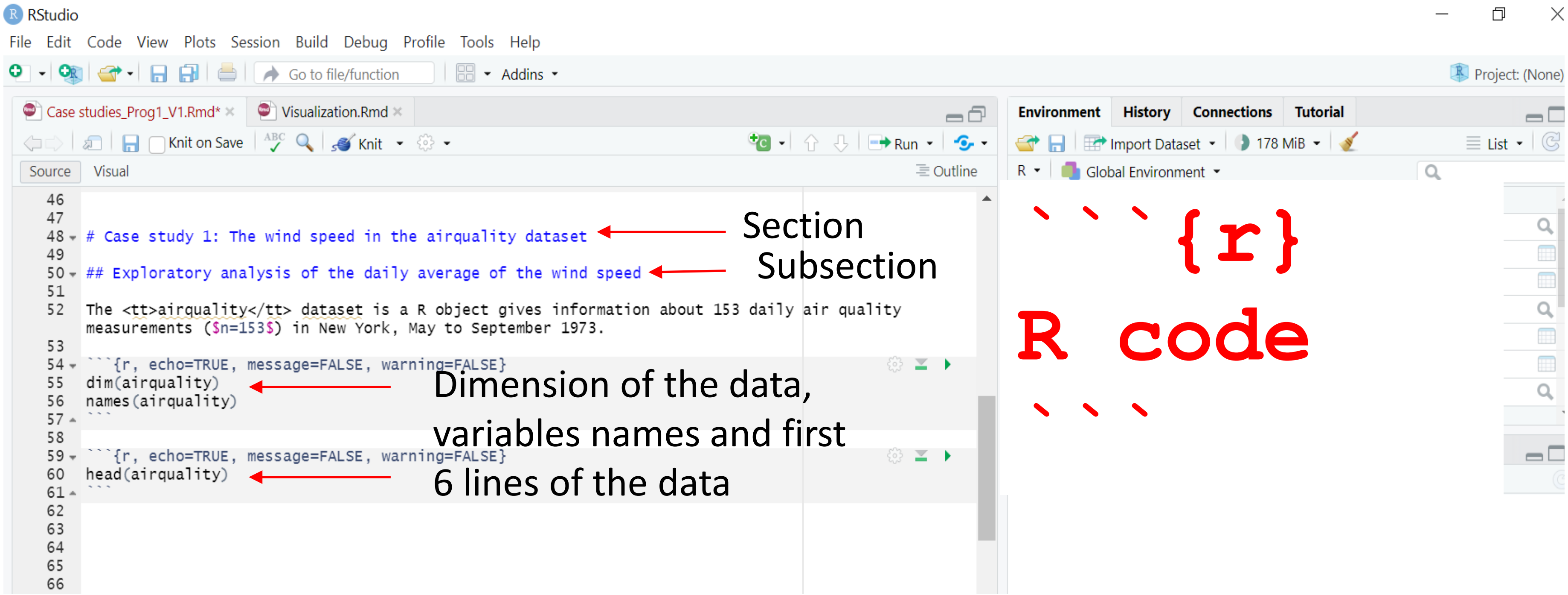
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

Document setup.

Many R packages, not all needed.

The Rmd program



~~~

Case study 1: The wind speed in the airquality dataset

Exploratory analysis of the daily average of the wind speed

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.

```{r, echo=TRUE, message=FALSE, warning=FALSE}
dim(airquality)
names(airquality)
```

```{r, echo=TRUE, message=FALSE, warning=FALSE}
head(airquality)
```
~~~

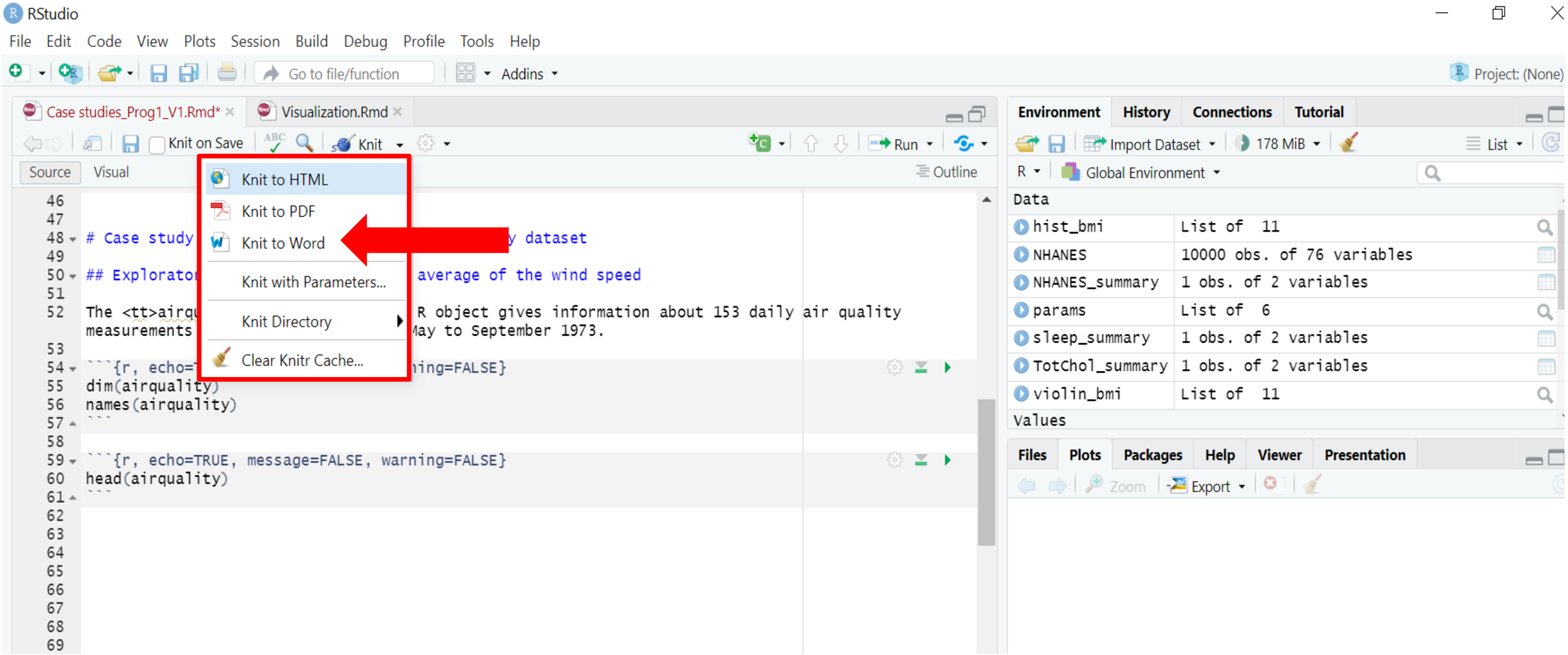
Section

Subsection

Dimension of the data, variables names and first 6 lines of the data

~~~ 
```{r}
R code
```
~~~

Choose the output



~~~
# Case study                              y dataset

## Explorator                   average of the wind speed

The <tt>airqu                   R object gives information about 153 daily air quality
measurements                   May to September 1973.

```{r, echo=T                   ning=FALSE}
dim(airquality)
names(airquality)
~~~

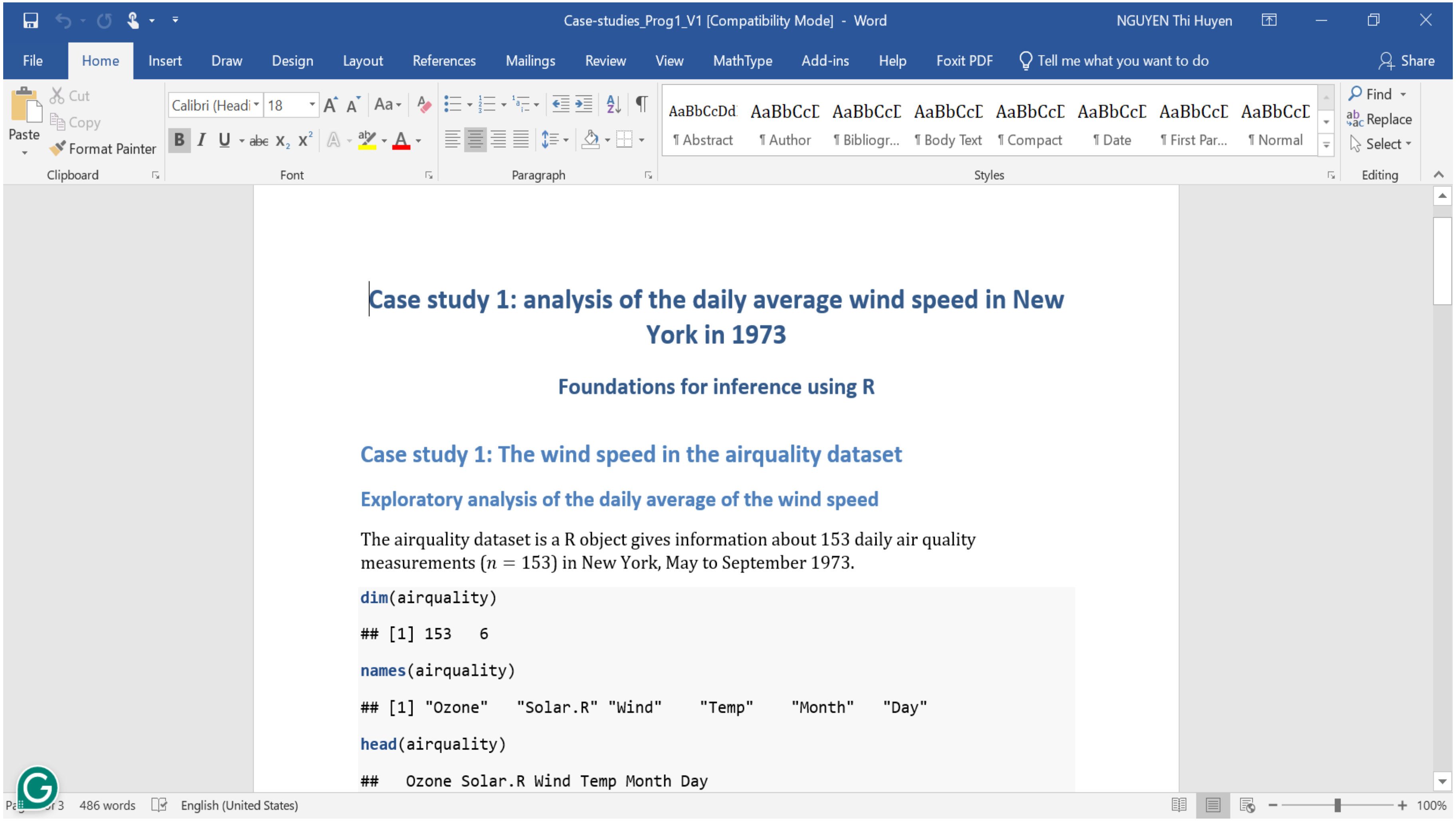
```{r, echo=TRUE, message=FALSE, warning=FALSE}
head(airquality)
```
```

Data

| | |
|---|---|
| hist_bmi | List of 11 |
| NHANES | 10000 obs. of 76 variables |
| NHANES_summary | 1 obs. of 2 variables |
| params | List of 6 |
| sleep_summary | 1 obs. of 2 variables |
| TotChol_summary | 1 obs. of 2 variables |
| violin_bmi | List of 11 |

Values
```

The Word document output



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

names(airquality)

## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"

head(airquality)

##   Ozone Solar.R Wind Temp Month Day
```

The `airquality` data

Part 3

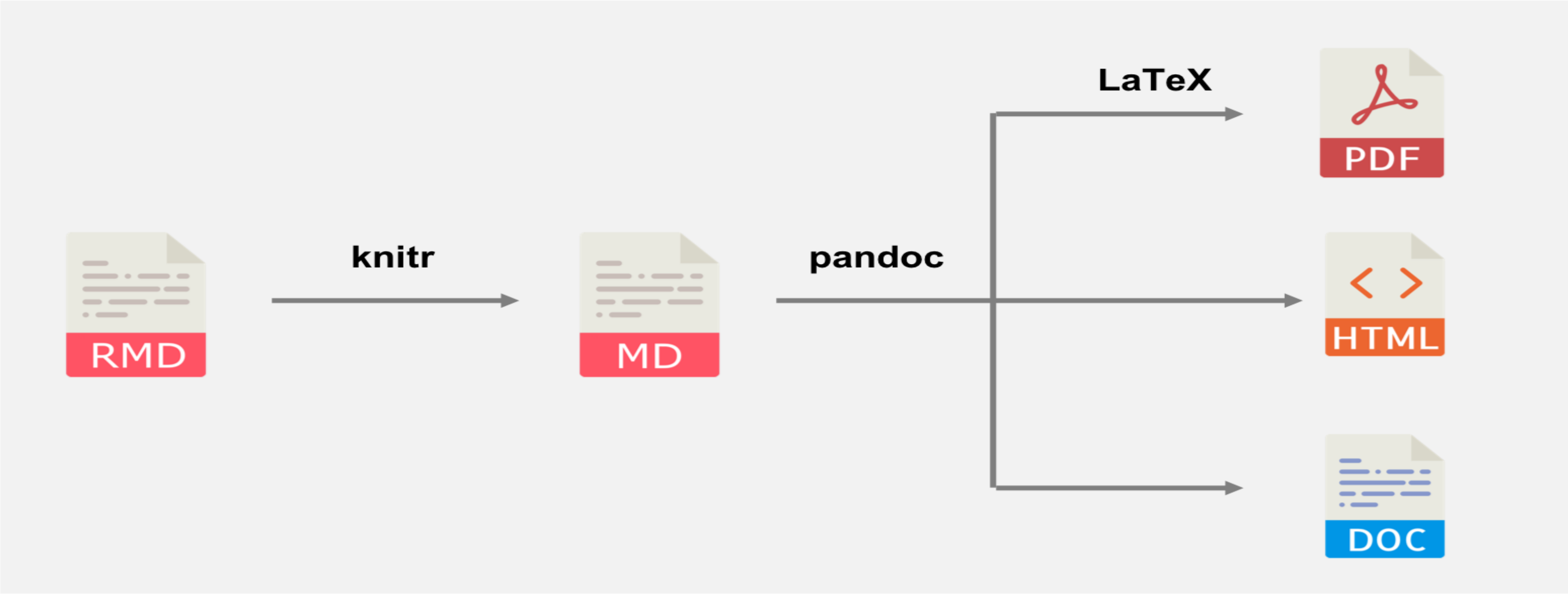
How to produce the Word output ?

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

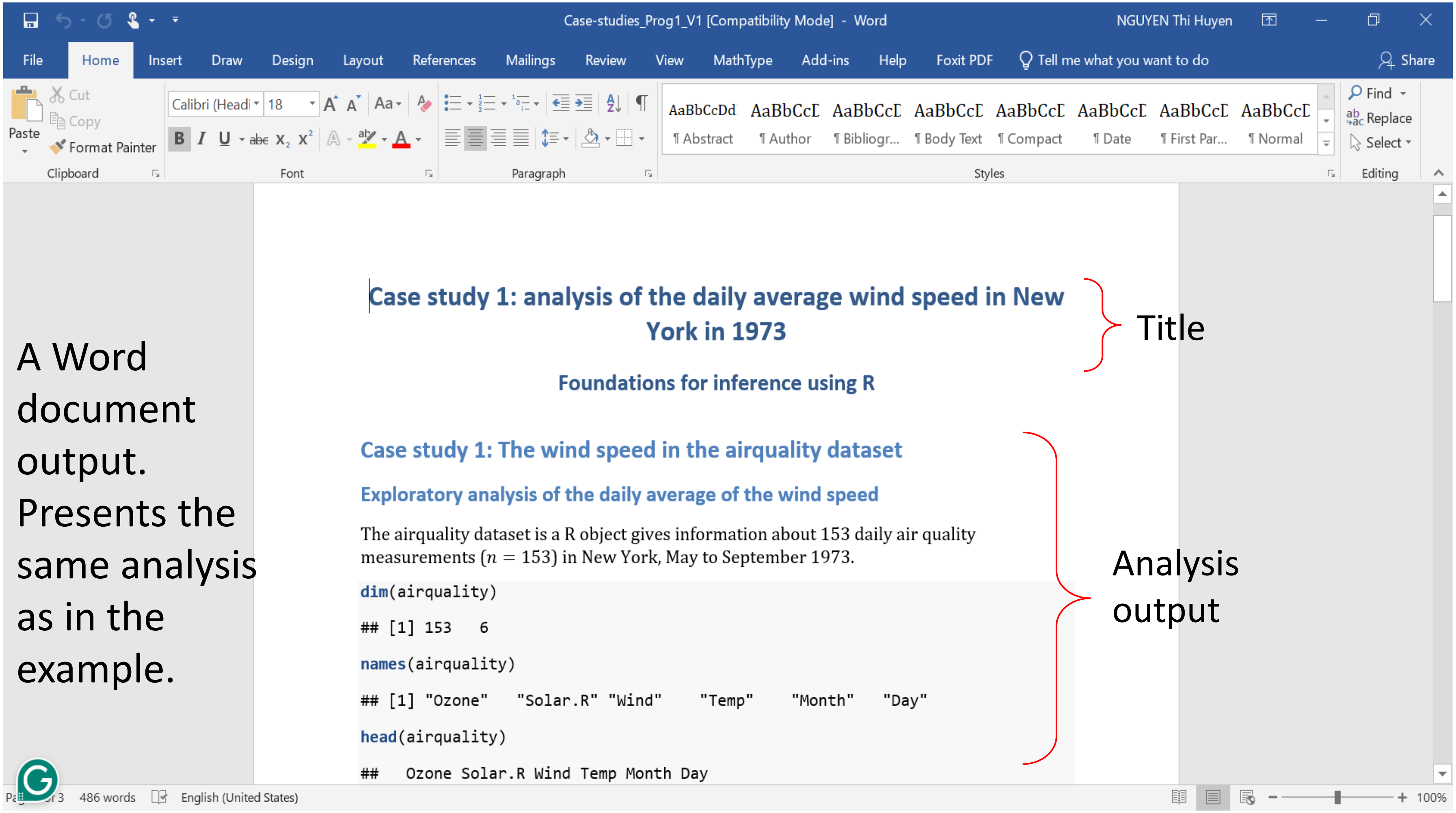
- Analyses → high quality report.
- Rmarkdown – Different dynamic and statistic formats (html, pdf, word, books, dashboard, e.t.c).



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.



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

Title

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

names(airquality)

## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"

head(airquality)

##   Ozone Solar.R Wind Temp Month Day
```

Analysis output

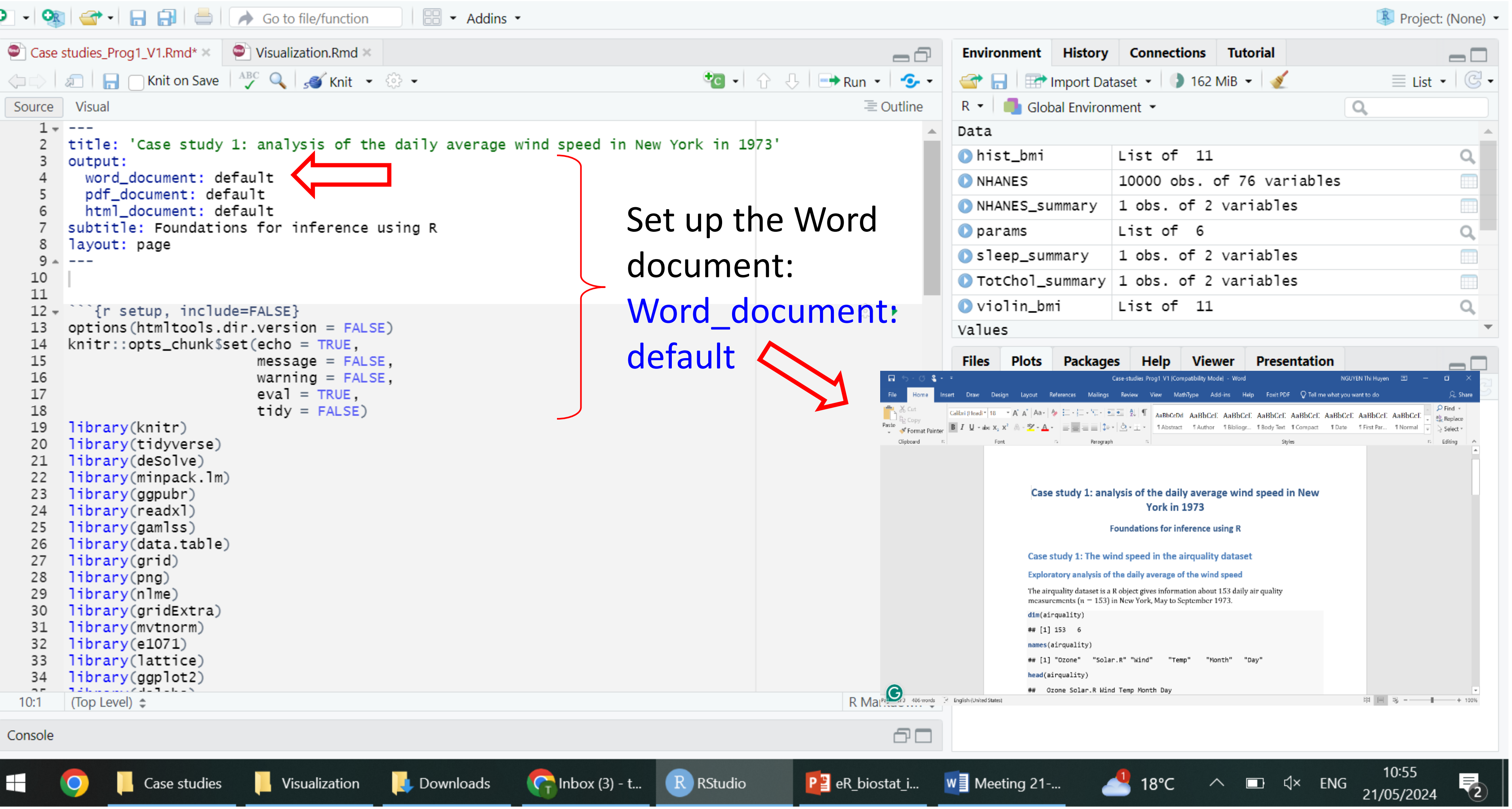
Part 3.1:
How to set up the Word file?

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)
~~~

Data

hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

Set up the Word document:

Word_document: default

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

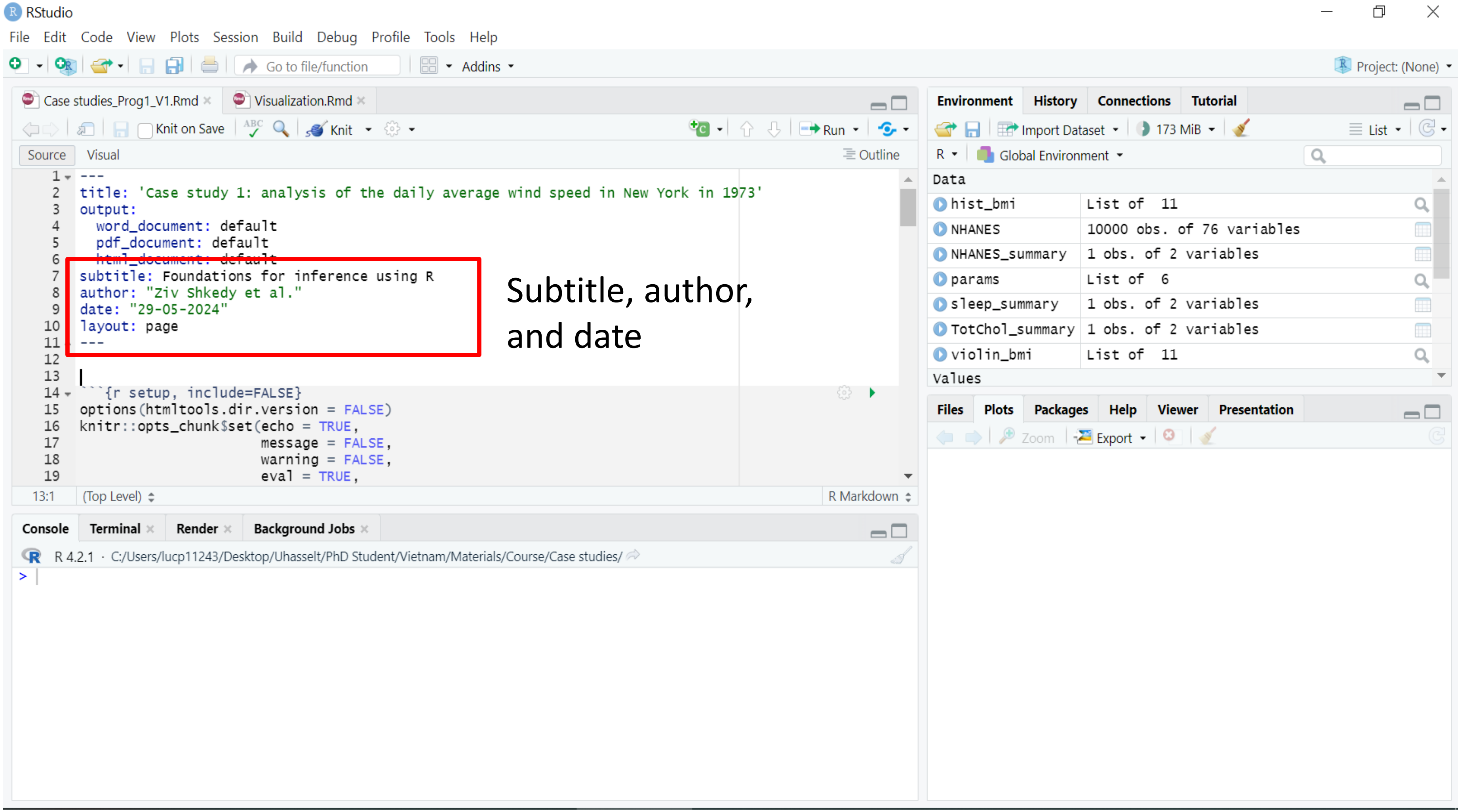
names(airquality)

## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"

head(airquality)

##   Ozone Solar.R Wind Temp Month Day
```

Titles, authors and dates



~~~
---
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
author: "Ziv Shkedy et al."
date: "29-05-2024"
layout: page
---

```{r setup, include=FALSE}
options(htmltools.dir.version = FALSE)
knitr::opts_chunk$set(echo = TRUE,
                      message = FALSE,
                      warning = FALSE,
                      eval = TRUE,
~~~

Subtitle, author, and date

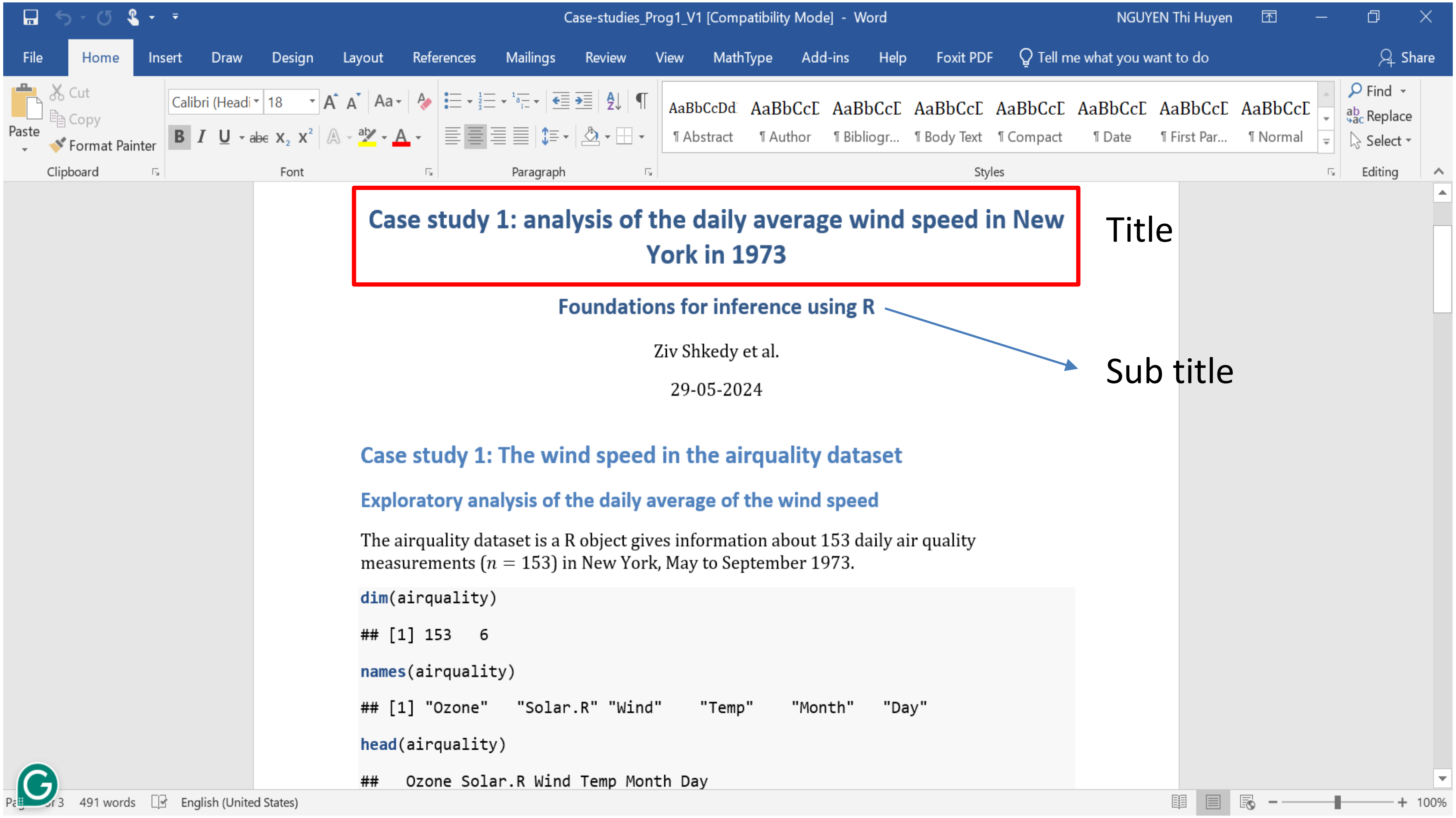
Data

hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

```
R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/
>
```

Titles, authors and dates



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

Title

Foundations for inference using R

Sub title

Ziv Shkedy et al.

29-05-2024

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

names(airquality)

## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"

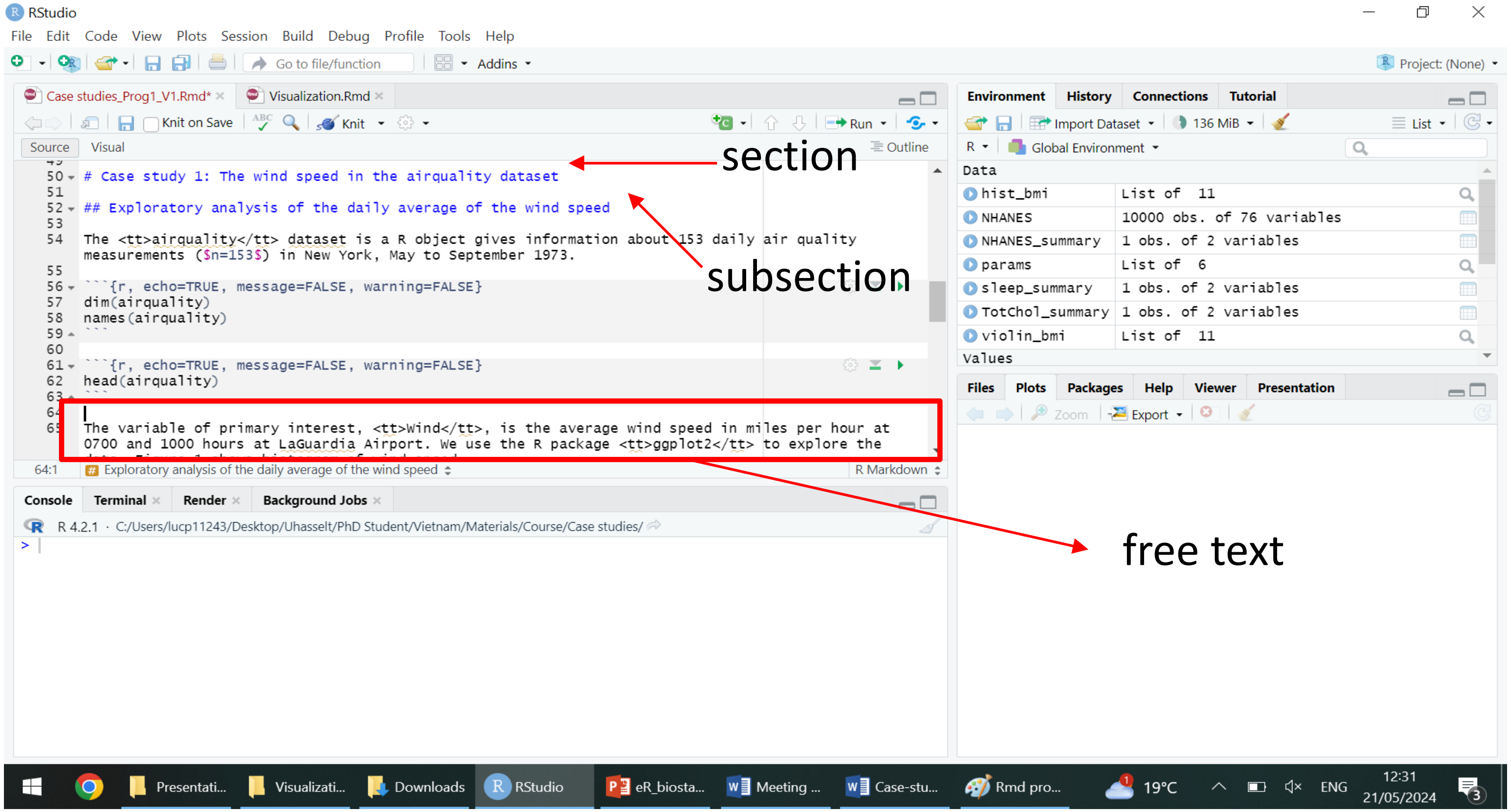
head(airquality)

##   Ozone Solar.R Wind Temp Month Day
```

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Part 3.2:
The Word document and the Rmd program in details.

Section, subsection, subsubsection



~~~
# Case study 1: The wind speed in the airquality dataset

## Exploratory analysis of the daily average of the wind speed

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.

```{r, echo=TRUE, message=FALSE, warning=FALSE}
dim(airquality)
names(airquality)
```

```{r, echo=TRUE, message=FALSE, warning=FALSE}
head(airquality)
```

The variable of primary interest, <tt>Wind</tt>, is the average wind speed in miles per hour at
0700 and 1000 hours at LaGuardia Airport. We use the R package <tt>ggplot2</tt> to explore the
~~~

64:1 Exploratory analysis of the daily average of the wind speed R Markdown

Data

hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

```
R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/
>
```

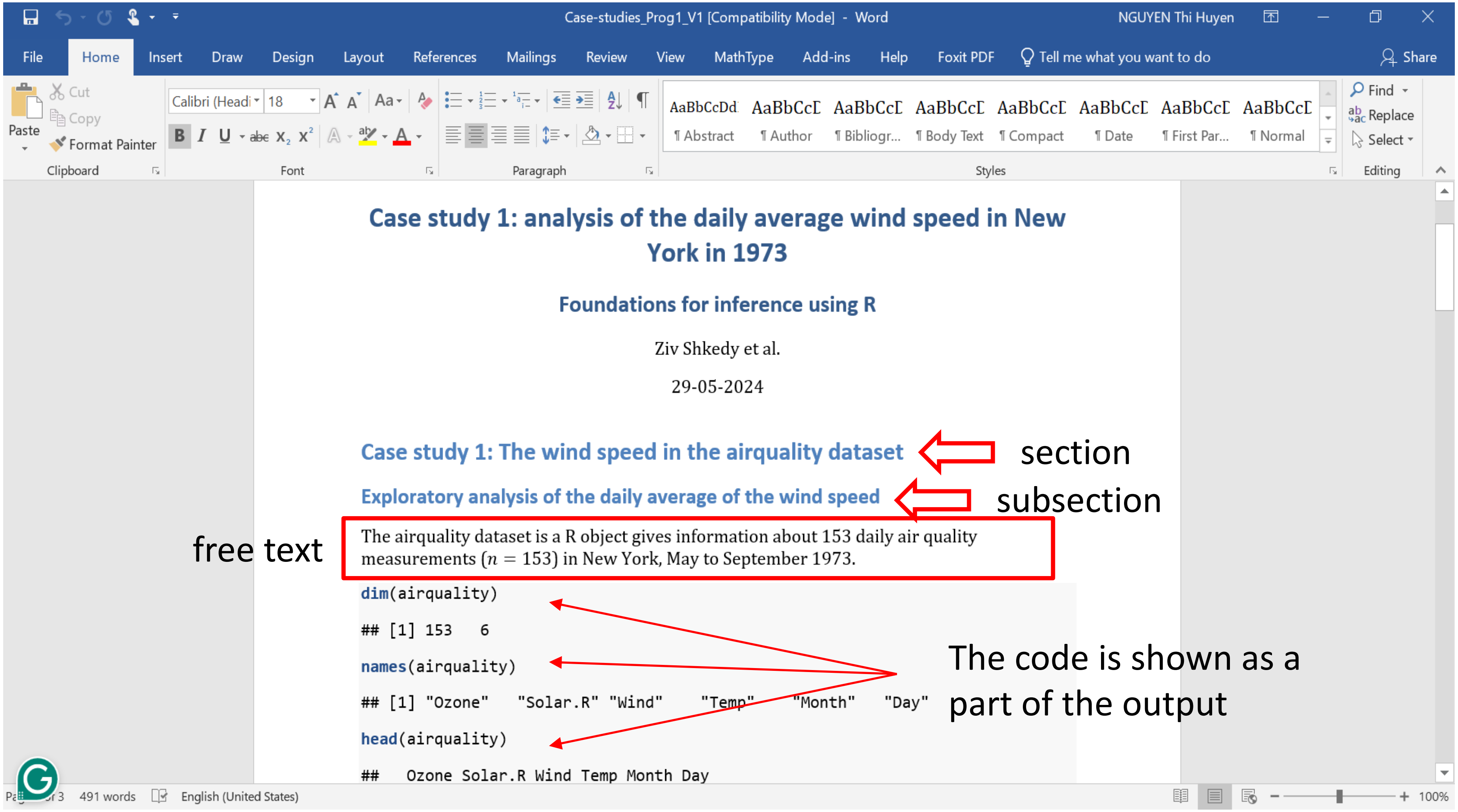
section

subsection

free text

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

Section, subsection, subsubsection



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

Foundations for inference using R

Ziv Shkedy et al.

29-05-2024

Case study 1: The wind speed in the airquality dataset

section

Exploratory analysis of the daily average of the wind speed

subsection

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

names(airquality)

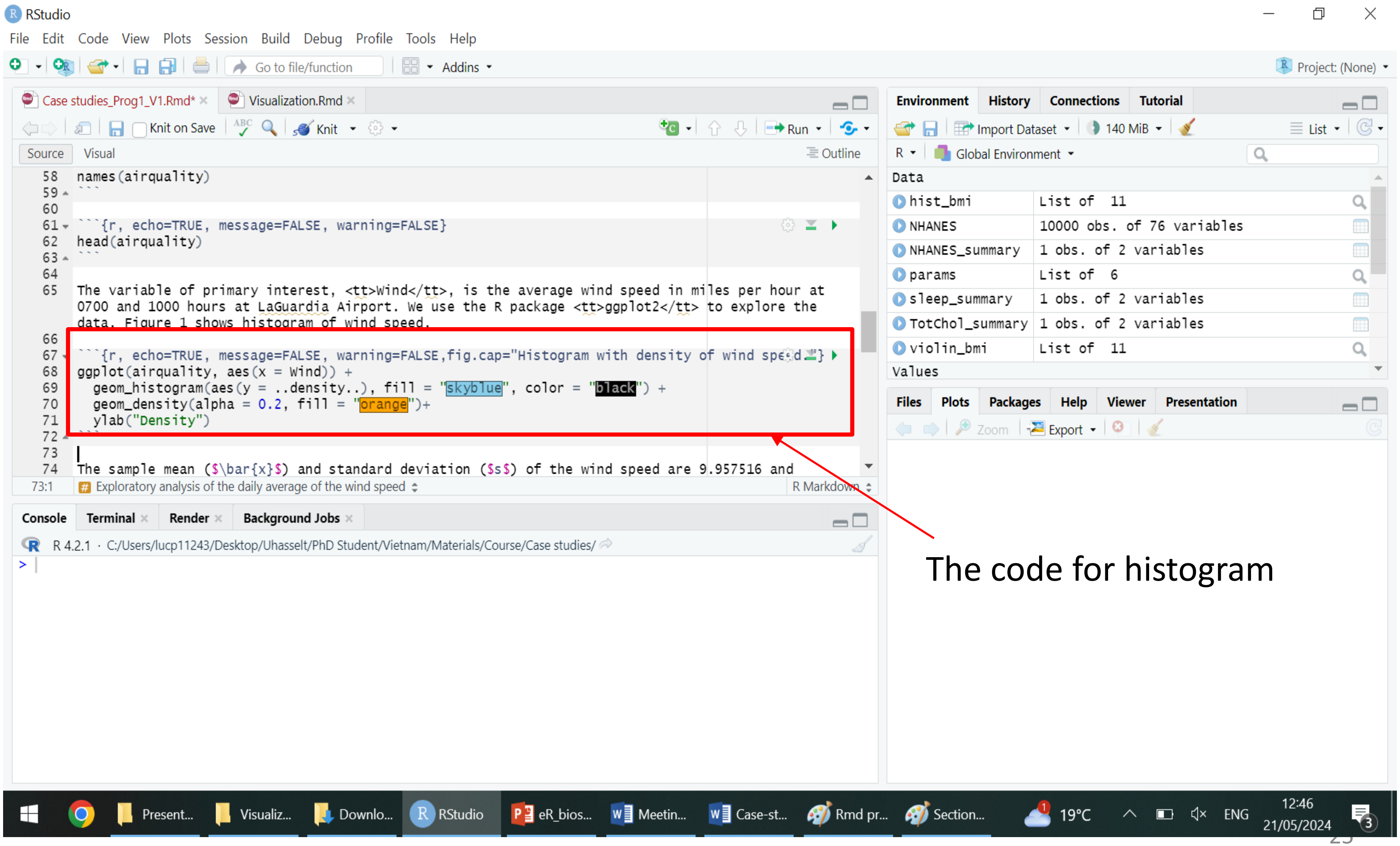
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"

head(airquality)

##   Ozone Solar.R Wind Temp Month Day
```

The code is shown as a part of the output

Code in the Rmd file



~~~~
names(airquality)
```

```{r, echo=TRUE, message=FALSE, warning=FALSE}
head(airquality)
```

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

```{r, echo=TRUE, message=FALSE, warning=FALSE,fig.cap="Histogram with density of wind speed"}
ggplot(airquality, aes(x = Wind)) +
  geom_histogram(aes(y = ..density..), fill = "skyblue", color = "black") +
  geom_density(alpha = 0.2, fill = "orange")+
  ylab("Density")
```

The sample mean ($\bar{x}$) and standard deviation ($s$) of the wind speed are 9.957516 and
~~~~

73:1 Exploratory analysis of the daily average of the wind speed

Data

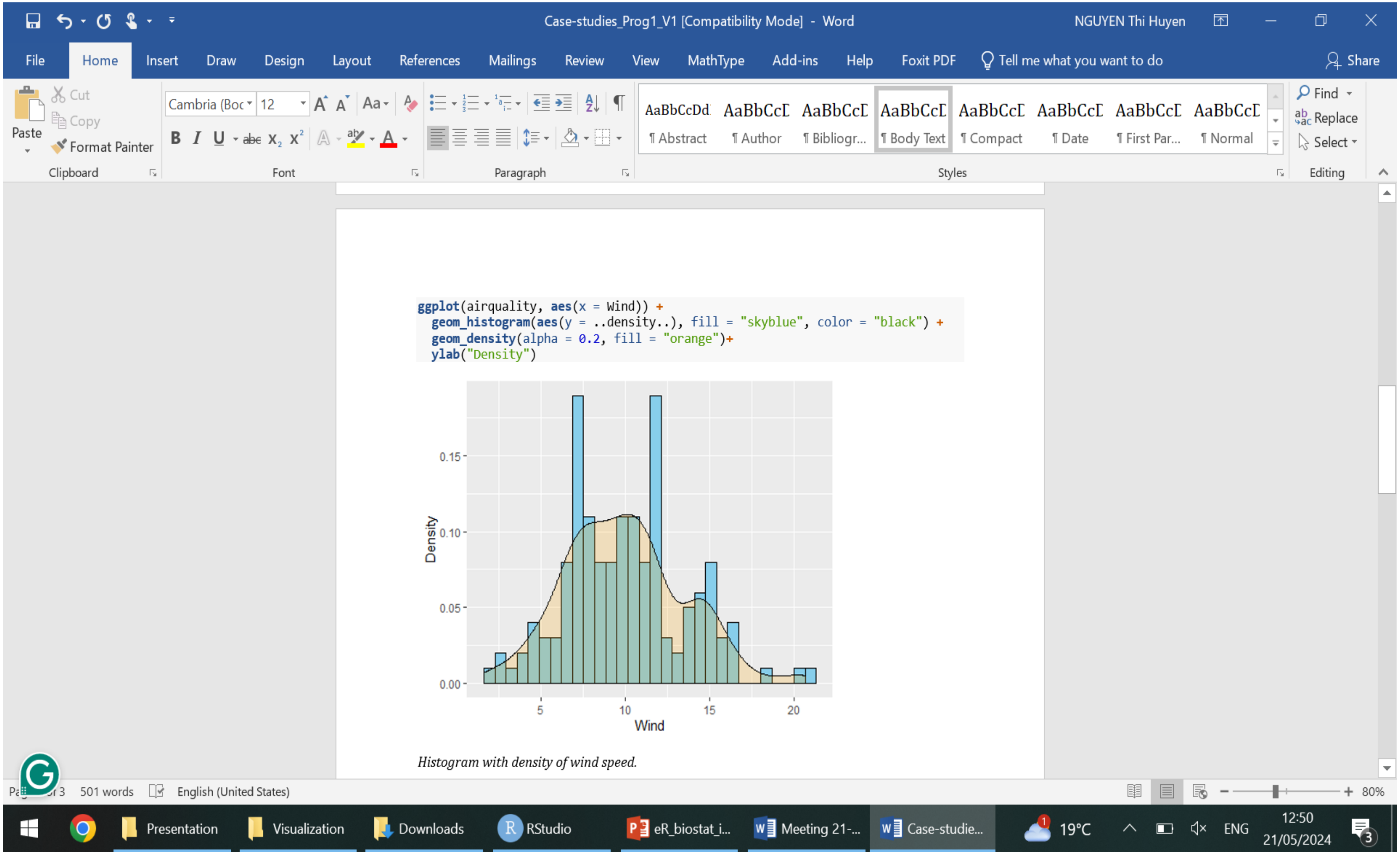
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

```
R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/
>
```

The code for histogram

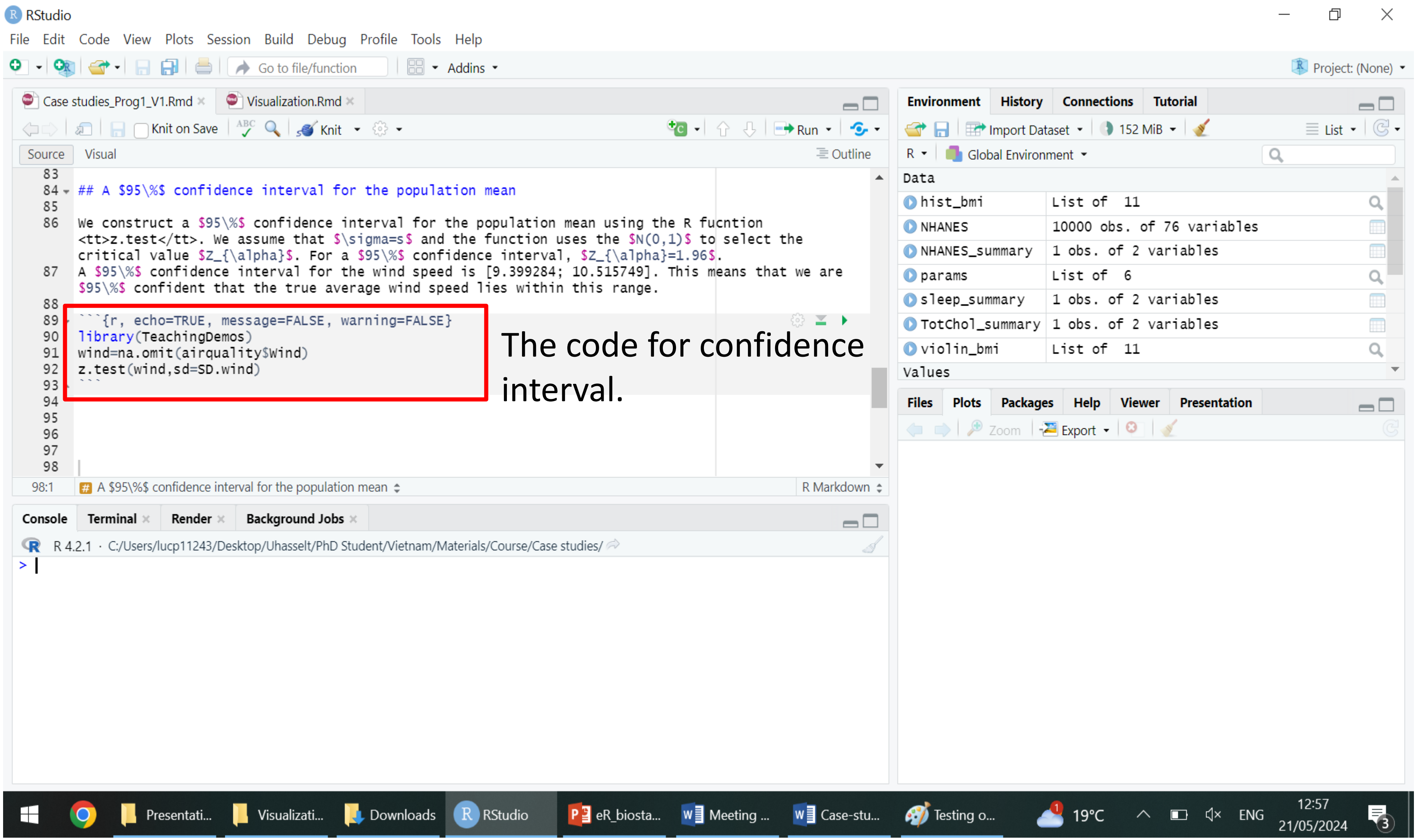
The output in the Word file



```
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.

Code in the Rmd file



~~~~
## A $95\%$ confidence interval for the population mean

We construct a $95\%$ confidence interval for the population mean using the R fucntion <tt>z.test</tt>. We assume that $\sigma=s$ and the function uses the $N(0,1)$ to select the critical value $Z_{\alpha}$. For a $95\%$ confidence interval, $Z_{\alpha}=1.96$.
A $95\%$ confidence interval for the wind speed is [9.399284; 10.515749]. This means that we are $95\%$ confident that the true average wind speed lies within this range.

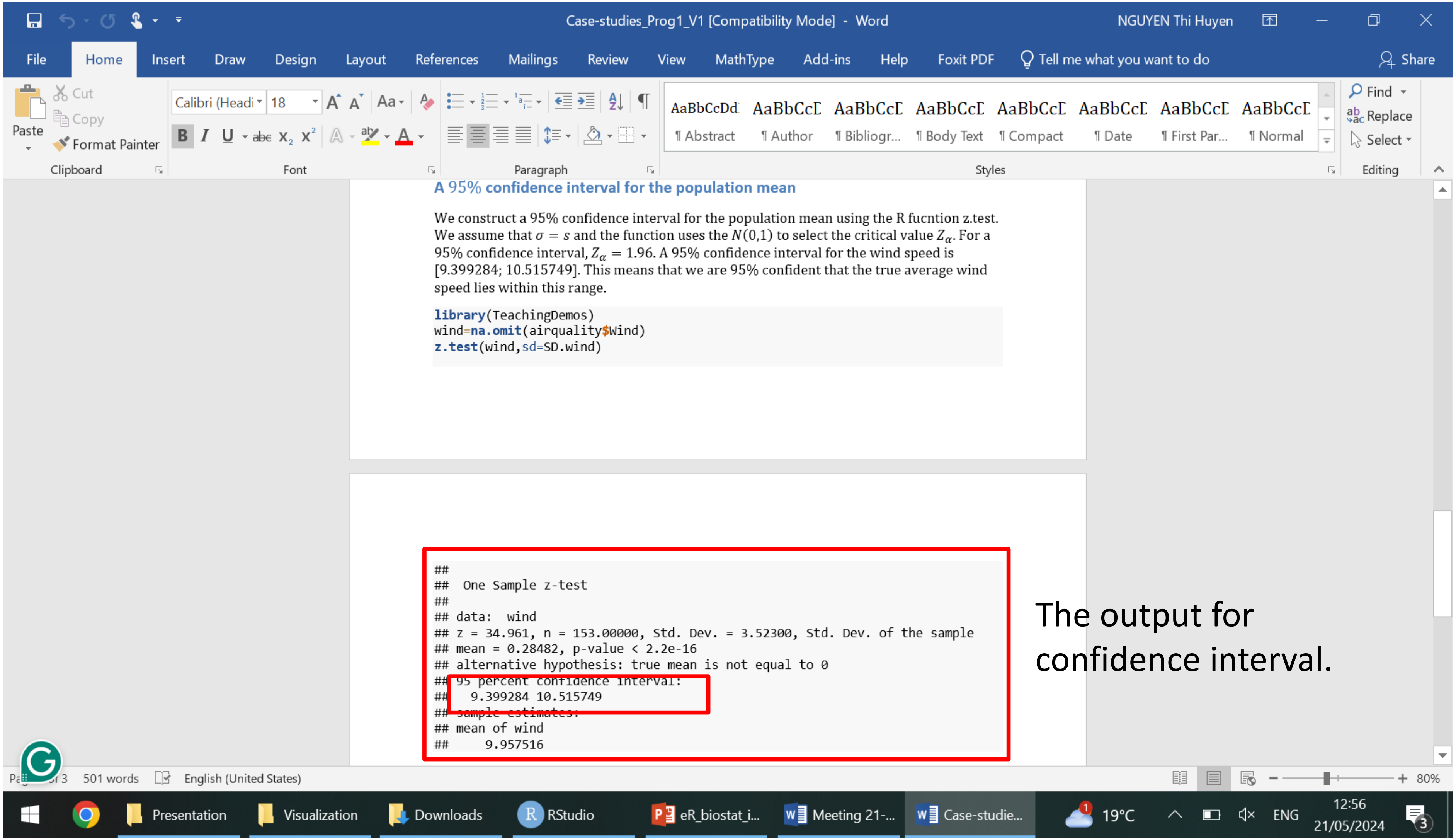
```{r, echo=TRUE, message=FALSE, warning=FALSE}
library(TeachingDemos)
wind=na.omit(airquality$Wind)
z.test(wind,sd=SD.wind)
```
~~~~

The code for confidence interval.

```
R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/
> 
```

Data	
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11
Values	

The output in the Word file



A 95% confidence interval for the population mean

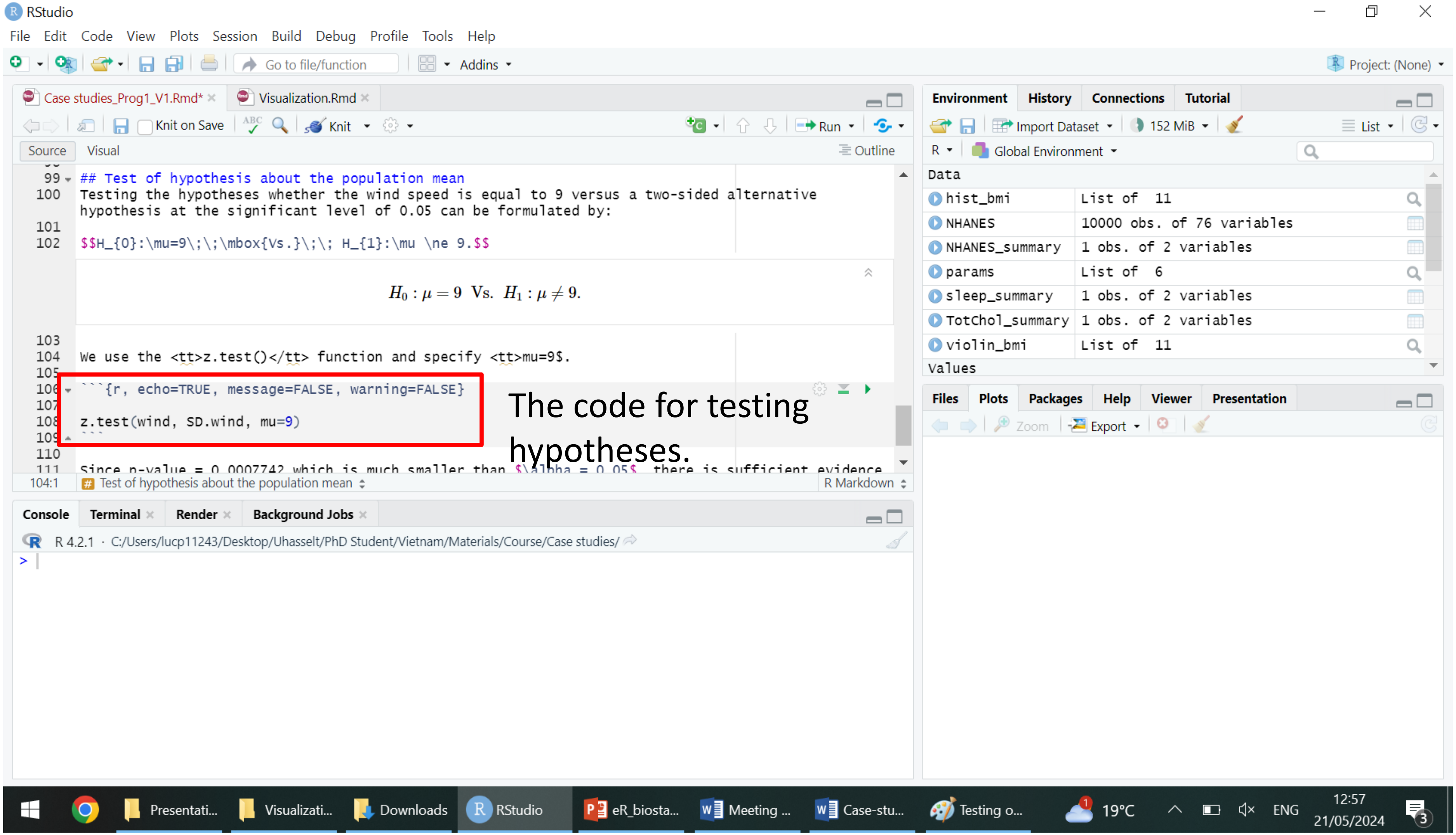
We construct a 95% confidence interval for the population mean using the R fucntion z.test. We assume that $\sigma = s$ and the function uses the $N(0,1)$ to select the critical value Z_α. For a 95% confidence interval, $Z_\alpha = 1.96$. A 95% confidence interval for the wind speed is [9.399284; 10.515749]. This means that we are 95% confident that the true average wind speed lies within this range.

```
library(TeachingDemos)
wind=na.omit(airquality$Wind)
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
```

The output for confidence interval.

Code in the Rmd file



```
## Test of hypothesis about the population mean
Testing the hypotheses whether the wind speed is equal to 9 versus a two-sided alternative
hypothesis at the significant level of 0.05 can be formulated by:

$$H_{0}:\mu=9\;\;\mbox{Vs.}\;\; H_{1}:\mu \ne 9.$$
```

$$H_0 : \mu = 9 \;\; \text{Vs.} \;\; H_1 : \mu \ne 9.$$

~~~
We use the <tt>z.test()</tt> function and specify <tt>mu=9$.

```{r, echo=TRUE, message=FALSE, warning=FALSE}

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

Since p-value = 0.0007742 which is much smaller than $\alpha = 0.05$, there is sufficient evidence
~~~

Data

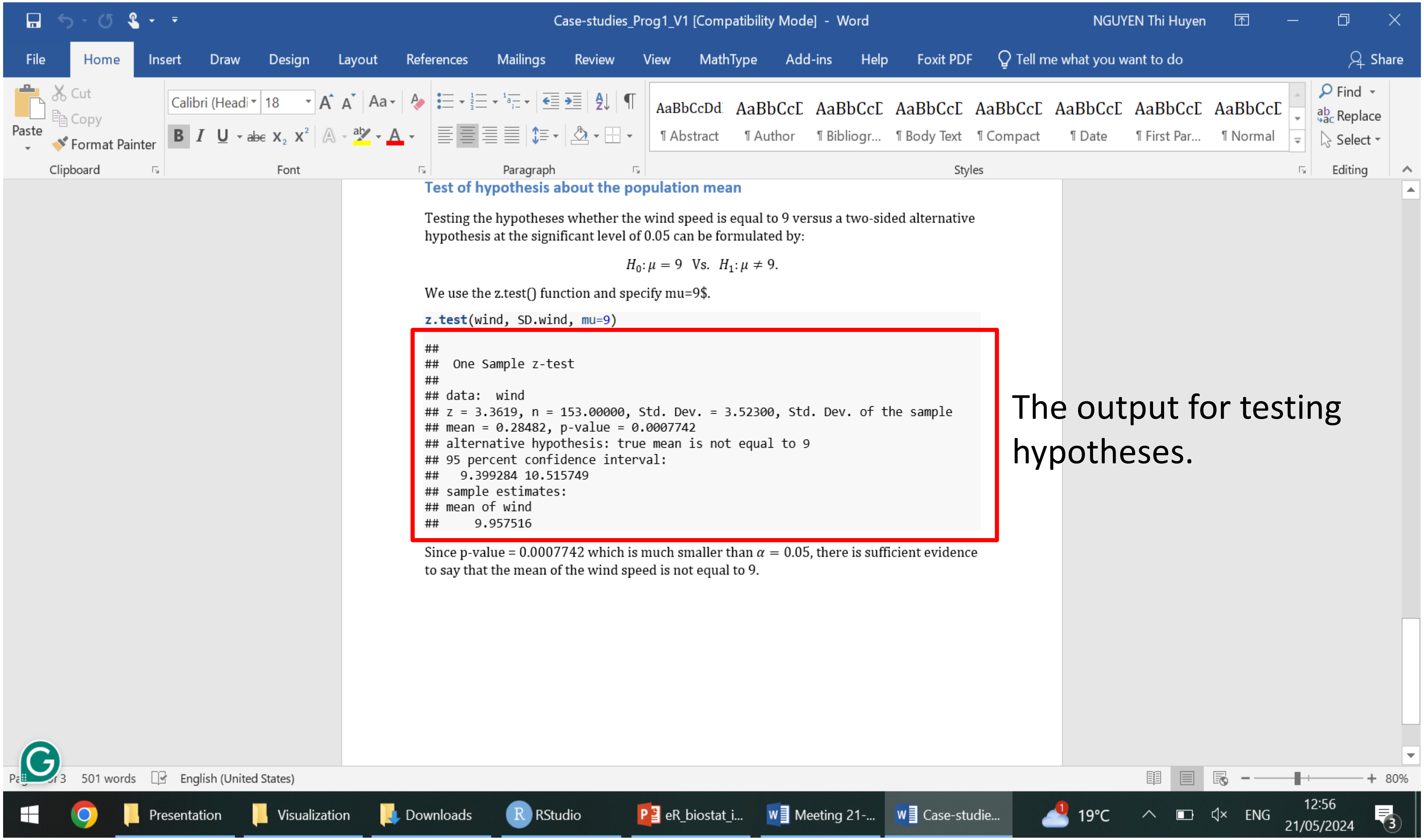
hist_bmi	List of 11
NHANES	10000 obs. of 76 variables
NHANES_summary	1 obs. of 2 variables
params	List of 6
sleep_summary	1 obs. of 2 variables
TotChol_summary	1 obs. of 2 variables
violin_bmi	List of 11

Values

```
R 4.2.1 · C:/Users/lucp11243/Desktop/Uhasselt/PhD Student/Vietnam/Materials/Course/Case studies/
>
```

The code for testing hypotheses.

The output in the Word file



Test of hypothesis about the population mean

Testing the hypotheses whether the wind speed is equal to 9 versus a two-sided alternative hypothesis at the significant level of 0.05 can be formulated by:

$$H_0: \mu = 9 \quad \text{Vs.} \quad H_1: \mu \neq 9.$$

We use the z.test() function and specify mu=9$.

```
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
```

Since p-value = 0.0007742 which is much smaller than $\alpha = 0.05$, there is sufficient evidence to say that the mean of the wind speed is not equal to 9.

The output for testing hypotheses.

Discussion

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