

# R for Newbies II - Data Analysis and Reporting

Carsten Lange

2021-03-25



# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Data Processing and Cleaning</b>	<b>7</b>
<b>3</b>	<b>Visualization</b>	<b>9</b>
3.1	Including Figures from Your Computer . . . . .	9
3.2	Scatter Plot . . . . .	9
3.3	Box Plot . . . . .	9
3.4	Histogram Plot . . . . .	9
<b>4</b>	<b>Regression Results</b>	<b>13</b>
<b>5</b>	<b>Final Words</b>	<b>15</b>



# Chapter 1

## Introduction

This book is not about math. However, everybody knows this formula:

$$c^2 = a^2 + b^2 \tag{1.1}$$

Equation (1.1) is a **very important** equations. At the end you have to calculate the square root of  $\sqrt{c^2}$

But as I said the book is not about math. It is about this workshop, which covers the following:

1. Tips and Tricks to Install R (Part I):
2. Tidy Verse and Tidy Models
  - Data Cleaning and Preparation (Part I and II)
  - Data Visualization (Part I and II)
  - Data Analysis (Part I and II)
3. Make a book in PDF format and as a web site about this workshop with *R-BookDown*

This is only a 30 minute book and consequently the content is not exactly deep.

A good book about R *BookDown* can be found here (Xie, 2015)

Material for this talk: [Click here](#).

A book that is not at all related to R can be found here (Lange, 1995)



## Chapter 2

# Data Processing and Cleaning

The original dataset *WageData* has 526 records but some are not complete. We can see this already when looking at the Table 2.1, which shows the first 10 records:

Using the code below we can clean and pre-process the data:

```
library(tidyverse)
library(tidymodels)
library(skimr)
library(readxl)
WageDataOrg <- read_excel("WageDataModifRTalk.xlsx",
                          sheet = "DataClean")

NOrig=nrow(WageDataOrg)
RecipeC1=recipe(Wage~Educ+Exper+Tenure+NonWhite+Sex, data=WageDataOrg)%>%
  step_dummy(all_nominal()) %>%
  step_meanimpute(Educ) %>%
  step_medianimpute(NonWhite_Yes) %>%
  step_knnimpute(Tenure) %>%
  step_naomit(Sex_Male) %>%
  prep()

WageData=juice(RecipeC1)

DataInfoNew=skim(WageData)

N=nrow(WageData)
```

Table 2.1: The Original Data

Wage	Educ	Exper	Tenure	NonWhite	Sex	Married	NumDep	Smsa	Region
3.10	11	2	0	NA	Female	No	2	Yes	West
3.24	12	22	2	No	Female	Yes	3	Yes	West
3.00	11	2	0	No	Male	No	2	No	West
6.00	8	44	28	No	Male	Yes	0	Yes	West
5.30	12	7	2	No	Male	Yes	1	No	West
8.75	16	9	8	No	Male	Yes	0	Yes	West
11.25	18	15	7	No	Male	No	0	Yes	West
5.00	12	5	NA	No	Female	No	0	Yes	West
3.60	NA	26	4	No	Female	No	2	Yes	West
18.18	17	22	21	No	NA	Yes	0	Yes	West

Table 2.2: The Cleaned and Processed Data

Educ	Exper	Tenure	Wage	NonWhite_Yes	Sex_Male
11.00000	2	0.0	3.10	0	0
12.00000	22	2.0	3.24	0	0
11.00000	2	0.0	3.00	0	1
8.00000	44	28.0	6.00	0	1
12.00000	7	2.0	5.30	0	1
16.00000	9	8.0	8.75	0	1
18.00000	15	7.0	11.25	0	1
12.00000	5	2.4	5.00	0	0
12.56489	26	4.0	3.60	0	0
16.00000	8	2.0	6.25	0	0

The new dataset *WageData* has 523 records and all are complete. We can see this when looking at the Table 2.2, which shows the first 10 records:



## Chapter 3

# Visualization

### 3.1 Including Figures from Your Computer



Figure 3.1: The R Logo

### 3.2 Scatter Plot

Figure 3.2 shows a scatter plot:

### 3.3 Box Plot

Figure 3.3 shows a box plot:

### 3.4 Histogram Plot

Figure 3.4 shows a histogram:

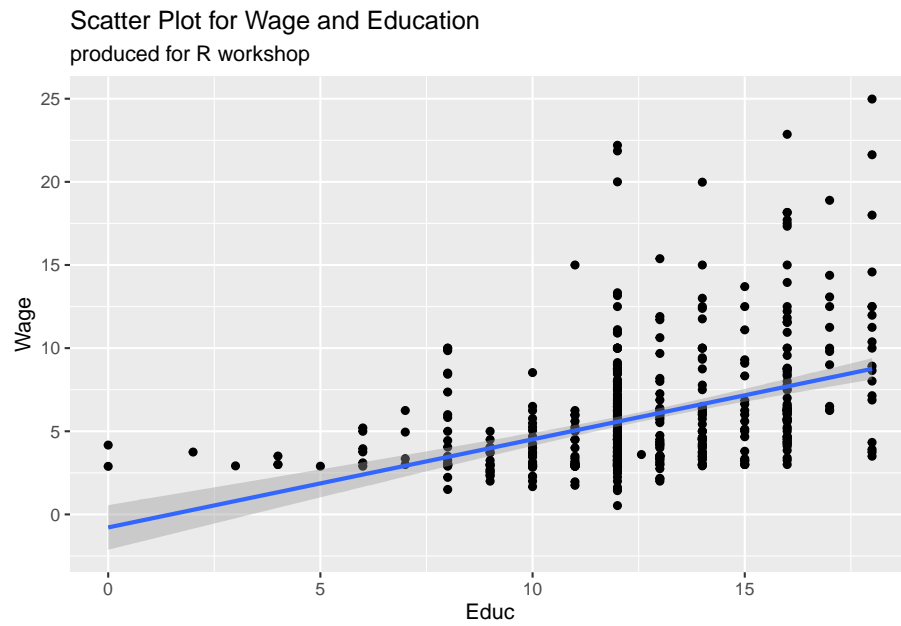


Figure 3.2: Scatter

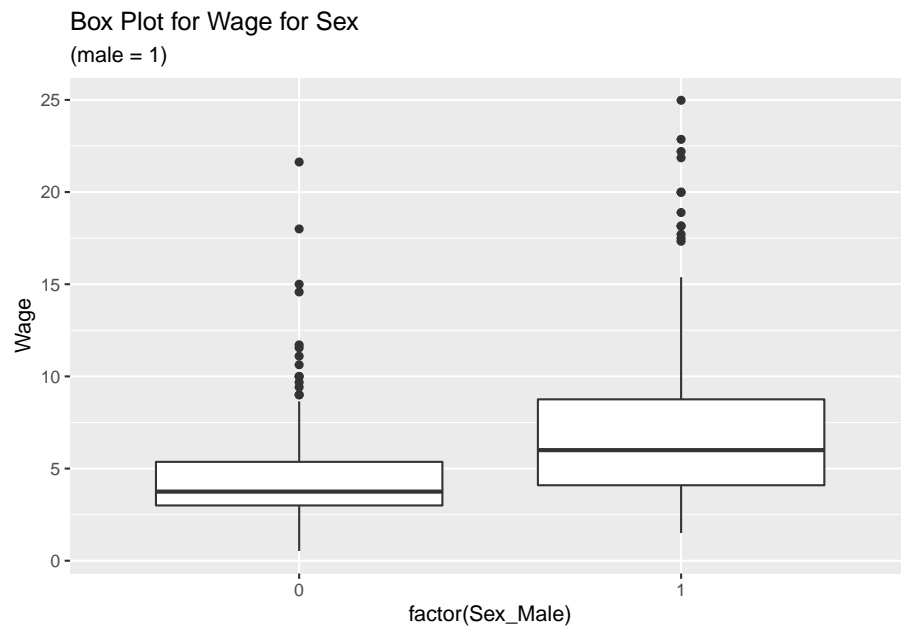


Figure 3.3: Box Plot

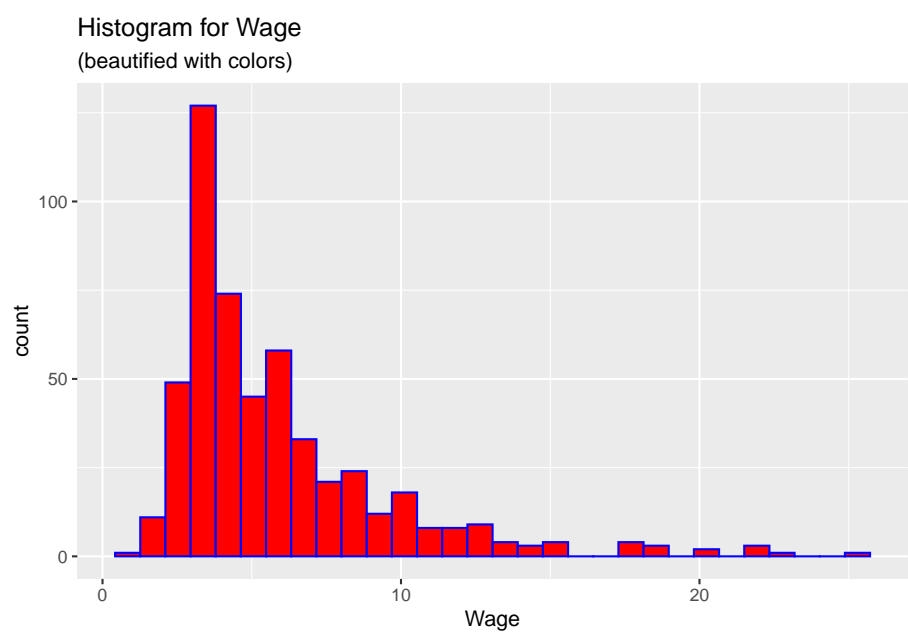


Figure 3.4: Box Plot



## Chapter 4

# Regression Results

Here is the code for the two regressions:

```
ModelLm0=lm(Wage~., data=WageData)
```

```
ModelLmFinal=lm(Wage~.-NonWhite_Yes, data=WageData)
```

Here is the result for the first regression:

term	estimate	std.error	statistic	p.value
(Intercept)	-3.2794868	0.7094184	-4.6227823	0.0000048
Educ	0.5630134	0.0494946	11.3752503	0.0000000
Exper	0.0262211	0.0115492	2.2703818	0.0235951
Tenure	0.1367860	0.0211980	6.4527798	0.0000000
NonWhite_Yes	-0.0995114	0.4255195	-0.2338585	0.8151874
Sex_Male	1.8218327	0.2647127	6.8823018	0.0000000



## Chapter 5

# Final Words

We have finished a nice book.





# Bibliography

Lange, C. (1995). *Seigniorage: Eine theoretische und empirische Analyse des staatlichen Geldschöpfungsgewinnes*. Duncker und Humblod.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.