Mid-Term Data Analysis Project Template - Coding

Vinh Do

10/23/2022

# IV. Appendix/Appendices.

**Remove the following set of code before your generate the file**

## Code

install.packages("readr")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)

install.packages("ggplot2")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)

install.packages("gapminder")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)

install.packages("gridExtra")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)

install.packages("dplyr")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)

install.packages("tidyverse")

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)

library(readr)  
library(ggplot2)  
library(gapminder)  
library(gridExtra)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':  
##   
## combine

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ tibble 3.1.8 ✔ stringr 1.4.1  
## ✔ tidyr 1.2.1 ✔ forcats 0.5.2  
## ✔ purrr 0.3.5   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::combine() masks gridExtra::combine()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

creditdata <- read\_csv("GermanCredit.csv")

## Rows: 1000 Columns: 32  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (32): OBS#, CHK\_ACCT, DURATION, HISTORY, NEW\_CAR, USED\_CAR, FURNITURE, R...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

creditdata

## # A tibble: 1,000 × 32  
## `OBS#` CHK\_ACCT DURATION HISTORY NEW\_CAR USED\_CAR FURNITURE RADIO/T…¹ EDUCA…²  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 0 6 4 0 0 0 1 0  
## 2 2 1 48 2 0 0 0 1 0  
## 3 3 3 12 4 0 0 0 0 1  
## 4 4 0 42 2 0 0 1 0 0  
## 5 5 0 24 3 1 0 0 0 0  
## 6 6 3 36 2 0 0 0 0 1  
## 7 7 3 24 2 0 0 1 0 0  
## 8 8 1 36 2 0 1 0 0 0  
## 9 9 3 12 2 0 0 0 1 0  
## 10 10 1 30 4 1 0 0 0 0  
## # … with 990 more rows, 23 more variables: RETRAINING <dbl>, AMOUNT <dbl>,  
## # SAV\_ACCT <dbl>, EMPLOYMENT <dbl>, INSTALL\_RATE <dbl>, MALE\_DIV <dbl>,  
## # MALE\_SINGLE <dbl>, MALE\_MAR\_or\_WID <dbl>, `CO-APPLICANT` <dbl>,  
## # GUARANTOR <dbl>, PRESENT\_RESIDENT <dbl>, REAL\_ESTATE <dbl>,  
## # PROP\_UNKN\_NONE <dbl>, AGE <dbl>, OTHER\_INSTALL <dbl>, RENT <dbl>,  
## # OWN\_RES <dbl>, NUM\_CREDITS <dbl>, JOB <dbl>, NUM\_DEPENDENTS <dbl>,  
## # TELEPHONE <dbl>, FOREIGN <dbl>, RESPONSE <dbl>, and abbreviated variable …

# Data Visualization 1  
creditdata$EDUCATION <-replace(creditdata$EDUCATION, creditdata$EDUCATION ==1, 'Yes')  
creditdata$EDUCATION <-replace(creditdata$EDUCATION, creditdata$EDUCATION ==0, 'No')  
creditdata$RESPONSE <-replace(creditdata$RESPONSE, creditdata$RESPONSE ==1, 'Good Credit Rating')  
creditdata$RESPONSE <-replace(creditdata$RESPONSE, creditdata$RESPONSE ==0, 'Bad Credit Rating')  
creditdata$HISTORY <-replace(creditdata$HISTORY, creditdata$HISTORY == 0, 'No credits taken')  
creditdata$HISTORY <-replace(creditdata$HISTORY, creditdata$HISTORY == 1, 'All credits at this bank paid back duly')  
creditdata$HISTORY <-replace(creditdata$HISTORY, creditdata$HISTORY == 2, 'Existing credits paid back duly until now')  
creditdata$HISTORY <-replace(creditdata$HISTORY, creditdata$HISTORY == 3, 'Delay in paying off in the past')  
creditdata$HISTORY <-replace(creditdata$HISTORY, creditdata$HISTORY == 4, 'Critical account')  
creditdata$EMPLOYMENT <-replace(creditdata$EMPLOYMENT, creditdata$EMPLOYMENT == 0, 'Unemployed')  
creditdata$EMPLOYMENT <-replace(creditdata$EMPLOYMENT, creditdata$EMPLOYMENT == 1, '<1 year')  
creditdata$EMPLOYMENT <-replace(creditdata$EMPLOYMENT, creditdata$EMPLOYMENT == 2, '1−3 years')  
creditdata$EMPLOYMENT <-replace(creditdata$EMPLOYMENT, creditdata$EMPLOYMENT == 3, '4−6 years')  
creditdata$EMPLOYMENT <-replace(creditdata$EMPLOYMENT, creditdata$EMPLOYMENT == 4, '7 years or more')  
creditdata$JOB <-replace(creditdata$JOB, creditdata$JOB == 0 ,'Unemployed')  
creditdata$JOB <-replace(creditdata$JOB, creditdata$JOB == 1 ,'Unskilled')  
creditdata$JOB <-replace(creditdata$JOB, creditdata$JOB == 2 ,'Skilled employee')  
creditdata$JOB <-replace(creditdata$JOB, creditdata$JOB == 3 ,'Management')  
creditdata$CHK\_ACCT <-replace(creditdata$CHK\_ACCT, creditdata$CHK\_ACCT == 0 ,'<0 DM')  
creditdata$CHK\_ACCT <-replace(creditdata$CHK\_ACCT, creditdata$CHK\_ACCT == 1 ,'0−200 DM')  
creditdata$CHK\_ACCT <-replace(creditdata$CHK\_ACCT, creditdata$CHK\_ACCT == 2 ,' >200 DM')  
creditdata$CHK\_ACCT <-replace(creditdata$CHK\_ACCT, creditdata$CHK\_ACCT == 3 ,'No checking account')  
creditdata$SAV\_ACCT <-replace(creditdata$SAV\_ACCT, creditdata$SAV\_ACCT == 0 ,'<100 DM')  
creditdata$SAV\_ACCT <-replace(creditdata$SAV\_ACCT, creditdata$SAV\_ACCT == 1 ,'101−500 DM')  
creditdata$SAV\_ACCT <-replace(creditdata$SAV\_ACCT, creditdata$SAV\_ACCT == 2 ,'501−1000 DM')  
creditdata$SAV\_ACCT <-replace(creditdata$SAV\_ACCT, creditdata$SAV\_ACCT == 3 ,'>1000 DM')  
creditdata$SAV\_ACCT <-replace(creditdata$SAV\_ACCT, creditdata$SAV\_ACCT == 4 ,'No saving account')

creditdata$RESPONSE <- as.factor(creditdata$RESPONSE)  
creditdata$HISTORY <- as.factor(creditdata$HISTORY)  
creditdata$EMPLOYMENT <- as.factor(creditdata$EMPLOYMENT)  
creditdata$JOB <- as.factor(creditdata$JOB)  
creditdata$EDUCATION <- as.factor(creditdata$EDUCATION)  
creditdata$CHK\_ACCT <- as.factor(creditdata$CHK\_ACCT)  
creditdata$SAV\_ACCT <- as.factor(creditdata$SAV\_ACCT)

cd <- creditdata %>%   
 select('OBS#', HISTORY, EDUCATION,EMPLOYMENT, JOB,AMOUNT, DURATION, INSTALL\_RATE, AGE, NUM\_CREDITS,JOB,SAV\_ACCT, CHK\_ACCT, RESPONSE)  
cd

## # A tibble: 1,000 × 13  
## `OBS#` HISTORY EDUCA…¹ EMPLO…² JOB AMOUNT DURAT…³ INSTA…⁴ AGE NUM\_C…⁵  
## <dbl> <fct> <fct> <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 Critical a… No 7 year… Skil… 1169 6 4 67 2  
## 2 2 Existing c… No 1−3 ye… Skil… 5951 48 2 22 1  
## 3 3 Critical a… Yes 4−6 ye… Unsk… 2096 12 2 49 1  
## 4 4 Existing c… No 4−6 ye… Skil… 7882 42 2 45 1  
## 5 5 Delay in p… No 1−3 ye… Skil… 4870 24 3 53 2  
## 6 6 Existing c… Yes 1−3 ye… Unsk… 9055 36 2 35 1  
## 7 7 Existing c… No 7 year… Skil… 2835 24 3 53 1  
## 8 8 Existing c… No 1−3 ye… Mana… 6948 36 2 35 1  
## 9 9 Existing c… No 4−6 ye… Unsk… 3059 12 2 61 1  
## 10 10 Critical a… No Unempl… Mana… 5234 30 4 28 2  
## # … with 990 more rows, 3 more variables: SAV\_ACCT <fct>, CHK\_ACCT <fct>,  
## # RESPONSE <fct>, and abbreviated variable names ¹​EDUCATION, ²​EMPLOYMENT,  
## # ³​DURATION, ⁴​INSTALL\_RATE, ⁵​NUM\_CREDITS

### table 1 - The Number and Proportion of Credit Rating by Education

jobplace <- cd %>%   
 group\_by(EDUCATION, RESPONSE) %>%   
 summarise(N = n()) %>%   
 mutate(Percentage = round(N / sum(N) \* 100 , 3))  
  
  
knitr::kable(jobplace, format = "markdown")

| EDUCATION | RESPONSE | N | Percentage |
| --- | --- | --- | --- |
| No | Bad Credit Rating | 278 | 29.263 |
| No | Good Credit Rating | 672 | 70.737 |
| Yes | Bad Credit Rating | 22 | 44.000 |
| Yes | Good Credit Rating | 28 | 56.000 |

### table 2 The Number and Proportion of Saving Account by Credit Rating

account <- cd %>%   
 group\_by(RESPONSE, SAV\_ACCT) %>%   
 summarise(savingacct = n()) %>%   
 mutate(Percentage = round(savingacct / sum(savingacct) \* 100 , 3))  
  
knitr::kable(account, format = "markdown")

| RESPONSE | SAV\_ACCT | savingacct | Percentage |
| --- | --- | --- | --- |
| Bad Credit Rating | <100 DM | 217 | 72.333 |
| Bad Credit Rating | >1000 DM | 6 | 2.000 |
| Bad Credit Rating | 101−500 DM | 34 | 11.333 |
| Bad Credit Rating | 501−1000 DM | 11 | 3.667 |
| Bad Credit Rating | No saving account | 32 | 10.667 |
| Good Credit Rating | <100 DM | 386 | 55.143 |
| Good Credit Rating | >1000 DM | 42 | 6.000 |
| Good Credit Rating | 101−500 DM | 69 | 9.857 |
| Good Credit Rating | 501−1000 DM | 52 | 7.429 |
| Good Credit Rating | No saving account | 151 | 21.571 |

### table 3 - Bad Credit Rating Account by Employment Years

EmploymentBadCredit <- cd %>%   
 filter(RESPONSE == "Bad Credit Rating") %>%   
 group\_by(RESPONSE, EMPLOYMENT) %>%   
 summarize(Number = n()) %>%   
 mutate(Percentage = round(Number / sum(Number) \* 100 , 3)) %>%   
 arrange(EMPLOYMENT)  
  
knitr::kable(EmploymentBadCredit, format = "markdown")

| RESPONSE | EMPLOYMENT | Number | Percentage |
| --- | --- | --- | --- |
| Bad Credit Rating | <1 year | 70 | 23.333 |
| Bad Credit Rating | 1−3 years | 104 | 34.667 |
| Bad Credit Rating | 4−6 years | 39 | 13.000 |
| Bad Credit Rating | 7 years or more | 64 | 21.333 |
| Bad Credit Rating | Unemployed | 23 | 7.667 |

### table 4 - Good Credit Rating Account by Employment Years

EmploymentGoodCredit <- cd %>%   
 filter(RESPONSE == "Good Credit Rating") %>%   
 group\_by(RESPONSE, EMPLOYMENT) %>%   
 summarize(Number = n()) %>%   
 mutate(Percentage = round(Number / sum(Number) \* 100 , 3)) %>%   
 arrange(EMPLOYMENT)  
  
knitr::kable(EmploymentGoodCredit, format = "markdown")

| RESPONSE | EMPLOYMENT | Number | Percentage |
| --- | --- | --- | --- |
| Good Credit Rating | <1 year | 102 | 14.571 |
| Good Credit Rating | 1−3 years | 235 | 33.571 |
| Good Credit Rating | 4−6 years | 135 | 19.286 |
| Good Credit Rating | 7 years or more | 189 | 27.000 |
| Good Credit Rating | Unemployed | 39 | 5.571 |

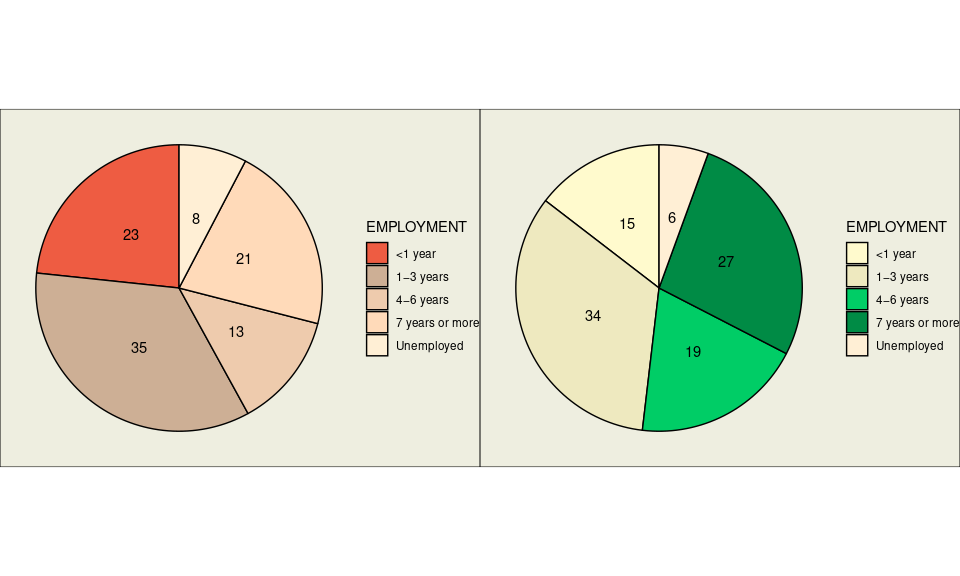
## Including Plots

### Figure 1 - Bad Credit Rating Account and Good Credit Rating Account by Employment Years

figure1 <- ggplot(data = EmploymentBadCredit ,mapping = aes(x ="", y= Percentage, fill = EMPLOYMENT, warning = FALSE)) +  
 geom\_bar(stat="identity", width=1, color = "black") +  
 geom\_text(aes(label = round(Percentage)),  
 position = position\_stack(vjust = 0.5),  
 show.legend = FALSE) +  
 coord\_polar("y", start=0) +  
 scale\_fill\_manual(values = c("tomato2","peachpuff3", "peachpuff2", "peachpuff1", "papayawhip")) +  
 theme\_void() +  
 theme( plot.background = element\_rect(fill = "ivory2"))

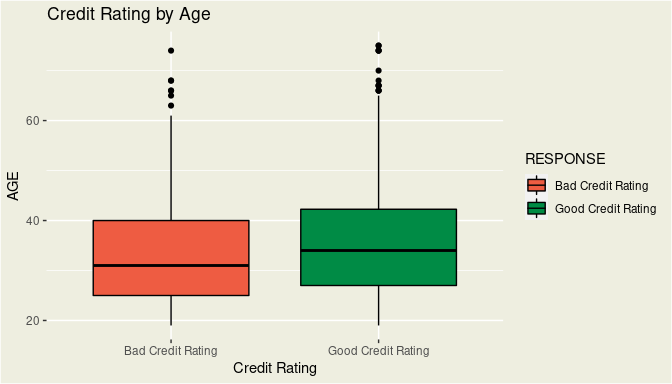
figure2 <- ggplot(data = EmploymentGoodCredit ,mapping = aes(x ="", y= Percentage, fill = EMPLOYMENT)) +  
 geom\_col(stat="identity", width=1, color = "black") +  
 geom\_text(aes(label = round(Percentage)),  
 position = position\_stack(vjust = 0.5),  
 show.legend = FALSE) +  
 coord\_polar("y", start=0) +  
 scale\_fill\_manual(values = c("lemonchiffon1", "lemonchiffon2", "springgreen3","springgreen4", "papayawhip")) +  
 theme\_void() +  
 theme(plot.background = element\_rect(fill = "ivory2"))

grid.arrange(figure1, figure2, nrow = 1)



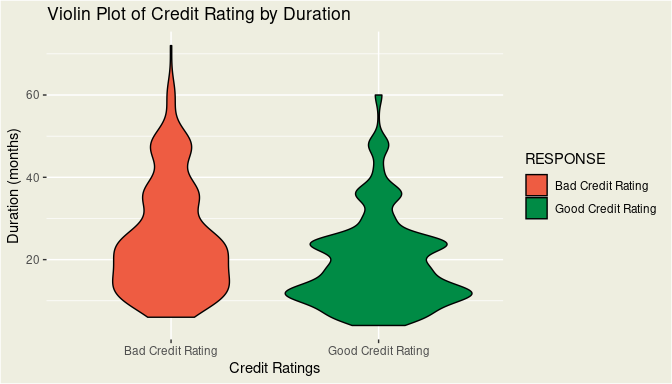
### Figure 2 - The Total Credit Rating by Age

ggplot(data = cd, mapping = aes(x = AGE, y = RESPONSE, fill = RESPONSE)) +  
 geom\_boxplot(color = "black") +  
 coord\_flip() +  
 scale\_fill\_manual(values = c("tomato2", "springgreen4")) +  
 labs(title = "Credit Rating by Age", x = "AGE", y = "Credit Rating") +  
 theme(legend.background = element\_rect(fill = "ivory2"),  
 panel.background = element\_rect(fill = "ivory2"),  
 plot.background = element\_rect(fill = "ivory2"))



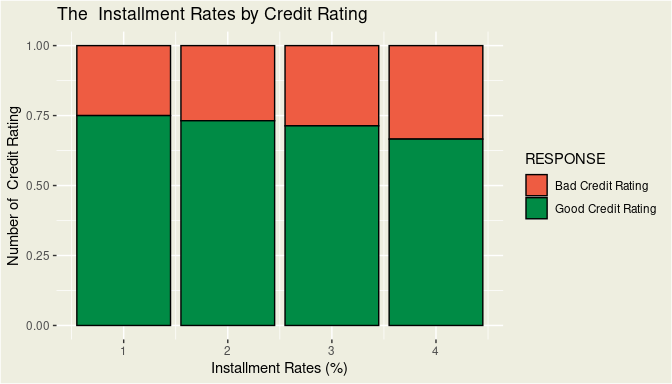
### Figure 3 - Violin Plot of Credit Rating by Duration

ggplot(data = cd ,mapping = aes(x = DURATION,y = RESPONSE , fill = RESPONSE)) +  
 geom\_violin(color = "black") + geom\_jitter(width = 0.001, alpha = 0.001) +  
 scale\_fill\_manual(values = c("tomato2", "springgreen4")) +  
 coord\_flip() +  
 labs(title = "Violin Plot of Credit Rating by Duration", y = "Credit Ratings", x = "Duration (months)") +  
 theme(legend.background = element\_rect(fill = "ivory2"),  
 panel.background = element\_rect(fill = "ivory2"),  
 plot.background = element\_rect(fill = "ivory2"))



### Figure 4 - The Credit Rating Accounts by Installment Rate

ggplot(data = cd,mapping = aes(x = INSTALL\_RATE, fill = RESPONSE)) +  
 geom\_bar(stat= "count", position = "fill", color = "black") +  
 scale\_fill\_manual(values = c("Bad Credit Rating" = "tomato2","Good Credit Rating" = "springgreen4")) +  
 labs(title = "The Installment Rates by Credit Rating ", x = "Installment Rates (%)", y = "Number of Credit Rating") +  
 theme(legend.background = element\_rect(fill = "ivory2"),  
 panel.background = element\_rect(fill = "ivory2"),  
 plot.background = element\_rect(fill = "ivory2"))



— End of Instructions —