Analysis of Online Shopping Data

Kultigin Bozdemir

3/31/2020

Introduction

A data data set about online shopping is provided the by the lecturer for the class practices. In the following sections, basic statistical concepts and R tool will be used to analyze the mentioned data.

Setting the Environment

As the reader would notice, R environment is used to analyze the data. Furthermore, R Markdown has been choosen to present the analysis along with the relevant R codes. First step is to set up the directory and to load the tidyverse library which is a must for R analysis.

```
knitr::opts_knit$set(root.dir = getwd())
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
getwd()
```

```
## [1] "/Users/kultiginbozdemir/GitHub/online_shop"
```

Exploration and the Transformation of the Data

The following code imports the data into R. And it immediately will be transformed into a data frame, whose first rows are below printed.

```
df<-read.delim2("iw_customer.txt")
head(df)
```

```
##   owner customerNo salutation firstname surname postcode        city
## 1    IW  KNR000001       Frau     Abcde Hijklmn    65232 Taunusstein
## 2    IW  KNR000002       Frau     Abcde Hijklmn    26904   B\xf6rger
## 3    IW  KNR000003       Frau     Abcde Hijklmn    78333    Stockach
## 4    IW  KNR000011       Frau     Abcde Hijklmn    79618 Rheinfelden
## 5    IW  KNR000020       Frau     Abcde Hijklmn    64625    Bensheim
## 6    IW  KNR000022       Herr     Abcde Hijklmn    45138       Essen
##          street         eMail newsletter              birthdate   riskID
## 1 Opqrst-Street mail@mail.com          1 1968-01-07 00:00:00.000 69918055
## 2 Opqrst-Street mail@mail.com          0 1978-10-10 00:00:00.000 92843675
## 3 Opqrst-Street mail@mail.com          1 1967-04-29 00:00:00.000 11272894
## 4 Opqrst-Street mail@mail.com          1 1974-10-21 00:00:00.000 86364865
## 5 Opqrst-Street mail@mail.com          1 1969-03-23 00:00:00.000 79987284
```

```
## 6 Opqrst-Street mail@mail.com          1 1957-06-04 00:00:00.000 87651195
##   credit creditLimit
## 1    400           1
## 2   1000           1
## 3      0           2
## 4    400           1
## 5    500           1
## 6    400           1
```

However, calling a summary function is usually necessary to figure out the general structure and content of the data. From those two pieces of analysis, it is assessed that many of columns are to be excluded for further analysis for the sake of simplicity.

```
summary(df)
```

```
##  owner           customerNo     salutation    firstname         surname
##  IW:247065   KNR000001:     1   Frau:204666   Abcde:247065   Hijklmn:247065
##              KNR000002:     1   Herr: 42399
##              KNR000003:     1
##              KNR000011:     1
##              KNR000020:     1
##              KNR000022:     1
##              (Other)  :247059
##     postcode                 city                  street
##  14532  :   357   Berlin       :  7506   Opqrst-Street:247065
##  61440  :   345   Hamburg      :  6501
##  76829  :   329   M\xfcnchen   :  5942
##  94315  :   289   K\xf6ln      :  3281
##  33378  :   278   D\xfcsseldorf:  2719
##  40489  :   269   Frankfurt    :  2527
##  (Other):245198   (Other)      :218589
##            eMail          newsletter                      birthdate
##  mail@mail.com:247065   Min.   :0.0000   1970-01-01 00:00:00.000:   168
##                         1st Qu.:0.0000   1969-08-05 00:00:00.000:   142
##                         Median :1.0000   1964-03-17 00:00:00.000:   134
##                         Mean   :0.6884   1970-08-14 00:00:00.000:   132
##                         3rd Qu.:1.0000   1970-06-13 00:00:00.000:   130
##                         Max.   :1.0000   1968-01-16 00:00:00.000:   129
##                                          (Other)                :246230
##      riskID             credit        creditLimit
##  Min.   :       1   Min.   :   0.0   Min.   :1.000
##  1st Qu.:24202294   1st Qu.: 400.0   1st Qu.:1.000
##  Median :49366895   Median : 500.0   Median :1.000
##  Mean   :49373379   Mean   : 688.4   Mean   :1.012
##  3rd Qu.:75136884   3rd Qu.: 800.0   3rd Qu.:1.000
##  Max.   :99998896   Max.   :2000.0   Max.   :2.000
##
```

The following code drops those columns. It is necessary to keep anonymous information to drive some conclusions to improve the business, whereas the individual information is discarded such as names, email addresses etc. However, the birtdate is kept because it will be soon transformed into "age" information which might be useful for future analysis.

```
drops <- c("riskID","eMail", "newsletter",  "owner",
           "creditLimit", "street","firstname","surname", "owner","customerNo")
df<-df[ , !(names(df) %in% drops)]
```

Now, the "age" column is created from the "birthdate" and present time information. The age is roughly calculated by dividing the total days by 365 for the sake of the simplicity. However, one final adjusment has to made to improve it. "salution" column needs to be transformed into "gender" column. Then finally, the shrunken version of the data is printed again.

```
df$birthdate<-as.Date(df$birthdate)
df$age<- as.integer((Sys.Date()-df$birthdate)/365)
df$gender<-ifelse ( df$salutation=="Frau", "Female", "Male" )
df$gender<-as.factor(df$gender)
drops <- c("salutation", "birthdate")
df<-df[ , !(names(df) %in% drops)]
head(df)
```

```
##   postcode        city credit age gender
## 1    65232 Taunusstein    400  52 Female
## 2    26904   B\xf6rger   1000  41 Female
## 3    78333    Stockach      0  52 Female
## 4    79618 Rheinfelden    400  45 Female
## 5    64625    Bensheim    500  51 Female
## 6    45138       Essen    400  62   Male
```

Now, some basic business questions can be answered from this dataframe, such as; 1. Is there any correlation between age and credit? 2. Which age group or gender consumes more in which city? 3. Which region is promising more? (from postal code)

Descriptive Analysis

The anwering the questions which are given above can be defined in statistical descriptive analysis, although they are here a few of many business questions.

The following code yields the correlation result, which shows a slight or weak correlation between "age" and "credit".

```
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
corr<-cor(df$age,df$credit)
print(corr)
```

```
## [1] 0.1160203
```

The initial number of postcodes identifies the regions in Germany. (Joyce 2020) The shopping numbers of each region are seen below. Number 4, which covers mostly NRW is busiest region in Germany.

```
table(substr(df$postcode,1,1))
```

```
##
##     0     1     2     3     4     5     6     7     8     9     A
```

```
## 11315 13823 27673 28589 29710 26439 25152 31217 29296 23850      1
```

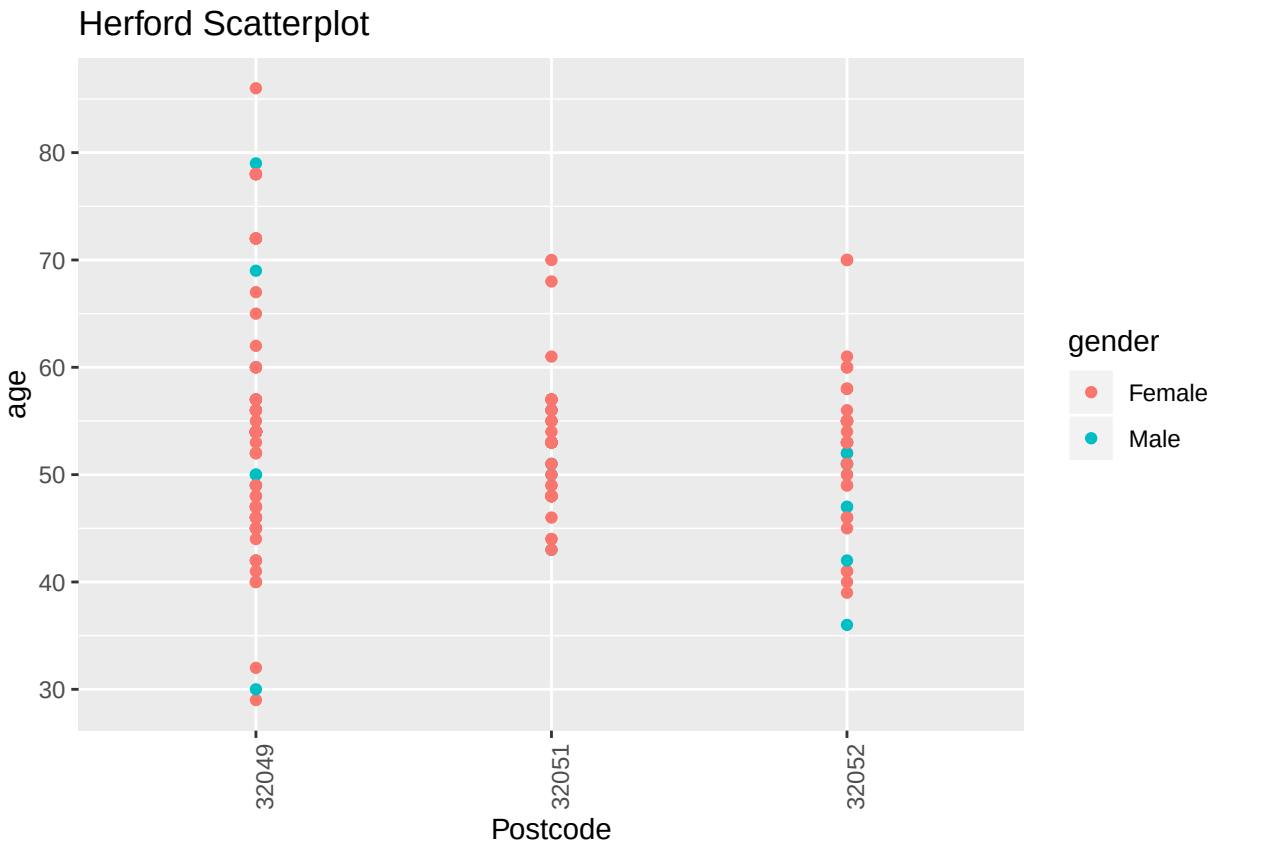
To see the number of shoppings per city, summary function can be used for the city column, instead of plotting all cities that would give less information due to large number of cities in germany besides the size of the data. Most promising 5 cities are printed below followed by the lesast ones.

```
summary(df$city)[1:5]
```

```
##        Berlin       Hamburg    M\xfcnchen       K\xf6ln D\xfcsseldorf
##          7506          6501          5942          3281          2719
```

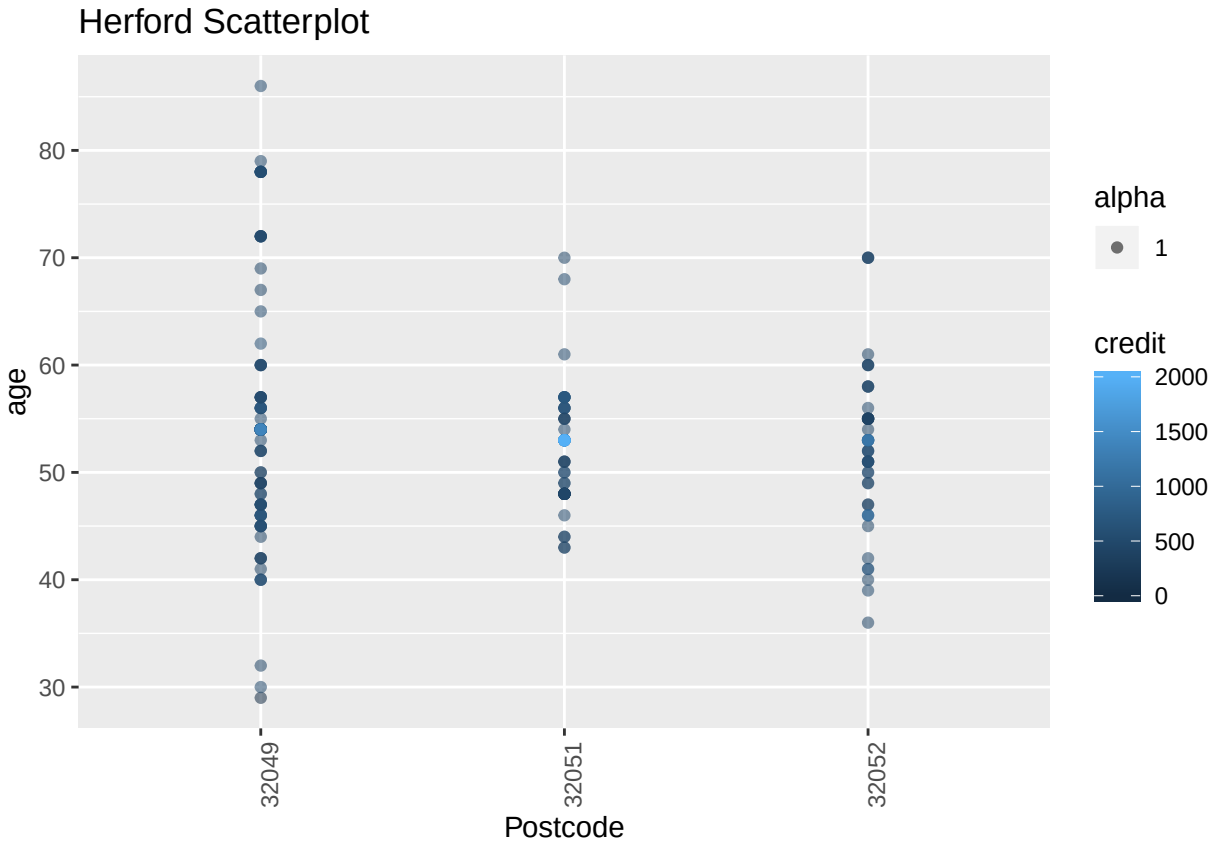
City Herford, which has a small shopping numbers has been selected to plot the number of transactions since plotting large cities gives less information due to the comlexity. On the x axis, the postal neigbourhoods , on the y axis the ages are plotted, while the colors are representing the gender thanks to ggplot library. A conclusion from the plot can be driven that there is no male customer in 32051 postal zone. Similarly, there are elderly customers (above 65) in zone 32049.

```
#Herford
herford<-subset(df, df$city=="Herford")
gg<-ggplot(herford, aes(x=postcode, y=age, color=gender)) +
        geom_point() + labs(title="Herford Scatterplot", x="Postcode", y="age")
gg+theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Next, the credit dimension is printed in the same plot instead of gender in the plot below. The plot shows, most of customers have less credit limits in this city.

```
gg<-ggplot(herford, aes(x=postcode, y=age, alpha=1, color=credit)) +
        geom_point() + labs(title="Herford Scatterplot", x="Postcode", y="age")
gg+theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Some important decisions can be made on the information derived from the plots above, which are asked at he beginning this section. For example which gender and which age group are living in which part of the city. Therefore, the products can be customized accordingly.

Conclusion

In this assignment, some of basic methods in R have been used in R Markdown to show the results accompanied with corresponding R codes. Some of very basic business questions have been answered in the text. Those are mainly the age, credit and gender profile of customers in different regions, cities or city zones. Such kind of analysis help the desion making process to improve the profitability of the business.

Bibliography

Joyce, Paul. 2020. "Using Bibtex: A Short Guide." 2020. http://joycep.myweb.port.ac.uk/abinitio/chap11-14.html.