Assignment 1

Pranay Kumar Kodeboyina

2023-01-29

install.packages(“tidyverse”) install.packages(“pillar”) install.packages (“gmodels”); install.packages(“MASS”) library (gmodels);

# Data set source :https://www.kaggle.com/datasets/rajeshrampure/black-friday-sale  
  
#Printing the data sets using head and tail command to get the first 10 observations  
  
#setting the working directory to the Assignment Folder  
setwd("/Users/kodeboyina/Documents/Kent State/Sem1/Fundamentals of ML/Assigment1")  
  
#Loading Black Friday sales csv data Import the data set into R  
Friday\_data <- read.csv("data/train.csv")  
  
#Loading the first 10 rows data to print  
head(Friday\_data, 10)

## User\_ID Product\_ID Gender Age Occupation City\_Category  
## 1 1000001 P00069042 F 0-17 10 A  
## 2 1000001 P00248942 F 0-17 10 A  
## 3 1000001 P00087842 F 0-17 10 A  
## 4 1000001 P00085442 F 0-17 10 A  
## 5 1000002 P00285442 M 55+ 16 C  
## 6 1000003 P00193542 M 26-35 15 A  
## 7 1000004 P00184942 M 46-50 7 B  
## 8 1000004 P00346142 M 46-50 7 B  
## 9 1000004 P0097242 M 46-50 7 B  
## 10 1000005 P00274942 M 26-35 20 A  
## Stay\_In\_Current\_City\_Years Marital\_Status Product\_Category\_1  
## 1 2 0 3  
## 2 2 0 1  
## 3 2 0 12  
## 4 2 0 12  
## 5 4+ 0 8  
## 6 3 0 1  
## 7 2 1 1  
## 8 2 1 1  
## 9 2 1 1  
## 10 1 1 8  
## Product\_Category\_2 Product\_Category\_3 Purchase  
## 1 NA NA 8370  
## 2 6 14 15200  
## 3 NA NA 1422  
## 4 14 NA 1057  
## 5 NA NA 7969  
## 6 2 NA 15227  
## 7 8 17 19215  
## 8 15 NA 15854  
## 9 16 NA 15686  
## 10 NA NA 7871

#Loading the Last 10 rows to print  
tail(Friday\_data, 10)

## User\_ID Product\_ID Gender Age Occupation City\_Category  
## 191 1000035 P00285442 M 46-50 1 C  
## 192 1000035 P00121142 M 46-50 1 C  
## 193 1000035 P00228242 M 46-50 1 C  
## 194 1000035 P00129542 M 46-50 1 C  
## 195 1000035 P00309542 M 46-50 1 C  
## 196 1000035 P00106742 M 46-50 1 C  
## 197 1000036 P00102642 M 26-35 3 B  
## 198 1000036 P00127442 M 26-35 3 B  
## 199 1000036 P00345642 M 26-35 3 B  
## 200 1000036 P00357842 M 26-35 3 B  
## Stay\_In\_Current\_City\_Years Marital\_Status Product\_Category\_1  
## 191 4+ 1 8  
## 192 4+ 1 8  
## 193 4+ 1 8  
## 194 4+ 1 8  
## 195 4+ 1 5  
## 196 4+ 1 3  
## 197 0 0 4  
## 198 0 0 1  
## 199 0 0 1  
## 200 0 0 5  
## Product\_Category\_2 Product\_Category\_3 Purchase  
## 191 NA NA 6139  
## 192 NA NA 5897  
## 193 NA NA 3990  
## 194 14 NA 6153  
## 195 14 NA 5319  
## 196 5 NA 10767  
## 197 8 9 2832  
## 198 2 15 8066  
## 199 15 16 15812  
## 200 NA NA 6967

# Providing the statistics for the Quantiative and Qualitative variables  
  
# Reading the data type for the purchase varible  
class(Friday\_data$Purchase)

## [1] "integer"

# Mean for the purchase amount Quantitative Variable  
print(paste(mean(Friday\_data$Purchase), " - Mean purchase amount in $ for black Friday sale"))

## [1] "9259.2 - Mean purchase amount in $ for black Friday sale"

print(paste(median(Friday\_data$Purchase, na.rm = FALSE), " - Median amount in $ for black Friday sale"))

## [1] "8002.5 - Median amount in $ for black Friday sale"

#Minimum Purchase amount in the Black Friday sale  
min(Friday\_data$Purchase)

## [1] 584

#Maximum Purchase amount in the Black Friday sale  
max(Friday\_data$Purchase)

## [1] 19672

# Displaying the counts for few Variables and glimse of the data set  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

#Counting the gender variable  
dplyr:: count(Friday\_data, Gender)

## Gender n  
## 1 F 68  
## 2 M 132

#Glimse function provide the summary of the data with the data type and categories  
glimpse(Friday\_data)

## Rows: 200  
## Columns: 12  
## $ User\_ID <int> 1000001, 1000001, 1000001, 1000001, 1000002…  
## $ Product\_ID <chr> "P00069042", "P00248942", "P00087842", "P00…  
## $ Gender <chr> "F", "F", "F", "F", "M", "M", "M", "M", "M"…  
## $ Age <chr> "0-17", "0-17", "0-17", "0-17", "55+", "26-…  
## $ Occupation <int> 10, 10, 10, 10, 16, 15, 7, 7, 7, 20, 20, 20…  
## $ City\_Category <chr> "A", "A", "A", "A", "C", "A", "B", "B", "B"…  
## $ Stay\_In\_Current\_City\_Years <chr> "2", "2", "2", "2", "4+", "3", "2", "2", "2…  
## $ Marital\_Status <int> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0…  
## $ Product\_Category\_1 <int> 3, 1, 12, 12, 8, 1, 1, 1, 1, 8, 5, 8, 8, 1,…  
## $ Product\_Category\_2 <int> NA, 6, NA, 14, NA, 2, 8, 15, 16, NA, 11, NA…  
## $ Product\_Category\_3 <int> NA, 14, NA, NA, NA, NA, 17, NA, NA, NA, NA,…  
## $ Purchase <int> 8370, 15200, 1422, 1057, 7969, 15227, 19215…

#Cross Tabulation for the Sample space across differnt age groups vs Gender and displaying the row and column percentages  
  
  
library(gmodels)  
CrossTable(Friday\_data$Gender, Friday\_data$Age, prop.r=TRUE, prop.c=TRUE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 200   
##   
##   
## | Friday\_data$Age   
## Friday\_data$Gender | 0-17 | 18-25 | 26-35 | 36-45 | 46-50 | 51-55 | 55+ | Row Total |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## F | 4 | 25 | 12 | 23 | 0 | 4 | 0 | 68 |   
## | 0.381 | 7.370 | 3.459 | 5.888 | 9.520 | 1.103 | 1.700 | |   
## | 0.059 | 0.368 | 0.176 | 0.338 | 0.000 | 0.059 | 0.000 | 0.340 |   
## | 0.250 | 0.581 | 0.200 | 0.561 | 0.000 | 0.571 | 0.000 | |   
## | 0.020 | 0.125 | 0.060 | 0.115 | 0.000 | 0.020 | 0.000 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## M | 12 | 18 | 48 | 18 | 28 | 3 | 5 | 132 |   
## | 0.196 | 3.796 | 1.782 | 3.033 | 4.904 | 0.568 | 0.876 | |   
## | 0.091 | 0.136 | 0.364 | 0.136 | 0.212 | 0.023 | 0.038 | 0.660 |   
## | 0.750 | 0.419 | 0.800 | 0.439 | 1.000 | 0.429 | 1.000 | |   
## | 0.060 | 0.090 | 0.240 | 0.090 | 0.140 | 0.015 | 0.025 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 16 | 43 | 60 | 41 | 28 | 7 | 5 | 200 |   
## | 0.080 | 0.215 | 0.300 | 0.205 | 0.140 | 0.035 | 0.025 | |   
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

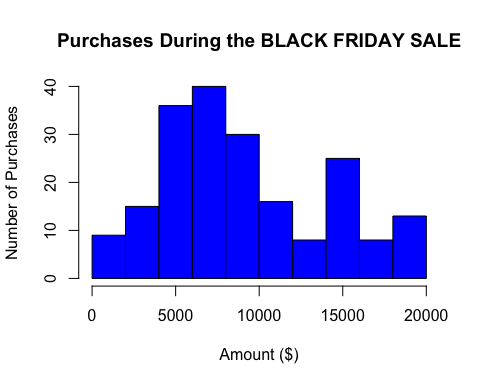
#Transformation of the categorical Age Variable to Double using mutate function by considering the midpoints as factors for age  
  
Friday\_data %>% mutate(Age\_1 = recode(Age,   
"0-17" = 17,  
"18-25" = 21.5,  
"26-35" = 30.5,  
"36-45" = 40.5,  
"46-50" = 48,  
"51-55" = 53,  
"55+" = 55,  
.default = 0.0)) %>%  
#displaying the age for the below variable  
 select(User\_ID, Age, Age\_1)

## User\_ID Age Age\_1  
## 1 1000001 0-17 17.0  
## 2 1000001 0-17 17.0  
## 3 1000001 0-17 17.0  
## 4 1000001 0-17 17.0  
## 5 1000002 55+ 55.0  
## 6 1000003 26-35 30.5  
## 7 1000004 46-50 48.0  
## 8 1000004 46-50 48.0  
## 9 1000004 46-50 48.0  
## 10 1000005 26-35 30.5  
## 11 1000005 26-35 30.5  
## 12 1000005 26-35 30.5  
## 13 1000005 26-35 30.5  
## 14 1000005 26-35 30.5  
## 15 1000006 51-55 53.0  
## 16 1000006 51-55 53.0  
## 17 1000006 51-55 53.0  
## 18 1000006 51-55 53.0  
## 19 1000007 36-45 40.5  
## 20 1000008 26-35 30.5  
## 21 1000008 26-35 30.5  
## 22 1000008 26-35 30.5  
## 23 1000008 26-35 30.5  
## 24 1000008 26-35 30.5  
## 25 1000008 26-35 30.5  
## 26 1000009 26-35 30.5  
## 27 1000009 26-35 30.5  
## 28 1000009 26-35 30.5  
## 29 1000009 26-35 30.5  
## 30 1000010 36-45 40.5  
## 31 1000010 36-45 40.5  
## 32 1000010 36-45 40.5  
## 33 1000010 36-45 40.5  
## 34 1000010 36-45 40.5  
## 35 1000010 36-45 40.5  
## 36 1000010 36-45 40.5  
## 37 1000010 36-45 40.5  
## 38 1000010 36-45 40.5  
## 39 1000010 36-45 40.5  
## 40 1000010 36-45 40.5  
## 41 1000010 36-45 40.5  
## 42 1000010 36-45 40.5  
## 43 1000010 36-45 40.5  
## 44 1000010 36-45 40.5  
## 45 1000010 36-45 40.5  
## 46 1000010 36-45 40.5  
## 47 1000010 36-45 40.5  
## 48 1000011 26-35 30.5  
## 49 1000011 26-35 30.5  
## 50 1000011 26-35 30.5  
## 51 1000012 26-35 30.5  
## 52 1000012 26-35 30.5  
## 53 1000013 46-50 48.0  
## 54 1000013 46-50 48.0  
## 55 1000013 46-50 48.0  
## 56 1000014 36-45 40.5  
## 57 1000015 26-35 30.5  
## 58 1000015 26-35 30.5  
## 59 1000015 26-35 30.5  
## 60 1000015 26-35 30.5  
## 61 1000015 26-35 30.5  
## 62 1000015 26-35 30.5  
## 63 1000015 26-35 30.5  
## 64 1000015 26-35 30.5  
## 65 1000015 26-35 30.5  
## 66 1000016 36-45 40.5  
## 67 1000016 36-45 40.5  
## 68 1000017 51-55 53.0  
## 69 1000017 51-55 53.0  
## 70 1000017 51-55 53.0  
## 71 1000018 18-25 21.5  
## 72 1000018 18-25 21.5  
## 73 1000018 18-25 21.5  
## 74 1000018 18-25 21.5  
## 75 1000018 18-25 21.5  
## 76 1000018 18-25 21.5  
## 77 1000018 18-25 21.5  
## 78 1000018 18-25 21.5  
## 79 1000018 18-25 21.5  
## 80 1000018 18-25 21.5  
## 81 1000018 18-25 21.5  
## 82 1000018 18-25 21.5  
## 83 1000018 18-25 21.5  
## 84 1000018 18-25 21.5  
## 85 1000018 18-25 21.5  
## 86 1000019 0-17 17.0  
## 87 1000019 0-17 17.0  
## 88 1000019 0-17 17.0  
## 89 1000019 0-17 17.0  
## 90 1000019 0-17 17.0  
## 91 1000019 0-17 17.0  
## 92 1000019 0-17 17.0  
## 93 1000019 0-17 17.0  
## 94 1000019 0-17 17.0  
## 95 1000019 0-17 17.0  
## 96 1000019 0-17 17.0  
## 97 1000019 0-17 17.0  
## 98 1000021 18-25 21.5  
## 99 1000022 18-25 21.5  
## 100 1000022 18-25 21.5  
## 101 1000022 18-25 21.5  
## 102 1000022 18-25 21.5  
## 103 1000022 18-25 21.5  
## 104 1000022 18-25 21.5  
## 105 1000022 18-25 21.5  
## 106 1000022 18-25 21.5  
## 107 1000022 18-25 21.5  
## 108 1000022 18-25 21.5  
## 109 1000022 18-25 21.5  
## 110 1000022 18-25 21.5  
## 111 1000022 18-25 21.5  
## 112 1000022 18-25 21.5  
## 113 1000023 36-45 40.5  
## 114 1000023 36-45 40.5  
## 115 1000023 36-45 40.5  
## 116 1000023 36-45 40.5  
## 117 1000023 36-45 40.5  
## 118 1000023 36-45 40.5  
## 119 1000023 36-45 40.5  
## 120 1000023 36-45 40.5  
## 121 1000023 36-45 40.5  
## 122 1000023 36-45 40.5  
## 123 1000023 36-45 40.5  
## 124 1000023 36-45 40.5  
## 125 1000024 26-35 30.5  
## 126 1000024 26-35 30.5  
## 127 1000024 26-35 30.5  
## 128 1000025 18-25 21.5  
## 129 1000025 18-25 21.5  
## 130 1000025 18-25 21.5  
## 131 1000026 26-35 30.5  
## 132 1000026 26-35 30.5  
## 133 1000026 26-35 30.5  
## 134 1000026 26-35 30.5  
## 135 1000026 26-35 30.5  
## 136 1000026 26-35 30.5  
## 137 1000026 26-35 30.5  
## 138 1000026 26-35 30.5  
## 139 1000026 26-35 30.5  
## 140 1000026 26-35 30.5  
## 141 1000026 26-35 30.5  
## 142 1000026 26-35 30.5  
## 143 1000027 26-35 30.5  
## 144 1000027 26-35 30.5  
## 145 1000027 26-35 30.5  
## 146 1000027 26-35 30.5  
## 147 1000027 26-35 30.5  
## 148 1000028 26-35 30.5  
## 149 1000028 26-35 30.5  
## 150 1000028 26-35 30.5  
## 151 1000028 26-35 30.5  
## 152 1000028 26-35 30.5  
## 153 1000029 36-45 40.5  
## 154 1000029 36-45 40.5  
## 155 1000029 36-45 40.5  
## 156 1000029 36-45 40.5  
## 157 1000030 36-45 40.5  
## 158 1000030 36-45 40.5  
## 159 1000030 36-45 40.5  
## 160 1000031 55+ 55.0  
## 161 1000031 55+ 55.0  
## 162 1000031 55+ 55.0  
## 163 1000031 55+ 55.0  
## 164 1000032 26-35 30.5  
## 165 1000033 46-50 48.0  
## 166 1000033 46-50 48.0  
## 167 1000033 46-50 48.0  
## 168 1000033 46-50 48.0  
## 169 1000033 46-50 48.0  
## 170 1000033 46-50 48.0  
## 171 1000033 46-50 48.0  
## 172 1000033 46-50 48.0  
## 173 1000033 46-50 48.0  
## 174 1000033 46-50 48.0  
## 175 1000033 46-50 48.0  
## 176 1000033 46-50 48.0  
## 177 1000033 46-50 48.0  
## 178 1000033 46-50 48.0  
## 179 1000033 46-50 48.0  
## 180 1000034 18-25 21.5  
## 181 1000034 18-25 21.5  
## 182 1000034 18-25 21.5  
## 183 1000034 18-25 21.5  
## 184 1000034 18-25 21.5  
## 185 1000034 18-25 21.5  
## 186 1000034 18-25 21.5  
## 187 1000034 18-25 21.5  
## 188 1000034 18-25 21.5  
## 189 1000034 18-25 21.5  
## 190 1000035 46-50 48.0  
## 191 1000035 46-50 48.0  
## 192 1000035 46-50 48.0  
## 193 1000035 46-50 48.0  
## 194 1000035 46-50 48.0  
## 195 1000035 46-50 48.0  
## 196 1000035 46-50 48.0  
## 197 1000036 26-35 30.5  
## 198 1000036 26-35 30.5  
## 199 1000036 26-35 30.5  
## 200 1000036 26-35 30.5

#convert character vector to numeric vector  
Age\_numeric <- as.double(Friday\_data$Age\_1)   
  
#convert character vector to numeric vector for occupation  
Occupation\_numeric <- as.numeric(Friday\_data$Occupation)   
  
#Displaing the Numeric data  
Occupation\_numeric[1:10]

## [1] 10 10 10 10 16 15 7 7 7 20

#Histogram indicating the puchase amount with the frequency of purchases  
hist(Friday\_data$Purchase,  
 main="Purchases During the BLACK FRIDAY SALE",  
 xlab="Amount ($)", ylab = "Number of Purchases",  
 col="blue")



#Scatter Plot indicating the Occupation vs the Purchase amount   
plot(Friday\_data$Occupation, Friday\_data$Purchase,main="Purchase Amount with Occupation",  
 xlab = "Occupation",  
 ylab = "Purchase Amount in $")

