DPA Assignment

Group 11

13/12/2019

## Members -

Aveenash Jayaraman 19PGDM021 Paridhi Agarwal 19PGDM043 Shweta Kurian 19PGDM062 Souvik Samanta 19PGDM064

# Question 1 : Importing Data 1 —————————————————-

library(readxl)  
  
excel\_sheets("C:/Users/souvi/Documents/R/IM507/DPA assignment.xlsx")

## [1] "Data1" "Data2" "Data4" "Description"

data1 <- read\_excel("C:/Users/souvi/Documents/R/IM507/DPA assignment.xlsx", sheet = "Data1")

# Question 2 : Understanding the data ——————————————-

head(data1)

## # A tibble: 6 x 13  
## ID Age Experience Income `ZIP Code` Family CCAvg Education Mortgage  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 59 35 8 92691 4 0.7 1 91  
## 2 2 57 32 55 90717 4 2.1 1 108  
## 3 3 46 20 85 95617 2 0.4 3 0  
## 4 4 25 1 62 94720 4 0 1 229  
## 5 5 53 23 65 95054 4 2 3 0  
## 6 6 38 12 138 95112 2 0 1 0  
## # ... with 4 more variables: `Securities Account` <dbl>, `CD  
## # Account` <dbl>, Online <dbl>, CreditCard <dbl>

dim(data1)

## [1] 1000 13

names(data1)

## [1] "ID" "Age" "Experience"   
## [4] "Income" "ZIP Code" "Family"   
## [7] "CCAvg" "Education" "Mortgage"   
## [10] "Securities Account" "CD Account" "Online"   
## [13] "CreditCard"

class(data1)

## [1] "tbl\_df" "tbl" "data.frame"

summary(data1)

## ID Age Experience Income   
## Min. : 1.0 Min. :24.00 Min. :-1.00 Min. : 8.0   
## 1st Qu.: 250.8 1st Qu.:35.00 1st Qu.:10.00 1st Qu.: 38.0   
## Median : 500.5 Median :45.00 Median :20.00 Median : 64.0   
## Mean : 500.5 Mean :45.35 Mean :20.12 Mean : 74.9   
## 3rd Qu.: 750.2 3rd Qu.:55.50 3rd Qu.:30.00 3rd Qu.:101.0   
## Max. :1000.0 Max. :67.00 Max. :42.00 Max. :201.0   
## NA's :5   
## ZIP Code Family CCAvg Education   
## Min. :90005 Min. :1.000 Min. :0.000 Min. :1.000   
## 1st Qu.:91711 1st Qu.:1.000 1st Qu.:0.600 1st Qu.:1.000   
## Median :93353 Median :2.000 Median :1.500 Median :2.000   
## Mean :93148 Mean :2.371 Mean :1.859 Mean :1.859   
## 3rd Qu.:94702 3rd Qu.:4.000 3rd Qu.:2.500 3rd Qu.:3.000   
## Max. :96651 Max. :4.000 Max. :9.000 Max. :3.000   
## NA's :3   
## Mortgage Securities Account CD Account Online   
## Min. : 0.0 Min. :0.000 Min. :0.000 Min. :0.000   
## 1st Qu.: 0.0 1st Qu.:0.000 1st Qu.:0.000 1st Qu.:0.000   
## Median : 0.0 Median :0.000 Median :0.000 Median :1.000   
## Mean : 58.5 Mean :0.119 Mean :0.063 Mean :0.602   
## 3rd Qu.:104.0 3rd Qu.:0.000 3rd Qu.:0.000 3rd Qu.:1.000   
## Max. :601.0 Max. :1.000 Max. :1.000 Max. :1.000   
## NA's :3   
## CreditCard   
## Min. :0.000   
## 1st Qu.:0.000   
## Median :0.000   
## Mean :0.289   
## 3rd Qu.:1.000   
## Max. :1.000   
##

# Question 3 : Structure of the dataset —————————————–

str(data1)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 1000 obs. of 13 variables:  
## $ ID : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : num 59 57 46 25 53 38 63 44 50 28 ...  
## $ Experience : num 35 32 20 1 23 12 38 19 24 3 ...  
## $ Income : num 8 55 85 62 65 138 103 99 152 25 ...  
## $ ZIP Code : num 92691 90717 95617 94720 95054 ...  
## $ Family : num 4 4 2 4 4 2 1 3 1 2 ...  
## $ CCAvg : num 0.7 2.1 0.4 0 2 0 2.5 3.5 7.3 0.9 ...  
## $ Education : num 1 1 3 1 3 1 1 3 1 3 ...  
## $ Mortgage : num 91 108 0 229 0 0 0 357 0 140 ...  
## $ Securities Account: num 1 0 0 0 0 1 0 0 0 0 ...  
## $ CD Account : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Online : num 0 1 0 1 1 1 1 0 1 1 ...  
## $ CreditCard : num 0 0 0 0 0 0 0 0 0 0 ...

# Question 4 : Conversion for proper datatype ———————————–

data1$Education <- factor(data1$Education,levels = c("1","2","3"), labels = c("Undergrad", "Graduate", "Advanced"))  
data1$`Securities Account` <- factor(data1$`Securities Account`,levels = c("0","1"), labels = c("Don't Have","Have"))  
data1$`CD Account` <- factor(data1$`CD Account`,levels = c("0","1"), labels = c("Don't Have","Have"))  
data1$Online <- factor(data1$Online,levels = c("0","1"), labels = c("Don't Have","Have"))  
data1$CreditCard <- factor(data1$CreditCard,levels = c("0","1"), labels = c("Don't Have","Have"))

# Question 5 : Joining Data 1 and Data 2 and creating Data 3——————–

data2 <- read\_excel("C:/Users/souvi/Documents/R/IM507/DPA assignment.xlsx", sheet = "Data2")  
  
names(data1)

## [1] "ID" "Age" "Experience"   
## [4] "Income" "ZIP Code" "Family"   
## [7] "CCAvg" "Education" "Mortgage"   
## [10] "Securities Account" "CD Account" "Online"   
## [13] "CreditCard"

names(data2)

## [1] "ID" "Personal Loan"

data3 <- merge(data1,data2, by.x = "ID", by.y = "ID")  
head(data3)

## ID Age Experience Income ZIP Code Family CCAvg Education Mortgage  
## 1 1 59 35 8 92691 4 0.7 Undergrad 91  
## 2 2 57 32 55 90717 4 2.1 Undergrad 108  
## 3 3 46 20 85 95617 2 0.4 Advanced 0  
## 4 4 25 1 62 94720 4 0.0 Undergrad 229  
## 5 5 53 23 65 95054 4 2.0 Advanced 0  
## 6 6 38 12 138 95112 2 0.0 Undergrad 0  
## Securities Account CD Account Online CreditCard Personal Loan  
## 1 Have Don't Have Don't Have Don't Have 0  
## 2 Don't Have Don't Have Have Don't Have 0  
## 3 Don't Have Don't Have Don't Have Don't Have 0  
## 4 Don't Have Don't Have Have Don't Have 0  
## 5 Don't Have Don't Have Have Don't Have 0  
## 6 Have Don't Have Have Don't Have 0

# Converting new column to a factor -

data3$`Personal Loan` <- factor(data3$`Personal Loan`,levels = c("0","1"), labels = c("Don't Have","Have"))

# Question 6 : Joining Data 3 and Data 4 and creating Data 5——————–

data4 <- read\_excel("C:/Users/souvi/Documents/R/IM507/DPA assignment.xlsx", sheet = "Data4")  
  
data4$Education <- factor(data4$Education,levels = c("1","2","3"), labels = c("Undergrad", "Graduate", "Advanced"))  
data4$`Securities Account` <- factor(data4$`Securities Account`,levels = c("0","1"), labels = c("Don't Have","Have"))  
data4$`CD Account` <- factor(data4$`CD Account`,levels = c("0","1"), labels = c("Don't Have","Have"))  
data4$Online <- factor(data4$Online,levels = c("0","1"), labels = c("Don't Have","Have"))  
data4$CreditCard <- factor(data4$CreditCard,levels = c("0","1"), labels = c("Don't Have","Have"))  
data4$`Personal Loan` <- factor(data4$`Personal Loan`,levels = c("0","1"), labels = c("Don't Have","Have"))  
  
  
names(data3)

## [1] "ID" "Age" "Experience"   
## [4] "Income" "ZIP Code" "Family"   
## [7] "CCAvg" "Education" "Mortgage"   
## [10] "Securities Account" "CD Account" "Online"   
## [13] "CreditCard" "Personal Loan"

names(data4)

## [1] "ID" "Age" "Experience"   
## [4] "Income" "ZIP Code" "Family"   
## [7] "CCAvg" "Education" "Mortgage"   
## [10] "Securities Account" "CD Account" "Online"   
## [13] "CreditCard" "Personal Loan"

dim(data3)

## [1] 1000 14

dim(data4)

## [1] 2300 14

data5 <- rbind(data3,data4)  
  
dim(data5)

## [1] 3300 14

# Question 7 : Exploration of data5 ——————————————–

head(data5)

## ID Age Experience Income ZIP Code Family CCAvg Education Mortgage  
## 1 1 59 35 8 92691 4 0.7 Undergrad 91  
## 2 2 57 32 55 90717 4 2.1 Undergrad 108  
## 3 3 46 20 85 95617 2 0.4 Advanced 0  
## 4 4 25 1 62 94720 4 0.0 Undergrad 229  
## 5 5 53 23 65 95054 4 2.0 Advanced 0  
## 6 6 38 12 138 95112 2 0.0 Undergrad 0  
## Securities Account CD Account Online CreditCard Personal Loan  
## 1 Have Don't Have Don't Have Don't Have Don't Have  
## 2 Don't Have Don't Have Have Don't Have Don't Have  
## 3 Don't Have Don't Have Don't Have Don't Have Don't Have  
## 4 Don't Have Don't Have Have Don't Have Don't Have  
## 5 Don't Have Don't Have Have Don't Have Don't Have  
## 6 Have Don't Have Have Don't Have Don't Have

dim(data5)

## [1] 3300 14

nrow(data5)

## [1] 3300

ncol(data5)

## [1] 14

names(data5)

## [1] "ID" "Age" "Experience"   
## [4] "Income" "ZIP Code" "Family"   
## [7] "CCAvg" "Education" "Mortgage"   
## [10] "Securities Account" "CD Account" "Online"   
## [13] "CreditCard" "Personal Loan"

class(data5)

## [1] "data.frame"

str(data5)

## 'data.frame': 3300 obs. of 14 variables:  
## $ ID : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : num 59 57 46 25 53 38 63 44 50 28 ...  
## $ Experience : num 35 32 20 1 23 12 38 19 24 3 ...  
## $ Income : num 8 55 85 62 65 138 103 99 152 25 ...  
## $ ZIP Code : num 92691 90717 95617 94720 95054 ...  
## $ Family : num 4 4 2 4 4 2 1 3 1 2 ...  
## $ CCAvg : num 0.7 2.1 0.4 0 2 0 2.5 3.5 7.3 0.9 ...  
## $ Education : Factor w/ 3 levels "Undergrad","Graduate",..: 1 1 3 1 3 1 1 3 1 3 ...  
## $ Mortgage : num 91 108 0 229 0 0 0 357 0 140 ...  
## $ Securities Account: Factor w/ 2 levels "Don't Have","Have": 2 1 1 1 1 2 1 1 1 1 ...  
## $ CD Account : Factor w/ 2 levels "Don't Have","Have": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Online : Factor w/ 2 levels "Don't Have","Have": 1 2 1 2 2 2 2 1 2 2 ...  
## $ CreditCard : Factor w/ 2 levels "Don't Have","Have": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Personal Loan : Factor w/ 2 levels "Don't Have","Have": 1 1 1 1 1 1 1 2 1 1 ...

summary(data5)

## ID Age Experience Income   
## Min. : 1.0 Min. :23.00 Min. :-3.00 Min. : 8.00   
## 1st Qu.: 825.8 1st Qu.:35.00 1st Qu.:10.00 1st Qu.: 39.00   
## Median :1650.5 Median :45.00 Median :20.00 Median : 63.00   
## Mean :1650.5 Mean :45.27 Mean :20.07 Mean : 73.75   
## 3rd Qu.:2475.2 3rd Qu.:55.00 3rd Qu.:30.00 3rd Qu.: 98.00   
## Max. :3300.0 Max. :67.00 Max. :43.00 Max. :224.00   
## NA's :5   
## ZIP Code Family CCAvg Education   
## Min. :90005 Min. :1.000 Min. : 0.000 Undergrad:1391   
## 1st Qu.:91902 1st Qu.:1.000 1st Qu.: 0.700 Graduate : 920   
## Median :93555 Median :2.000 Median : 1.500 Advanced : 989   
## Mean :93184 Mean :2.376 Mean : 1.914   
## 3rd Qu.:94609 3rd Qu.:3.000 3rd Qu.: 2.600   
## Max. :96651 Max. :4.000 Max. :10.000   
## NA's :3   
## Mortgage Securities Account CD Account Online   
## Min. : 0.00 Don't Have:2956 Don't Have:3101 Don't Have:1306   
## 1st Qu.: 0.00 Have : 344 Have : 199 Have :1994   
## Median : 0.00   
## Mean : 57.38   
## 3rd Qu.:102.00   
## Max. :635.00   
## NA's :3   
## CreditCard Personal Loan   
## Don't Have:2329 Don't Have:2984   
## Have : 971 Have : 316   
##   
##   
##   
##   
##

# Question 8 : Check for missing values —————————————-

# Total missing values in the data frame -  
  
sum(is.na(data5))

## [1] 11

# Missing values in variables -  
  
sapply(data5, function(x)sum(is.na(x)))

## ID Age Experience   
## 0 5 0   
## Income ZIP Code Family   
## 0 0 0   
## CCAvg Education Mortgage   
## 3 0 3   
## Securities Account CD Account Online   
## 0 0 0   
## CreditCard Personal Loan   
## 0 0

# Age, CCAvg, Mortgage have missing values  
  
# Replacing missing values by the means of the particular variable -  
  
data5$Age[is.na(data5$Age)] <- mean(data5$Age, na.rm = T)  
data5$CCAvg[is.na(data5$CCAvg)] <- mean(data5$CCAvg, na.rm = T)  
data5$Mortgage[is.na(data5$Mortgage)] <- mean(data5$Mortgage, na.rm = T)  
  
sum(is.na(data5))

## [1] 0

# Question 9 : Creating Age group variable ————————————-

summary(data5$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 23.00 35.00 45.00 45.27 55.00 67.00

# Min age = 23 ; max age = 67 Let us take 5 age groups like-  
# Group A : 20+ to <= 30  
# Group B : 30+ to <= 40  
# Group C : 40+ to <= 50  
# Group D : 50+ to <= 60  
# Group E : 60+  
  
data5$Age\_group <- cut(data5$Age,c(20,30,40,50,60,Inf),labels = c("A","B","C","D","E"))

# Question 10 : Creating income group variable ———————————

summary(data5$Income)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8.00 39.00 63.00 73.75 98.00 224.00

# Min age = 8 ; max age = 224 Let us take 5 age groups like-  
# Group A : 0 to <= 50  
# Group B : 50+ to <= 100  
# Group C : 100+ to <= 150  
# Group D : 150+ to <= 200  
# Group E : 200+  
  
data5$income\_group <- cut(data5$Income,c(0,50,100,150,200,Inf),labels = c("A","B","C","D","E"))

# Question 11 : Listing customer id, Mortgage details of the customers who belong to age group (61-65), have family size 3 or above, have education level = 3 and have accepted the loan offer———————————————-

library(dplyr)

##   
## Attaching package: 'dplyr'

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

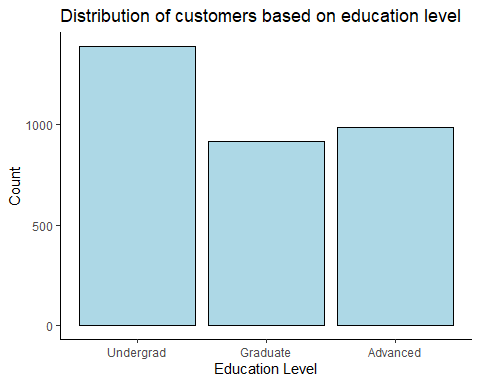
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data5 %>% filter(Age>=61,Age<=65,Family>=3,Education=="Advanced",`Personal Loan`=="Have") %>% select(ID,Mortgage)

## ID Mortgage  
## 1 207 0  
## 2 1803 522  
## 3 1907 0  
## 4 2156 0  
## 5 2334 0  
## 6 2342 351

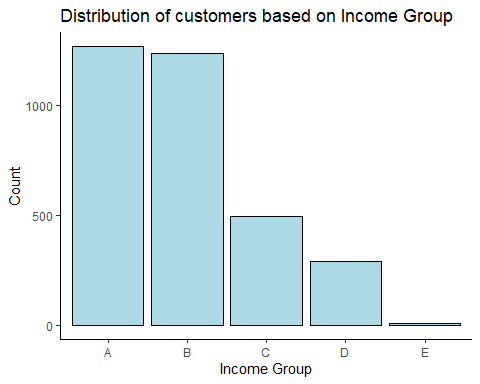
# Question 12 : Creating a graph to display the distribution of customers based on education level————————————————————–

library(ggplot2)  
  
ggplot(data5, aes(Education))+  
 geom\_bar(color="black", fill = "light blue") +   
 xlab("Education Level") +  
 ylab("Count")+  
 ggtitle("Distribution of customers based on education level")+  
 theme\_classic()



# Question 13 : Creating a graph to display the distribution of customers based on income group—————————————————————–

ggplot(data5, aes(income\_group))+  
 geom\_bar(color="black", fill = "light blue") +   
 xlab("Income Group") +  
 ylab("Count")+  
 ggtitle("Distribution of customers based on Income Group")+  
 theme\_classic()



# Question 14 : Creating a graph to display the distribution of customers who accepted or rejected the loan offer based on family size————————

ggplot(data5, aes(Family, fill=`Personal Loan`))+  
 geom\_bar(color="black") +   
 xlab("Family size") +  
 ylab("Count")+  
 ggtitle(" Distribution of customers\n accepted loan based on Family size")+  
 theme\_classic()

