Read\_WriteData

Paul

7/10/2020

#Introduction This is the starting point in R Markdown

x<-1:10

#Read data from github

#Basic data manipulation

library(skimr) # Gives the general overview of the data  
skim(new\_data)

Data summary

|  |  |
| --- | --- |
| Name | new\_data |
| Number of rows | 2509 |
| Number of columns | 28 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 23 |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| account\_type | 0 | 1.00 | 4 | 19 | 0 | 7 | 0 |
| district | 0 | 1.00 | 10 | 10 | 0 | 3 | 0 |
| urban | 0 | 1.00 | 5 | 5 | 0 | 2 | 0 |
| gender | 0 | 1.00 | 4 | 6 | 0 | 2 | 0 |
| highest\_grade\_completed | 256 | 0.90 | 4 | 11 | 0 | 15 | 0 |
| mm\_account\_cancelled | 0 | 1.00 | 2 | 3 | 0 | 2 | 0 |
| prefer\_cash | 49 | 0.98 | 2 | 3 | 0 | 3 | 0 |
| mm\_trust | 180 | 0.93 | 2 | 3 | 0 | 3 | 0 |
| mm\_account\_telco | 783 | 0.69 | 9 | 29 | 0 | 7 | 0 |
| mm\_account\_telco\_main | 1657 | 0.34 | 9 | 9 | 0 | 3 | 0 |
| v234 | 886 | 0.65 | 2 | 3 | 0 | 3 | 0 |
| agent\_trust | 1051 | 0.58 | 2 | 3 | 0 | 3 | 0 |
| v236 | 1964 | 0.22 | 2 | 3 | 0 | 2 | 0 |
| v237 | 472 | 0.81 | 2 | 3 | 0 | 2 | 0 |
| v238 | 462 | 0.82 | 2 | 3 | 0 | 2 | 0 |
| v240 | 462 | 0.82 | 2 | 3 | 0 | 2 | 0 |
| v241 | 527 | 0.79 | 2 | 3 | 0 | 2 | 0 |
| v242 | 551 | 0.78 | 2 | 3 | 0 | 2 | 0 |
| v243 | 472 | 0.81 | 2 | 3 | 0 | 2 | 0 |
| v244 | 1984 | 0.21 | 2 | 3 | 0 | 2 | 0 |
| v245 | 472 | 0.81 | 2 | 3 | 0 | 2 | 0 |
| v246 | 462 | 0.82 | 2 | 3 | 0 | 2 | 0 |
| mm\_account | 0 | 1.00 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| hhid | 0 | 1 | 1597.33 | 351.45 | 1001.00 | 1290.00 | 1593.00 | 1905.00 | 2205.00 | ▇▇▇▇▇ |
| weight | 0 | 1 | 443.16 | 475.81 | 14.58 | 196.72 | 296.21 | 511.74 | 4812.17 | ▇▁▁▁▁ |
| account\_num | 0 | 1 | 1.78 | 0.89 | 1.00 | 1.00 | 2.00 | 2.00 | 6.00 | ▇▂▁▁▁ |
| age | 0 | 1 | 37.67 | 13.62 | 18.00 | 27.00 | 35.00 | 46.00 | 97.00 | ▇▆▃▁▁ |
| hh\_members | 0 | 1 | 4.71 | 1.99 | 1.00 | 3.00 | 5.00 | 6.00 | 18.00 | ▇▇▁▁▁ |

#use of summary, head, tail  
head(new\_data) #first 6rows of dataset

## # A tibble: 6 x 28  
## hhid weight account\_num account\_type district urban gender age hh\_members  
## <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl>  
## 1 1001 146. 1 SACCO Accou~ Distric~ Urban male 32 1  
## 2 1001 146. 2 VSLA Account Distric~ Urban male 32 1  
## 3 1002 123. 1 Mobile Money Distric~ Rural male 32 4  
## 4 1002 123. 2 Bank Account Distric~ Rural male 32 4  
## 5 1002 123. 3 VSLA Account Distric~ Rural male 32 4  
## 6 1003 760. 1 Mobile Money Distric~ Urban male 30 8  
## # ... with 19 more variables: highest\_grade\_completed <chr>,  
## # mm\_account\_cancelled <chr>, prefer\_cash <chr>, mm\_trust <chr>,  
## # mm\_account\_telco <chr>, mm\_account\_telco\_main <chr>, v234 <chr>,  
## # agent\_trust <chr>, v236 <chr>, v237 <chr>, v238 <chr>, v240 <chr>,  
## # v241 <chr>, v242 <chr>, v243 <chr>, v244 <chr>, v245 <chr>, v246 <chr>,  
## # mm\_account <chr>

tail(new\_data) #last 6 rows of dataset

## # A tibble: 6 x 28  
## hhid weight account\_num account\_type district urban gender age hh\_members  
## <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl>  
## 1 2203 103. 1 Mobile Money Distric~ Rural female 18 7  
## 2 2203 103. 2 Bank Account Distric~ Rural female 18 7  
## 3 2204 496. 1 Mobile Money Distric~ Rural female 58 6  
## 4 2204 496. 2 Bank Account Distric~ Rural female 58 6  
## 5 2205 667. 1 Mobile Money Distric~ Rural female 23 7  
## 6 2205 667. 2 Bank Account Distric~ Rural female 23 7  
## # ... with 19 more variables: highest\_grade\_completed <chr>,  
## # mm\_account\_cancelled <chr>, prefer\_cash <chr>, mm\_trust <chr>,  
## # mm\_account\_telco <chr>, mm\_account\_telco\_main <chr>, v234 <chr>,  
## # agent\_trust <chr>, v236 <chr>, v237 <chr>, v238 <chr>, v240 <chr>,  
## # v241 <chr>, v242 <chr>, v243 <chr>, v244 <chr>, v245 <chr>, v246 <chr>,  
## # mm\_account <chr>

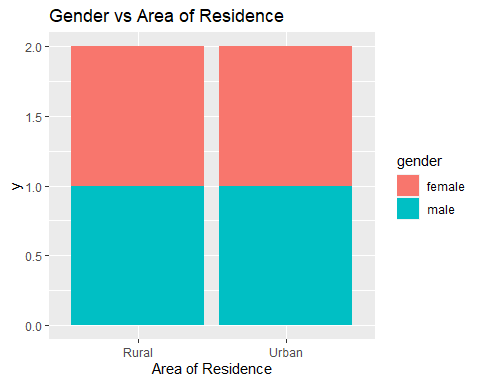
summary(new\_data) # genera; summary of the data

## hhid weight account\_num account\_type   
## Min. :1001 Min. : 14.58 Min. :1.000 Length:2509   
## 1st Qu.:1290 1st Qu.: 196.72 1st Qu.:1.000 Class :character   
## Median :1593 Median : 296.21 Median :2.000 Mode :character   
## Mean :1597 Mean : 443.16 Mean :1.775   
## 3rd Qu.:1905 3rd Qu.: 511.74 3rd Qu.:2.000   
## Max. :2205 Max. :4812.16 Max. :6.000   
## district urban gender age   
## Length:2509 Length:2509 Length:2509 Min. :18.00   
## Class :character Class :character Class :character 1st Qu.:27.00   
## Mode :character Mode :character Mode :character Median :35.00   
## Mean :37.67   
## 3rd Qu.:46.00   
## Max. :97.00   
## hh\_members highest\_grade\_completed mm\_account\_cancelled  
## Min. : 1.000 Length:2509 Length:2509   
## 1st Qu.: 3.000 Class :character Class :character   
## Median : 5.000 Mode :character Mode :character   
## Mean : 4.714   
## 3rd Qu.: 6.000   
## Max. :18.000   
## prefer\_cash mm\_trust mm\_account\_telco mm\_account\_telco\_main  
## Length:2509 Length:2509 Length:2509 Length:2509   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## v234 agent\_trust v236 v237   
## Length:2509 Length:2509 Length:2509 Length:2509   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## v238 v240 v241 v242   
## Length:2509 Length:2509 Length:2509 Length:2509   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## v243 v244 v245 v246   
## Length:2509 Length:2509 Length:2509 Length:2509   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## mm\_account   
## Length:2509   
## Class :character   
## Mode :character   
##   
##   
##

##Gender vs Urban summary(frequencies and percentages)  
#ctrl + shift + m - to insert the pipe  
new\_data %>% count(urban, gender)

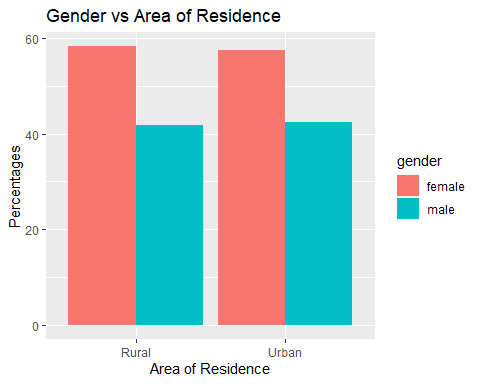
## # A tibble: 4 x 3  
## urban gender n  
## <chr> <chr> <int>  
## 1 Rural female 1144  
## 2 Rural male 819  
## 3 Urban female 314  
## 4 Urban male 232

#visualize the data using columns  
library(ggplot2)  
new\_data %>% count(urban, gender) %>%   
 ggplot(aes(x=urban, y=1, fill=gender))+  
 geom\_col()+   
 xlab("Area of Residence")+  
 #Ylab("Percentages")+  
 ggtitle("Gender vs Area of Residence")



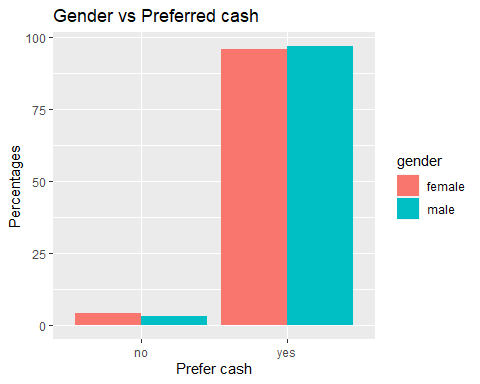
#visualize the data using a bar graph  
new\_data %>% group\_by(urban, gender)%>%   
summarise(count=n())%>%   
 mutate(percent1=(count/sum(count))\*100)%>%   
ggplot(aes(x=urban, y=percent1, fill=gender))+  
 geom\_bar(stat="identity",position="dodge")+   
 xlab("Area of Residence")+  
 ylab("Percentages")+  
 ggtitle("Gender vs Area of Residence")

## `summarise()` regrouping output by 'urban' (override with `.groups` argument)

 The number of female and male respondents in rural areas were 58.28 and 57.51 percent while those in urban areas were 41.72 and 42.49 percent respectively.

#Summarize the data and plot a bar graph  
#to drop the nas  
new\_data %>% group\_by(gender, prefer\_cash)%>%  
 na\_if(-97) %>%   
 drop\_na(gender, prefer\_cash) %>%   
summarise(count=n())%>%   
 mutate(percent2=(count/sum(count))\*100) %>%   
ggplot(aes(x=prefer\_cash, y=percent2, fill=gender))+  
 geom\_bar(stat="identity",position="dodge")+   
 xlab("Prefer cash")+  
 ylab("Percentages")+  
 ggtitle("Gender vs Preferred cash")

## `summarise()` regrouping output by 'gender' (override with `.groups` argument)



The percentage number of female who had a preferred cash were 95.79 while those without were only 4.21. On the other hand, the percentage number of males with a preferred cash was 96.71 while those without were only 3.19.

#Use of Mutate and rename in the dataset

library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

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

#Mutate creates a new column in the data set  
new\_data %>% mutate(hh\_age=age\*hh\_members)

## # A tibble: 2,509 x 29  
## hhid weight account\_num account\_type district urban gender age hh\_members  
## <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl>  
## 1 1001 146. 1 SACCO Accou~ Distric~ Urban male 32 1  
## 2 1001 146. 2 VSLA Account Distric~ Urban male 32 1  
## 3 1002 123. 1 Mobile Money Distric~ Rural male 32 4  
## 4 1002 123. 2 Bank Account Distric~ Rural male 32 4  
## 5 1002 123. 3 VSLA Account Distric~ Rural male 32 4  
## 6 1003 760. 1 Mobile Money Distric~ Urban male 30 8  
## 7 1003 760. 2 Bank Account Distric~ Urban male 30 8  
## 8 1004 434. 1 Mobile Money Distric~ Rural male 68 4  
## 9 1005 303. 1 None Distric~ Rural female 28 2  
## 10 1006 1053. 1 Mobile Money Distric~ Rural female 36 7  
## # ... with 2,499 more rows, and 20 more variables:  
## # highest\_grade\_completed <chr>, mm\_account\_cancelled <chr>,  
## # prefer\_cash <chr>, mm\_trust <chr>, mm\_account\_telco <chr>,  
## # mm\_account\_telco\_main <chr>, v234 <chr>, agent\_trust <chr>, v236 <chr>,  
## # v237 <chr>, v238 <chr>, v240 <chr>, v241 <chr>, v242 <chr>, v243 <chr>,  
## # v244 <chr>, v245 <chr>, v246 <chr>, mm\_account <chr>, hh\_age <dbl>

#Rename a column in the data set  
 new\_data %>% rename(c("education\_level"= "highest\_grade\_completed"))

## # A tibble: 2,509 x 28  
## hhid weight account\_num account\_type district urban gender age hh\_members  
## <dbl> <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl>  
## 1 1001 146. 1 SACCO Accou~ Distric~ Urban male 32 1  
## 2 1001 146. 2 VSLA Account Distric~ Urban male 32 1  
## 3 1002 123. 1 Mobile Money Distric~ Rural male 32 4  
## 4 1002 123. 2 Bank Account Distric~ Rural male 32 4  
## 5 1002 123. 3 VSLA Account Distric~ Rural male 32 4  
## 6 1003 760. 1 Mobile Money Distric~ Urban male 30 8  
## 7 1003 760. 2 Bank Account Distric~ Urban male 30 8  
## 8 1004 434. 1 Mobile Money Distric~ Rural male 68 4  
## 9 1005 303. 1 None Distric~ Rural female 28 2  
## 10 1006 1053. 1 Mobile Money Distric~ Rural female 36 7  
## # ... with 2,499 more rows, and 19 more variables: education\_level <chr>,  
## # mm\_account\_cancelled <chr>, prefer\_cash <chr>, mm\_trust <chr>,  
## # mm\_account\_telco <chr>, mm\_account\_telco\_main <chr>, v234 <chr>,  
## # agent\_trust <chr>, v236 <chr>, v237 <chr>, v238 <chr>, v240 <chr>,  
## # v241 <chr>, v242 <chr>, v243 <chr>, v244 <chr>, v245 <chr>, v246 <chr>,  
## # mm\_account <chr>

#slice your data

new\_data2<-new\_data %>% select(gender,age,hh\_members) %>% group\_by(gender) %>%   
summarise\_all(mean)