Data Exploration and Visualization

Cleaning Dataset Task

Done by Group 3 members - Submitted on 12-Apr-2023

Mohammad Awad

Alaa *

Anas *

Ibrahim *

Loading Libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## Warning: package 'lubridate' was built under R version 4.1.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(ggplot2)
library(ggthemes)
library(tidyr)
```

Importing data

```
df <- read_csv('Project Data - Uncleaned.csv')
```

```
## Rows: 1211 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): Region, Gender, Marital_Status, Employment, Rent, Loans, Smoking, H...
## dbl (9): ID, Age, BMI, Education, HH_Income, Diabetes_Duration, CVD, HbA1c, ...
##
## 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.
```

```
glimpse(df)
```

```
## Rows: 1,211
## Columns: 17
## $ ID                    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ Region                <chr> "WA", "EA", "EA", "WZ", "EA", "WZ", "EA", "SZ", ~
## $ Age                   <dbl> 36, 48, 46, 65, 48, 46, 58, 65, 39, 47, 69, 68, ~
## $ BMI                   <dbl> 27.9, 30.2, 28.6, 34.9, 29.2, 24.9, 31.6, 25.1, ~
## $ Gender                <chr> "Female", "Female", "Female", "Female", "Female"~
## $ Marital_Status        <chr> "Married", "Married", "Married", "Married", "Mar~
## $ Education             <dbl> 2, 1, 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 2, 3, 2, 2, ~
## $ Employment            <chr> "Yes", "No", "No", "No", "Yes", "No", "Yes", "No~
## $ HH_Income             <dbl> 3, 1, 1, 2, 2, 2, 2, 1, 2, 4, 3, 2, 2, 1, 1, 1, ~
## $ Rent                  <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",~
## $ Loans                 <chr> "No", "No", "No", "Yes", "No", "Yes", "No", "No"~
## $ Smoking               <chr> "Yes", "No", "Yes", "Yes", "Yes", "No", "Yes", "~
## $ Diabetes_Duration     <dbl> 43, 97, 68, 119, 61, 98, 42, 154, 35, 109, 182, ~
## $ Hypertension_category <chr> "Normal", "First-Grade", "First-Grade", "Second-~
## $ CVD                   <dbl> 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, ~
## $ HbA1c                 <dbl> 7.2, 8.9, 9.8, 9.7, 10.4, 8.7, 8.9, 6.8, 9.2, 8.~
## $ Uncontrolled          <dbl> 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
```

Convert to factor the categorical variables

```
df$Region <- as.factor(df$Region)
df$Gender <- as.factor(df$Gender)
df$Marital_Status <- as.factor(df$Marital_Status)
df$Education <- factor(df$Education, levels = c("1","2","3"), ordered = TRUE)
df$Employment <- as.factor(df$Employment)

df$HH_Income <- factor(df$HH_Income, levels = c("1","2","3","4","5"), ordered = TRUE)

df$Rent <- as.factor(df$Rent)
df$Loans <- as.factor(df$Loans)
df$Smoking <- as.factor(df$Smoking)
df$Hypertension_category <- as.factor(df$Hypertension_category)
df$CVD <- as.factor(df$CVD)
df$Uncontrolled <- as.factor(df$Uncontrolled)

glimpse(df)
```

```
## Rows: 1,211
## Columns: 17
## $ ID                    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ Region                <fct> WA, EA, EA, WZ, EA, WZ, EA, SZ, WZ, EA, SZ, SZ, ~
## $ Age                   <dbl> 36, 48, 46, 65, 48, 46, 58, 65, 39, 47, 69, 68, ~
## $ BMI                   <dbl> 27.9, 30.2, 28.6, 34.9, 29.2, 24.9, 31.6, 25.1, ~
## $ Gender                <fct> Female, Female, Female, Female, Female, Female, ~
## $ Marital_Status        <fct> Married, Married, Married, Married, Married, Mar~
## $ Education             <ord> 2, 1, 1, 1, 2, 1, 2, 1, 1, 2, 1, 1, 2, 3, 2, 2, ~
## $ Employment            <fct> Yes, No, No, No, Yes, No, Yes, No, No, Yes, No, ~
## $ HH_Income             <ord> 3, 1, 1, 2, 2, 2, 2, 1, 2, 4, 3, 2, 2, 1, 1, 1, ~
## $ Rent                  <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, No, No, ~
## $ Loans                 <fct> No, No, No, Yes, No, Yes, No, No, Yes, No, No, N~
## $ Smoking               <fct> Yes, No, Yes, Yes, Yes, No, Yes, Yes, No, Yes, N~
## $ Diabetes_Duration     <dbl> 43, 97, 68, 119, 61, 98, 42, 154, 35, 109, 182, ~
## $ Hypertension_category <fct> Normal, First-Grade, First-Grade, Second-Grade, ~
## $ CVD                   <fct> 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, ~
## $ HbA1c                 <dbl> 7.2, 8.9, 9.8, 9.7, 10.4, 8.7, 8.9, 6.8, 9.2, 8.~
## $ Uncontrolled          <fct> 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
```

Checking that there is no duplicate lines

```
nrow(df) == nrow(distinct(df))
```

```
## [1] TRUE
```

Starting with summary for all features

```
summary(df)
```

```
##        ID         Region        Age             BMI            Gender
##  Min.   :   1.0   EA:416   Min.   : 3.60   Min.   :  2.20   Female:667
##  1st Qu.: 303.5   SZ:393   1st Qu.:50.00   1st Qu.: 25.90   Male  :544
##  Median : 606.0   WA:145   Median :56.00   Median : 28.30
##  Mean   : 606.0   WZ:257   Mean   :54.71   Mean   : 29.32
##  3rd Qu.: 908.5            3rd Qu.:59.00   3rd Qu.: 31.50
##  Max.   :1211.0            Max.   :79.00   Max.   :254.00
##   Marital_Status Education  Employment HH_Income  Rent     Loans     Smoking
##  Divorced:135    1   :486   No :738    1:120     No :413   No :759   No :731
##  M       :  1    2   :578   Yes:473    2:650     Yes:798   Yes:452   Yes:480
##  Married :962    3   :145              3:211
##  Single  : 61    NA's:  2              4:181
##  Widowed : 52                          5: 49
##
##  Diabetes_Duration      Hypertension_category CVD         HbA1c
##  Min.   :  6.00    First-Grade     :376       0:794   Min.   : 5.90
##  1st Qu.: 37.00    Normal          :232       1:417   1st Qu.: 7.40
##  Median : 62.00    Pre-hypertention:479               Median : 8.30
##  Mean   : 65.97    Prehypertention :  1               Mean   : 8.43
##  3rd Qu.: 85.00    Second-Grade    :110               3rd Qu.: 9.20
##  Max.   :860.00    Third-Grade     : 13               Max.   :13.20
##  Uncontrolled
##  0:859
##  1:352
##
##
##
##
```

From the summary below is noted:

- Age has a min value of 3.6 while metadata mention adults.

- BMI has a min value of 2.2 and the maximum of 254 both are beyond normal range.

- For marital status most of the data showing as Maried there is 1 M category , will convert it to Maried.

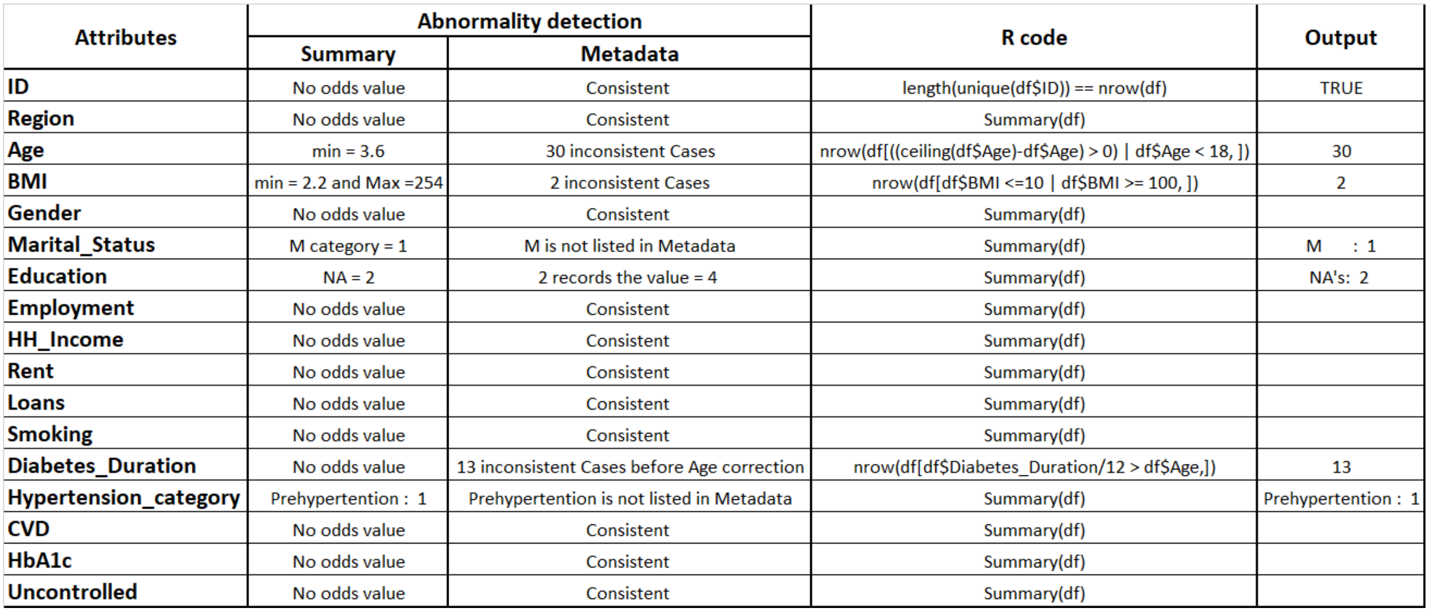
- Education has 2 NA’s values which they actually refer to as 4 in the uncleaned data.

- There is 1 record that should be called Prehypertention to be Pre-hypertention.

- The Diabetes_Duration has a max vaule of 860 which is 71.667, we need to check and compare it against age of that record, or if any other records that are greater than Age.

Below table is a comparison between the metadata provided in the TOR and the summary generated using R.

```
knitr::include_graphics("SummaryvsMetadata.png")
```



Attributes	Abnormality detection		R code	Output
	Summary	Metadata		
ID	No odds value	Consistent	length(unique(df$ID)) == nrow(df)	TRUE
Region	No odds value	Consistent	Summary(df)	
Age	min = 3.6	30 inconsistent Cases	nrow(df[((ceiling(df$Age)-df$Age) > 0) | df$Age < 18,])	30
BMI	min = 2.2 and Max =254	2 inconsistent Cases	nrow(df[df$BMI <=10 \| df$BMI >= 100,])	2
Gender	No odds value	Consistent	Summary(df)	
Marital_Status	M category = 1	M is not listed in Metadata	Summary(df)	M : 1
Education	NA = 2	2 records the value = 4	Summary(df)	NA's: 2
Employment	No odds value	Consistent	Summary(df)	
HH_Income	No odds value	Consistent	Summary(df)	
Rent	No odds value	Consistent	Summary(df)	
Loans	No odds value	Consistent	Summary(df)	
Smoking	No odds value	Consistent	Summary(df)	
Diabetes_Duration	No odds value	13 inconsistent Cases before Age correction	nrow(df[df$Diabetes_Duration/12 > df$Age,])	13
Hypertension_category	Prehypertention : 1	Prehypertention is not listed in Metadata	Summary(df)	Prehypertention : 1
CVD	No odds value	Consistent	Summary(df)	
HbA1c	No odds value	Consistent	Summary(df)	
Uncontrolled	No odds value	Consistent	Summary(df)	

The following changes were made after checking with Dr. Osama regarding the correct approach to deal with the mentioned issues.

To print the lines that have NA values in the education , we can fill them with the value of category 3

```
df[!complete.cases(df), ]
```

```
## # A tibble: 2 x 17
##      ID Region   Age   BMI Gender Marital_Status Education Employment HH_Income
##   <dbl> <fct>  <dbl> <dbl> <fct>  <fct>          <ord>     <fct>      <ord>
## 1  1032 WZ        47  26.7 Male   Married        <NA>      No         2
## 2  1036 WZ        52  23.4 Male   Married        <NA>      Yes        3
## # i 8 more variables: Rent <fct>, Loans <fct>, Smoking <fct>,
## #   Diabetes_Duration <dbl>, Hypertension_category <fct>, CVD <fct>,
## #   HbA1c <dbl>, Uncontrolled <fct>
```

```
df <- df %>%
  replace_na( list(Education = "3"))
summary(df$Education)
```

```
##   1   2   3 
## 486 578 147
```

To correct the Marital status of the category M to be Married

```
df$Marital_Status <- gsub("M", "Married", df$Marital_Status)
df$Marital_Status <- gsub("Marriedarried", "Married", df$Marital_Status)
df$Marital_Status <- as.factor(df$Marital_Status)
summary(df$Marital_Status)
```

```
## Divorced  Married   Single  Widowed 
##      135      963       61       52
```

To correct the Hypertension_category of Prehypertention to Pre-hypertention and then transforming it into ordinal categories

```
df$Hypertension_category <- gsub("Prehypertention", "Pre-hypertention", df$Hypertension_categor
y)
df$Hypertension_category <- factor(df$Hypertension_category, levels= c("Normal", "Pre-hypertenti
on", "First-Grade", "Second-Grade", "Third-Grade"), ordered=TRUE)
summary(df$Hypertension_category)
```

```
##           Normal Pre-hypertention      First-Grade     Second-Grade 
##              232              480              376              110 
##      Third-Grade 
##               13
```

To correct the BMI values of 2.2 and 254

```
df[(df$BMI > 56 | df$BMI < 10),] # to get the ID of those records
```

```
## # A tibble: 2 x 17
##      ID Region   Age   BMI Gender Marital_Status Education Employment HH_Income
##   <dbl> <fct>  <dbl> <dbl> <fct>  <fct>          <ord>     <fct>      <ord>    
## 1   207 WZ        35 254   Male   Married        2         Yes        2        
## 2   884 WZ        55   2.2 Male   Married        1         Yes        1        
## # i 8 more variables: Rent <fct>, Loans <fct>, Smoking <fct>,
## #   Diabetes_Duration <dbl>, Hypertension_category <ord>, CVD <fct>,
## #   HbA1c <dbl>, Uncontrolled <fct>
```

```
df$BMI[df$ID == 207] <- df$BMI[df$ID == 207]/10
df$BMI[df$ID == 207]
```

```
## [1] 25.4
```

```
df$BMI[df$ID == 884] <- df$BMI[df$ID == 884]*10
df$BMI[df$ID == 884]
```

```
## [1] 22
```

```
summary(df$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.20   25.90   28.30   29.15   31.50   55.60 
```

To correct the Age values that are less 18 (since the data is mentioning adults)

```
df[df$Age <= 20,]
```

```
## # A tibble: 30 x 17
##       ID Region   Age   BMI Gender Marital_Status Education Employment HH_Income
##    <dbl> <fct>  <dbl> <dbl> <fct>  <fct>          <ord>     <fct>      <ord>    
##  1    42 EA       6.3  23.7 Female Married        2         No         1        
##  2    43 EA       6.8  28.7 Male   Married        1         No         1        
##  3    44 SZ       6.9  26.8 Female Married        1         No         2        
##  4    45 EA       4.5  34.6 Male   Divorced       2         No         2        
##  5    46 EA       5.3  26.9 Female Married        1         Yes        3        
##  6    47 EA       4.8  25.7 Male   Married        2         Yes        2        
##  7    48 WA       5.5  34.8 Female Divorced       2         No         2        
##  8    49 WA       3.9  42.9 Female Married        2         No         4        
##  9    50 SZ       5.8  32.8 Female Married        1         No         4        
## 10    51 EA       4.7  42.9 Male   Married        2         No         2        
## # i 20 more rows
## # i 8 more variables: Rent <fct>, Loans <fct>, Smoking <fct>,
## #   Diabetes_Duration <dbl>, Hypertension_category <ord>, CVD <fct>,
## #   HbA1c <dbl>, Uncontrolled <fct>
```

```
df$Age <- ifelse(df$Age < 20, df$Age * 10, df$Age)
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   35.00   50.00   56.00   55.91   59.00   79.00 
```

The last thing we want to check is to ensure that Diabetes_Duration is less than the Age of the patience as we have a max value of 860 which is in years 71.67

```
df$Diabetes_Duration_years <- df$Diabetes_Duration/12.0
summary(df[,c("Age", "Diabetes_Duration_years")])
```

```
##       Age        Diabetes_Duration_years
##  Min.   :35.00   Min.   : 0.500
##  1st Qu.:50.00   1st Qu.: 3.083
##  Median :56.00   Median : 5.167
##  Mean   :55.91   Mean   : 5.497
##  3rd Qu.:59.00   3rd Qu.: 7.083
##  Max.   :79.00   Max.   :71.667
```

```
df[df$Diabetes_Duration_years >= df$Age, ]
```

```
## # A tibble: 1 x 18
##      ID Region   Age   BMI Gender Marital_Status Education Employment HH_Income
##   <dbl> <fct>  <dbl> <dbl> <fct>  <fct>          <ord>     <fct>      <ord>
## 1   533 WA        59  23.3  Female Married        2         Yes        4
## # i 9 more variables: Rent <fct>, Loans <fct>, Smoking <fct>,
## #   Diabetes_Duration <dbl>, Hypertension_category <ord>, CVD <fct>,
## #   HbA1c <dbl>, Uncontrolled <fct>, Diabetes_Duration_years <dbl>
```

From the code above we can see that the Diabetes_Duration = 71.667 years and the person age is 59 which is incorrect, this value should be divided by 10

The id = 533

```
df$Diabetes_Duration[df$ID == 533] <- df$Diabetes_Duration[df$ID == 533]/10
df[df$ID == 533,]
```

```
## # A tibble: 1 x 18
##      ID Region   Age   BMI Gender Marital_Status Education Employment HH_Income
##   <dbl> <fct>  <dbl> <dbl> <fct>  <fct>          <ord>     <fct>      <ord>
## 1   533 WA        59  23.3 Female Married        2         Yes        4
## # i 9 more variables: Rent <fct>, Loans <fct>, Smoking <fct>,
## #   Diabetes_Duration <dbl>, Hypertension_category <ord>, CVD <fct>,
## #   HbA1c <dbl>, Uncontrolled <fct>, Diabetes_Duration_years <dbl>
```

Data is now ready, just dropping the columns created and we can save it into CSV

```
df <- subset(df, select = -Diabetes_Duration_years)
write_csv(df, "Project Data-cleand-group3.csv")
```