# Numerical Variable Markdown

Everett

2024-07-03

## R Markdown Numerical Variable Analysis

The first chunk of code will set the working directory and load the tidyverse library which will be used for the variable analysis. Then the dataset will be limited to only the variables that are of the numeric type.

```
setwd("C:/Users/escra/OneDrive/Documents/Job Stuff/DA Project Logistic Regression/Dataset")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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 become errors
```

```
df <- read_csv("dataset.csv")
```

```
## Rows: 9709 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): Income_type, Education_type, Family_status, Housing_type, Occupati...
## dbl (15): ID, Gender, Own_car, Own_property, Work_phone, Phone, Email, Unemp...
##
## 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.
```

```
head(df)
```

```
## # A tibble: 6 x 20
##        ID Gender Own_car Own_property Work_phone Phone Email Unemployed
##     <dbl>  <dbl>   <dbl>        <dbl>      <dbl> <dbl> <dbl>      <dbl>
## 1 5008804      1       1            1          1     0     0          0
## 2 5008806      1       1            1          0     0     0          0
## 3 5008808      0       0            1          0     1     1          0
## 4 5008812      0       0            1          0     0     0          1
```

```
## 5 5008815      1       1            1          1     1     1          0
## 6 5008819      1       1            1          0     0     0          0
## # i 12 more variables: Num_children <dbl>, Num_family <dbl>,
## #   Account_length <dbl>, Total_income <dbl>, Age <dbl>, Years_employed <dbl>,
## #   Income_type <chr>, Education_type <chr>, Family_status <chr>,
## #   Housing_type <chr>, Occupation_type <chr>, Target <dbl>
```

```
df_quant <- df[, c("Num_children", "Num_family", "Account_length", "Total_income", "Age", "Years_employ
```

### Variable Analysis for the Num_children Variable

A numerical feature that indicates how many children an applicant has. A greater number of children means that a customer will have more expenses.

```
mean(df_quant[["Num_children"]])
```

```
## [1] 0.4228036
```

```
median(df_quant[["Num_children"]])
```

```
## [1] 0
```

```
max(df_quant[["Num_children"]])
```

```
## [1] 19
```

```
min(df_quant[["Num_children"]])
```

```
## [1] 0
```

```
sd(df_quant[["Num_children"]])
```

```
## [1] 0.7670189
```

### Variable Analysis for the Num_family Variable

A numerical feature that indicates the total number of family members an applicant has. More family members shows that a customer has more expenses.

```
mean(df_quant[["Num_family"]])
```

```
## [1] 2.182614
```

```
median(df_quant[["Num_family"]])
```

```
## [1] 2
```

```r
max(df_quant[["Num_family"]])
```

```
## [1] 20
```

```r
min(df_quant[["Num_family"]])
```

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

```r
sd(df_quant[["Num_family"]])
```

```
## [1] 0.9329182
```

### Variable Analysis for the Account_length Variable

A numerical feature that indicates the length of a customers account with the bank. Longer account lengths indicates a customer's stability with the financial institution.

```r
mean(df_quant[["Account_length"]])
```

```
## [1] 27.27006
```

```r
median(df_quant[["Account_length"]])
```

```
## [1] 26
```

```r
max(df_quant[["Account_length"]])
```

```
## [1] 60
```

```r
min(df_quant[["Account_length"]])
```

```
## [1] 0
```

```r
sd(df_quant[["Account_length"]])
```

```
## [1] 16.64806
```

### Variable Analysis for the Total_income Variable

A numerical variable showing the total income of a customer. Higher levels of income show more stability and financial security.

```r
mean(df_quant[["Total_income"]])
```

```
## [1] 181228.2
```

```
median(df_quant[["Total_income"]])
```

```
## [1] 157500
```

```
max(df_quant[["Total_income"]])
```

```
## [1] 1575000
```

```
min(df_quant[["Total_income"]])
```

```
## [1] 27000
```

```
sd(df_quant[["Total_income"]])
```

```
## [1] 99277.31
```

### Variable Analysis for the Age Variable

A numerical variable that shows the age of an individual in years. This variable is correct to multiple decimal points. Older people may represent more financial security. In the US, it is important to understand that age can't be used to discriminate when making lending decisions but can be considered when the applicant is entering into a binding contract (CFPB).

```
mean(df_quant[["Age"]])
```

```
## [1] 43.78409
```

```
median(df_quant[["Age"]])
```

```
## [1] 42.74147
```

```
max(df_quant[["Age"]])
```

```
## [1] 68.86384
```

```
min(df_quant[["Age"]])
```

```
## [1] 20.50419
```

```
sd(df_quant[["Age"]])
```

```
## [1] 11.62577
```

### Variable Analysis for the Years_employed Variable

A numerical variable that indicates the number of years a customer has been employed. Greater years of employment can indicate greater savings and more reliability.

```r
mean(df_quant[["Years_employed"]])
```

```
## [1] 5.66473
```

```r
median(df_quant[["Years_employed"]])
```

```
## [1] 3.761884
```

```r
max(df_quant[["Years_employed"]])
```

```
## [1] 43.02073
```

```r
min(df_quant[["Years_employed"]])
```

```
## [1] 0
```

```r
sd(df_quant[["Years_employed"]])
```

```
## [1] 6.342241
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.