Wrangling and Visualizing Data with R

Abu Ali

2024-08-09

The Titanic Dataset

we will use R packages to explore the Titanic dataset and visualize key patterns and insights, and their relations to the survival rate of the passengers.

You can download the the titanic dataset here: https://www.kaggle.com/datasets/yasserh/titanic-dataset

LOADING THE DATASET:

```
library(readxl)
titanic_ds <- read_excel("titanic_ds.xls")
```

```
## Warning: Coercing text to numeric in M1306 / R1306C13: '328'
```

```
str(titanic_ds)
```

```
## tibble [1,309 x 14] (S3: tbl_df/tbl/data.frame)
##  $ pclass   : num [1:1309] 1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : num [1:1309] 1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : chr [1:1309] "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison
##  $ sex      : chr [1:1309] "female" "male" "female" "male" ...
##  $ age      : num [1:1309] 29 0.917 2 30 25 ...
##  $ sibsp    : num [1:1309] 0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : num [1:1309] 0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : chr [1:1309] "24160" "113781" "113781" "113781" ...
##  $ fare     : num [1:1309] 211 152 152 152 152 ...
##  $ cabin    : chr [1:1309] "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ embarked : chr [1:1309] "S" "S" "S" "S" ...
##  $ boat     : chr [1:1309] "2" "11" NA NA ...
##  $ body     : num [1:1309] NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: chr [1:1309] "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterv
```

CLEANING THE DATASET:

```
library(Amelia)
```

1. Check for missing values

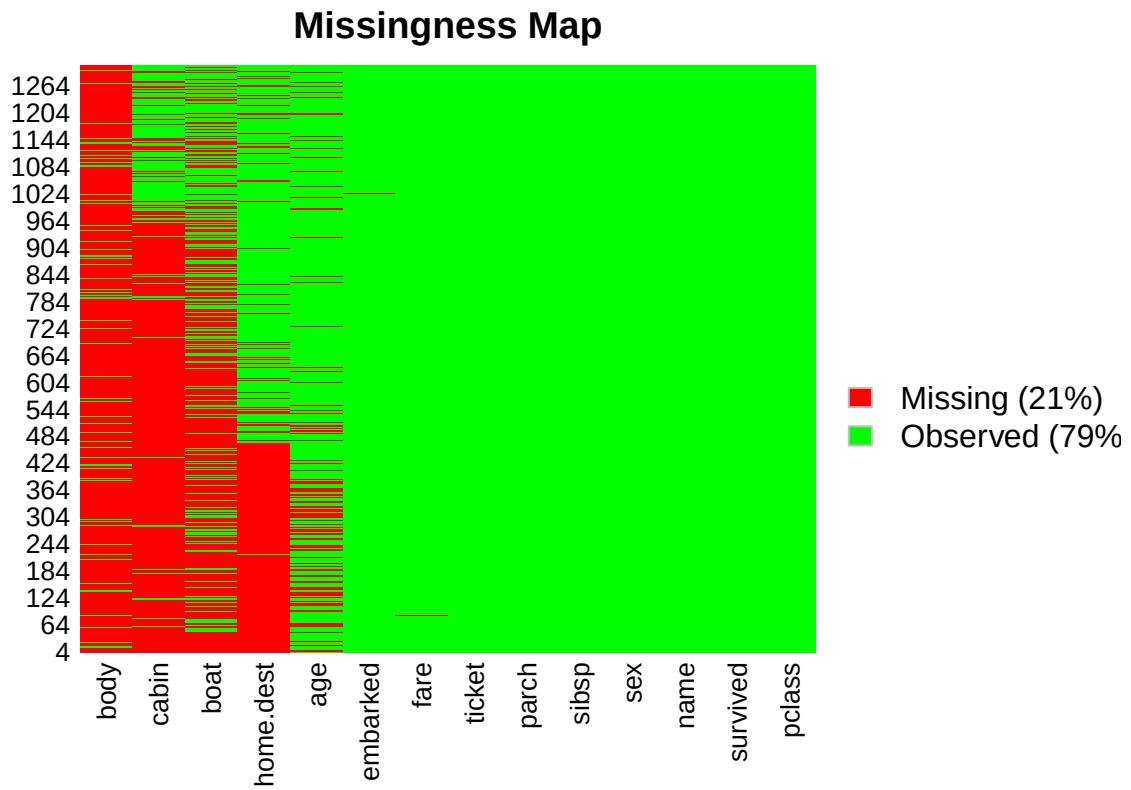
```
## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.2, built: 2024-04-10)
## ## Copyright (C) 2005-2024 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(titanic_ds, col = c("red", "green"))
```

```
## Warning: Unknown or uninitialised column: 'arguments'.
## Unknown or uninitialised column: 'arguments'.

## Warning: Unknown or uninitialised column: 'imputations'.
```



Note that you can add `echo = FALSE` parameter to the code chunk to prevent printing of the R code that generated the plot.

```
library(tidyverse)
```

2. Select relevant columns for the analysis

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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
```

```
selected_titanic <- titanic_ds %>%
  select (age, pclass, sex, survived, embarked, home.dest, fare, parch, sibsp)
```

```
selected_titanic$FamilySize <- selected_titanic$sibsp + selected_titanic$parch + 1
str(selected_titanic)
```

3. Merge columns parch and sibsp to create a new column, FamilySize

```
## tibble [1,309 x 10] (S3: tbl_df/tbl/data.frame)
##  $ age       : num [1:1309] 29 0.917 2 30 25 ...
##  $ pclass    : num [1:1309] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : chr [1:1309] "female" "male" "female" "male" ...
##  $ survived  : num [1:1309] 1 1 0 0 0 1 1 0 1 0 ...
##  $ embarked  : chr [1:1309] "S" "S" "S" "S" ...
##  $ home.dest : chr [1:1309] "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterv
##  $ fare      : num [1:1309] 211 152 152 152 152 ...
##  $ parch     : num [1:1309] 0 2 2 2 2 0 0 0 0 0 ...
##  $ sibsp     : num [1:1309] 0 1 1 1 1 0 1 0 2 0 ...
##  $ FamilySize: num [1:1309] 1 4 4 4 4 1 2 1 3 1 ...
```

```
selected_titanic$FareCategory <- cut(selected_titanic$fare,
                                     breaks = c(0, 10, 20, 50, 100, Inf),
                                     labels = c("Lowest", "Lower Middle",
                                          "Upper Middle", "Higher", "Highest"))
str(selected_titanic)
```

4. Categorize the fare column and assign label to each category

```
## tibble [1,309 x 11] (S3: tbl_df/tbl/data.frame)
##  $ age         : num [1:1309] 29 0.917 2 30 25 ...
##  $ pclass      : num [1:1309] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex         : chr [1:1309] "female" "male" "female" "male" ...
##  $ survived    : num [1:1309] 1 1 0 0 0 1 1 0 1 0 ...
##  $ embarked    : chr [1:1309] "S" "S" "S" "S" ...
##  $ home.dest   : chr [1:1309] "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chest
##  $ fare        : num [1:1309] 211 152 152 152 152 ...
```

```
##  $ parch       : num [1:1309] 0 2 2 2 2 0 0 0 0 0 ...
##  $ sibsp       : num [1:1309] 0 1 1 1 1 0 1 0 2 0 ...
##  $ FamilySize  : num [1:1309] 1 4 4 4 4 1 2 1 3 1 ...
##  $ FareCategory: Factor w/ 5 levels "Lowest","Lower Middle",..: 5 5 5 5 5 3 4 NA 4 3 ...
```

```
selected_titanic <- selected_titanic %>%
  select(-fare, -parch, -sibsp)
```

5. Remove the columns that are being merged to form new columns

```
selected_titanic <- selected_titanic %>%
  mutate(
    survived = ifelse(survived == 0, "No", "Yes"),
    age = ifelse(age >= 18, "Adult", "Child"),
    pclass = case_when(
      pclass == 1 ~ "1st",
      pclass == 2 ~ "2nd",
      pclass == 3 ~ "3rd"
    ),

    embarked = case_when(
      embarked == "C" ~ "Cherbourg",
      embarked == "Q" ~ "Queenstown",
      embarked == "S" ~ "Southampton"
    )
  )
```

6. Change the values of columns pclass, survived, and embarked

```
selected_titanic <- selected_titanic %>%
  rename(
    Class = pclass,
    Destination = home.dest
  )
```

7. Change the name of column pclass to Class, and home.dest to Destination

```
selected_titanic <- selected_titanic %>%
  rename_all(~str_to_title(.))
```

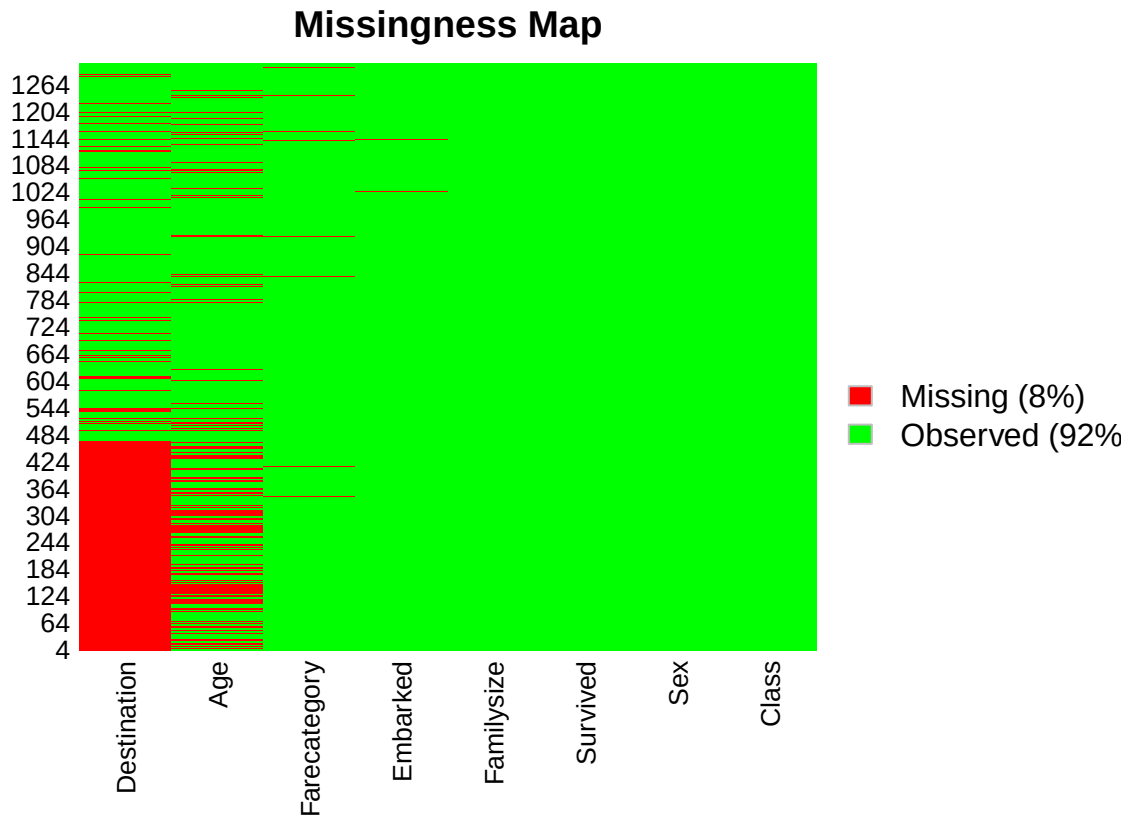
8. Capitalize the initials of all the columns name

```
missmap(selected_titanic, col = c("red", "green"))
```

9. Check for missing values again

```
## Warning: Unknown or uninitialised column: ‘arguments‘.
## Unknown or uninitialised column: ‘arguments‘.
```

```
## Warning: Unknown or uninitialised column: ‘imputations‘.
```

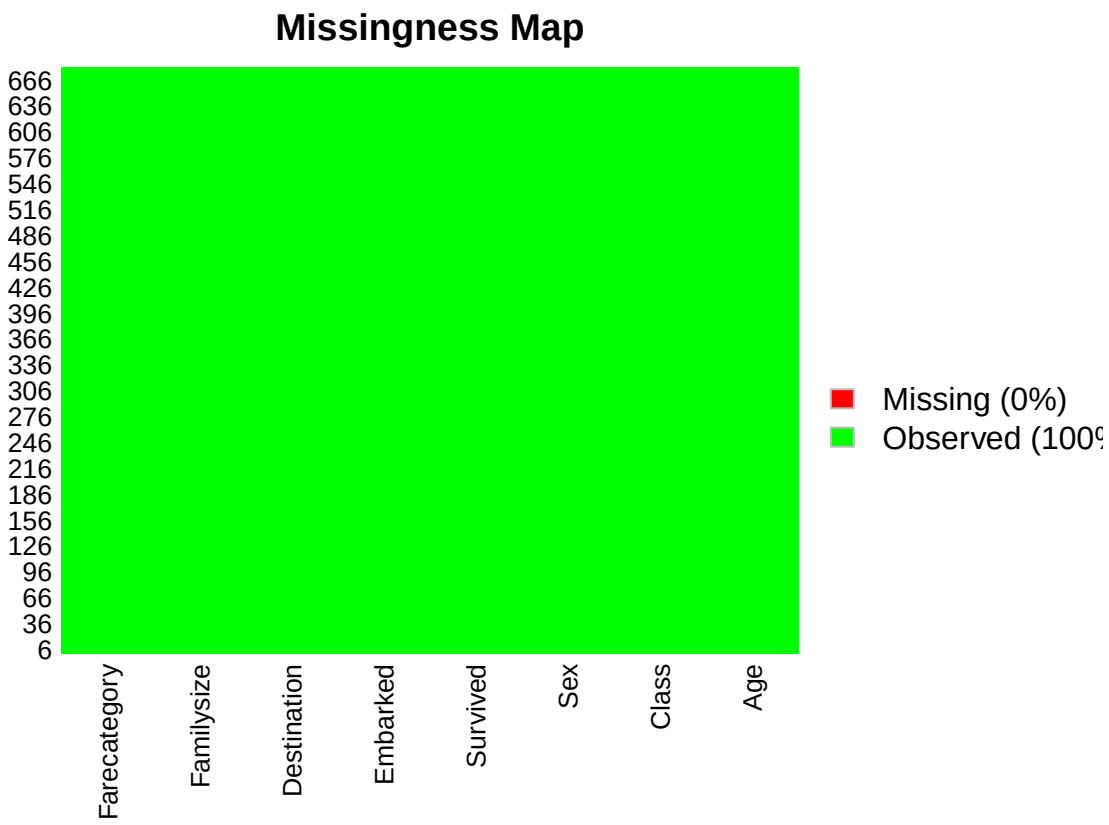



```
selected_titanic <- drop_na(selected_titanic)
```

10. Drop all the missing values from the dataset

```
## Warning: Unknown or uninitialised column: ‘arguments‘.
## Unknown or uninitialised column: ‘arguments‘.
```

```
## Warning: Unknown or uninitialised column: ‘imputations‘.
```



EXPLORING THE DATASET: