Customer dropout membership*

Pedro Sobreiro, Javier Berrocal, Domingos Martinho, José Garcia Alonso

Extremadura University

25 julho, 2021

Abstract

Abstract of the article. Here we can place more info.

1 Introduction

Research idea:

-

Context: An organization membership located in Portugal. The organization offers an annual membership for the members, the service subscription has several payment options:

- Men with a annual fee of 10€
- Women annual fee of 6€

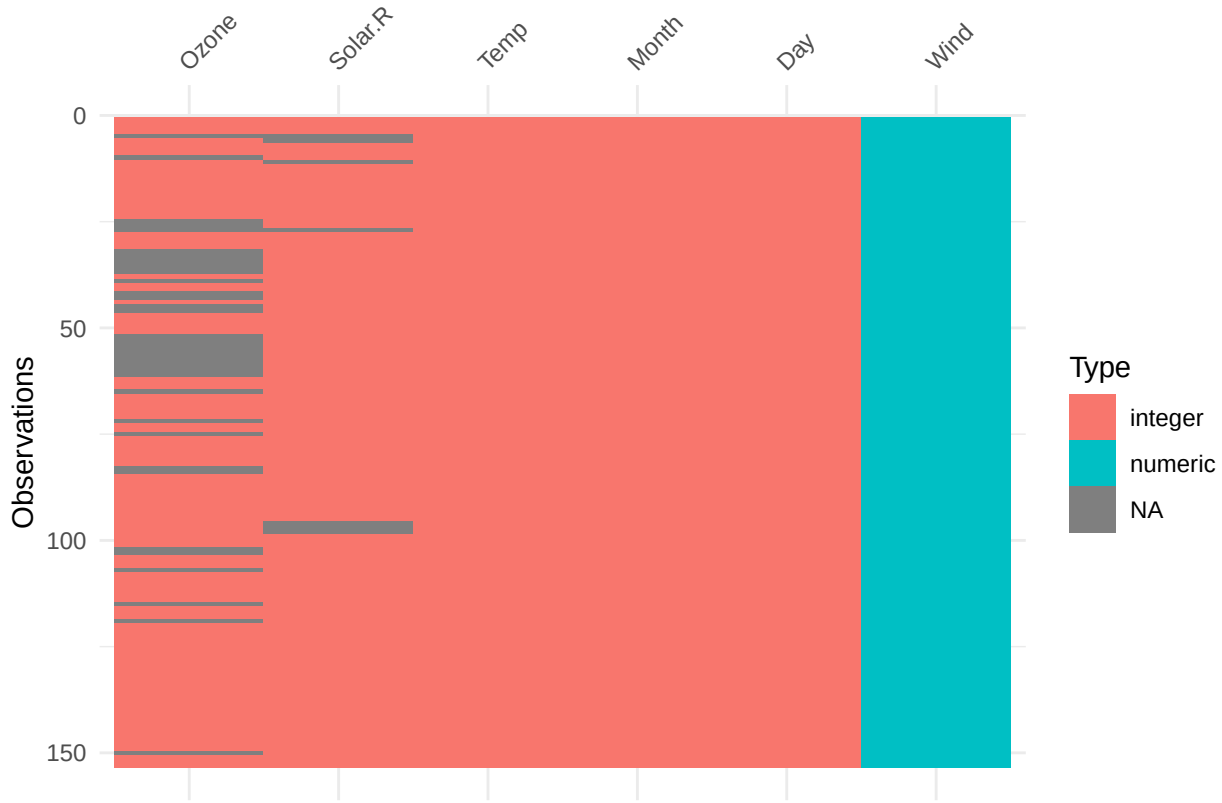
*Corresponding address: sobreiro@esdrm.ipsantarem.pt. The current template adapts part of the Rmd code by Paul C. Bauer, Mannheim Centre for European Social Research.

- Correspondent fee 6€
- Retired fee 5€
- Student fee 2.5€
- under-14 fee 1€

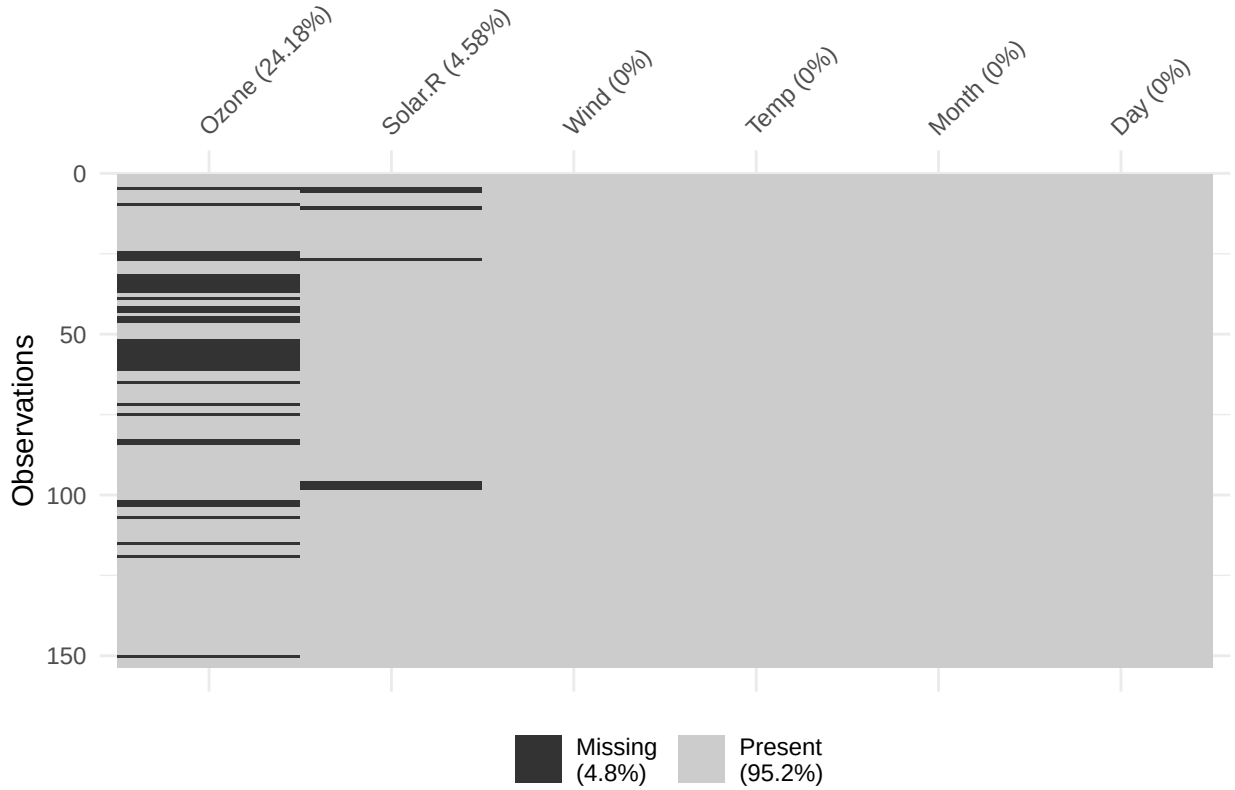
```
library(dplyr)
library(dlookr)
library(ggplot2)

#eda_report(nlswork,output_dir  =
#   "C:/Users/mangelo.EEG/Documents/GitHub/prjs/reports/",
#   output_file = "eda_report.pdf")

## The data

names(airquality)
```

```
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

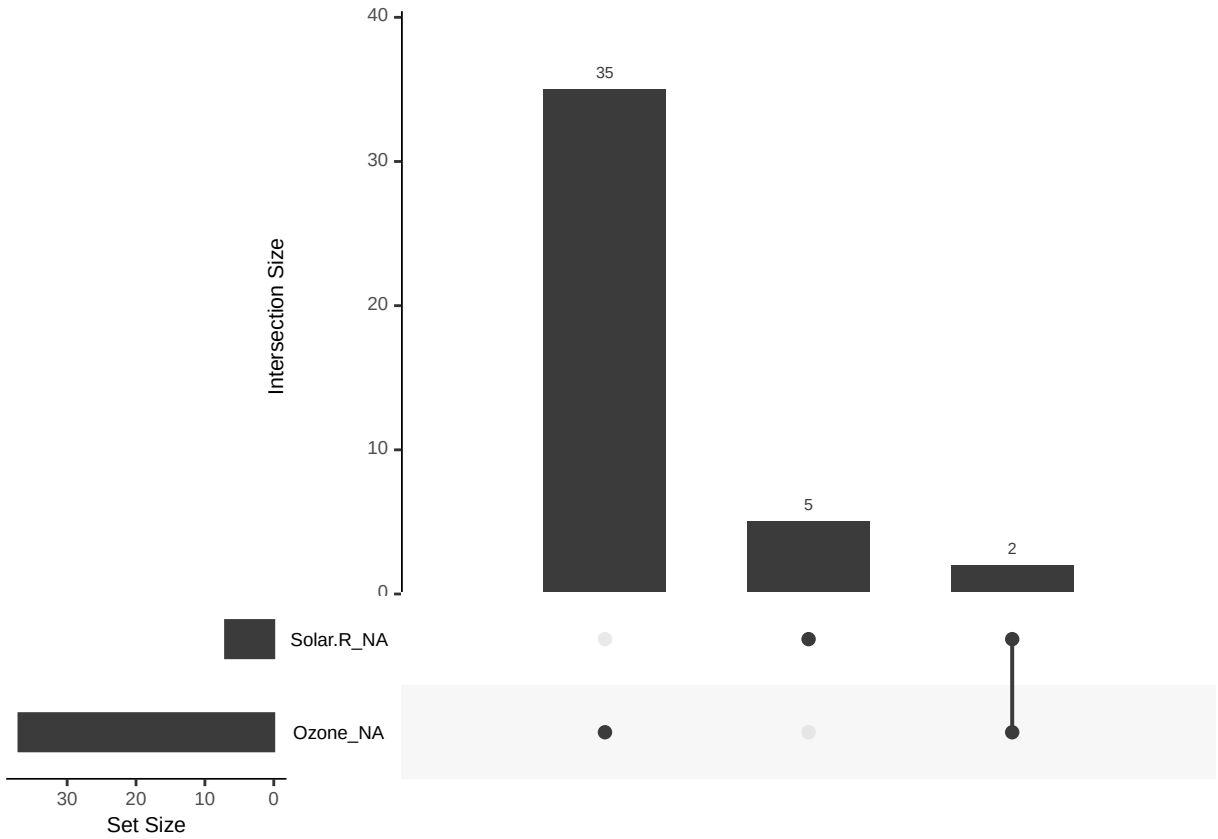
```
#summary(nlswork)

## Missing values
library(visdat)
vis_dat(airquality)
```



```
library(naniar)
vis_miss(airquality)
```



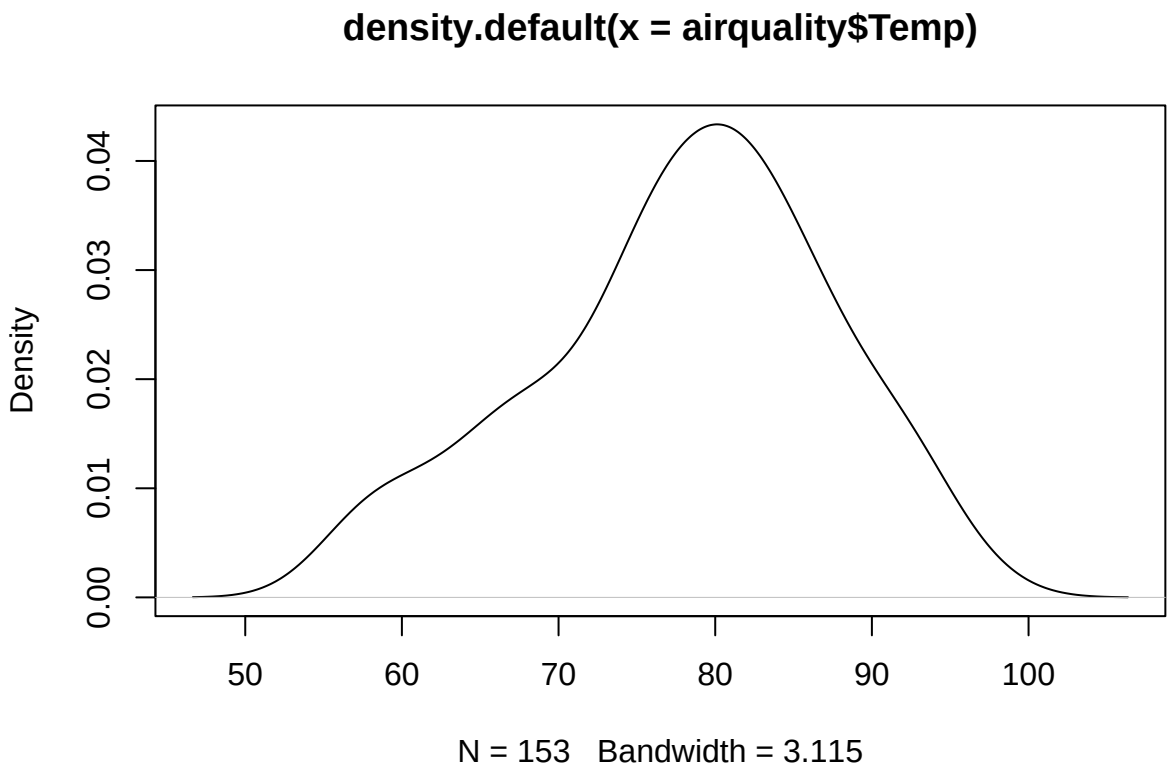

```
gg_miss_upset(airquality)
```



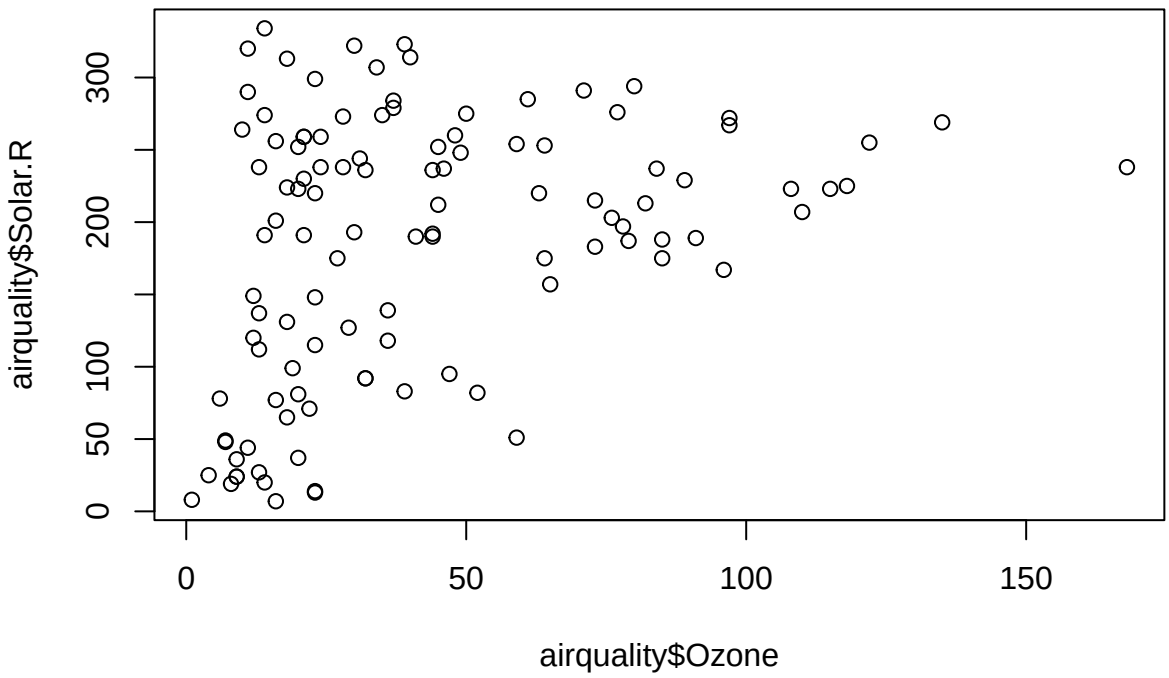

```
## GRAPHS
dplyr::glimpse(cars$Ozone)
```

```
##  NULL
```

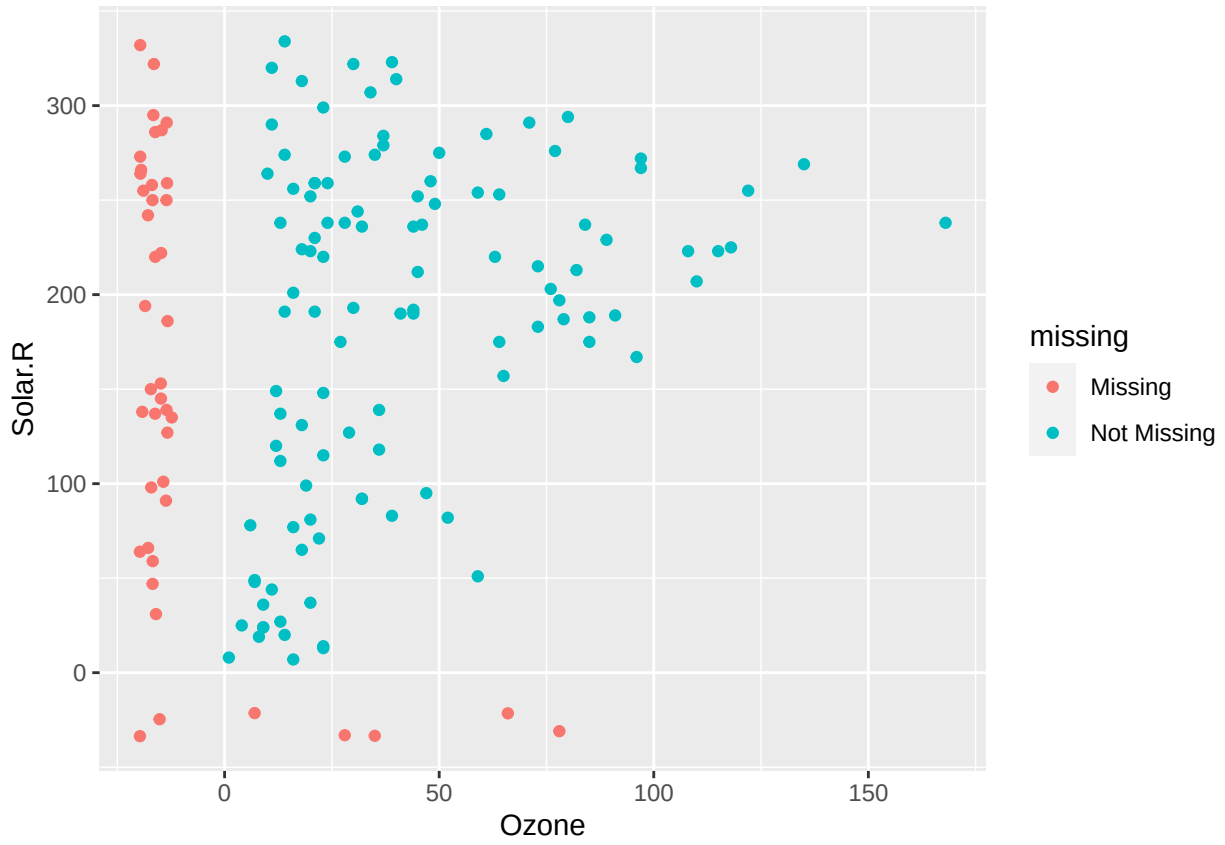
```
d <- density(airquality$Temp)
plot(d)
```



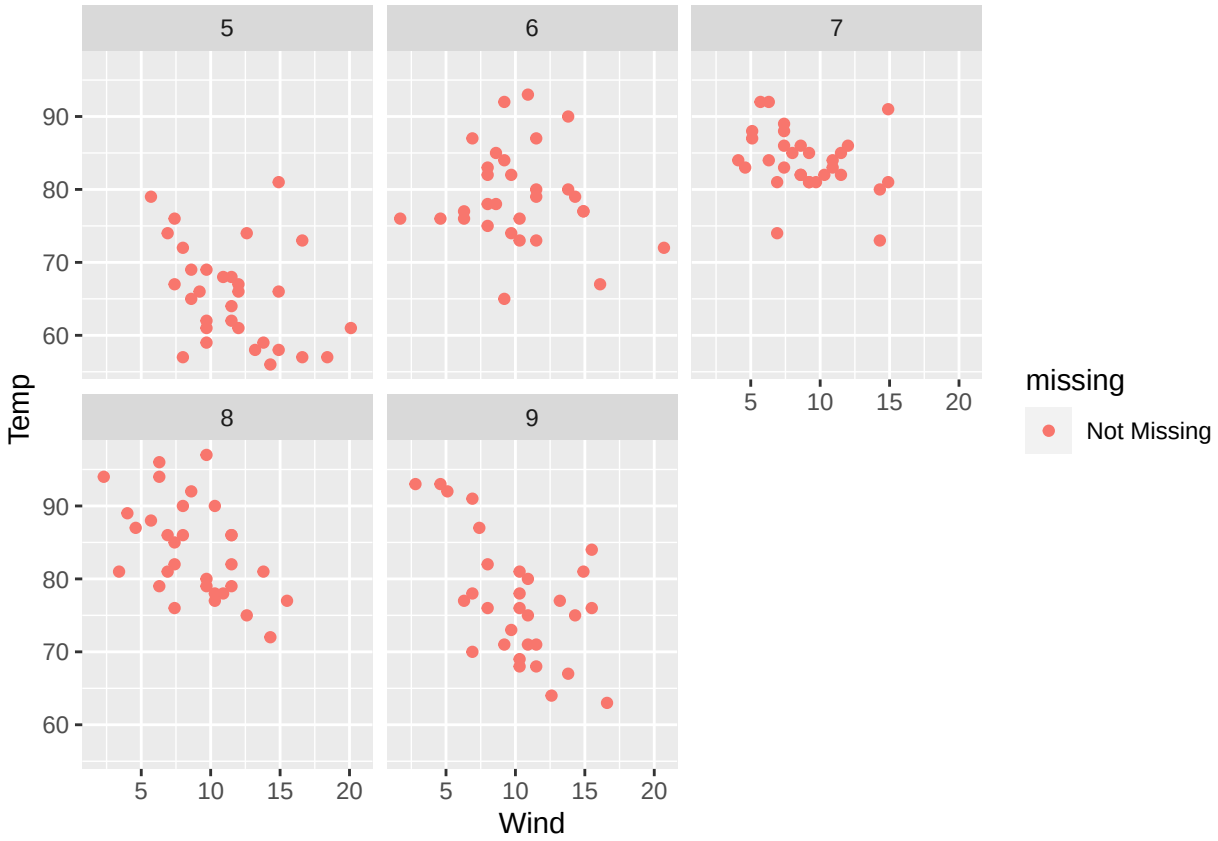

```
plot(airquality$Ozone, airquality$Solar.R)
```




```
ggplot(airquality, aes(x = Ozone, y = Solar.R)) +
geom_miss_point()
```




```
ggplot(airquality, aes(x = Wind, y = Temp)) +
geom_miss_point() +
facet_wrap(vars(Month))
```




```
stats <- summary(airquality$Temp)
stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   72.00   79.00   77.88   85.00   97.00
```

```
describe(airquality)
```

```
## # A tibble: 6 x 26
##   variable     n    na   mean    sd se_mean   IQR skewness kurtosis   p00   p01
##   <chr>    <int> <int>  <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl>
## 1 Ozone      116    37  42.1  33.0    3.06   45.2  1.24       1.29    1    4.3
## 2 Solar.R    146     7 186.   90.1    7.45  143   -0.428     -0.968   7   10.2
## 3 Wind       153     0   9.96  3.52   0.285   4.1  0.348      0.111   1.7  2.56
## 4 Temp       153     0  77.9   9.47   0.765  13   -0.378     -0.404  56   57
## 5 Month      153     0   6.99  1.42   0.115   2   -0.00239   -1.30    5    5
## 6 Day        153     0  15.8   8.86   0.717  15    0.00265   -1.20    1    1
## # ... with 15 more variables: p05 <dbl>, p10 <dbl>, p20 <dbl>, p25 <dbl>,
## #   p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>, p75 <dbl>,
## #   p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>
```

2 Experimental Results

2.1 Data description

```
##  [1] "Sócio"               "dataAdesao"          "ano"
##  [4] "dataNascimento"      "idade"               "sexo"
##  [7] "estadoCivil"         "categoria"           "quotaMensal"
## [10] "profissao"           "codPostal"           "ultimaQuota"
## [13] "ultimoPagamento"     "valorTotal"          "totalJogos"
## [16] "jogosEpoca"          "diasUltimoPagamento" "mesesUP"
## [19] "abandonou"           "anosSocio"           "idaEstadio"
## [22] "escaloesTotalJogos"  "mes"
```

```
##  [1] "num_socio"                "dt_inscription"
##  [3] "year"                     "birth_date"
##  [5] "age"                      "sex"
##  [7] "marital_status"           "category"
##  [9] "monthly_fee"              "occupation"
## [11] "zip_code"                 "dt_last_invoice"
## [13] "dt_last_payment"          "total_amount"
## [15] "total_matches"            "season_matches"
## [17] "days_since_last_payment"  "months_since_last_payment"
## [19] "dropout"                  "years_membership"
## [21] "stadium_access"           "quart_stadium_entries"
## [23] "inscription_month"
```

```
## tibble [25,316 x 14] (S3: tbl_df/tbl/data.frame)
##  $ year                     : num [1:25316] 1944 1944 1945 1945 1945 ...
##  $ age                      : num [1:25316] 83 88 73 97 97 91 88 95 88 78 ...
##  $ sex                      : chr [1:25316] "M" "M" "M" "M" ...
##  $ marital_status           : chr [1:25316] "casado" "solteiro" "nao definido" "casad
##  $ monthly_fee              : num [1:25316] 10 10 10 5 10 5 5 5 10 10 ...
##  $ total_amount             : num [1:25316] 1906 1906 1553 790 1466 ...
##  $ total_matches            : num [1:25316] 0 0 0 0 0 20 74 0 154 0 ...
##  $ season_matches           : num [1:25316] 0 0 0 0 0 0 0 0 6 0 ...
##  $ months_since_last_payment: num [1:25316] 3 3 36 8 35 4 41 40 4 2 ...
##  $ dropout                  : num [1:25316] 0 0 1 0 1 0 1 1 0 0 ...
##  $ years_membership         : num [1:25316] 74 74 73 73 73 73 73 73 73 72 ...
##  $ stadium_access           : num [1:25316] 0 0 0 0 0 1 1 0 1 0 ...
##  $ quart_stadium_entries    : chr [1:25316] "ate 1" "ate 1" "ate 1" "ate 1" ...
##  $ inscription_month        : num [1:25316] 10 10 8 9 9 12 1 1 2 4 ...
```

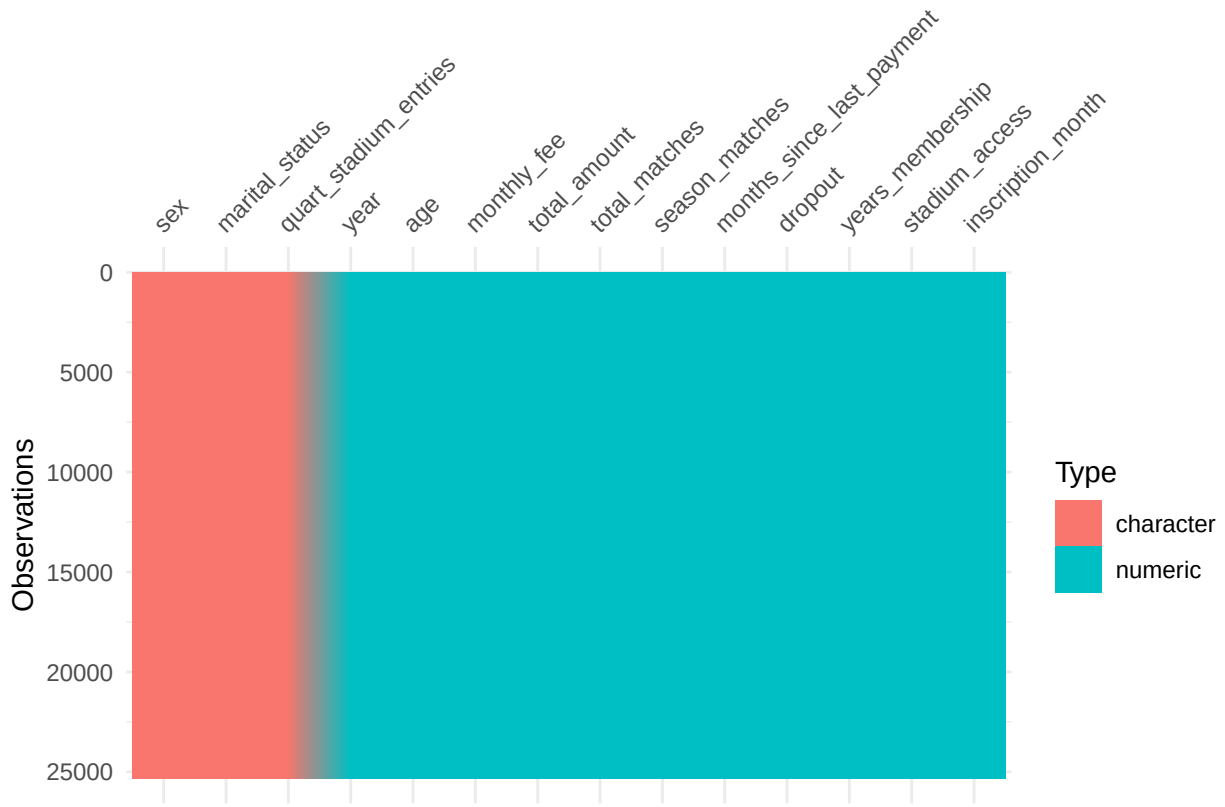


Table **??** shows data's summary statistics.[1] `stargazer()` is and excellent solution to export outputs.

Teste 1

```
#from pysurvival.utils.display import correlation_matrix
import pandas as pd
import numpy as np

col = ['sex','marital_status','quart_stadium_entries']

df_members = r.df_members #copy r dataframe to python

df_members = pd.get_dummies(df_members, columns=col,drop_first=True)

# Creating the time and event columns
time_column = 'years_membership'
event_column = 'dropout'
```

[1] You can reference the table as **??**.

Table 1: Summary statistics

Characteristic	N = 25,316
year	2,007 (11)
age	27 (20)
sex	
F	32%
M	68%
marital_status	
casado	20%
nao definido	30%
outro	2.0%
solteiro	48%
monthly_fee	
0	<0.1%
1	32%
2.5	28%
5	3.4%
6	12%
10	24%
total_amount	316 (494)
total_matches	27 (46)
season_matches	2.2 (4.1)
months_since_last_payment	19 (32)
dropout	22%
years_membership	11 (11)
stadium_access	40%
quart_stadium_entries	
1 a 21	10%
21 a 56	9.8%
56 a 105	10.0%
ate 1	60%
mais 105	10.0%
inscription_month	6.9 (3.4)

[1] Mean (SD); %

```
# Extracting the features
features = np.setdiff1d(df_members.columns, [time_column, event_column] ).tolist()


#correlation_matrix(df_members, figure_size=(10,10), text_fontsize=6)
```

The average age in our data is 27.3.

3 Tables

R Markdown PDF is now able to produce good tables with our output. For `stargazer` the label is contained in the function, while for `kable` it's contained in the chunk name.

3.1 stargazer(): Summary and regression tables

Table 2 reports regression outputs. Name the models as you can refer to their names in the text (M1, M2, M3).

```
library(stargazer)
model1 <- lm(speed ~ dist, data = cars)
model2 <- lm(speed ~ dist, data = cars)
model3 <- lm(dist ~ speed, data = cars)
stargazer(model1, model2, model3,
          title = "Regression table with stargazer",
          label = "tab2",
          table.placement = "h",
          column.labels = c("M1", "M2", "M3"),
          model.numbers = FALSE,
          header = FALSE)
```

4 Figures

4.1 Graphs with R

You can insert figures like this. One would like to produce and insert them on the fly in the `.rmd` file. Figure **??** is such an example.

```
plot(cars$speed, cars$dist)
```

However, in some cases it does not work.

Table 2: Regression table with stargazer

	Dependent variable:		
	speed		dist
	M1	M2	M3
dist	0.166***	0.166***	
	(0.017)	(0.017)	
speed			3.932***
			(0.416)
Constant	8.284***	8.284***	−17.579**
	(0.874)	(0.874)	(6.758)
Observations	50	50	50
R^2	0.651	0.651	0.651
Adjusted R^2	0.644	0.644	0.644
Residual Std. Error (df = 48)	3.156	3.156	15.380
F Statistic (df = 1; 48)	89.567***	89.567***	89.567***
Note:			*p<0.1; **p<0.05; ***p<0.01

4.2 Example: ggplot2 graphs

See the `ggplot2` output reported in Figure **??**.

```
mtcars$cyl <- as.factor(mtcars$cyl) # Convert cyl to factor
library(ggplot2)
ggplot(mtcars, aes(x = wt, y = mpg, shape = cyl)) + geom_point() +
  labs(x = "Weight (lb/1000)", y = "Miles/(US) gallon",
       shape = "Number of \n Cylinders") + theme_classic()
```

4.3 Another example using Plotly

With `Plotly` we can produce interactive graphs which play well, for example, once can embedded in html webpages (drop by here for an example). One can insert this type of graphs in R Markdown PDF using `Orca` (it generates static images from Plotly graphs). Go here to check how to install it. See Figure 3 for an example.

```
library(plotly)
p <- plot_ly(cars, type = "scatter", mode = "markers",
        x = ~speed,
        y = ~dist)
```

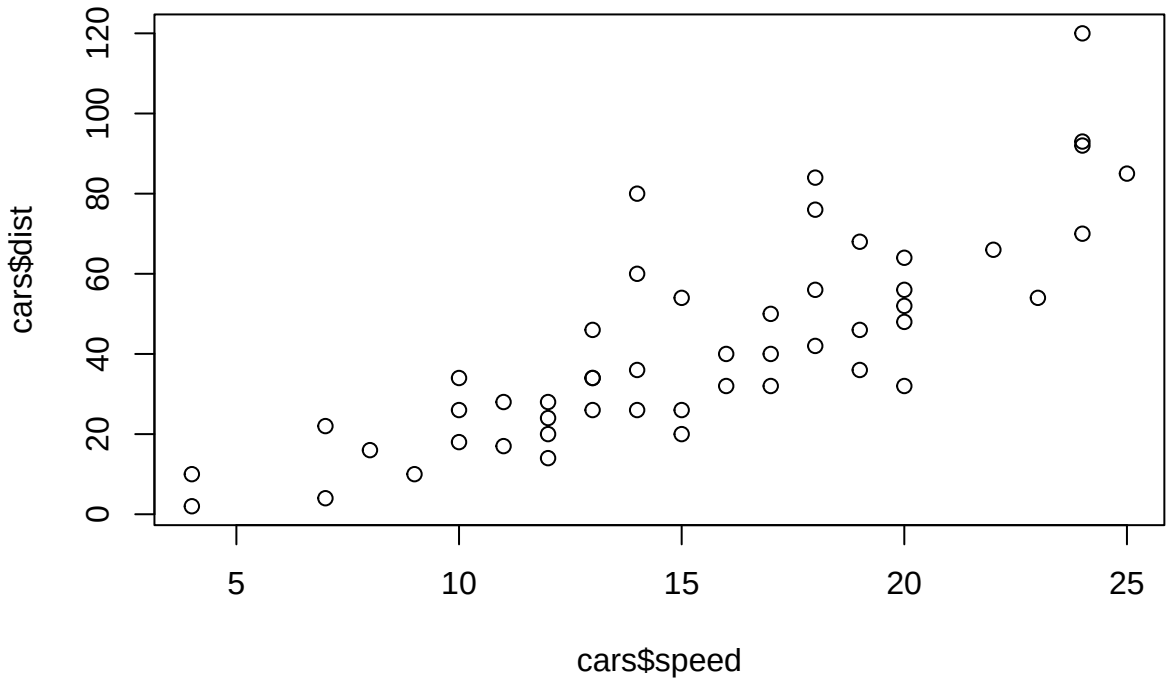


Figure 1: Scatterplot of Speed and Distance

(#fig:Figures 1, fig-1)

```
#Sys.setenv('MAPBOX_TOKEN' = '12123123') # set arbitrary token
# Lets create a value for example


media <- mean(cars$speed)
```

The criminal rate is 15.4%o.

5 Miguel's tests

5.1 R

Example of an equation

$$\int_0^{2\pi} \sin x \, dx$$

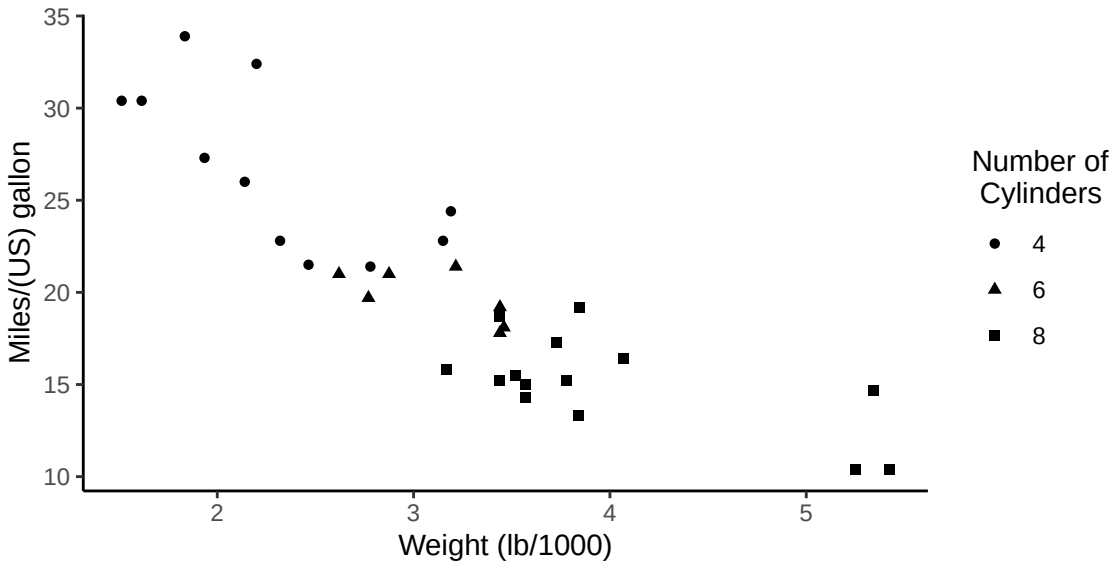


Figure 2: Miles per gallon according to the weight

Example of a matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

or

$$f\left(k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k} \tag{1}$$

See Equation (1).

$$y_{ijt} = \beta x_{ijt} + \eta_i + \gamma_j + \lambda_t + \varepsilon_{ijt} \tag{2}$$

```
library(stargazer)
stargazer(cars,
          title = "Summary 24",
          label = "tab24",
          table.placement = "ht",
          header = FALSE)
```

Table 3: Summary 24

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
speed	50	15.400	5.288	4	12	19	25
dist	50	42.980	25.769	2	26	56	120

6 Final remarks

Check the replication package for Bonhomme, Lamadon and Manresa (2019): https://github.com/tlamadon/blm-replicate

References

Appendix: Chunk options

6.1 Software versioning

6.1.1 R

```
cat(paste("#", capture.output(sessionInfo()), "\n", collapse  = ""))
```

```
## # R version 4.1.0 (2021-05-18)
## # Platform: x86_64-w64-mingw32/x64 (64-bit)
## # Running under: Windows 10 x64 (build 19043)
## #
## # Matrix products: default
## #
## # locale:
## # [1] LC_COLLATE=Portuguese_Portugal.1252  LC_CTYPE=Portuguese_Portugal.1252
## # [3] LC_MONETARY=Portuguese_Portugal.1252 LC_NUMERIC=C
## # [5] LC_TIME=Portuguese_Portugal.1252
## #
## # attached base packages:
## # [1] stats     graphics  grDevices utils     datasets  methods   base
## #
## # other attached packages:
## #  [1] plotly_4.9.4.1   kableExtra_1.3.4 gtsummary_1.4.2  readxl_1.3.1
## #  [5] stargazer_5.2.2  naniar_0.6.1     visdat_0.5.3     ggplot2_3.3.4
## #  [9] dlookr_0.4.5     dplyr_1.0.7
## #
## # loaded via a namespace (and not attached):
## #   [1] webshot_0.5.2       RColorBrewer_1.1-2  httr_1.4.2
## #   [4] UpSetR_1.4.0        tools_4.1.0         backports_1.2.1
## #   [7] utf8_1.2.1          R6_2.5.0            rpart_4.1-15
## #  [10] lazyeval_0.2.2      Hmisc_4.5-0         nortest_1.0-4
## #  [13] colorspace_2.0-1    nnet_7.3-16         withr_2.4.2
## #  [16] tidyselect_1.1.1    gridExtra_2.3       curl_4.3.1
## #  [19] compiler_4.1.0      extrafontdb_1.0     cli_2.5.0
## #  [22] rvest_1.0.0         gt_0.3.0            htmlTable_2.2.1
## #  [25] xml2_1.3.2          sandwich_3.0-1      labeling_0.4.2
## #  [28] bookdown_0.22       scales_1.1.1        checkmate_2.0.0
## #  [31] mvtnorm_1.1-2       proxy_0.4-26        RcmdrMisc_2.7-1
```

```
## #  [34] systemfonts_1.0.2   stringr_1.4.0       digest_0.6.27
## #  [37] foreign_0.8-81      rmarkdown_2.9       svglite_2.0.0
## #  [40] rio_0.5.27          base64enc_0.1-3     jpeg_0.1-8.1
## #  [43] pkgconfig_2.0.3     htmltools_0.5.1.1   extrafont_0.17
## #  [46] highr_0.9           htmlwidgets_1.5.3   rlang_0.4.11
## #  [49] rstudioapi_0.13     prettydoc_0.4.1     farver_2.1.0
## #  [52] generics_0.1.0      jsonlite_1.7.2      zoo_1.8-9
## #  [55] crosstalk_1.1.1     zip_2.2.0           car_3.0-11
## #  [58] magrittr_2.0.1      Formula_1.2-4       Matrix_1.3-3
## #  [61] Rcpp_1.0.6          munsell_0.5.0       fansi_0.5.0
## #  [64] reticulate_1.20     abind_1.4-5         gdtools_0.2.3
## #  [67] partykit_1.2-13     lifecycle_1.0.0     stringi_1.6.1
## #  [70] yaml_2.2.1          inum_1.0-4          carData_3.0-4
## #  [73] MASS_7.3-54         plyr_1.8.6          grid_4.1.0
## #  [76] hrbrthemes_0.8.0    forcats_0.5.1       crayon_1.4.1
## #  [79] lattice_0.20-44     haven_2.4.1         splines_4.1.0
## #  [82] hms_1.1.0           knitr_1.33          pillar_1.6.1
## #  [85] glue_1.4.2          evaluate_0.14       latticeExtra_0.6-29
## #  [88] broom.helpers_1.3.0 data.table_1.14.0   png_0.1-7
## #  [91] vctrs_0.3.8         Rttf2pt1_1.3.8      cellranger_1.1.0
## #  [94] tidyr_1.1.3         gtable_0.3.0        purrr_0.3.4
## #  [97] xfun_0.24           openxlsx_4.2.4      libcoin_1.0-8
## # [100] e1071_1.7-7         class_7.3-19        survival_3.2-11
## # [103] viridisLite_0.4.0   tibble_3.1.2        cluster_2.1.2
## # [106] corrplot_0.89       ellipsis_0.3.2
```

```
  # or use message() instead of cat()
```

Figure 3: Example: export a Plotly figure using ‘orca‘