Домашнее задание

# Задачи

## Задача 1

Загрузите данные из файла travel-times.csv и сохраните их в датафрейм travel.

*Ответ:*

setwd("C:/Users/alexandra\_c/Desktop/Уроки R/Практические задания/08 Работа с пропущенными значениями в R/08-7 Практическая работа")  
travel <- read.csv("travel-times.csv")

## Задача 2

Удалите из датафрейма travel столбец Comments.

*Ответ:*

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.6 v purrr 0.3.4   
## v tibble 3.1.8 v dplyr 1.0.10  
## v tidyr 1.2.1 v stringr 1.4.1   
## v readr 2.1.2 v forcats 0.5.2   
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

travel <- travel %>% select(-Comments)

## Задача 3

Замените в датафрейме travel все значения "" и "-" на NA.

**Подсказка.** Пример кода для замены значения 99 на NA:

dat[dat == 99] <- NA

*Ответ:*

travel[travel == '' | travel == '-' | travel == 'NA'] <- NA

## Задача 4

Посчитайте количество ячеек с пропущенными значениями в датафрейме travel. Посчитайте количество строк, содержащих пропущенные значения, в датафрейме travel. Сравните результаты. Можно ли считать, что каждая строка содержит ровно одно пропущенное значение? Приведите ответ на вопрос в файле с домашним заданием.

*Ответ:*

sum(is.na(travel)) # кол-во ячеек с пропущ.значениями

## [1] 19

sum(!complete.cases(travel)) # кол-во строк с пропущ.значениями

## [1] 19

rowSums(is.na(travel[!complete.cases(travel),])) # Да, одна срока содержит ровно одно пропущ.значение

## 1 2 3 4 5 6 7 8 35 36 37 38 39 40 41 42 203 204 205   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

## Задача 5

Используя библиотеку VIM, постройте две визуализации для выявления закономерностей в пропущенных значениях. Проинтерпретируйте полученные результаты: определите, в каких столбцах встречаются пропущенные значения, поясните, наблюдаются ли какие-то закономерности в пропущенных значениях. Предложите возможные причины появления пропусков в данных. Приведите ответы на вопросы в файле с домашним заданием.

*Ответ:*

library(VIM)

## Загрузка требуемого пакета: colorspace

## Загрузка требуемого пакета: grid

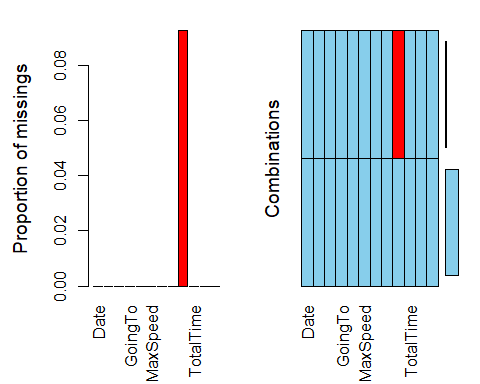
## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

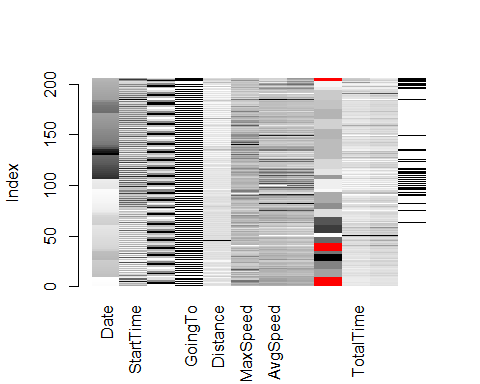
##   
## Присоединяю пакет: 'VIM'

## Следующий объект скрыт от 'package:datasets':  
##   
## sleep

aggr(travel)



matrixplot(travel)



a <- colSums(is.na(travel[!complete.cases(travel),]))  
a[a == max(a)]

## FuelEconomy   
## 19

# пропущен.значения находятся только в столбце FluelEconomy  
# честно сказать не наблюдаю какой-либо закономерности, кроме того что этих данных не хватает в период поездок в 2012 году.  
# Возможно в этот период времени перестали вести учет топлива

## Задача 6

Вычислите среднее и медианное значение по столбцам, в которых встречаются пропущенные значения. Если среднее значение сильно отличается от медианного, заполните пропуски медианным значением по столбцу, если нет — средним значением по столбцу.

*Ответ:*

str(travel)

## 'data.frame': 205 obs. of 12 variables:  
## $ Date : chr "1/6/2012" "1/6/2012" "1/4/2012" "1/4/2012" ...  
## $ StartTime : chr "16:37" "08:20" "16:17" "07:53" ...  
## $ DayOfWeek : chr "Friday" "Friday" "Wednesday" "Wednesday" ...  
## $ GoingTo : chr "Home" "GSK" "Home" "GSK" ...  
## $ Distance : num 51.3 51.6 51.3 49.2 51.1 ...  
## $ MaxSpeed : num 127 130 127 132 136 ...  
## $ AvgSpeed : num 78.3 81.8 82 74.2 83.4 84.5 82.9 77.5 80.9 70.6 ...  
## $ AvgMovingSpeed: num 84.8 88.9 85.8 82.9 88.1 88.8 87.3 85.9 88.3 78.1 ...  
## $ FuelEconomy : chr NA NA NA NA ...  
## $ TotalTime : num 39.3 37.9 37.5 39.8 36.8 36.8 37.2 37.9 39.3 43.5 ...  
## $ MovingTime : num 36.3 34.9 35.9 35.6 34.8 35 35.3 34.3 36 39.3 ...  
## $ Take407All : chr "No" "No" "No" "No" ...

# Не удалось найти среднее значение и медиану, потому что тип данных в столбце FuelEconomy был строковый.  
# Необходимо изменить тип данных на числовой с плавающей точкой  
travel$FuelEconomy <- as.numeric(travel$FuelEconomy)  
# теперь найдем среднее значение и медиану  
mean(travel$FuelEconomy, na.rm = TRUE)

## [1] 8.690591

median(travel$FuelEconomy, na.rm = TRUE)

## [1] 8.52

# Значения не сильно отличаются, тогда мы можем заполнить пропуски средним значением  
travel <- travel %>%mutate(FuelEconomy = ifelse(is.na(FuelEconomy),  
 mean(FuelEconomy, na.rm = TRUE),  
 FuelEconomy))

## Задача 7

Загрузите данные из файла melb\_data.csv и сохраните их в датафрейм melb.

*Ответ:*

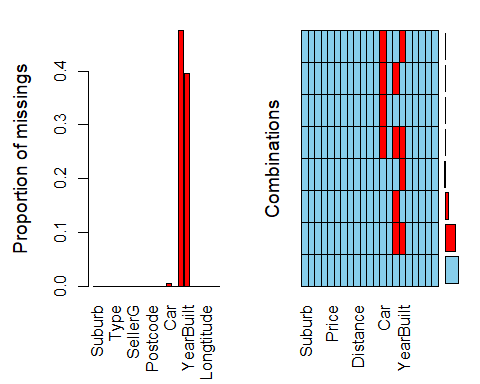
melb <- read.csv("melb\_data.csv")

## Задача 8

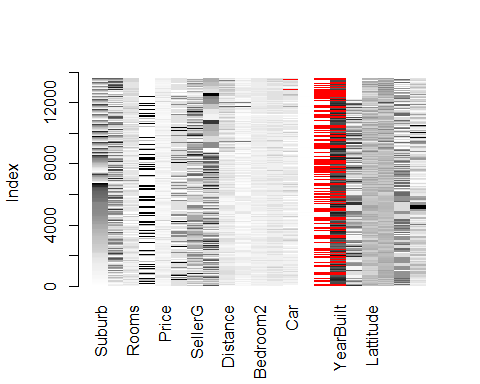
Используя библиотеку VIM, постройте две визуализации для выявления закономерностей в пропущенных значениях. Проинтерпретируйте полученные результаты: определите, в каких столбцах встречаются пропущенные значения, поясните, наблюдаются ли какие-то закономерности в пропущенных значениях. Предложите возможные причины появления пропусков в данных. Приведите ответы на вопросы в файле с домашним заданием.

*Ответ:*

aggr(melb)



matrixplot(melb)



# пропущен.значения находятся в столбцах Car, BuildingArea и YearBuild  
  
# В Car отсутствуют некоторые значения, в основном с типом недвижимости коттетдж,  
# возможно владельцы дома просто не предоставили инфомрацию о наличии парковочных мест.   
  
# Тоже применимо и для BuildingArea и YearBuild, возможно владельцы дома просто не предоставили инфомрацию.

## Задача 9

Заполните пропущенные значения в столбце Car с помощью метода, который берёт значение из ячейки над ячейкой с пропуском.

*Ответ:*

melb <- melb %>% fill(Car, .direction = "down")

### Задача 10

Заполните пропущенные значения в столбце BuildingArea средним значением по столбцу.

*Ответ:*

melb <- melb %>%mutate(BuildingArea = ifelse(is.na(BuildingArea),  
 mean(BuildingArea, na.rm = TRUE),  
 BuildingArea))

### Задача 11

Заполните пропущенные значения в столбце YearBuilt медианным значением по столбцу.

*Ответ:*

melb <- melb %>%mutate(YearBuilt = ifelse(is.na(YearBuilt),  
 median(YearBuilt, na.rm = TRUE),  
 YearBuilt))