

Task 5

Abhishek Murthy

2024-09-07

```
library(nycflights13)

## Warning: package 'nycflights13' was built under R version 4.3.3

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(tidyr)
library(mice)

## Warning: package 'mice' was built under R version 4.3.3

##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##   filter

## The following objects are masked from 'package:base':
##
##   cbind, rbind

library(VIM)

## Warning: package 'VIM' was built under R version 4.3.3

## Loading required package: colorspace

## Loading required package: grid
```

```

## VIM is ready to use.

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

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep

# Load the flights dataset
data("flights")

# View the structure of the dataset
str(flights)

## tibble [336,776 × 19] (S3: tbl_df/tbl/data.frame)
## $ year      : int [1:336776] 2013 2013 2013 2013 2013 2013 2013 2013 2013
2013 2013 ...
## $ month     : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
## $ day       : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
## $ dep_time  : int [1:336776] 517 533 542 544 554 554 555 557 557 558
...
## $ sched_dep_time: int [1:336776] 515 529 540 545 600 558 600 600 600 600
...
## $ dep_delay  : num [1:336776] 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
## $ arr_time   : int [1:336776] 830 850 923 1004 812 740 913 709 838 753
...
## $ sched_arr_time: int [1:336776] 819 830 850 1022 837 728 854 723 846 745
...
## $ arr_delay  : num [1:336776] 11 20 33 -18 -25 12 19 -14 -8 8 ...
## $ carrier    : chr [1:336776] "UA" "UA" "AA" "B6" ...
## $ flight     : int [1:336776] 1545 1714 1141 725 461 1696 507 5708 79
301 ...
## $ tailnum    : chr [1:336776] "N14228" "N24211" "N619AA" "N804JB" ...
## $ origin     : chr [1:336776] "EWR" "LGA" "JFK" "JFK" ...
## $ dest       : chr [1:336776] "IAH" "IAH" "MIA" "BQN" ...
## $ air_time   : num [1:336776] 227 227 160 183 116 150 158 53 140 138
...
## $ distance   : num [1:336776] 1400 1416 1089 1576 762 ...
## $ hour       : num [1:336776] 5 5 5 5 6 5 6 6 6 6 ...
## $ minute     : num [1:336776] 15 29 40 45 0 58 0 0 0 0 ...
## $ time_hour  : POSIXct[1:336776], format: "2013-01-01 05:00:00" "2013-
01-01 05:00:00" ...

# Summarize missing data
summary(flights)

```

```

##      year      month      day      dep_time
sched_dep_time
## Min.   :2013   Min.    : 1.000   Min.    : 1.00   Min.    : 1   Min.    :
106
## 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.:
906
## Median :2013   Median : 7.000   Median :16.00   Median :1401   Median
:1359
## Mean   :2013   Mean    : 6.549   Mean    :15.71   Mean    :1349   Mean
:1344
## 3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd
Qu.:1729
## Max.   :2013   Max.    :12.000   Max.    :31.00   Max.    :2400   Max.
:2359
##
##                                     NA's    :8255
##      dep_delay      arr_time      sched_arr_time      arr_delay
## Min.   : -43.00   Min.    : 1   Min.    : 1   Min.    : -86.000
## 1st Qu.: -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
## Median : -2.00   Median :1535   Median :1556   Median : -5.000
## Mean    : 12.64   Mean    :1502   Mean    :1536   Mean    : 6.895
## 3rd Qu.: 11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.: 14.000
## Max.    :1301.00   Max.    :2400   Max.    :2359   Max.    :1272.000
## NA's    :8255     NA's    :8713     NA's    :9430
##      carrier      flight      tailnum      origin
## Length:336776   Min.    : 1   Length:336776   Length:336776
## Class :character 1st Qu.: 553   Class :character Class :character
## Mode  :character Median :1496   Mode  :character Mode  :character
##                      Mean    :1972
##                      3rd Qu.:3465
##                      Max.    :8500
##
##      dest      air_time      distance      hour
## Length:336776   Min.    : 20.0   Min.    : 17   Min.    : 1.00
## Class :character 1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
## Mode  :character Median :129.0   Median : 872   Median :13.00
##                      Mean    :150.7   Mean    :1040   Mean    :13.18
##                      3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                      Max.    :695.0   Max.    :4983   Max.    :23.00
##                      NA's    :9430
##      minute      time_hour
## Min.    : 0.00   Min.    :2013-01-01 05:00:00.00
## 1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00.00
## Median :29.00   Median :2013-07-03 10:00:00.00
## Mean    :26.23   Mean    :2013-07-03 05:22:54.64
## 3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00.00
## Max.    :59.00   Max.    :2013-12-31 23:00:00.00
##

```

```

# Mean imputation for numeric columns
flights_imputed <- flights %>%

```

```

mutate(
  dep_time = ifelse(is.na(dep_time), mean(dep_time, na.rm = TRUE),
dep_time),
  arr_time = ifelse(is.na(arr_time), mean(arr_time, na.rm = TRUE),
arr_time)
)

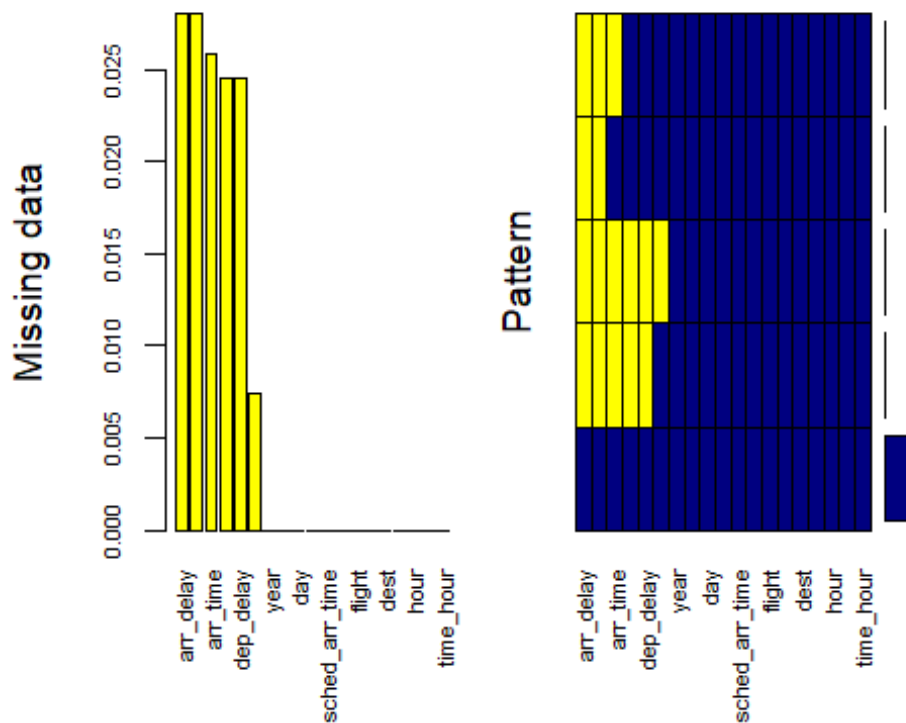
# Median imputation (another global method)
flights_imputed <- flights %>%
  mutate(
    dep_delay = ifelse(is.na(dep_delay), median(dep_delay, na.rm = TRUE),
dep_delay),
    arr_delay = ifelse(is.na(arr_delay), median(arr_delay, na.rm = TRUE),
arr_delay)
  )

# Impute missing values by grouping by the carrier (class-based method)
flights_imputed <- flights %>%
  group_by(carrier) %>%
  mutate(
    dep_time = ifelse(is.na(dep_time), mean(dep_time, na.rm = TRUE),
dep_time),
    arr_time = ifelse(is.na(arr_time), mean(arr_time, na.rm = TRUE),
arr_time)
  ) %>%
  ungroup()

# Visualize missing data
aggr(flights, col = c('navyblue', 'yellow'), numbers = TRUE, sortVars = TRUE,
labels = names(flights), cex.axis = .7, gap = 3, ylab = c("Missing data",
"Pattern"))

## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies

```



```
##
## Variables sorted by number of missings:
##   Variable      Count
##   arr_delay 0.028000808
##   air_time  0.028000808
##   arr_time  0.025871796
##   dep_time  0.024511842
##   dep_delay 0.024511842
##   tailnum   0.007458964
##   year      0.000000000
##   month     0.000000000
##   day       0.000000000
##   sched_dep_time 0.000000000
##   sched_arr_time 0.000000000
##   carrier      0.000000000
##   flight       0.000000000
##   origin       0.000000000
##   dest         0.000000000
##   distance     0.000000000
##   hour         0.000000000
##   minute       0.000000000
##   time_hour    0.000000000

# Compare distributions before and after imputation
par(mfrow = c(1, 2))
hist(flights$dep_time, main = "Before Imputation", col = "red")
```

Before Imputation

