DA 2

Abhishek Murthy

21BDS0064

2024-09-12

```
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)
```

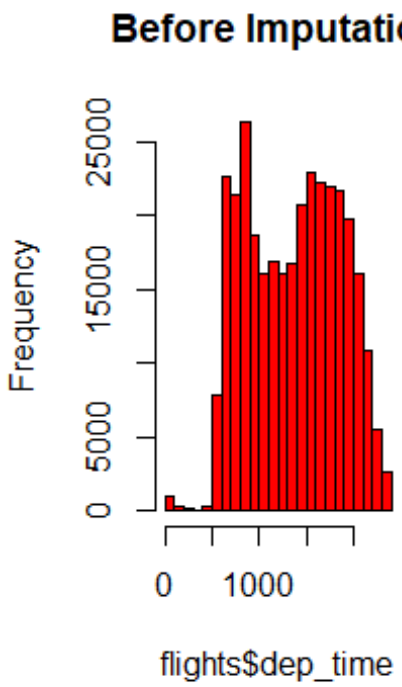
```
## Warning: package 'VIM' was built under R version 4.3.3
```

```
## Loading required package: colorspace
```

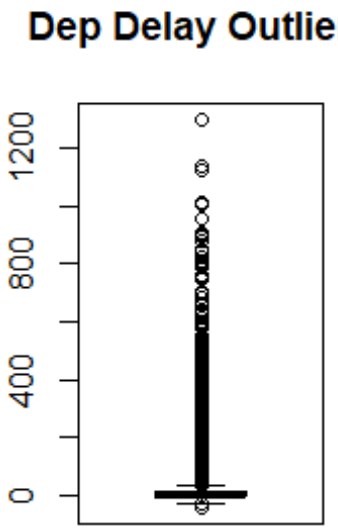
```
## Loading required package: grid
```

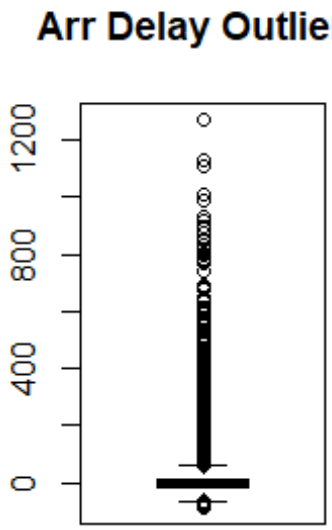
```
## VIM is ready to use.

## 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")

## OUTLIER DETECTION
detect_outliers_zscore <- function(x, threshold = 3) {
  z_scores <- (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
  return(which(abs(z_scores) > threshold))
}

dep_delay_clean <- flights$dep_delay[!is.na(flights$dep_delay)]
arr_delay_clean <- flights$arr_delay[!is.na(flights$arr_delay)]

dep_delay_outliers_z <- detect_outliers_zscore(dep_delay_clean)
arr_delay_outliers_z <- detect_outliers_zscore(arr_delay_clean)

cat("Number of outliers in dep_delay (Z-score):",
length(dep_delay_outliers_z), "\n")

## Number of outliers in dep_delay (Z-score): 7928

cat("Number of outliers in arr_delay (Z-score):",
length(arr_delay_outliers_z), "\n")

## Number of outliers in arr_delay (Z-score): 7285

detect_outliers_iqr <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  return(which(x < lower_bound | x > upper_bound))
}

dep_delay_outliers_iqr <- detect_outliers_iqr(dep_delay_clean)
arr_delay_outliers_iqr <- detect_outliers_iqr(arr_delay_clean)

cat("Number of outliers in dep_delay (IQR):", length(dep_delay_outliers_iqr),
"\n")
```

```
## Number of outliers in dep_delay (IQR): 43216

cat("Number of outliers in arr_delay (IQR):", length(arr_delay_outliers_iqr),
"\n")

## Number of outliers in arr_delay (IQR): 27880

# Boxplot to visualize outliers
par(mfrow = c(1, 2))
```




```
boxplot(dep_delay_clean, main = "Dep Delay Outliers", col = "lightblue")
boxplot(arr_delay_clean, main = "Arr Delay Outliers", col = "lightgreen")
```


Dep Delay Outliers
0
400
800
1200


Arr Delay Outliers
0
400
800
1200