Week 5

Heleine Fouda

2023-10-02

The data used in this assignment are from Numbersense, Kaiser Fung, McGraw Hill, 2013

Loading the data

```
flights_data<- data.frame(
  Cities= c("Los Angeles", "Phoenix", "San Diego", "San Francisco", "Seattle"),
  Alaska_on_time = c(497, 221,  212, 503, 1841),
  Alaska_delayed= c(62, 12, 20, 102, 305),
  Am_west_on_time = c(694, 4840, 383, 320, 201),
  Am_west_delayed = c(117,415, 65, 129, 61)
)
data
```

```
## function (..., list = character(), package = NULL, lib.loc = NULL,
##     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
## {
##     fileExt <- function(x) {
##         db <- grepl("\\.[^.]+\\.(gz|bz2|xz)$", x)
##         ans <- sub(".*\\.", "", x)
##         ans[db] <- sub(".*\\.([^.]+\\.)(gz|bz2|xz)$", "\\1\\2",
##             x[db])
##         ans
##     }
##     my_read_table <- function(...) {
##         lcc <- Sys.getlocale("LC_COLLATE")
##         on.exit(Sys.setlocale("LC_COLLATE", lcc))
##         Sys.setlocale("LC_COLLATE", "C")
##         read.table(...)
##     }
##     stopifnot(is.character(list))
##     names <- c(as.character(substitute(list(...))[-1L]), list)
##     if (!is.null(package)) {
##         if (!is.character(package))
##             stop("'package' must be a character vector or NULL")
##     }
##     paths <- find.package(package, lib.loc, verbose = verbose)
##     if (is.null(lib.loc))
##         paths <- c(path.package(package, TRUE), if (!length(package)) getwd(),
##             paths)
##     paths <- unique(normalizePath(paths[file.exists(paths)]))
##     paths <- paths[dir.exists(file.path(paths, "data"))]
##     dataExts <- tools:::.make_file_exts("data")
##     if (length(names) == 0L) {
##         db <- matrix(character(), nrow = 0L, ncol = 4L)
```

```
##         for (path in paths) {
##             entries <- NULL
##             packageName <- if (file_test("-f", file.path(path,
##                 "DESCRIPTION")))
##                 basename(path)
##             else "."
##             if (file_test("-f", INDEX <- file.path(path, "Meta",
##                 "data.rds"))) {
##                 entries <- readRDS(INDEX)
##             }
##             else {
##                 dataDir <- file.path(path, "data")
##                 entries <- tools::list_files_with_type(dataDir,
##                   "data")
##                 if (length(entries)) {
##                   entries <- unique(tools::file_path_sans_ext(basename(entries)))
##                   entries <- cbind(entries, "")
##                 }
##             }
##             if (NROW(entries)) {
##                 if (is.matrix(entries) && ncol(entries) == 2L)
##                   db <- rbind(db, cbind(packageName, dirname(path),
##                     entries))
##                 else warning(gettextf("data index for package %s is invalid and will be ignored",
##                   sQuote(packageName)), domain = NA, call. = FALSE)
##             }
##         }
##         colnames(db) <- c("Package", "LibPath", "Item", "Title")
##         footer <- if (missing(package))
##             paste0("Use ", sQuote(paste("data(package =", ".packages(all.available = TRUE))")),
##                 "\n", "to list the data sets in all *available* packages.")
##         else NULL
##         y <- list(title = "Data sets", header = NULL, results = db,
##             footer = footer)
##         class(y) <- "packageIQR"
##         return(y)
##     }
##     paths <- file.path(paths, "data")
##     for (name in names) {
##         found <- FALSE
##         for (p in paths) {
##             tmp_env <- if (overwrite)
##                 envir
##             else new.env()
##             if (file_test("-f", file.path(p, "Rdata.rds"))) {
##                 rds <- readRDS(file.path(p, "Rdata.rds"))
##                 if (name %in% names(rds)) {
##                   found <- TRUE
##                   if (verbose)
##                     message(sprintf("name=%s:\t found in Rdata.rds",
##                       name), domain = NA)
##                   thispkg <- sub(".*/([^/]*)/data$", "\\1", p)
##                   thispkg <- sub("_.*$", "", thispkg)
##                   thispkg <- paste0("package:", thispkg)
```

```
##                   objs <- rds[[name]]
##                   lazyLoad(file.path(p, "Rdata"), envir = tmp_env,
##                     filter = function(x) x %in% objs)
##                   break
##                 }
##                 else if (verbose)
##                   message(sprintf("name=%s:\t NOT found in names() of Rdata.rds, i.e.,\n\t%s\n",
##                     name, paste(names(rds), collapse = ",")),
##                     domain = NA)
##             }
##             if (file_test("-f", file.path(p, "Rdata.zip"))) {
##                 warning("zipped data found for package ", sQuote(basename(dirname(p))),
##                   ".\nThat is defunct, so please re-install the package.",
##                   domain = NA)
##                 if (file_test("-f", fp <- file.path(p, "filelist")))
##                   files <- file.path(p, scan(fp, what = "", quiet = TRUE))
##                 else {
##                   warning(gettextf("file 'filelist' is missing for directory %s",
##                     sQuote(p)), domain = NA)
##                   next
##                 }
##             }
##             else {
##                 files <- list.files(p, full.names = TRUE)
##             }
##             files <- files[grep(name, files, fixed = TRUE)]
##             if (length(files) > 1L) {
##                 o <- match(fileExt(files), dataExts, nomatch = 100L)
##                 paths0 <- dirname(files)
##                 paths0 <- factor(paths0, levels = unique(paths0))
##                 files <- files[order(paths0, o)]
##             }
##             if (length(files)) {
##                 for (file in files) {
##                   if (verbose)
##                     message("name=", name, ":\t file= ...", .Platform$file.sep,
##                       basename(file), "::\t", appendLF = FALSE,
##                       domain = NA)
##                   ext <- fileExt(file)
##                   if (basename(file) != paste0(name, ".", ext))
##                     found <- FALSE
##                   else {
##                     found <- TRUE
##                     zfile <- file
##                     zipname <- file.path(dirname(file), "Rdata.zip")
##                     if (file.exists(zipname)) {
##                       Rdatadir <- tempfile("Rdata")
##                       dir.create(Rdatadir, showWarnings = FALSE)
##                       topic <- basename(file)
##                       rc <- .External(C_unzip, zipname, topic,
##                         Rdatadir, FALSE, TRUE, FALSE, FALSE)
##                       if (rc == 0L)
##                         zfile <- file.path(Rdatadir, topic)
##                     }
```

```
##                     if (zfile != file)
##                       on.exit(unlink(zfile))
##                     switch(ext, R = , r = {
##                       library("utils")
##                       sys.source(zfile, chdir = TRUE, envir = tmp_env)
##                     }, RData = , rdata = , rda = load(zfile,
##                       envir = tmp_env), TXT = , txt = , tab = ,
##                       tab.gz = , tab.bz2 = , tab.xz = , txt.gz = ,
##                       txt.bz2 = , txt.xz = assign(name, my_read_table(zfile,
##                         header = TRUE, as.is = FALSE), envir = tmp_env),
##                       CSV = , csv = , csv.gz = , csv.bz2 = ,
##                       csv.xz = assign(name, my_read_table(zfile,
##                         header = TRUE, sep = ";", as.is = FALSE),
##                         envir = tmp_env), found <- FALSE)
##                   }
##                   if (found)
##                     break
##                 }
##                 if (verbose)
##                   message(if (!found)
##                     "*NOT* ", "found", domain = NA)
##             }
##             if (found)
##                 break
##         }
##         if (!found) {
##             warning(gettextf("data set %s not found", sQuote(name)),
##                 domain = NA)
##         }
##         else if (!overwrite) {
##             for (o in ls(envir = tmp_env, all.names = TRUE)) {
##                 if (exists(o, envir = envir, inherits = FALSE))
##                   warning(gettextf("an object named %s already exists and will not be overwritten",
##                     sQuote(o)))
##                 else assign(o, get(o, envir = tmp_env, inherits = FALSE),
##                   envir = envir)
##             }
##             rm(tmp_env)
##         }
##     }
##     invisible(names)
## }
## <bytecode: 0x7fdd19bded08>
## <environment: namespace:utils>
```

```
file_path <- "sample_data.csv"
write.csv(flights_data, file = file_path, row.names = FALSE)
file.exists(file_path)
```

```
## [1] TRUE
```

```
glimpse(flights_data)
```

```
## Rows: 5
## Columns: 5
```

```
## $ Cities          <chr> "Los Angeles", "Phoenix", "San Diego", "San Francisco"~
## $ Alaska_on_time  <dbl> 497, 221, 212, 503, 1841
## $ Alaska_delayed  <dbl> 62, 12, 20, 102, 305
## $ Am_west_on_time <dbl> 694, 4840, 383, 320, 201
## $ Am_west_delayed <dbl> 117, 415, 65, 129, 61
```

Let's add two new columns to the data set" arr_delays (delayed flights) and ontime_arr (flights that arrived on time):

```
flights_data |>
  mutate(
    arr_delays = Alaska_delayed + Am_west_delayed,

    ontime_arr = Alaska_on_time + Am_west_on_time)
```

```
##          Cities Alaska_on_time Alaska_delayed Am_west_on_time Am_west_delayed
## 1   Los Angeles            497             62             694             117
## 2       Phoenix            221             12            4840             415
## 3     San Diego            212             20             383              65
## 4 San Francisco            503            102             320             129
## 5       Seattle           1841            305             201              61
##   arr_delays ontime_arr
## 1        179       1191
## 2        427       5061
## 3         85        595
## 4        231        823
## 5        366       2042
```

Let's also push the new variables or columns to the front of the data frame using the .before = 1

```
flights_data |>
  mutate(
    arr_delays = Alaska_delayed + Am_west_delayed,

    ontime_arr = Alaska_on_time + Am_west_on_time,
    .before = 1)
```

```
##   arr_delays ontime_arr        Cities Alaska_on_time Alaska_delayed
## 1        179       1191   Los Angeles            497             62
## 2        427       5061       Phoenix            221             12
## 3         85        595     San Diego            212             20
## 4        231        823 San Francisco            503            102
## 5        366       2042       Seattle           1841            305
##   Am_west_on_time Am_west_delayed
## 1             694             117
## 2            4840             415
## 3             383              65
## 4             320             129
## 5             201              61
```

Let's push the new columns to the front of the data frame using the function relocate().

```
flights_data |>
  mutate(
    arr_delays = Alaska_delayed + Am_west_delayed,

    ontime_arr = Alaska_on_time + Am_west_on_time) |>
```

```
  relocate(arr_delays,  ontime_arr)
```

```
##   arr_delays ontime_arr        Cities Alaska_on_time Alaska_delayed
## 1        179       1191   Los Angeles            497             62
## 2        427       5061       Phoenix            221             12
## 3         85        595     San Diego            212             20
## 4        231        823 San Francisco            503            102
## 5        366       2042       Seattle           1841            305
##   Am_west_on_time Am_west_delayed
## 1             694             117
## 2            4840             415
## 3             383              65
## 4             320             129
## 5             201              61
```

```
flights_data2 <-flights_data |>
  mutate(
    arr_delays = Alaska_delayed + Am_west_delayed,

    ontime_arr = Alaska_on_time + Am_west_on_time) |>
  relocate(arr_delays,  ontime_arr)
```

Let's now compare the arrival delays for the two airlines: Alaska delays

```
flights_data2 |>
  group_by(Alaska_delayed) |>
  summarize(Avg_delay = mean(Alaska_delayed, na.rm = TRUE))
```

```
## # A tibble: 5 x 2
##   Alaska_delayed Avg_delay
##            <dbl>     <dbl>
## 1             12        12
## 2             20        20
## 3             62        62
## 4            102       102
## 5            305       305
```

Am_west_delays

```
flights_data2 |>
  group_by(Am_west_delayed) |>
  summarize(
    Avg_delay = mean(Am_west_delayed, na.rm = TRUE))
```

```
## # A tibble: 5 x 2
##   Am_west_delayed Avg_delay
##             <dbl>     <dbl>
## 1              61        61
## 2              65        65
## 3             117       117
## 4             129       129
## 5             415       415
```

Average delay for both airlines:

```
flights_data2 |>
  group_by(arr_delays) |>
```

```
  summarize(avg_delay = mean(arr_delays))
```

```
## # A tibble: 5 x 2
##   arr_delays avg_delay
##        <dbl>     <dbl>
## 1         85        85
## 2        179       179
## 3        231       231
## 4        366       366
## 5        427       427
```

Let's close with summary statistics for both airlines

```
flights_data2|>
  summary(flights_data2)
```

```
##    arr_delays      ontime_arr      Cities          Alaska_on_time
##  Min.   : 85.0   Min.   : 595   Length:5           Min.   : 212.0
##  1st Qu.:179.0   1st Qu.: 823   Class :character   1st Qu.: 221.0
##  Median :231.0   Median :1191   Mode  :character   Median : 497.0
##  Mean   :257.6   Mean   :1942                      Mean   : 654.8
##  3rd Qu.:366.0   3rd Qu.:2042                      3rd Qu.: 503.0
##  Max.   :427.0   Max.   :5061                      Max.   :1841.0
##  Alaska_delayed  Am_west_on_time Am_west_delayed
##  Min.   : 12.0   Min.   : 201    Min.   : 61.0
##  1st Qu.: 20.0   1st Qu.: 320    1st Qu.: 65.0
##  Median : 62.0   Median : 383    Median :117.0
##  Mean   :100.2   Mean   :1288    Mean   :157.4
##  3rd Qu.:102.0   3rd Qu.: 694    3rd Qu.:129.0
##  Max.   :305.0   Max.   :4840    Max.   :415.0
```

‘‘‘