Datasets and Data Frames with R

Bernardo Vimpi

9/13/2020

This is the first project from one of my analytics for Data Science class using R programming. Here we do some basic analysis of the nycflights13 dataset and also build a data frame.

Loading packages

```
library(tinytex)
library(nycflights13)
View(flights)
class(flights)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
#glimpse(flights)
```

(LC1.3) What does any ONE row in this flights dataset refer to?

- A. Data on an airline
- B. Data on a flight
- C. Data on an airport
- D. Data on multiple flights

Answer: Any of the ONE row in the flights dataset refers to the Data on a given flight. In other words, it refers to the answer in line B: Data on a flight.This is because each row refers to an observation of a given flight which also has the data on the same flight.

(LC1.4) What are some other examples in this dataset of categorical variables? What makes them different than quantitative variables?

Answer: In this dataset, there are four different categorical variables which are: carrier, tailnum, origin, and dest. The other variables are quantitative variables such as dep_time, sched_dep_time, arr_time, dist, hour, minute and others. In this dataset the factor that differentiates the categorical variables from the quantitative variable is the computer coding for the quantitative/numerical variables. For example, when we run the function glimpse (flights) we notice that the categorical variables are either or which are "integer", "double" which are coding terminology used for quantitative/numerical variables. On the other had, all categorical variables are of "chr" for character.

```
View(airports)
#glimpse(airports)
class(airports)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
library(nycflights13)
?airports
```

```
## starting httpd help server ... done
```

(LC1.5) What properties of each airport do the variables lat, lon, alt, tz, dst, and tzone describe in the airports data frame? Take your best guess.

Answer: This was a good tricky question, for me at least, but I took the best guess possible and I did the following:

I searched the airport dataset using "Help in R" and the ?airport to uncover the properties of each variables. Here are the results:

- A. lat and lon are the Latitude and Longitude, respectively. They correspond to the location of the airport.
- B. alt variable is the Altitude measured in feet.
- C. tz refers to the Timezone offset from GMT.
- D. dst is the Daylight savings time zone. This variable has different values: A is for "Standard US DST which starts on the second Sunday of March, ends on the first Sunday of November". U is for "Unknown" while N is for no dst.

(LC1.6) Provide the names of variables in a data frame with at least three variables in which one of them is an identification variable and the other two are not. In other words, create your own tidy data frame that matches these conditions.

Here I create a data frame of my top 5 favorite African soccer players in the last 10 years. I assign a unique player ID to each player, their first and last names, total career goals, and country or national origin. Please not this is not a real data. Its simply for practice.

```
player_id <- c( 001, 002,003, 004, 005)
first_name <- c("Samuel", "Didier", "Sadio", "Fabrice", "Mohamed")
last_name <- c( "Eto'o", "Drogba", "Mane", "Akwa", "Salah")
total_goals <- c( 201, 198, 203, 250, 259)
country <-c ("Cameroon", "Ivory Coast", "Senegal", "Angola", "Egypt")

players_df <- data.frame(Player_ID = player_id, First_Name= first_name, Last_Name=last_name, Goals= tota
players_df
```

```
##   Player_ID First_Name Last_Name Goals     Country
## 1         1     Samuel     Eto'o   201    Cameroon
## 2         2     Didier    Drogba   198 Ivory Coast
## 3         3      Sadio      Mane   203     Senegal
## 4         4    Fabrice      Akwa   250      Angola
## 5         5    Mohamed     Salah   259       Egypt
```

Thank you

Bernardo Vimpi