How to get started with R

Ursulina Kölbener

2023-12-17

Exercise 1

LC1.1

Repeat the earlier installation steps, but for the dplyr, nycflights13, and knitr packages. This will install the earlier mentioned dplyr package for data wrangling, the nycflights13 package containing data on all domestic flights leaving a NYC airport in 2013, and the knitr package for generating easy-to-read tables in R. We'll use these packages in the next section.

```
# install.packages("ggplot2")
# install.packages("dplyr")
# install.packages("nycflights13")
# install.packages("knitr")
```

LC1.2

"Load" the dplyr, nycflights13, and knitr packages as well by repeating the earlier steps.

```
library(dplyr)
library(nycflights13)
library(knitr)
```

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

```
View(flights)
```

Answer: B. Data on a flight

LC1.4

What are some other examples in this dataset of categorical variables? What makes them different than quantitative variables?

```
glimpse(flights)
```

```
## Rows: 336,776
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2~
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, ~
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, ~
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1~
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,~
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,~
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1~
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "~
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4~
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394~
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",~
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",~
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1~
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, ~
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6~
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0~
```

categorical variables: `carrier`, `tailnum`, `origin`, `dest`.

Categorical variables represent groups or labels, while quantitative variables represent measurable quantities with numbers. Categorical variables categorize data into groups, while quantitative variables are about numerical values or measurements.

Categorical variables represent categories or groups. They have finite, distinct values that classify data into groups or labels. These variables can be nominal, where the categories don't have an inherent order (like colors or types of fruits), or ordinal, where there's a specific order among the categories (like education level or socio-economic status).

On the other hand, quantitative variables represent measurable quantities and can be expressed numerically. These variables can be discrete, taking on specific integer values (like number of siblings), or continuous, representing a range of values within a certain interval (like height or weight).

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.

var	guess
lat	latitude
lon	longitude
alt	altitude
tz	timezone
dst	distance
tzone	timezone?

LC1.6

Provide the names of variables in a data frame with at least three variables where one of them is an identification variable and the other two are not. Further, create your own tidy data frame that matches these conditions.

```
View(airports)
df_airports <- data.frame(
  faa = airports$faa,
  alt = airports$alt,
  tz = airports$tz)

head(df_airports)
```

```
##   faa  alt tz
## 1 04G 1044 -5
## 2 06A  264 -6
## 3 06C  801 -6
## 4 06N  523 -5
## 5 09J   11 -5
## 6 0A9 1593 -5
```

LC1.7

Look at the help file for the airports data frame. Revise your earlier guesses about what the variables lat, lon, alt, tz, dst, and tzone each describe.

```
?airports
```

var	description
lat	latitude, Location of airport.
lon	longitude, Location of airport.
alt	Altitude, in feet.
tz	Timezone offset from GMT.
dst	Daylight savings time zone. A = Standard US DST: starts on the second Sunday of March, ends on the first Sunday of November. U = unknown. N = no dst.
tzone	IANA time zone, as determined by GeoNames webservice.

Exercise 2

This exercise is not (yet) about R and coding, but about everything we have learnt so far in this course. Think of a research topic you are interested in. Write it down. You might take up what you did in the previous take home exercise on hypotheses, if that's what interests you, or think about a possible seminar paper.

Topic: Measuring E-Government Maturity

Now here's the question:
What kind of data would you need? Write some sentences about it and describe why you think that would be useful data.

Service Accessibility Metrics: Data on the availability of government services online, including the number of services offered, their accessibility (24/7 availability), and the percentage of the population that can access these services. This data is crucial as it reflects the level of convenience and inclusivity offered by the e-government platform.

User Adoption Rates: Metrics on the number of users engaging with e-government services, their demographics, and frequency of use. Understanding user behavior helps in evaluating the acceptance and usability of digital services among the population.

Technology Infrastructure Data: Information on the technological infrastructure supporting e-government initiatives, such as internet penetration rates, broadband speed, and connectivity in urban and rural areas. This data is fundamental as it influences the reach and effectiveness of digital government services.

Transaction Volumes: Quantitative data on the volume and types of transactions conducted through e-government portals. This includes the number of online forms submitted, payments made, and inquiries handled digitally. It offers insights into the efficiency and effectiveness of the digital platforms.

Security and Privacy Metrics: Quantitative data on cybersecurity incidents, data breaches, and the response time to mitigate such incidents. Evaluating security measures helps gauge the trustworthiness and reliability of e-government systems.

Financial Data: Budget allocation and spending on e-government initiatives. Tracking financial data provides an overview of investment in digital infrastructure and helps in assessing the correlation between investment and service improvements.