Data analysis with R

Monika Avila Márquez, Ph.D.

February 28, 2024

Agenda

1. What is Data Analysis?
2. R packages for Data Analysis
3. Importing and Exporting data with R
4. Pre-analysis of the data
5. Data pre-processing
6. Exploratory data analysis
7. Model Development
8. Model Evaluation

What is Data Analysis?

Data analysis

1. Data analysis is not a new domain.
2. “Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.” John Tukey, Annals of Mathematical Statistics 1961

Data Analysis: the process

Data analysis process:

1. Define the problem.
2. Prepare and clean the data.
3. Conduct exploratory data analysis.
4. Build and evaluate the model.

The problem

The problem is a question we want to answer.

1. Is there data that allows us to answer our question?
2. What information do we need to answer our question?

The problem: an example

Do working mothers have children with higher birth weights in Uruguay?

1. Is there data on the birth weight of women's children? Is there data on pregnant women who work and do not work?
2. Are there differences between working and non-working women and their children's outcomes?
    2.1 Education, income, total number of prenatal visits, birthweight, etc.

Data source: *Amarante, V., Manacorda, M., Miguel, E., Vigorito, A. (2016). Do cash transfers improve birth outcomes? Evidence from matched vital statistics, and program and social security data. American Economic Journal: Economic Policy, 8(2), 1-43.*

Understanding the data set

1. Format of the data set (.csv, .dta, .xlsx).
2. Each row represents an observation.
3. Each column represents a feature, variable, or attribute.

R packages for Data Analysis

R packages used for Data Analysis

1. Data import and Management: readr of tidyverse
2. Data wrangling and transformation: dplyr, tidyr of tidyverse, mice, naniar.
3. Data visualization and exploration: dplyr of tidyverse, ggplot2.
4. Data modeling: command lm for linear regression.

Pipe operator %>%

Pipe operator is used to combine functions.

Importing and Exporting data with R

Importing data

1. Check the format of the data
2. File path of dataset
    2.1 Locally
    2.2 Hosted online (need of url)

Readr package

read_csv(file)

? read_csv

Process if dataset is on URL

1. Get url where the dataset is contained url<-
2. Download dataset locally download.file()
3. Untar the file before using read_csv (Not always needed) untar()
4. Read the dataset downloaded locally read_csv()
5. Print the data head() tail()

Export data

write_csv()

Pre-analysis of the data

Pre-analysis of the data

Main goals:

1. Understand data before analysis
2. Check variable data types
3. Check data distribution
4. Identify issues with the data

Data types

1. Use function glimpse(database)
2. Important to check if the data is of the correct type. Use function typeof(database$variablename)

Data distribution

1. Use summary() to get the statistical summary of the data.
2. Use group_by() to summarize data by groups defined in our dataset in our variable.

Some commands in Dplyr package

1. select(): select variables based on their names
2. filter(): filter observations based on values
3. summarize(): compute summary statistics
4. arrange(): reorder the rows
5. mutate(): create new variables

Data pre-processing

Data pre-processing

Also called data wrangling, or data cleaning.

Maps data from raw format to another format for analysis.

Main goals:

1. Dealing with missing values.
2. Data formatting.
3. Data normalization
4. Binning.
5. Turning categorical values to numerical variables.

Missing values

Missing values are represented by NA in R.

Dealing with missing values:

1. Check with data collection source.
2. Drop missing values:
    2.1 Drop the variable.
    2.2 Drop the data entry.
3. Replace the missing values:
    3.1 An average.
    3.2 Values of similar data points.

Treating missing values with R

1. is.na(): allows to know if it is a missing value.
2. Dropping missing values:
    2.1 Drop columns using: data[, c[n1.n2]]
    2.2 Drop missing values (drops rows): Use drop_na().
    2.3 Replace values: replace_na() or use mutate function.

Data formatting

Data formatting means bringing data to a common standard of expression which allows to make comparisons.

Goal: coherent data.

Correcting data types

1. Identify data types: sapply(data, typeof)
2. Convert data types: mutate_all(type.convert) %>% mutate_if(is.character,as.numeric)

Data normalization

Normalization allows to bring all data to the same range.

1. Comparison between futures.
2. Makes statistical analysis easier.
3. Lower computational burden.

Change categorical variables into Numeric Variables

A categorical variable takes one out of limited possible values, and it assigns each observation to a group. Thus, the values of the variable are coded as strings. Problem: models cannot accept strings as input. Solution:

1. Assign a dummy variable for each category that takes value 1 if the value is equal to the corresponding category and 0 otherwise.

2. Use dataframe_name %>% mutate(dummy=1) %>% spread(key=reporting$_a irline, value = dummy, fill = 0)% > %slice(1 : 5)

Exploratory data analysis

Descriptive statistical analysis

1. Continuous and discrete data.
    1.1 summarize()
    1.2 group_by(): before summarizing to obtain summary statistics by group.
2. Categorical data.
    2.1 count()

Graphical analysis

1. Continuous and discrete data.
    1.1 Boxplots: allows to see the shape of the distribution of a variable.
    1.2 Scatter plots: allows to see the relationship between two variables.
2. Categorical data.
    2.1 count()

Grouping data

1. Use $group_by()$:
   Categoriesofoneormorevariables.*Sortdataindescendingorder*.*Useaheatmapwithggplot*.

Model Development

Types of models

Linear regression model can be used to model:

- Linear relationships
- Nonlinear relationships
    1. Polynomial regression.

Model Evaluation

Frame Title

1. Prediction
2. Assessing the model: look at the behaviour of the residuals
3. Model evaluation
4. Overfitting and underfitting