Juan Manuel Moreno

# Pill Pandas

**FECHA**

## Introduction

In this pill you're going to solve some questions about a dataframe, for these purposes you'll need to make transformations in a pandas dataframe about transport links

## What are the main objectives in this project?

- Know how to create pandas dataframes
- Filter dataframe values
- How to manage NaN values and deal with bad data
- How to apply statistical functions and functions to resume the information of the data like group by
- Export results into csv files

## 1. General analysis

A database (in .csv - airports.csv ) from the web https://openflights.org/data.html will be used, the dataset contains the following information:

- **AirportID**: Identifier of each flight for an airport.
- **Name**: Name of the airport.
- **City**: City where the airport is located.
- **Country**:: Country or territory in which the airport is located.
- **IATA**: International Air Transport Association code, airport code.
- **ICAO**: International civil organization code, airport code.
- **Latitude**: Coordinate of the airport (latitude).
- **Longitude**: Airport coordinate (longitude).
- **Altitude**: Altitude of the airport (in feet).
- **Timezone**: Time zone.
- **DST**: Code referring to the continent (Daylight savings time). Europe (E), A (US/CANADA), S (South America), O (Australia), Z (New Zeeland), N (None), U (Unknown).
- **Tz**: Airport time zone. For example: (America/Los_Angeles).
- **Type**: Type of airport: airport, station, port, unknown.

- **Source**: Data source.

With the dataset information, do the following:

1. Loading of the dataset as a dataframe.
2. Shows the first 10 rows of the dataframe.
3. Get a statistical summary.
4. For this analysis we are not going to use the 'AirportID', 'Latitude', 'Longitude' and 'Altitude' columns, remove them from the dataframe.
5. Get a statistical summary again, how has the data changed?
6. On the statistical summary above it seems that in column TZ there is a rare value \N, check the proportion of them with value_counts.
7. Reload the dataset so that null values are correctly interpreted (repeat section 4, delete columns).
8. Checks the entire dataframe for null values.
9. Overwrites the null values of the IATA and ICAO columns with the value 'UNKNOWN'
10. Changes the type of the DST and TZ variables to categorical.
11. Obtain a statistical summary of the categorical variables.
12. Groups the dataframe by airport type, showing the type count.
13. Select the name of the cities whose airport type is "port"
14. Shows all the rows of the fields name of the airport, name of the country and, name of the city, whose country is Spain.
15. Shows the name of the country and the airport belonging to the city of Madrid and Barcelona. Are all the records from Spain?
16. Save the previous results in a csv called Madrid_Barcelona.csv

## 2. Project organization

You can use markdown or text comments to document the exercise.

## 3. Requirements

- You can do the exercise in whatever software that suppor notebook like VS code, jupyter notebook, jupyter lab, Databricks, google Colab, etc.

## 4. Development

All the development should be made by pandas functions in Python (also you can use more python libraries if as your convenience)

## 5. Deliverables

- A jupyter notebook file with the airports analysis

## 6. Resources

- Ways to filter pandas dataframe https://towardsdatascience.com/8-ways-to-filter-pandas-dataframes-d34ba585c1b8
- Pandas Tutorial https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html