Bocconi

Dataset Description

The dataset is sourced from the NYC Taxi and Limousine Commission. Each observation corresponds to a ride taken in a yellow taxi cab. Although data for billions of rides across several years are available, we will focus on a subset of data from May 2015. The **raw data** has been partially **cleaned and merged** with spatial data from the NYC Department of City Planning.

The data has been **filtered** to include only transactions paid for with a credit card. Variables include the hour and day of the week the cab was picked up, neighborhood codes for the neighborhood tabulation area where the cab was picked up, and the neighborhood where the passenger was dropped off.

The data can be **downloaded here** (`taxi_tidy.csv`).

Columns

A data dictionary of the raw dataset is available here. Below, you can find a description of the actual columns. The raw data and further documentation can be found at this link.

- `pickup_hour`: Hour of the pick-up
- `pickup_month`: Month of the pick-up (always May)
- `pickup_week`: Week of the pick-up
- `pickup_doy`: Day of the year of the pick-up (from 1 to 365)
- `pickup_wday`: Day of the week of the pick-up (from 1 to 7)
- `length_time`: Duration of the ride (in seconds)
- `pickup_BoroCode`: Macro-area of the pick-up
- `pickup_NTACode`: Micro-area of the pick-up (Neighborhood Tabulation Areas)
- `dropoff_BoroCode`: Macro-area of the drop-off
- `dropoff_NTACode`: Micro-area of the drop-off (Neighborhood Tabulation Areas)
- `pickup_longitude`: Longitude of the pick-up
- `pickup_latitude`: Latitude of the pick-up
- `dropoff_longitude`: Longitude of the drop-off
- `dropoff_latitude`: Latitude of the drop-off

- `vendor_id` : The TPEP provider that provided the record
- `passenger_count` : The number of passengers in the vehicle. This is a driver-entered value.
- `trip_distance` : The elapsed trip distance in miles reported by the taximeter
- `fare_amount` : The time-and-distance fare calculated by the meter
- `pair` : Combination of the variables `pickup_NTACode` and `dropoff_NTACode`

Homework Rules

You are required to create a **single, tidy dataset**.

- You will work in groups.
- You can and should **explore the data** (descriptive statistics, plots, etc.) to identify any problematic aspects.
- You need to submit the following files:
  - A Python notebook (`tidy_homework_group_name.ipynb`) that uses as input the (`taxi_tidy.csv`) file, perform the analysis and produces a clean dataset.
  - A comma-separated file containing the tidy dataset (`tidy_dataset_group_name.csv`) generated by the aforementioned Python notebook.