# Milestone 1 DE W23 Description

## Deadline - 26/10 @11.59pm.

The goal of this milestone is to load a csv file,perform exploratory data analysis with visualization, extract additional data,perform feature engineering and pre-process the data for downstream cases such as ML and data analysis.

The dataset you will be working on is NYC green taxis dataset. It contains records about trips conducted in NYC through green taxis. Dataset description can be found on the cms.

There are multiple datasets for this case study(a dataset for each month). You can find which dataset you are assigned to in the zip folder named 'Milestone 1' on the CMS. Download your dataset from *here*.

You will also find weight distribution and marking rubric for this milestone on CMS(same zip folder) which should help you understand what is precisely required from you and how to achieve the highest marks.

IMPORTANT NOTE regarding the tasks:

- Each and every task should be written as a function. Organizing the code as functions is extremely important. Your functions should be as dynamic as possible and able to handle different datasets (by that i mean different data but same schema, different months of the green taxi records).
- VERY IMPORTANT: The notebook will not be run, your notebook MUST have the output of any function or task already shown.

Milestone requirements (what is required from you) (Please note that you are not obliged to follow the same flow of tasks, but it has to be logical):

- Load the dataset

## EDA

- Explore the dataset and ask at least 5 questions to give you a better understanding of the data provided to you.
- Visualize the answer to these 5 questions.

## Cleaning the data

- Tidy up the column names, make sure there is no spaces
- Observe,comment on and handle inconsistent data.(i.e duplicates, irrelevant data,incorrect data,etc)
- Observe missing data and comment on why you believe it is missing(MCAR,MAR or MNAR).
- Handle missing data
- Observe and comment on outliers
- Handle outliers

IMPORTANT NOTE : With every change you are making to the data you need to comment on why you used this technique and how has it affected the data(by both showing the change in the data i.e change in number of rows/columns, change in distribution, etc. and commenting on it).

## Data transformation and feature engineering

- Add 2 new columns named 'Week number' and 'Date range' and discretize the data into weeks according to the dates.
    - Tip: Change the datatype of the date feature to datetime type instead of object.
- Encode any categorical feature(s) and comment on why you used this technique and how the data has changed.
- If exists , Identify feature(s) which need normalization and show your reasoning. Then choose a technique to normalize the feature(s) and comment on why you chose this technique.
- Add at least two more columns which adds more info to the dataset by evaluating specific feature(s). I.E( Column indicating whether the trip was on a weekend or not).

## Additional data extraction

- Add GPS coordinates for the cities/locations.
- For this task you can extract the GPS coordinates from an API or web scraping and integrate into your csv file as new features.
- Tip 1 - you can find the web scraping and data integration notebooks under 'additional resources' on the CMS useful.

- Tip 2 - If you are going to use an API make sure you do not make request for each existing row but rather group by the cities and get their respective coordinates. Making a request for each row is too inefficient and expensive.
- Tip 3 - Rather than running the code for calling the API each time you load the notebook, the first time you call the API save the results in a csv file and then you could you check if a csv file exists for the GPS coordinates, if so, load directly and don't call APi. Same applies for web scraping.

## Lookup table and load back into new csv file

- Create a lookup table

You will need to create a lookup table(csv) which will contain info about the original values for any feature/values you have imputed. This lookup table must be created programmatically and not hard coded(not done be hand).

For any imputation with arbitrary values or encoding done, you have to store what the value imputed or encoded represents in a new csv file. I.e if you impute a missing value with -1 or 100 you must have a csv file illustrating what -1 and 100 means. Or for instance, if you encode cities with 1,2,3,4,etc what each number represents must be shown in the new csv file.

Example of a lookup table

| Column name | Original value | Imputed Value |
|---|---|---|
| PU Location | Missing | -1 |
| Tip amount | nan/null | -1 |
| Payment type | Credit card | 1 |
| Payment type | Cash | 2 |

- Load the new dataset into a new csv file named `green_trip_data_{year}-{month}clean.csv`. replace year and month with the appropriate values.
- Load the lookup table to a csv file called `lookup_table_green_taxis.csv`

## Bonus

- Bonus: save the cleaned dataset as a parquet file instead of a csv file(Parquet file is a compressed file format).

## Submission guidelines

- Important: Project is Individual
- ONLY accepted format is IPYNB (Python notebooks)
- Name of the notebook MUST be `M1_MAJOR_ID_groupno_month-year`.
- Replace major,id and group no. with the appropriate values and replace month-year with the month and year of the dataset you are working on.
- I.e `M1_MET_12-3456_2_7-12`
- Upload your notebook, cleaned csv and lookup table to your M1 google drive folder(check the semester plan if you haven't already)

## Important notes regarding submission(read carefully):

- Each and every task should be written as a function. Organizing the code as functions is extremely important. Your functions should be as dynamic as possible and able to handle different datastes (by that i mean different data but same schema, different months of the green taxi records).
- DO NOT under any circumstance edit your drive folder after the deadline. Here it is again, DO NOT under any circumstance edit your drive folder after the deadline
- notebook and csv file names must be as instructed, not doing so will result in marks deduction
- VERY IMPORTANT: The notebook will not be run, your notebook MUST have the output of any function or task already shown.