Computer Science Department
CS672 – Introduction to Deep Learning (CRN: 72938)
Fall 2025

Project #1 / Due 10-Oct-2025

Performing Exploratory Data Analysis (EDA) on data is of paramount importance for every Data Scientist / Data Analyst. Exploratory Data Analysis is often used to uncover various **patterns** present in the data and to draw conclusions from it. EDA is the core part when it comes to developing a Machine Learning model. This takes place through analysis and visualization of the data which will be fed to the Machine Learning Model. A Machine Learning Model is as good as the training data - you must understand it if you want to understand your model.

Presentation and action phase

- **Communication of insights**: After performing your analysis, you must present your findings effectively. This includes summarizing key discoveries and communicating the story behind the data in a clear and compelling way.
- **Interpretation and evaluation**: Discuss the implications of your findings and interpret how your results address the original business problem.
- **Recommendations**: Based on your insights, propose specific, actionable next steps. This translates your findings into practical guidance that helps achieve the business objective.

Prior to commencing your efforts on coding, you must install the following libraries:

- pip install -q tensorflow_data_validation [visualization] (**)
  - https://pypi.org/project/tensorflow-data-validation/
- pip install apache-beam [interactive]
  - https://beam.apache.org/get-started/quickstart-py/
  - https://pypi.org/project/apache-beam/
- Install the GraphViz library
  - https://www.graphviz.org/download/

Perform an **explanatory data analysis** (**EDA**) on **NYC's Yellow Taxi Trip Records** from **2020**. Although there will be no need to build a model based on the data provided, you are asked to look for issues in the data and find correlation among the various variables to improve **ride time predictions**.

Create the training dataset on data based on March of 2020, and evaluation dataset on data based on May of 2020. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

March

- **Yellow Taxi Trip Records** (CSV)
- **Green Taxi Trip Records** (CSV)
- **For-Hire Vehicle Trip Records** (CSV)
- **High Volume For-Hire Vehicle Trip Records** (CSV)

April

- **Yellow Taxi Trip Records** (CSV)
- **Green Taxi Trip Records** (CSV)
- **For-Hire Vehicle Trip Records** (CSV)
- **High Volume For-Hire Vehicle Trip Records** (CSV)

May

- **Yellow Taxi Trip Records** (CSV)
- **Green Taxi Trip Records** (CSV)
- **For-Hire Vehicle Trip Records** (CSV)
- **High Volume For-Hire Vehicle Trip Records** (CSV)

March

- **Yellow Taxi Trip Records** (PARQUET)
- **Green Taxi Trip Records** (PARQUET)
- **For-Hire Vehicle Trip Records** (PARQUET)
- **High Volume For-Hire Vehicle Trip Records** (PARQUET)

April

- **Yellow Taxi Trip Records** (PARQUET)
- **Green Taxi Trip Records** (PARQUET)
- **For-Hire Vehicle Trip Records** (PARQUET)
- **High Volume For-Hire Vehicle Trip Records** (PARQUET)

May

- **Yellow Taxi Trip Records** (PARQUET)
- **Green Taxi Trip Records** (PARQUET)
- **For-Hire Vehicle Trip Records** (PARQUET)
- **High Volume For-Hire Vehicle Trip Records** (PARQUET)

Note: I understand NYC's portal with Yellow Cab's Trip Data has made an important change.
They no longer provide '.csv' files, instead 'parquet' file format is the dominant one.

TLC Trip Record Data

Yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

ATTENTION!

On 05/13/2022, we are making the following changes to trip record files:

1. All files will be stored in the PARQUET format. Please see the 'Working With PARQUET Format' under the Data Dictionaries and MetaData section.
2. Trip data will be published monthly (with two months delay) instead of bi-annually.
3. HVFHV files will now include 17 more columns (please see High Volume FHV Trips Dictionary for details). Additional columns will be added to the old files as well. The earliest date to include additional columns: February 2019.

There are two options to read a 'parquet' file within Python:
1_ Convert 'parquet' file to 'csv'.

Here is a link to convert a parquet file to csv file (within a Windows-10 machine):
https://phoenixnap.com/kb/install-spark-on-windows-10

2_ Read 'parquet' file via Pandas' read_parquet method.
e.g. df = pd.read_parquet('yellow_tripdata_2020-03.parquet', engine='fastparquet')
Make sure you have installed the 'fastparquet' library.

Write **Python** scripts to complete the following tasks along with their output. All work should be done and submitted in a single **Jupyter or Colab Notebook.**

1) Prep the data to be ready to be fed to a model.
   Look for missing, null, NaN records.
   Find outliers.
   Transform data – all entries should be numeric.
2) List all types of data, numeric, categorical, …
3) Perform EDA on data.
   Utilize both:
   - Classic approach in EDA (Pandas, Numpy libraries)
   - The TFDV (TensorFlow Data Validation) module with the powerful graphical statistics generated (Apache beam library…)

   Present dependencies and correlations among the various features in the data.
   List the most variables (Feature Importance) that will affect the target label.
4) Be aware of the time-window selection for the data.
   March 2020 was when COVID19 pandemic broke out in the US.
   Every industry and business initiatives were impacted drastically.
   Starting March 2020, the NYC Taxi industry has established a 'new normal'.
   << Extra Credit >>:
   - January 2020 data present the 'baseline' of what the NYC Taxi business used to be.
   - Compare the data of Jan-2020 vs Mar-2020.
   - Present your findings.

(**) Highly recommend having installed the whole gamma of TensorFlow's module.
Here is a 'base' list of them:

```
tensorboard                 2.6.0
tensorboard-data-server     0.6.1
tensorboard-plugin-wit      1.6.0
tensorflow                  2.6.0
tensorflow-data-validation  1.3.0
tensorflow-datasets         4.4.0
tensorflow-estimator        2.6.0
tensorflow-metadata         1.2.0
tensorflow-serving-api      2.6.0
```