# ETL Project Report

DataMeisters Team - Can Alaluf, Omotoyosi Odele

Extract:
**your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).**
Original data was gotten from 2 sources:

1. Kaggle as a **CSV** file for Uber rides data in Boston - https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices#cab_rides.csv
2. Kaggle as a **CSV** file for Boston weather between 2015 and 2018 - https://www.kaggle.com/willmichels/boston-weather-20152018-use-with-crimes-data

Transform:
**what data cleaning or transformation was required.**
The first Kaggle data was imported into a Python notebook on Jupyter Lab, using the *read_csv* command. We randomly selected 10000 rides as the master dataset had over 300,000 rides and it took quite bit of time for the following steps. We did further cleaning by selecting only the needed columns and assigned them to a new dataframe. We then converted the UNIX timestamps to date and added new columns to hold the Day, Month and Year extracted from the date, using the *split* function. Then finally concatenated the new dates to DDMMYY format.

We obtained the weather data from the boston-weather.csv where the dates were in DDMM(String)YY format. We first split the day month and the year and passed a dictionary to convert months in a string format to a integer format. Finally concatenated the same way as the uber_rides and exported it to a csv. The Precipitation column had a value "T" that we also cleaned since it was not an integer. Renamed "High(°F)" to High as SQL was not able to read the format.

Load:
**the final database, tables/collections, and why this was chosen.**
The final database to be used is PostgresSQL and this was chosen because after the data cleaning exercise, we were able to load our data into CSV files, which are easily readable in relational databases. We joined 2 dataframes on the date using an inner join.

Analyze:
Also, we are able to perform any further analysis like checking the difference in prices under different weather conditions for the same trips (.i.e. same departure to destination) or for similar distances. We could also look into the preference of request for specific service provider per location and per weather status. We could further compare the surge prices between providers Lyft and Uber when the precipitation is high and see which provider has a better price from one source to destination. It would be interesting to see which provider that one should go to if it is a rainy day, finding which provider would give the more affordable price. Its also possible to analyze on average how much more expensive it is per mile given the a degree change in  weather compared to average .