Homework-4

A new ridesharing company will start to service in New York. They want to analyze public taxi trip data in NYC. This company collects Yellow and Green Taxi data in order to find out answers for the following questions.

* Retrieve statistical information from datasets (warmup task)
* What is the most expensive root for each datasets?
* What is the top 5 busiest Zone for each datasets?
* What is the longest trip?
* Find the crowded Pickup and Drop-off zones for each hour.
* Find the Pickup and Drop-off count for each hour. After that draw two lineplot graphs for Pickup and Drop-off.
* (BONUS) What do other columns affect the tip column? (HINT: Correlation of tip amount to other columns)

We are expecting you to calculate results using PySpark in PyCharm. But only in a small part of the code you can use pandas to draw plot. In template.py file you will find the functions in which you will write your codes. You will start from main function to implement and then you will join two dataset separately with zone lookup in join\_look\_up\_with\_cities. Some functions will return dataframe but some of them just show the question’ results to console. For any other questions, you can send email to us. Good luck.

You are needed to work with New York Taxi data. The one needs to download the data from following URLs;

NYC Yellow Taxi Data – March: <https://d37ci6vzurychx.cloudfront.net/trip-data/yellow_tripdata_2021-03.parquet>

NYC Yellow Taxi Data Dictionary: <https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf>

NYC Green Taxi Data – March: <https://d37ci6vzurychx.cloudfront.net/trip-data/green_tripdata_2021-03.parquet>

NYC Green Taxi Data Dictionary: <https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf>

Taxi Zone Lookup Table: <https://d37ci6vzurychx.cloudfront.net/misc/taxi+_zone_lookup.csv>

References:

<https://sparkbyexamples.com/pyspark-tutorial/>

<https://spark.apache.org/docs/latest/api/python/reference/index.html>

Email for Questions: [tugtas@hotmail.com](mailto:tugtas@hotmail.com)

[merttbilginn@gmail.com](mailto:merttbilginn@gmail.com)

[boran.isl@hotmail.com](mailto:boran.isl@hotmail.com)