DA 331 - Big Data Analytics : Tools & Techniques

Lab 3

Instructor: Dr. Chiranjib Sur

Dataset link: https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data

Given the dataset, find the following:

1. Which drivers had the most number of passengers during weekends?
2. Which drivers had the most tip to fare ratio during weekdays?
3. Which drivers avoided tolls the most?
4. Which rate code had the most trips?
5. Which drivers rode the most best rate code?

You may need this to process the dates.

```
from pyspark.sql.functions import year, month, day
extracted_df = df.select(year("date").alias("year"), /
                month("date").alias("month"), day("date").alias("day"))
extracted df.show()
```

```
!gdown 1Sev85co2s8EnJvs5AisXMJRGjhOI1S3p
result1 = df1.unionByName(df2)
```

For Concatenation of two dataframes in spark.