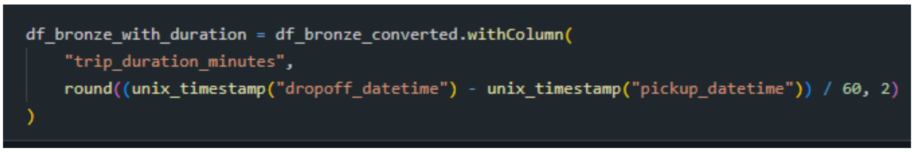
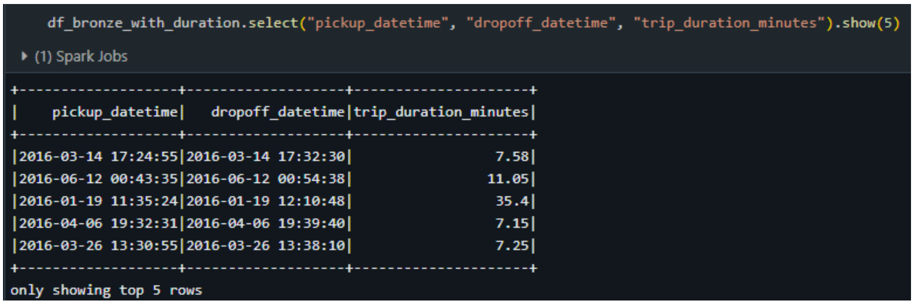
Ansh Ranjan
Databricks Case Study
EXERCISE 3: Data Transformation

TASK 1 and 2: Apply transformation to data (filtering, grouping, etc)

1. Creating a new column trip_duration as difference of pickup and dropoff time in minutes



```
df_bronze_with_duration = df_bronze_converted.withColumn(
    "trip_duration_minutes",
    round((unix_timestamp("dropoff_datetime") - unix_timestamp("pickup_datetime")) / 60, 2)
)
```

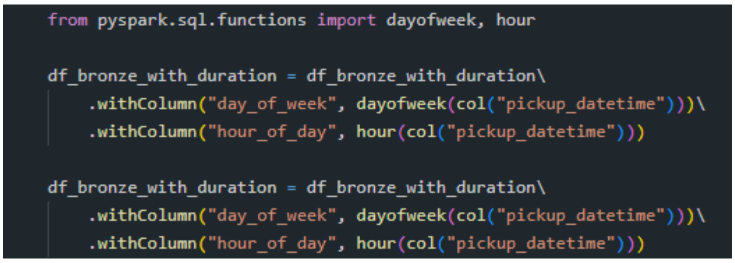


```
df_bronze_with_duration.select("pickup_datetime", "dropoff_datetime", "trip_duration_minutes").show(5)
```

(1) Spark Jobs

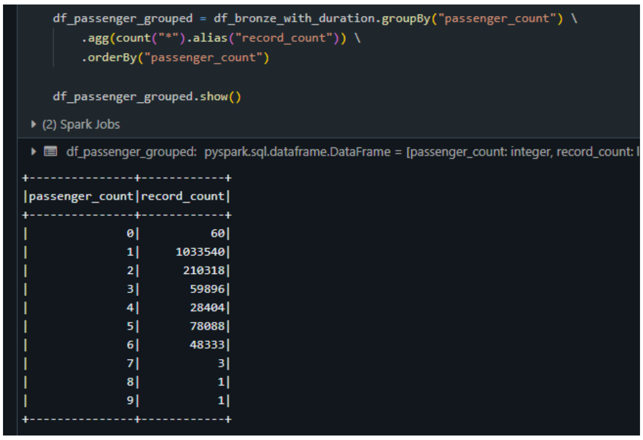
```
+-------------------+-------------------+---------------------+
|    pickup_datetime|   dropoff_datetime|trip_duration_minutes|
+-------------------+-------------------+---------------------+
|2016-03-14 17:24:55|2016-03-14 17:32:30|                 7.58|
|2016-06-12 00:43:35|2016-06-12 00:54:38|                11.05|
|2016-01-19 11:35:24|2016-01-19 12:10:48|                 35.4|
|2016-04-06 19:32:31|2016-04-06 19:39:40|                 7.15|
|2016-03-26 13:30:55|2016-03-26 13:38:10|                 7.25|
+-------------------+-------------------+---------------------+
only showing top 5 rows
```

2. Creating new columns day_od_week and Hour_of_day



```
from pyspark.sql.functions import dayofweek, hour

df_bronze_with_duration = df_bronze_with_duration\
    .withColumn("day_of_week", dayofweek(col("pickup_datetime")))\
    .withColumn("hour_of_day", hour(col("pickup_datetime")))

df_bronze_with_duration = df_bronze_with_duration\
    .withColumn("day_of_week", dayofweek(col("pickup_datetime")))\
    .withColumn("hour_of_day", hour(col("pickup_datetime")))
```

3. Finding number of trips per number of passengers



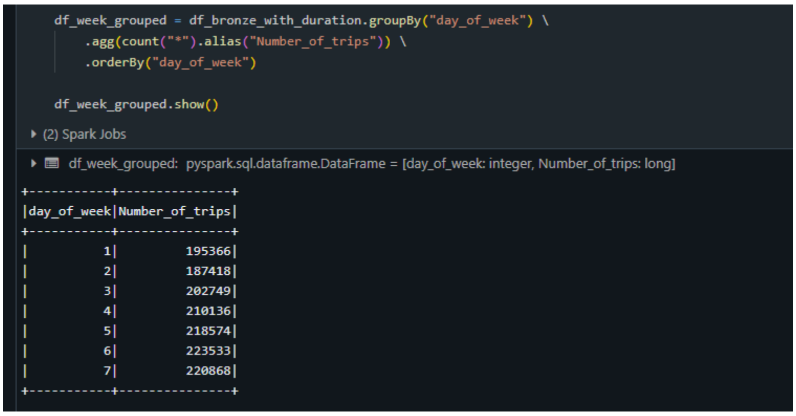
```
df_passenger_grouped = df_bronze_with_duration.groupBy("passenger_count") \
    .agg(count("*").alias("record_count")) \
    .orderBy("passenger_count")

df_passenger_grouped.show()
```

(2) Spark Jobs

df_passenger_grouped: pyspark.sql.dataframe.DataFrame = [passenger_count: integer, record_count: l

```
+---------------+------------+
|passenger_count|record_count|
+---------------+------------+
|              0|          60|
|              1|     1033540|
|              2|      210318|
|              3|       59896|
|              4|       28404|
|              5|       78088|
|              6|       48333|
|              7|           3|
|              8|           1|
|              9|           1|
+---------------+------------+
```

4. Grouping data by day of week to get number of trips made for each day of the week



```
df_week_grouped = df_bronze_with_duration.groupBy("day_of_week") \
    .agg(count("*").alias("Number_of_trips")) \
    .orderBy("day_of_week")

df_week_grouped.show()
```

(2) Spark Jobs

df_week_grouped: pyspark.sql.dataframe.DataFrame = [day_of_week: integer, Number_of_trips: long]

```
+-----------+---------------+
|day_of_week|Number_of_trips|
+-----------+---------------+
|          1|         195366|
|          2|         187418|
|          3|         202749|
|          4|         210136|
|          5|         218574|
|          6|         223533|
|          7|         220868|
+-----------+---------------+
```