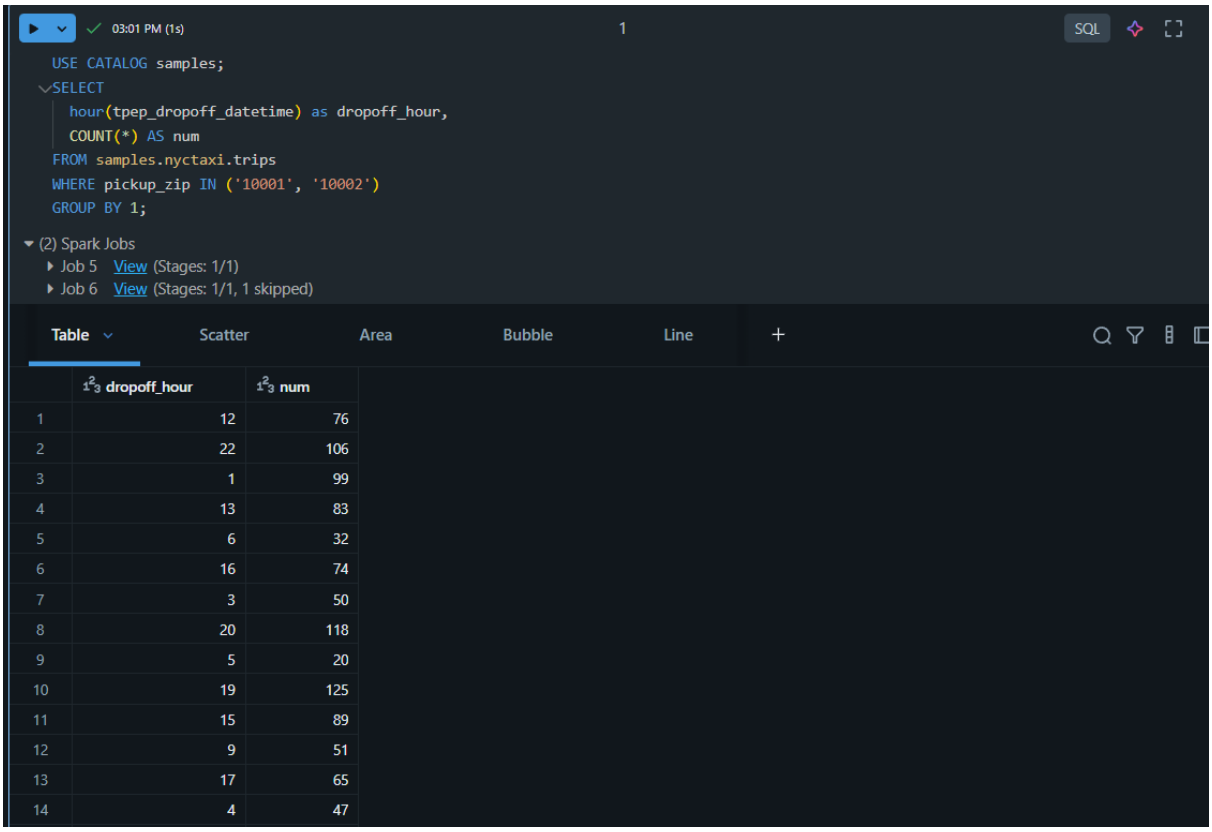
ASSIGNMENT – 2 (Azure Databricks)

TRAINEE NAME: Swathi Baskaran

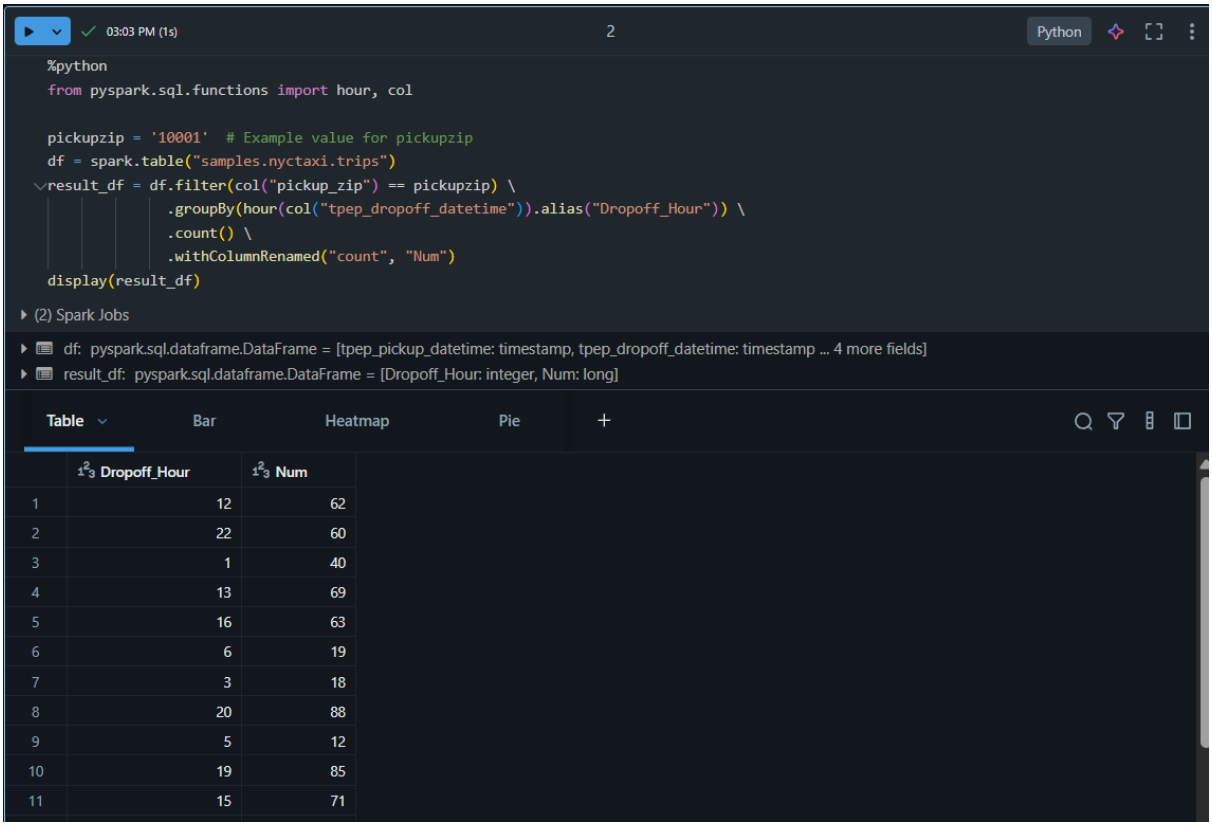
EDA Analysis



```
USE CATALOG samples;
SELECT
  hour(tpep_dropoff_datetime) as dropoff_hour,
  COUNT(*) AS num
FROM samples.nyctaxi.trips
WHERE pickup_zip IN ('10001', '10002')
GROUP BY 1;
```

(2) Spark Jobs
Job 5 View (Stages: 1/1)
Job 6 View (Stages: 1/1, 1 skipped)

	dropoff_hour	num
1	12	76
2	22	106
3	1	99
4	13	83
5	6	32
6	16	74
7	3	50
8	20	118
9	5	20
10	19	125
11	15	89
12	9	51
13	17	65
14	4	47



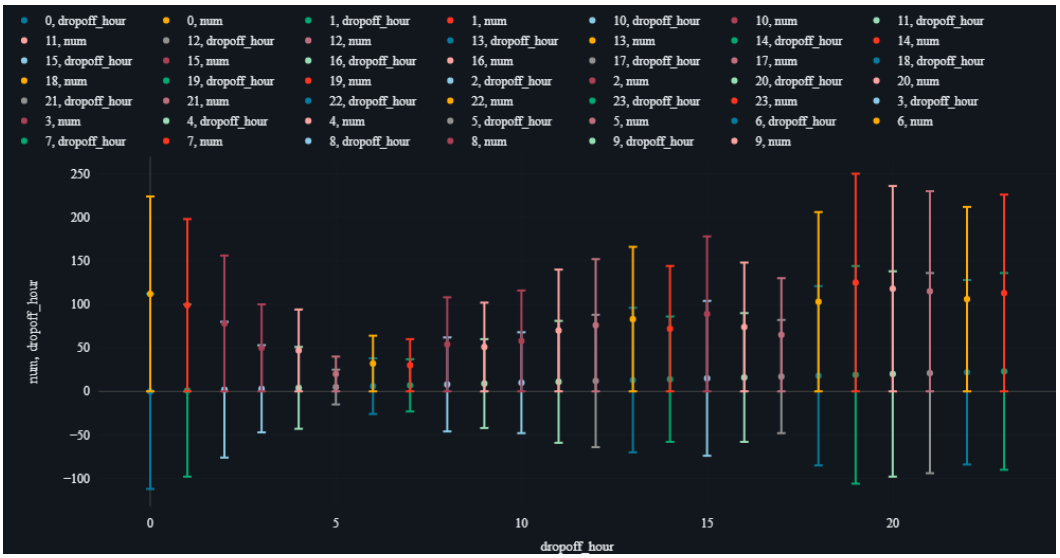
```
%python
from pyspark.sql.functions import hour, col

pickupzip = '10001'  # Example value for pickupzip
df = spark.table("samples.nyctaxi.trips")
result_df = df.filter(col("pickup_zip") == pickupzip) \
              .groupBy(hour(col("tpep_dropoff_datetime")).alias("Dropoff_Hour")) \
              .count() \
              .withColumnRenamed("count", "Num")
display(result_df)
```

(2) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [tpep_pickup_datetime: timestamp, tpep_dropoff_datetime: timestamp ... 4 more fields]
result_df: pyspark.sql.dataframe.DataFrame = [Dropoff_Hour: integer, Num: long]

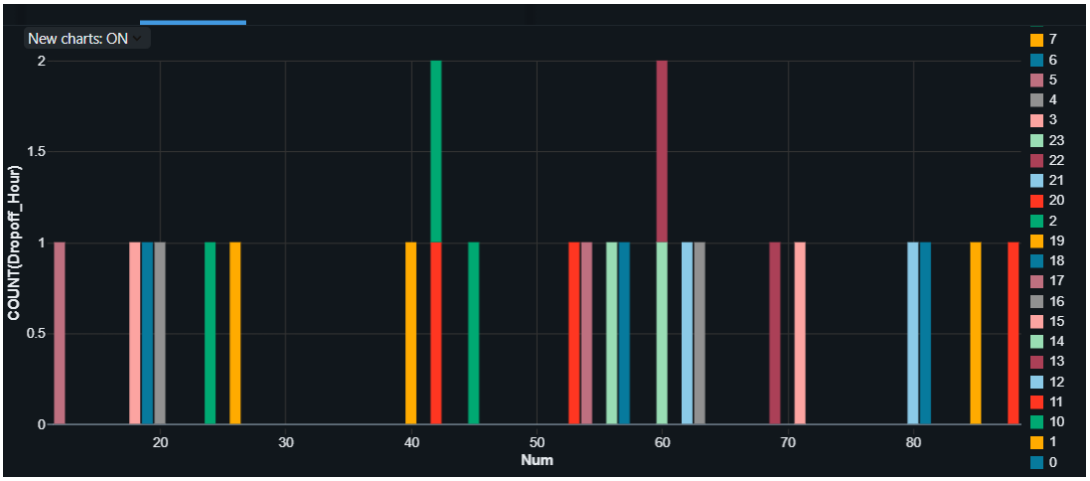
	Dropoff_Hour	Num
1	12	62
2	22	60
3	1	40
4	13	69
5	16	63
6	6	19
7	3	18
8	20	88
9	5	12
10	19	85
11	15	71

EDA Visualization

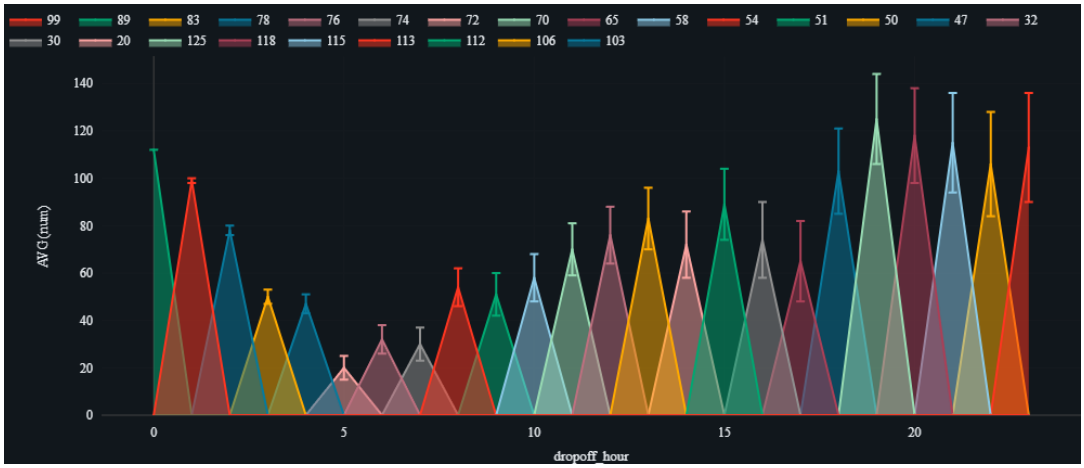
1. Scatter



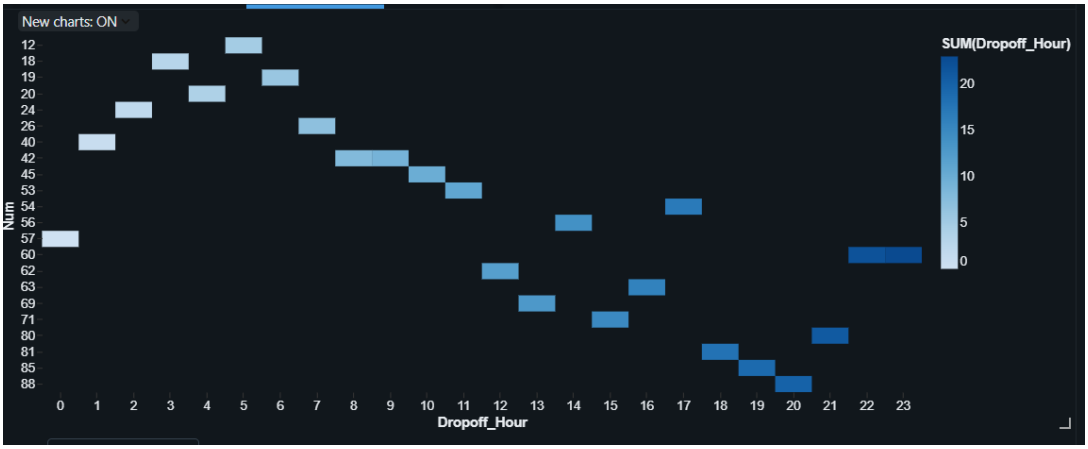
2. Bar



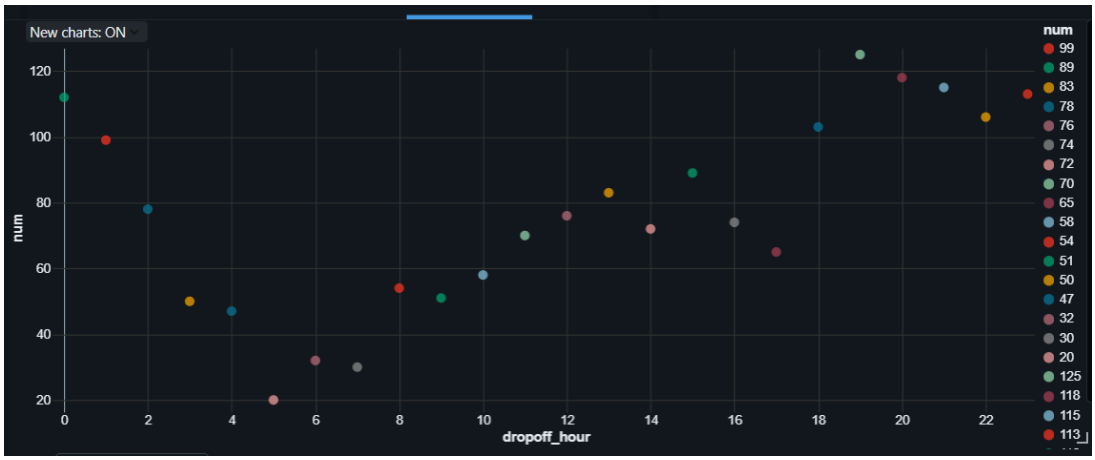
3. Area



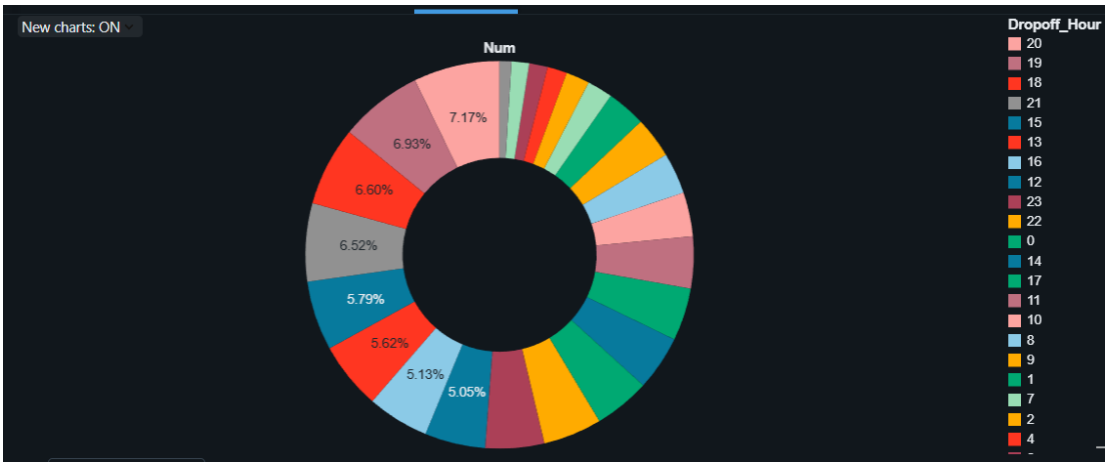
4. Heatmap



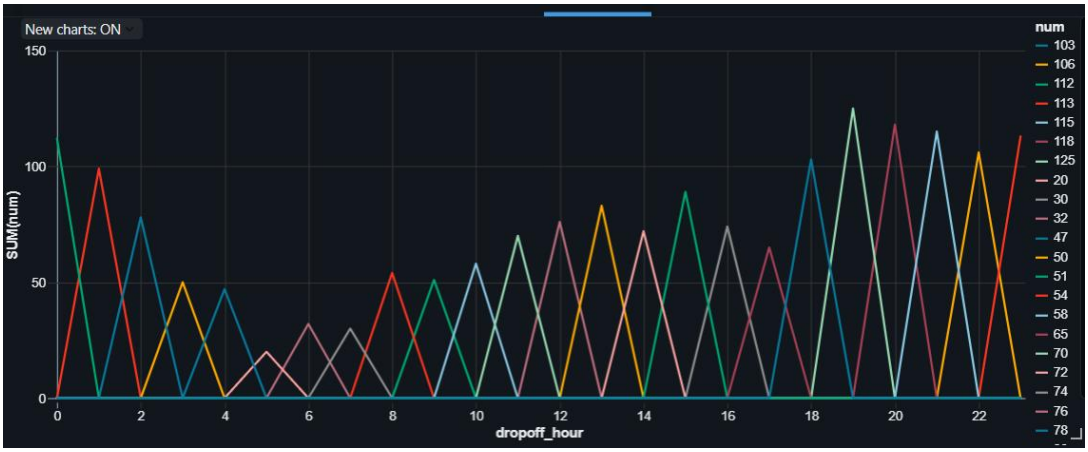
5. Bubble



6. Pie



7. Line


New charts: ON
num
103
106
112
113
115
118
125
20
30
32
47
50
51
54
58
65
70
72
74
76
78
150
100
50
0
SUM(num)
0
2
4
6
8
10
12
14
16
18
20
22
dropoff_hour