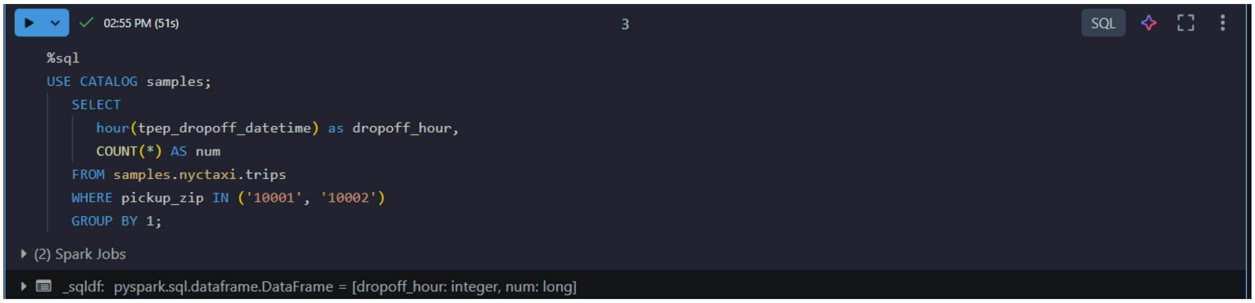
Tejashree G
tejashreeg.ai2021@dce.edu.in

EDA Assignment

SQL

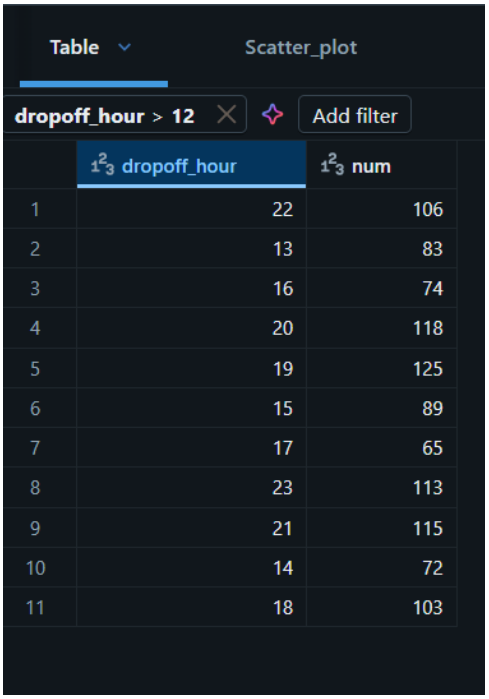


```
%sql
USE CATALOG samples;
  SELECT
    hour(tpep_dropoff_datetime) as dropoff_hour,
    COUNT(*) AS num
  FROM samples.nyctaxi.trips
  WHERE pickup_zip IN ('10001', '10002')
  GROUP BY 1;
```

```
(2) Spark Jobs
_sqldf:  pyspark.sql.dataframe.DataFrame = [dropoff_hour: integer, num: long]
```

Table output

dropoff_hours is filtered where dropoff_hours > 12



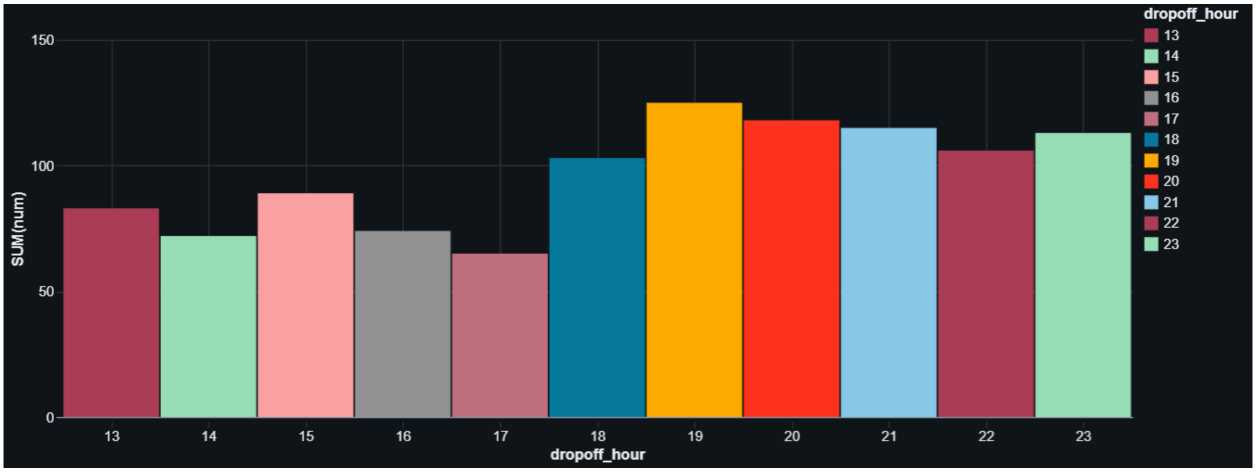
dropoff_hour > 12

	dropoff_hour	num
1	22	106
2	13	83
3	16	74
4	20	118
5	19	125
6	15	89
7	17	65
8	23	113
9	21	115
10	14	72
11	18	103

Visualizations

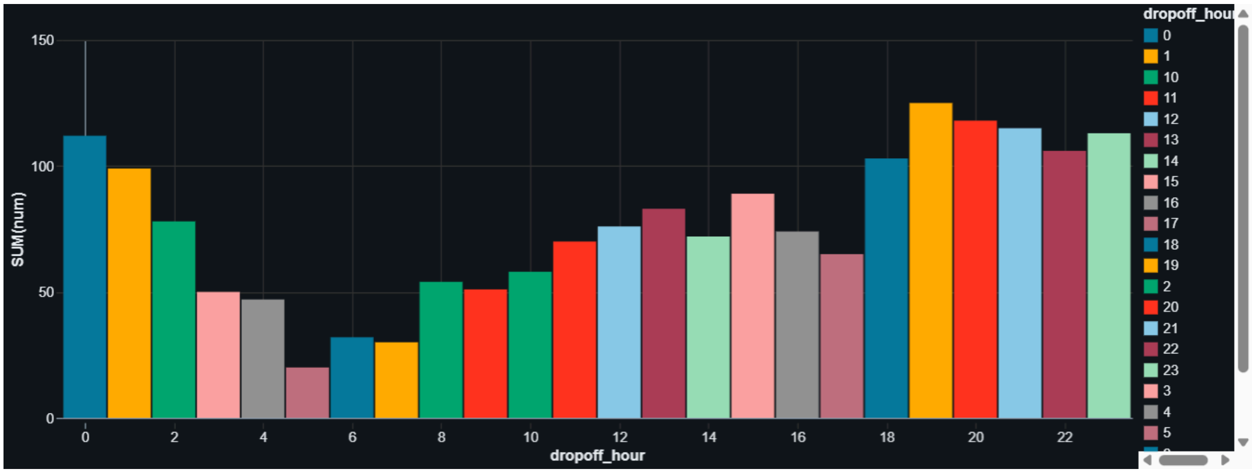
Bar Chart

- Visualization Type: Bar.
- X Column: dropoff_hour
- Y Column: num, showing the sum of drop-offs for each hour.
- Aggregation: Sum applied to num.
- Group By: dropoff_hour
- Filter: dropoff_hour > 12



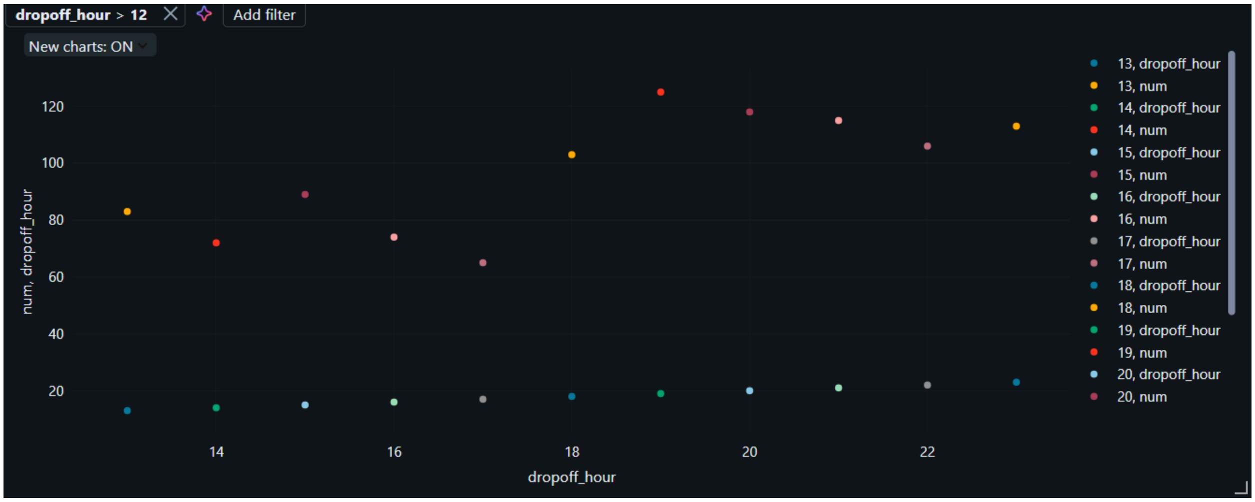
With Aggregation applied

- Visualization Type: Bar.
- X Column: dropoff_hour
- Y Column: num, showing the sum of drop-offs for each hour.
- Aggregation: Sum applied to num.
- Group By: dropoff_hour

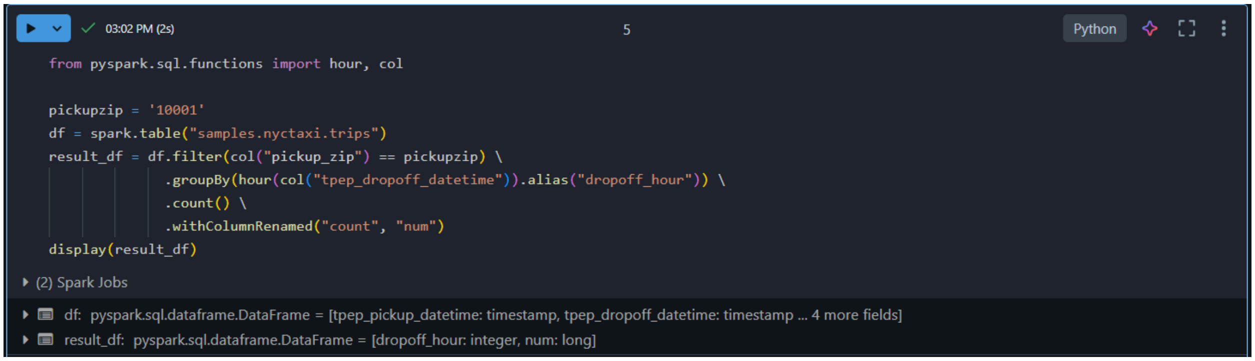


Scatter Plot

- Visualization Type: Scatter.
- X Column: dropoff_hour, representing the horizontal axis.
- Y Columns: num and dropoff_hour
- Group By: dropoff_hour, grouping points by hour
- Filter: dropoff_hours is filtered where dropoff_hours > 12



Python



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

pickupzip = '10001'
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]
```

Table output

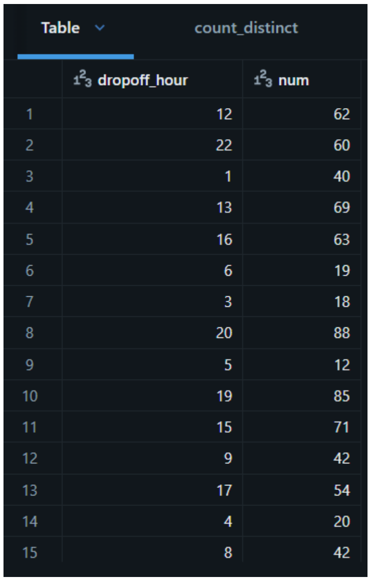


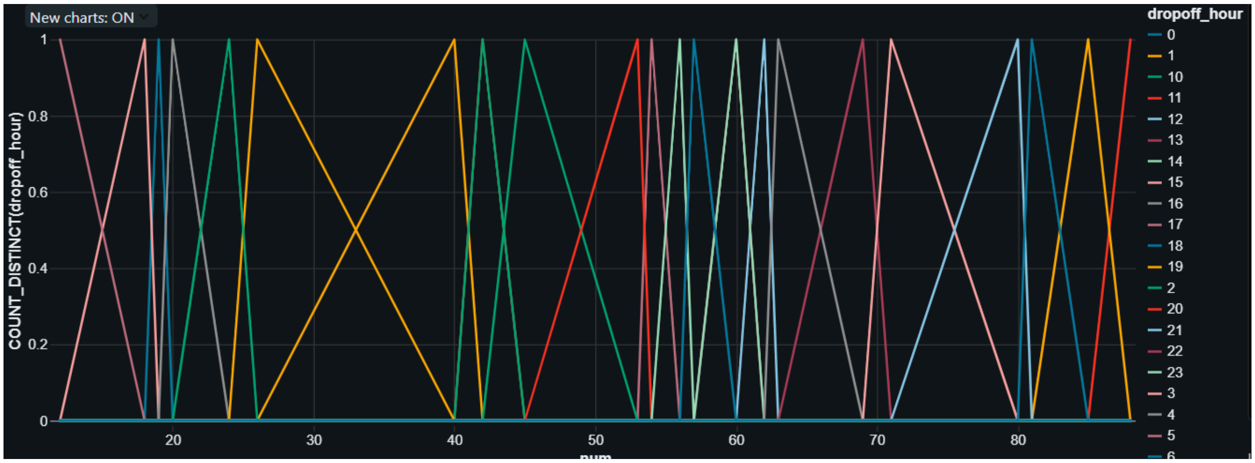
Table
count_distinct

	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
12	9	42
13	17	54
14	4	20
15	8	42

Visualizations

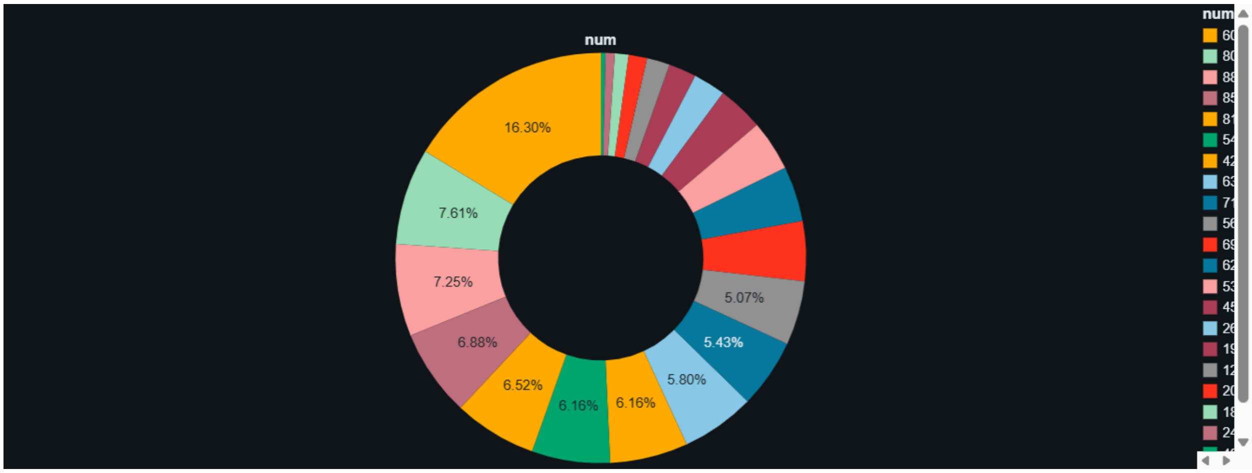
Line Chart

- Visualization Type: Line.
- X Column: num, representing the count of drop-offs on the horizontal axis.
- Y Column: dropoff_hour, showing the count of distinct drop-off hours for each num value.
- Aggregation: Count distinct applied to dropoff_hour
- Group By: dropoff_hour



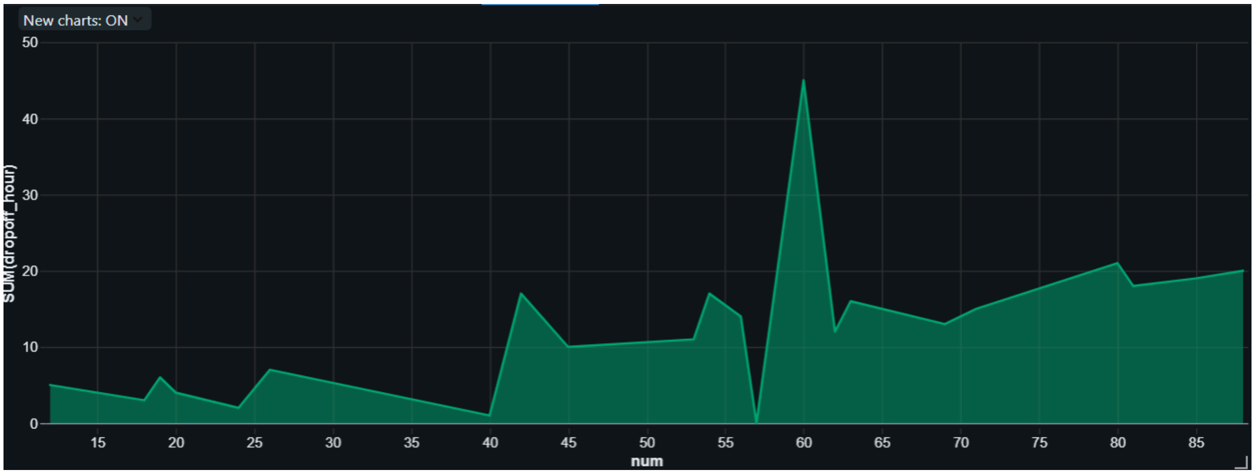
Pie Chart

- Visualization Type: Pie.
- X Column: num
- Y Column: dropoff_hour
- Aggregation: Sum of dropoff_hour
- Direction: Counterclockwise.



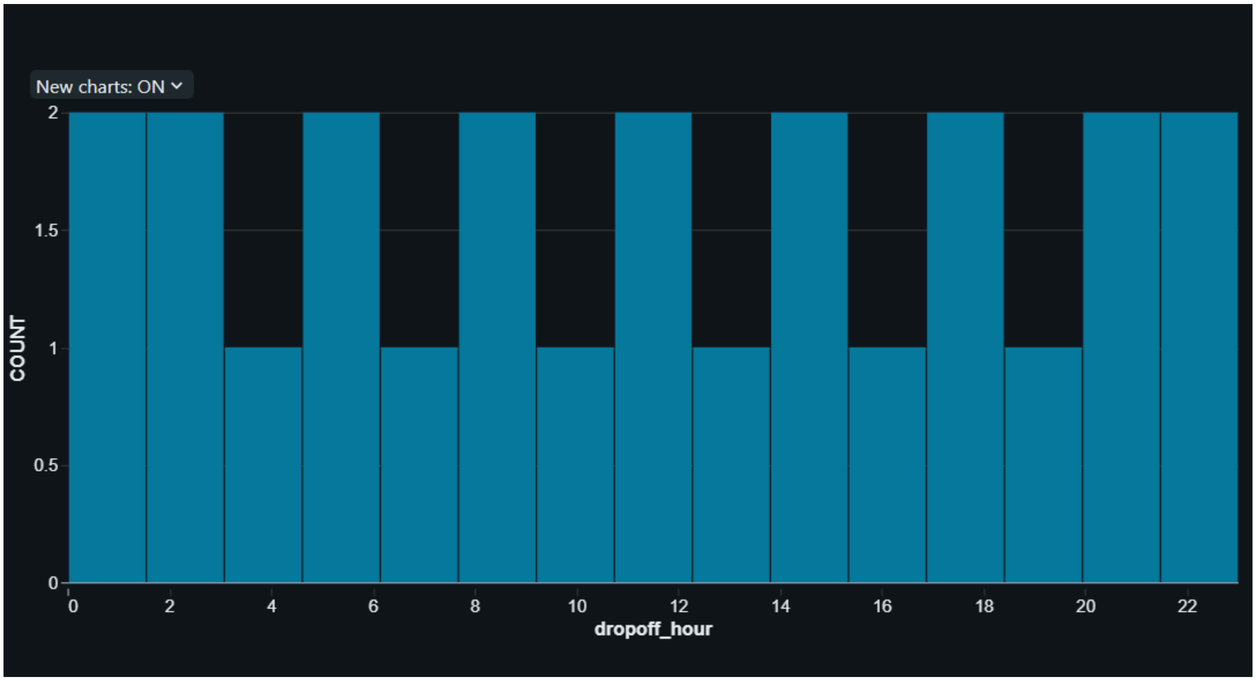
Area Chart

- Visualization Type: Area
- X Column: num
- Y Column: dropoff_hour
- Aggregation: Sum of dropoff_hour



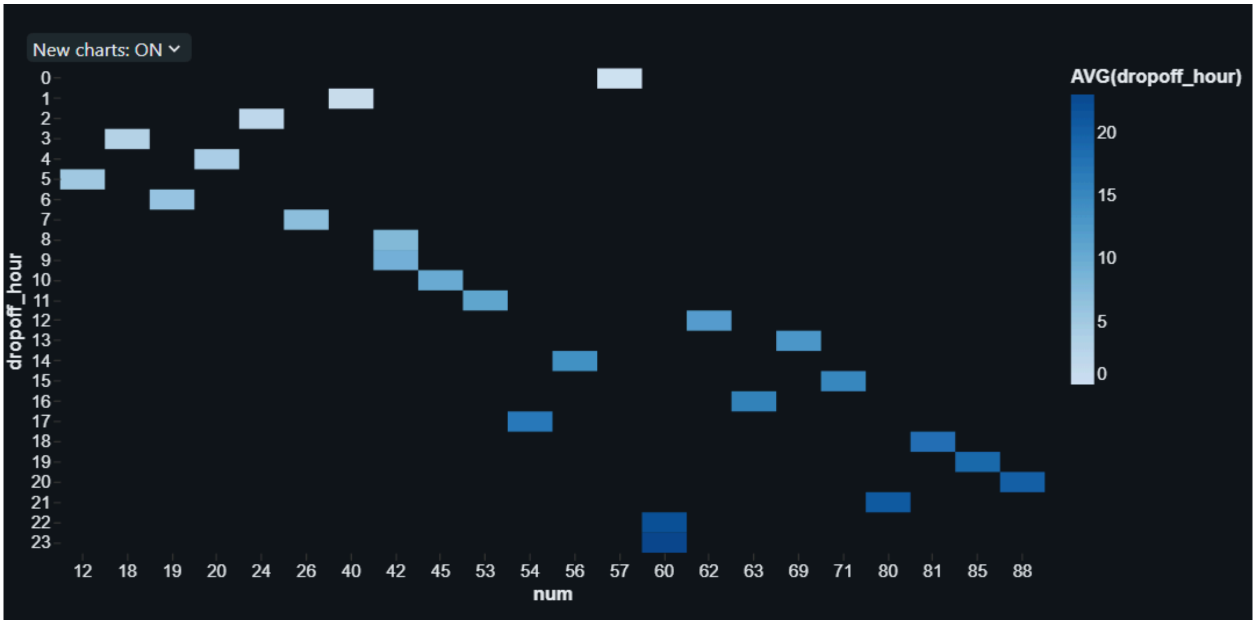
Histogram

- Visualization Type:Histogram
- X Column: dropoff_hour
- Number of Bins: 15



Heatmap

- Visualization Type: Heatmap.
- X Column: num, representing the count of drop-offs on the horizontal axis.
- Y Column: dropoff_hour, representing the hours of drop-off on the vertical axis.
- Color Column: dropoff_hour, with color intensity based on the average of dropoff_hour
- Aggregation: Average applied to dropoff_hour



New charts: ON