

EDA Visualization

Building Queries and Visualizations

-Bavatharani S

SQL

```
USE CATALOG samples;
```

```
SELECT
```

```
hour(tped_dropoff_datetime) as dropoff_hour,
```

```
COUNT(*) AS num
```

```
FROM samples.nyctaxi.trips
```

```
WHERE pickup_zip IN ('10001', '10002')
```

```
GROUP BY 1;
```

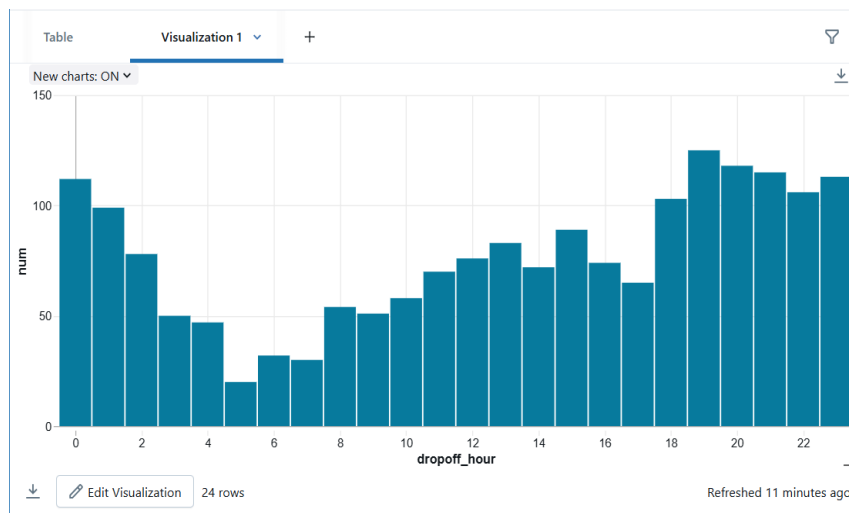
	1 ² ₃ dropoff_hour	1 ² ₃ num
1	15	89
2	20	118
3	21	115
4	22	106
5	17	65
6	1	99
7	23	113
8	4	47
9	13	83
10	8	54
11	6	32
12	3	50
13	16	74
14	11	70
15	0	112

↓

▼

24 rows | 3.77s runtime

Bar Chart Visualization



Python

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

pickupzip = '10001' # Example value for pickupzip

df = spark.table("samples.nyctaxi.trips")

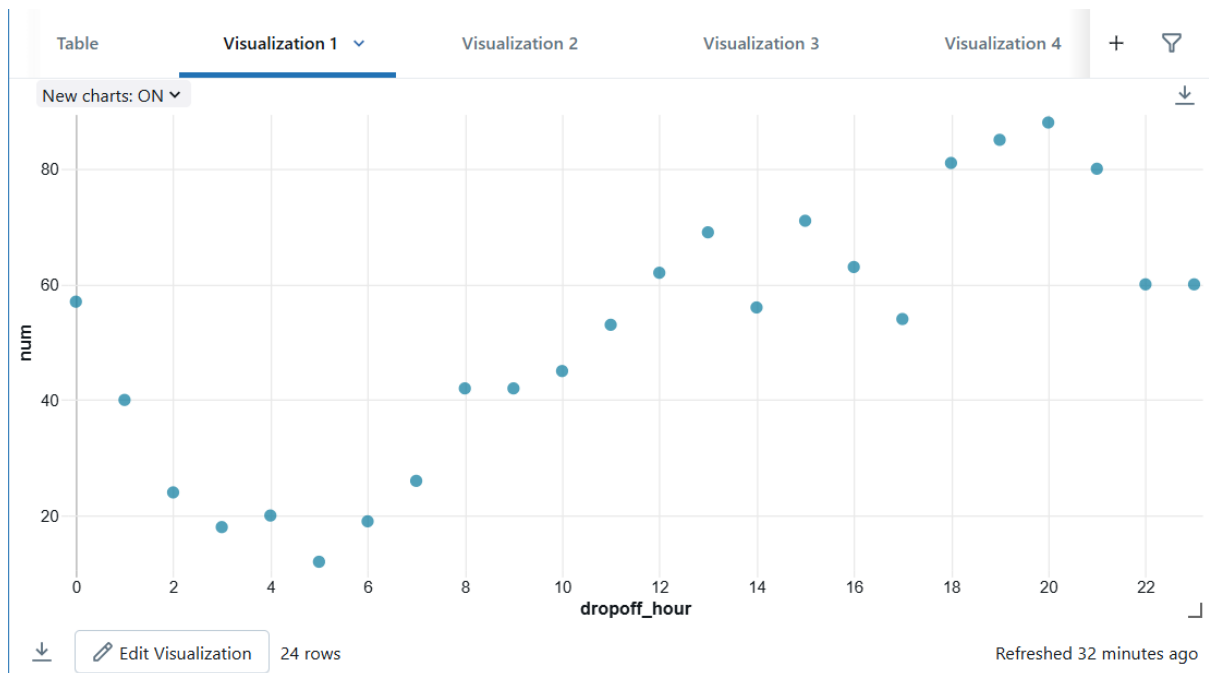
result_df = df.filter(col("pickup_zip") == pickupzip) \
.groupBy(hour(col("tpep_dropoff_datetime")).alias("dropoff_hour")) \
.count() \

.withColumnRenamed("count", "num")

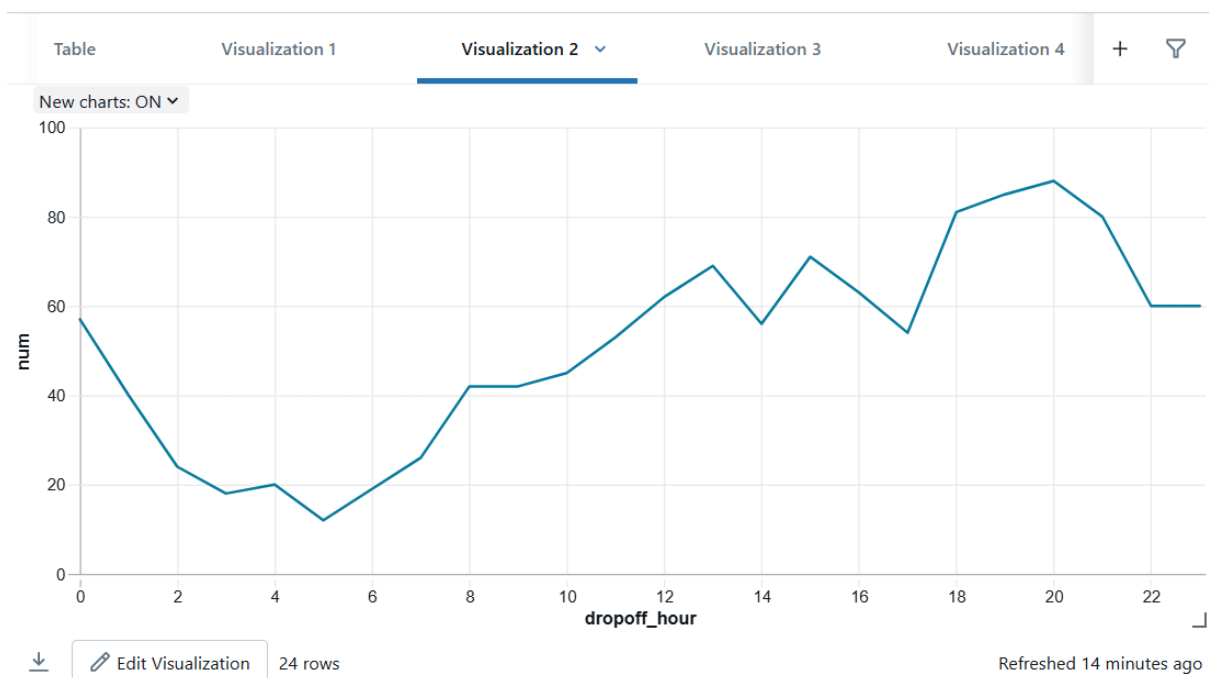
display(result_df)
```

	dropoff_hour	num
1	15	71
2	20	88
3	21	80
4	22	60
5	17	54
6	1	40
7	23	60
8	4	20
9	13	69
10	8	42
11	6	19
12	3	18
13	16	63
14	11	53
15	0	57

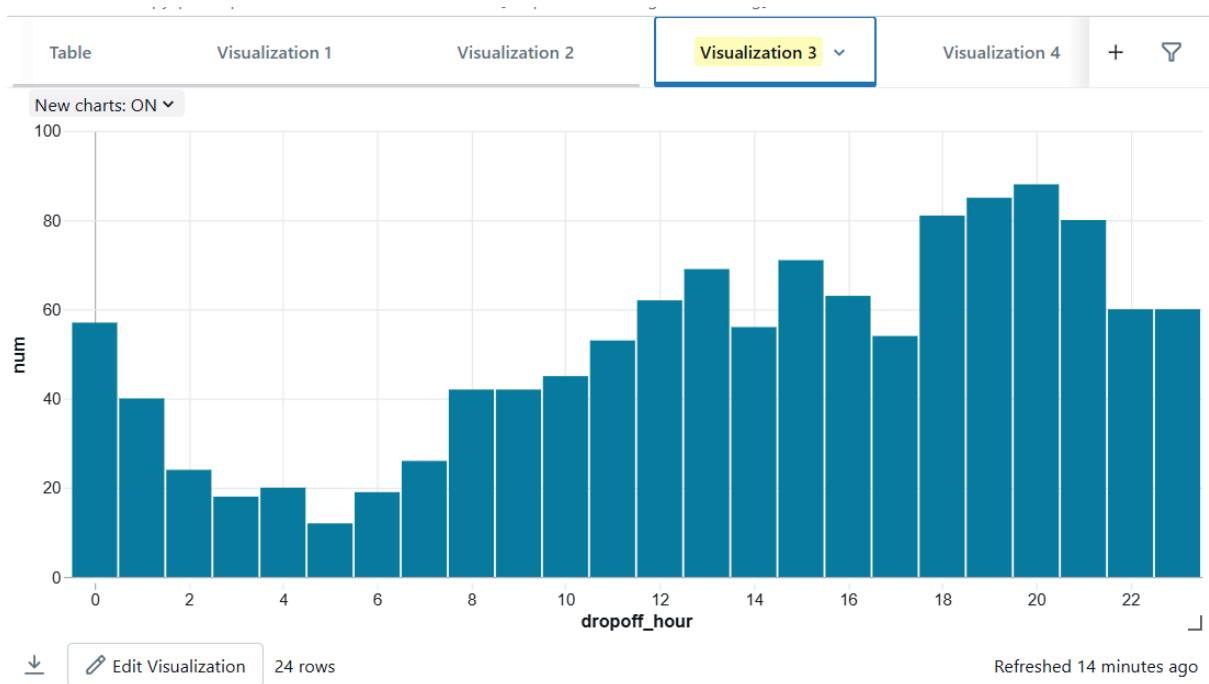
Scatter Plot



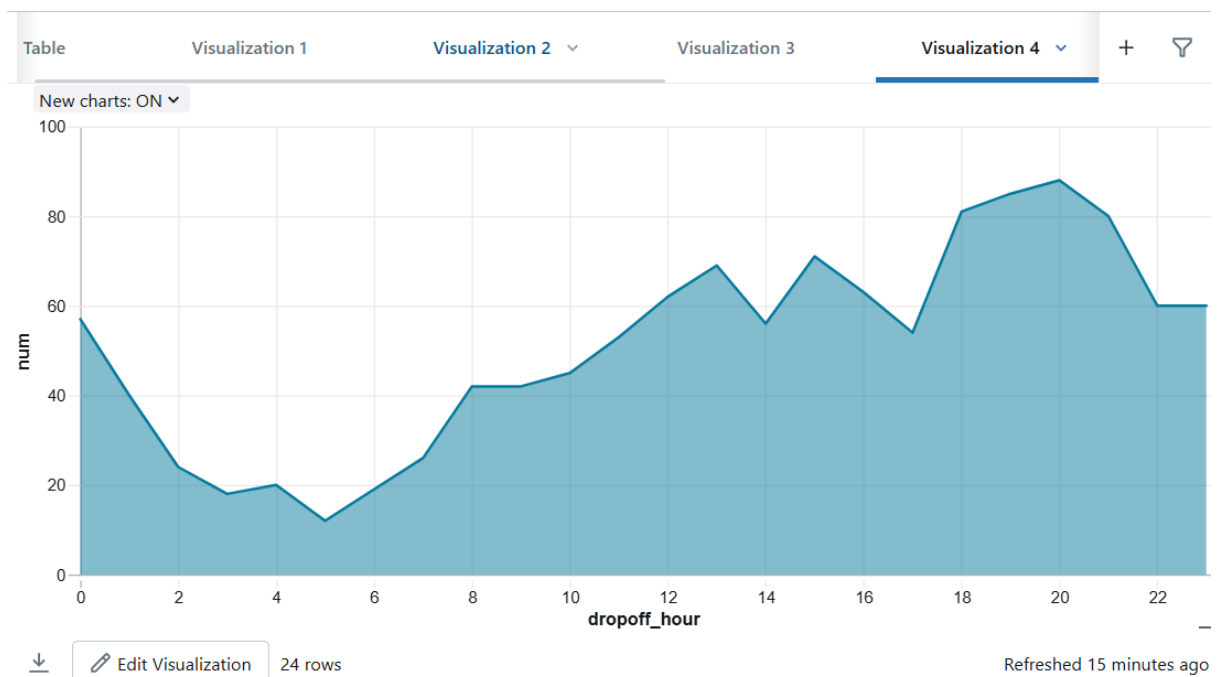
Line Chart



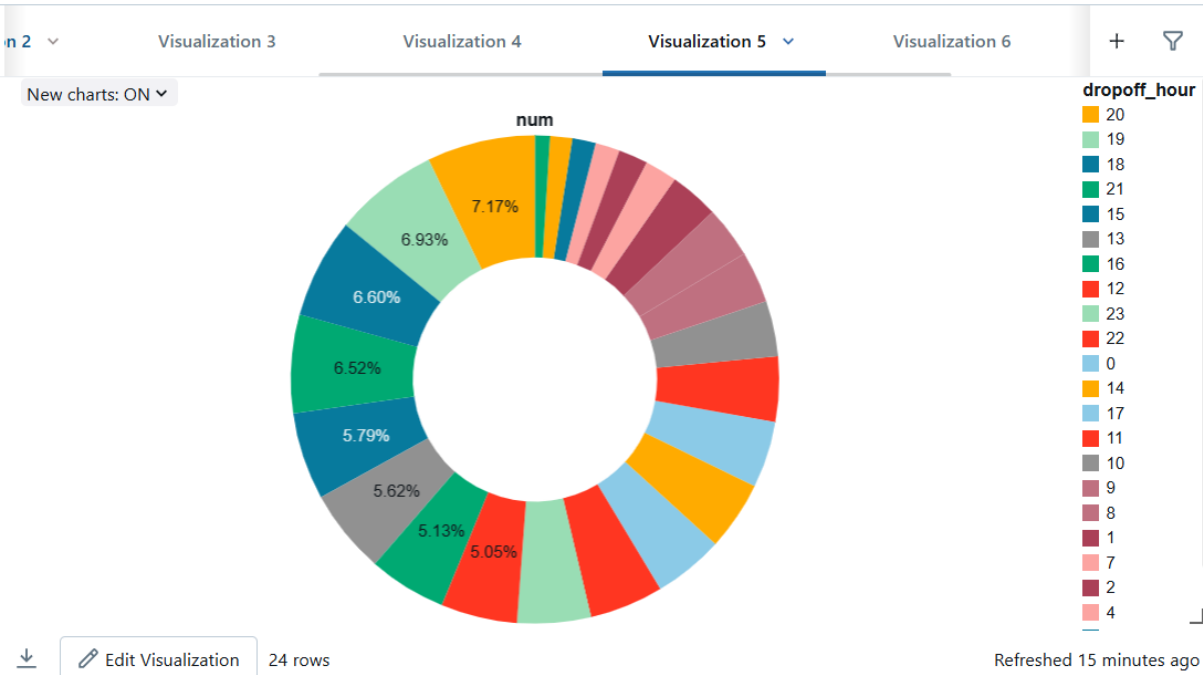
Bar Chart



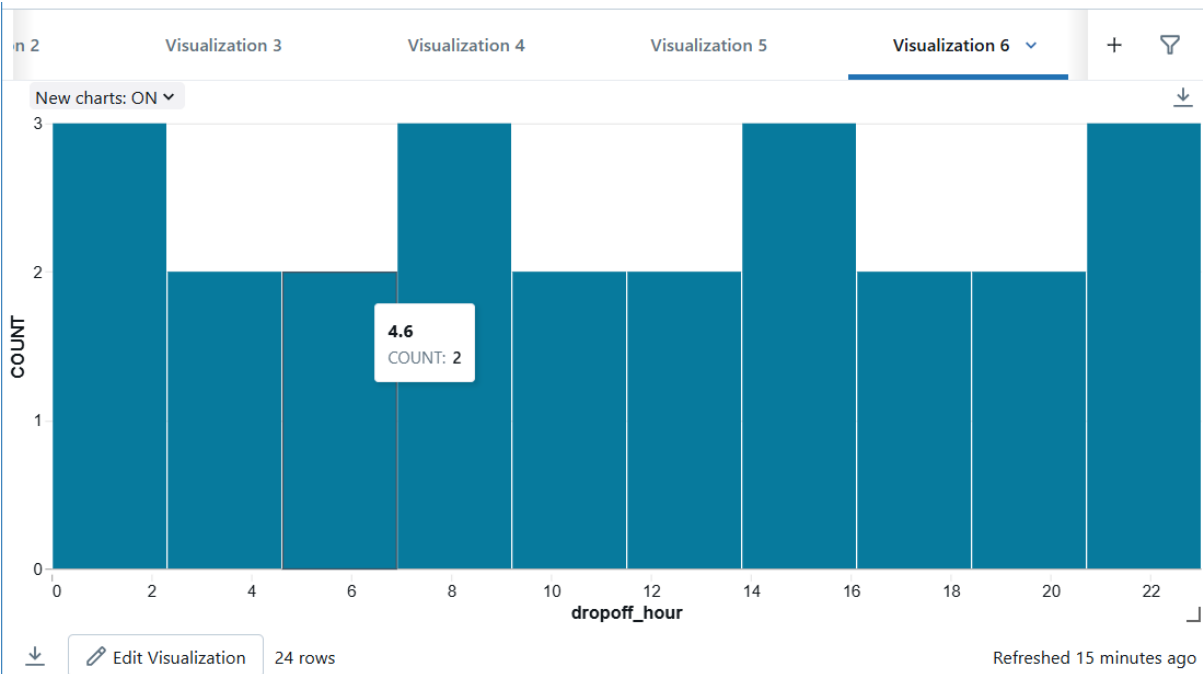
Area Chart



Pie Chart



Histogram



Box Chart

