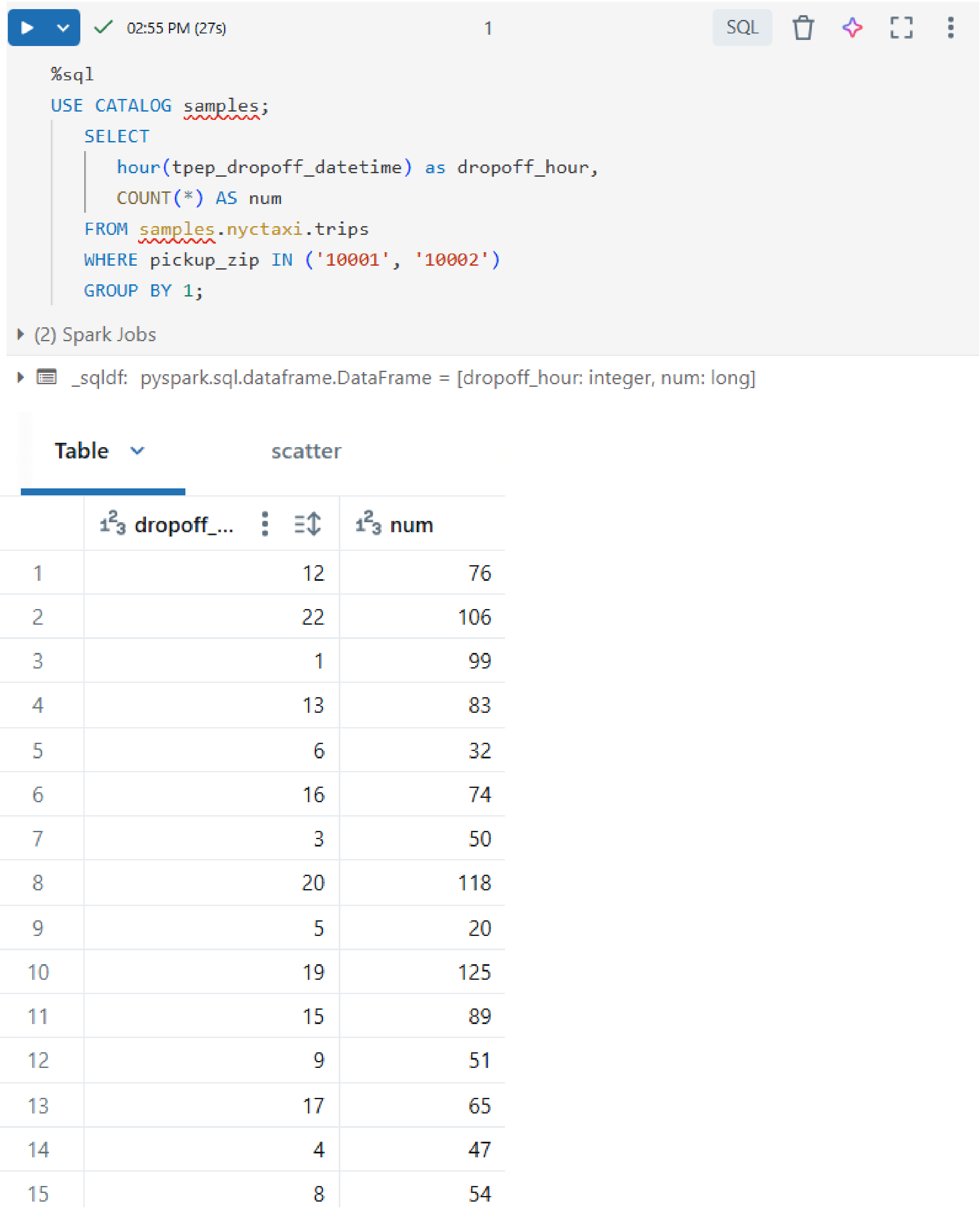
Assignment 6 - EDA and Visualization

Generate a result set to visualize - PRACTICE

SQL Query

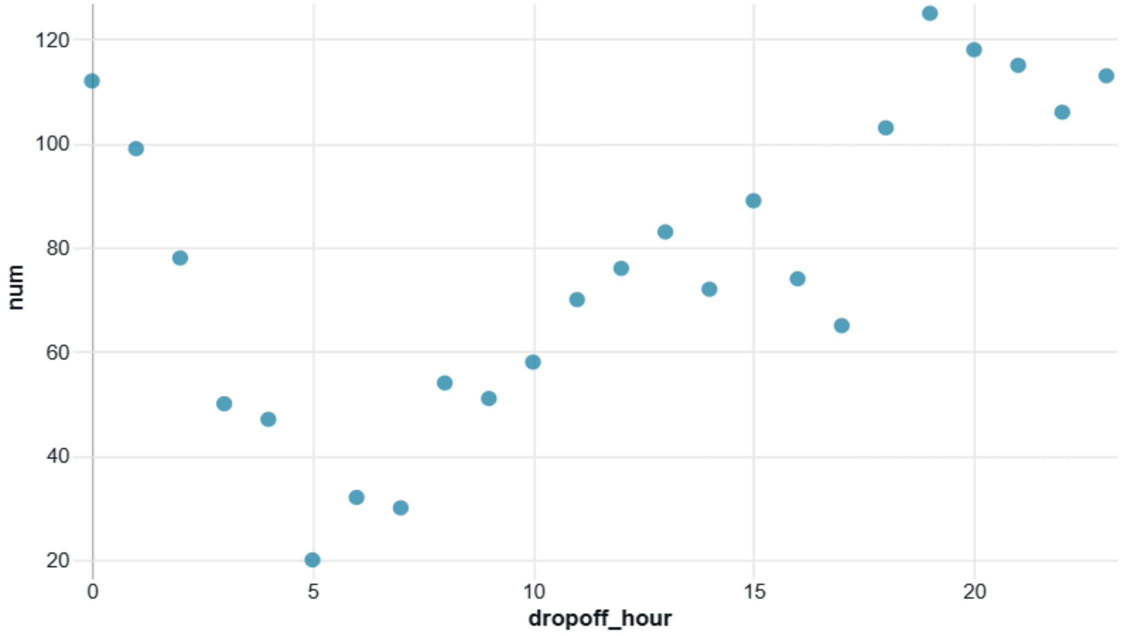


```
%sql
USE CATALOG samples;
  SELECT
    hour(tpep_dropoff_datetime) as dropoff_hour,
    COUNT(*) AS num
  FROM samples.nyctaxi.trips
  WHERE pickup_zip IN ('10001', '10002')
  GROUP BY 1;
```

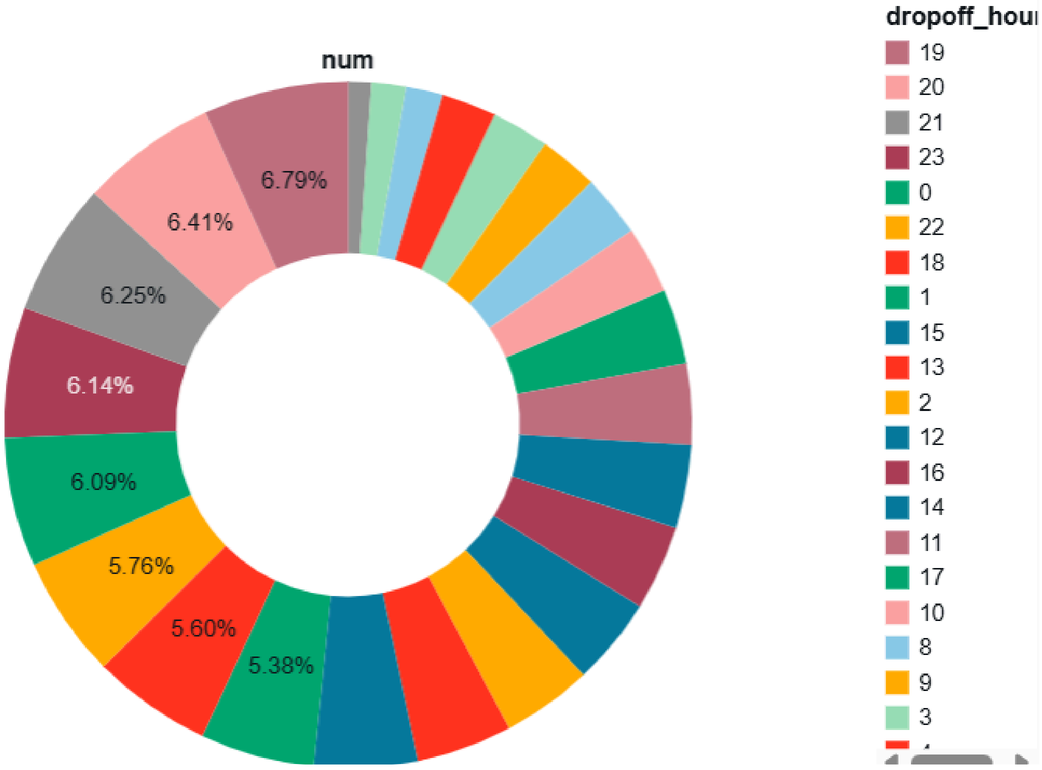
```
_sqldf:  pyspark.sql.dataframe.DataFrame = [dropoff_hour: integer, num: long]
```

	dropoff_...	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
15	8	54

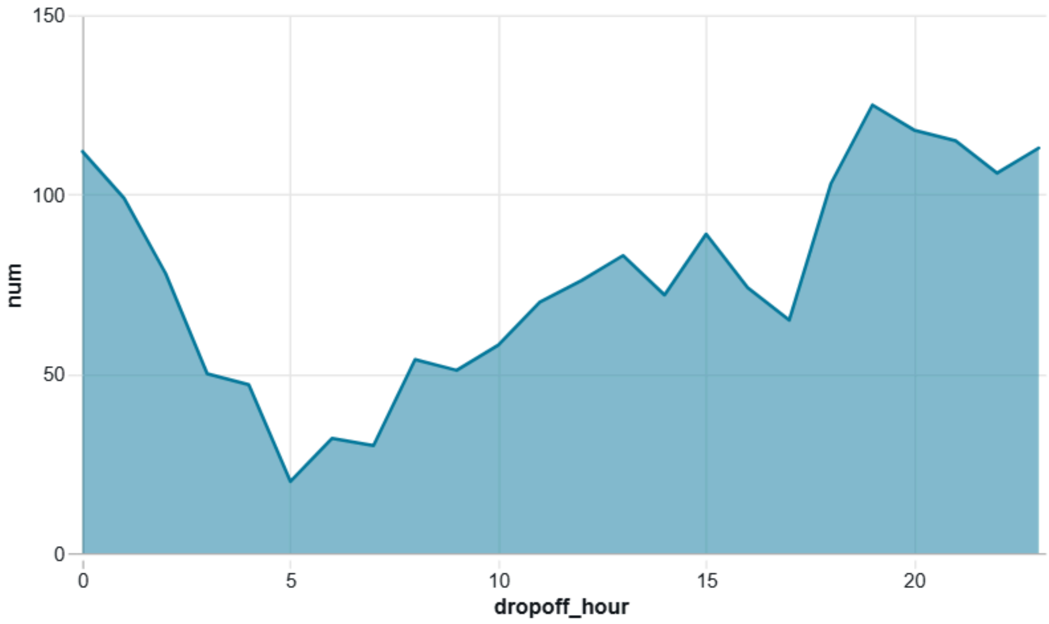
SCATTER PLOT



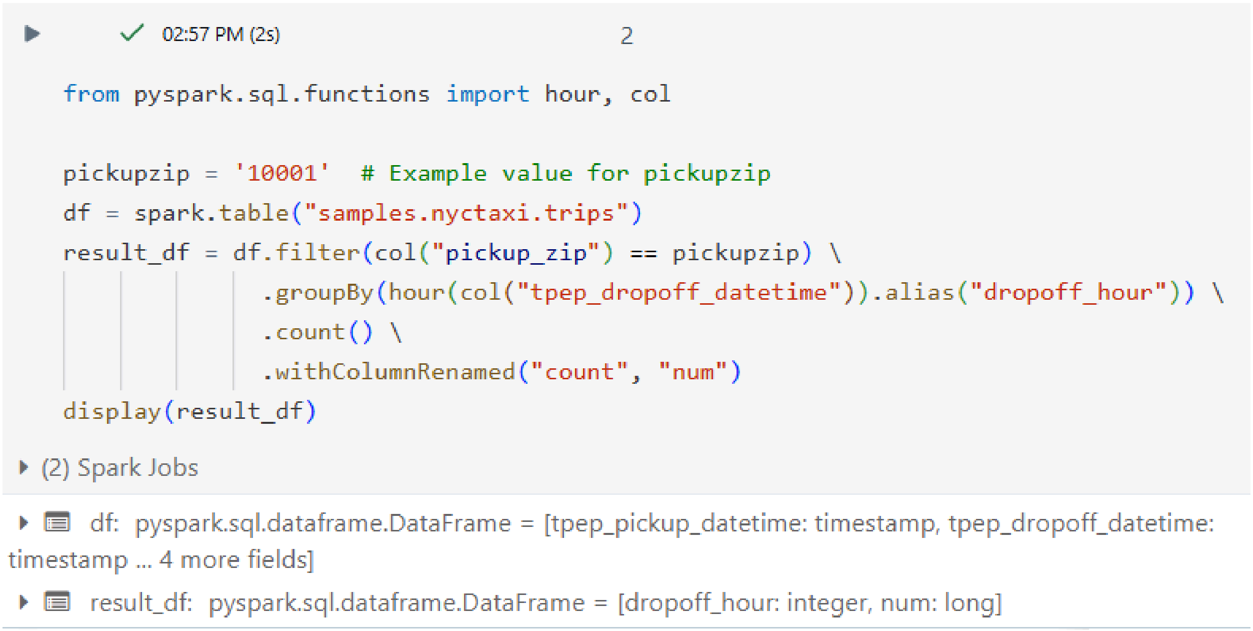
PIE CHART - MIN ( )



AREA CHART



PYTHON CODE :



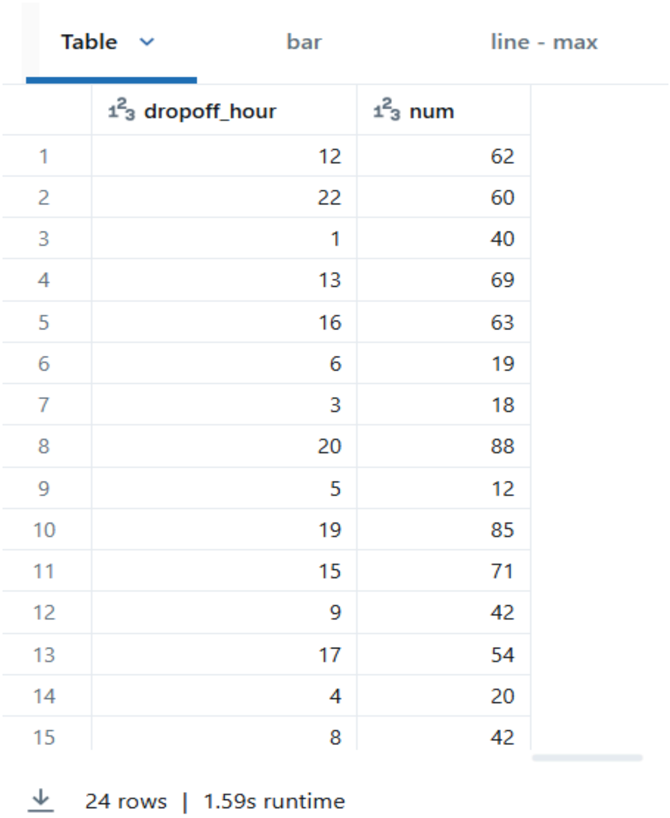
```
from pyspark.sql.functions import hour, col

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]

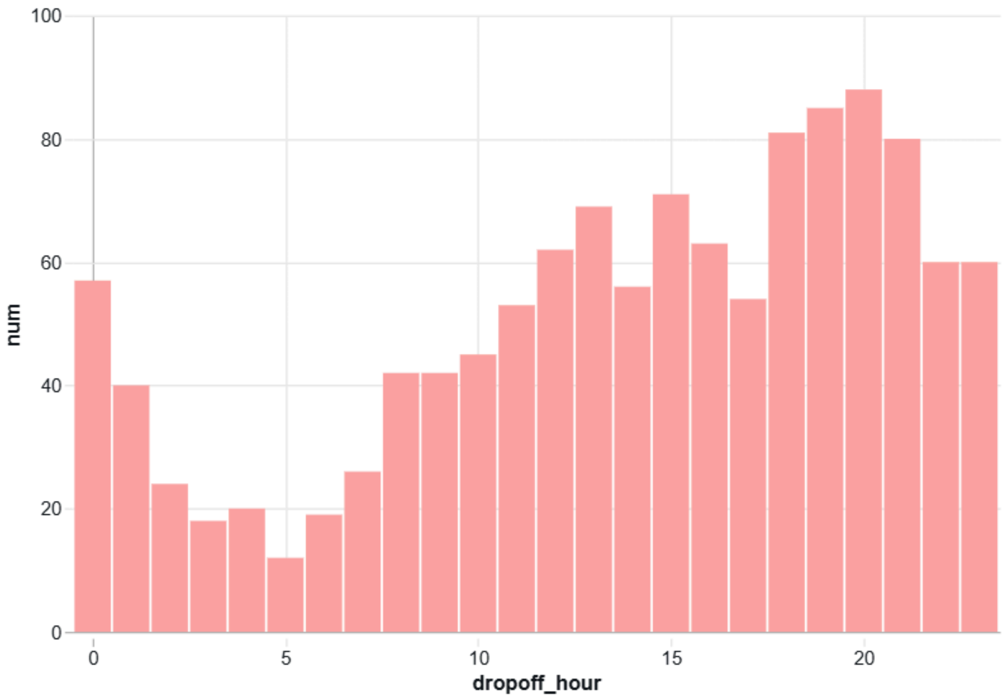
result_df: pyspark.sql.dataframe.DataFrame = [dropoff_hour: integer, num: long]



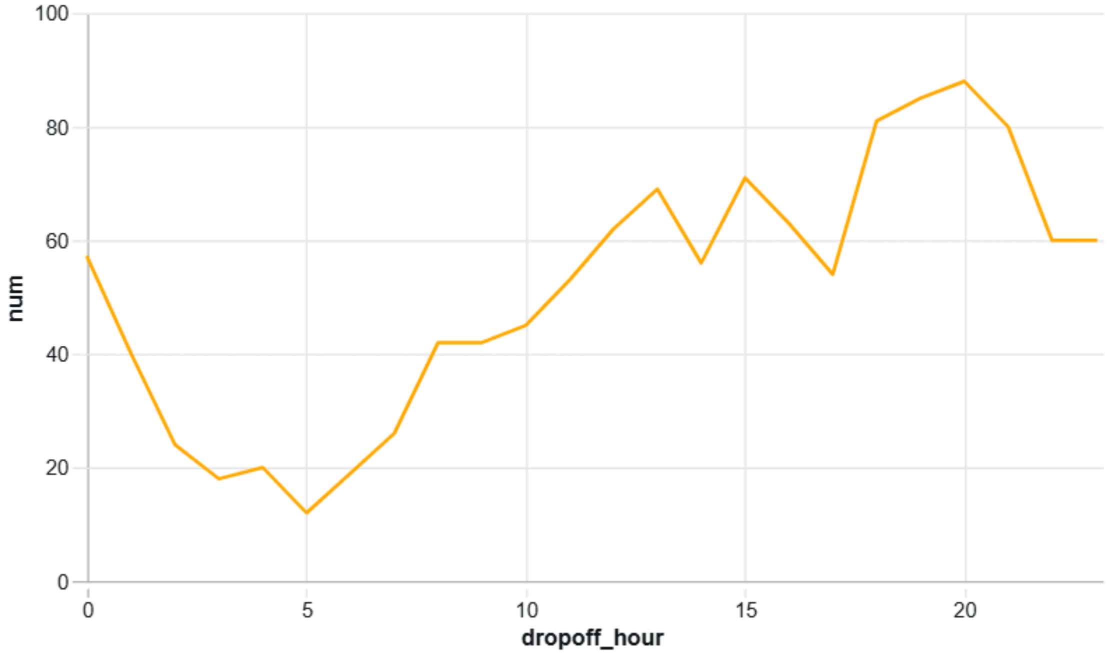
	dropoff_hour	num
1	12	62
2	22	60
3	1	40
4	13	69
5	16	63
6	6	19
7	3	18
8	20	88
9	5	12
10	19	85
11	15	71
12	9	42
13	17	54
14	4	20
15	8	42

24 rows | 1.59s runtime

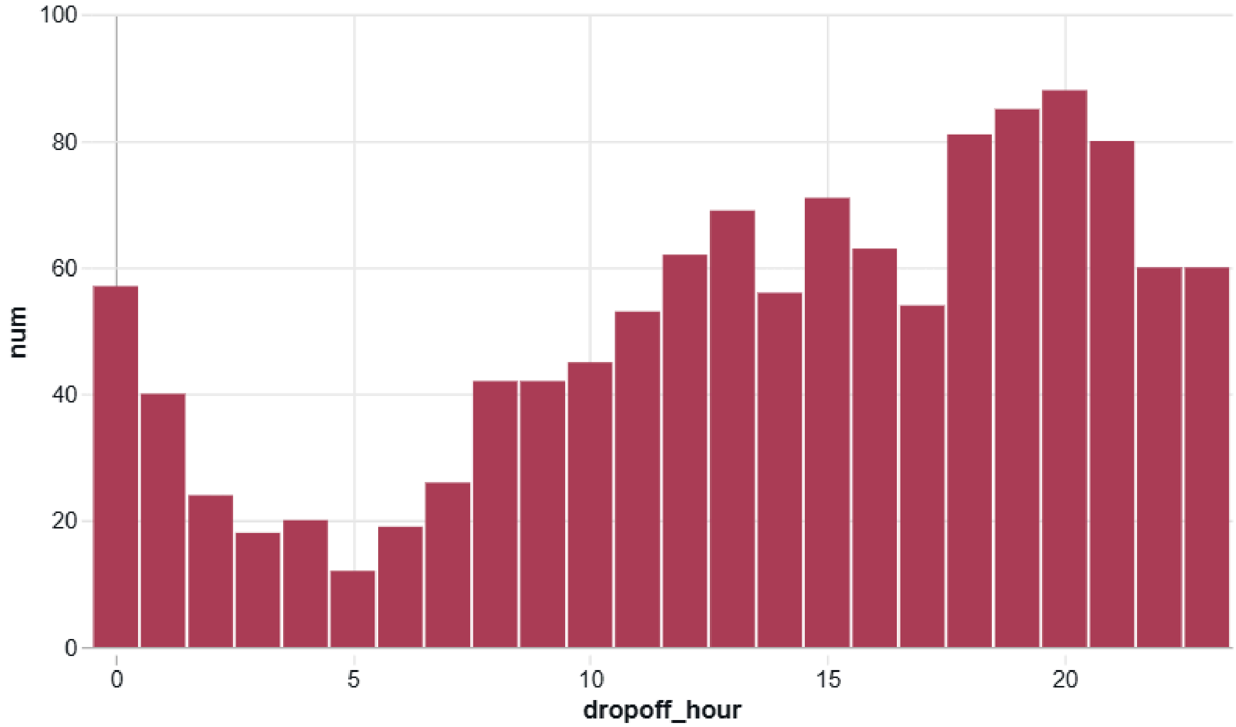
BAR GRAPH



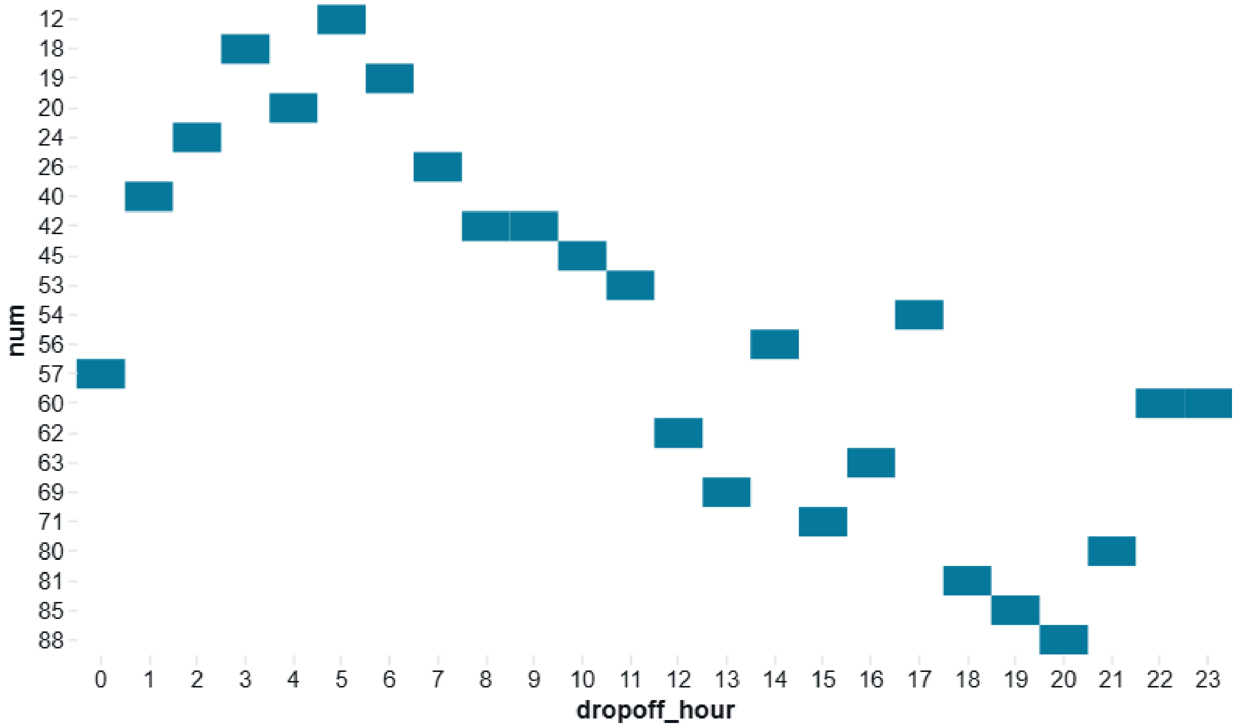
LINE CHART - MAX ( )



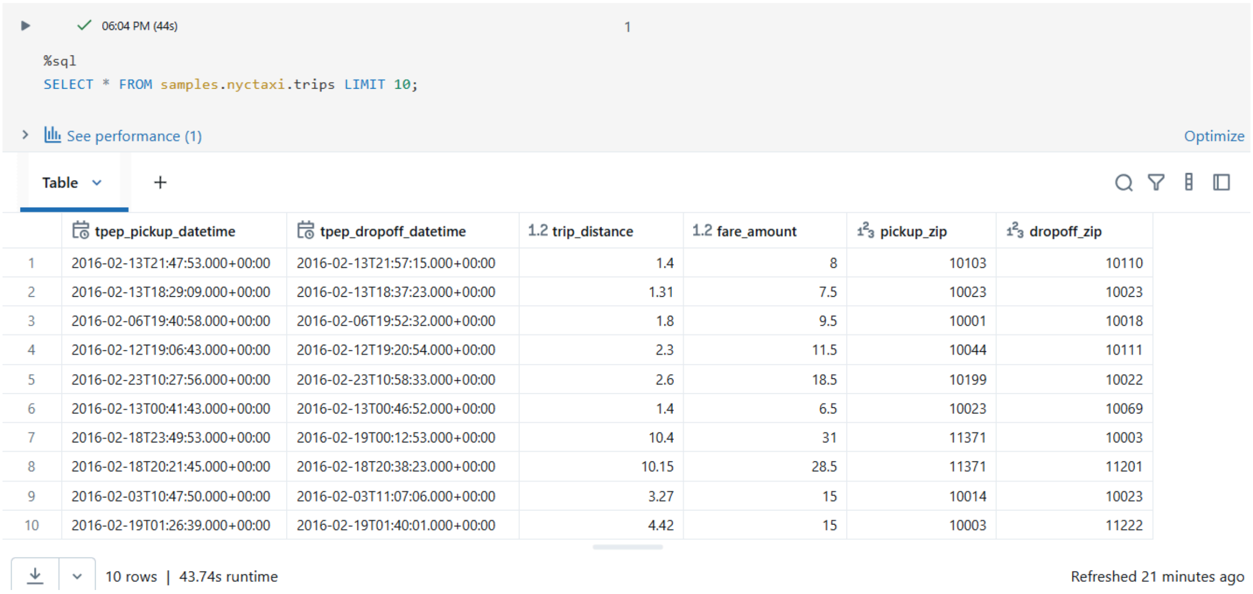
COMBO CHART - WITH RED COLOUR



HEAP CHART



QUESTIONS :



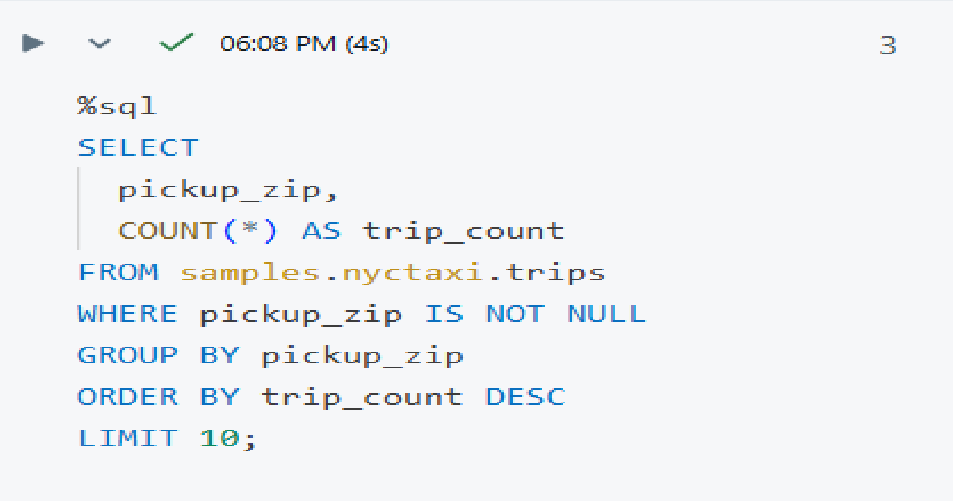
```
%sql
SELECT * FROM samples.nyctaxi.trips LIMIT 10;
```

	tpep_pickup_datetime	tpep_dropoff_datetime	1.2 trip_distance	1.2 fare_amount	pickup_zip	dropoff_zip
1	2016-02-13T21:47:53.000+00:00	2016-02-13T21:57:15.000+00:00	1.4	8	10103	10110
2	2016-02-13T18:29:09.000+00:00	2016-02-13T18:37:23.000+00:00	1.31	7.5	10023	10023
3	2016-02-06T19:40:58.000+00:00	2016-02-06T19:52:32.000+00:00	1.8	9.5	10001	10018
4	2016-02-12T19:06:43.000+00:00	2016-02-12T19:20:54.000+00:00	2.3	11.5	10044	10111
5	2016-02-23T10:27:56.000+00:00	2016-02-23T10:58:33.000+00:00	2.6	18.5	10199	10022
6	2016-02-13T00:41:43.000+00:00	2016-02-13T00:46:52.000+00:00	1.4	6.5	10023	10069
7	2016-02-18T23:49:53.000+00:00	2016-02-19T00:12:53.000+00:00	10.4	31	11371	10003
8	2016-02-18T20:21:45.000+00:00	2016-02-18T20:38:23.000+00:00	10.15	28.5	11371	11201
9	2016-02-03T10:47:50.000+00:00	2016-02-03T11:07:06.000+00:00	3.27	15	10014	10023
10	2016-02-19T01:26:39.000+00:00	2016-02-19T01:40:01.000+00:00	4.42	15	10003	11222

10 rows | 43.74s runtime

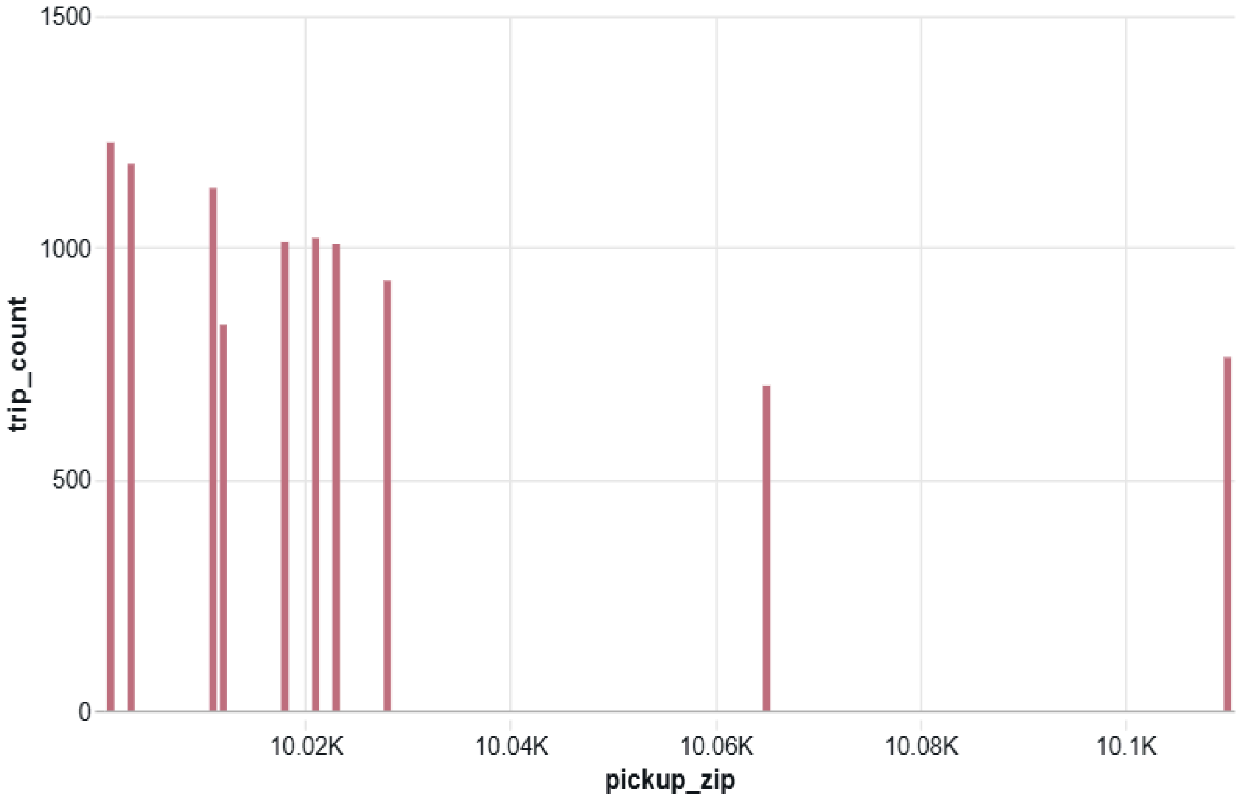
Refreshed 21 minutes ago

1. Top 10 Pickup Zip Codes by Trip Count

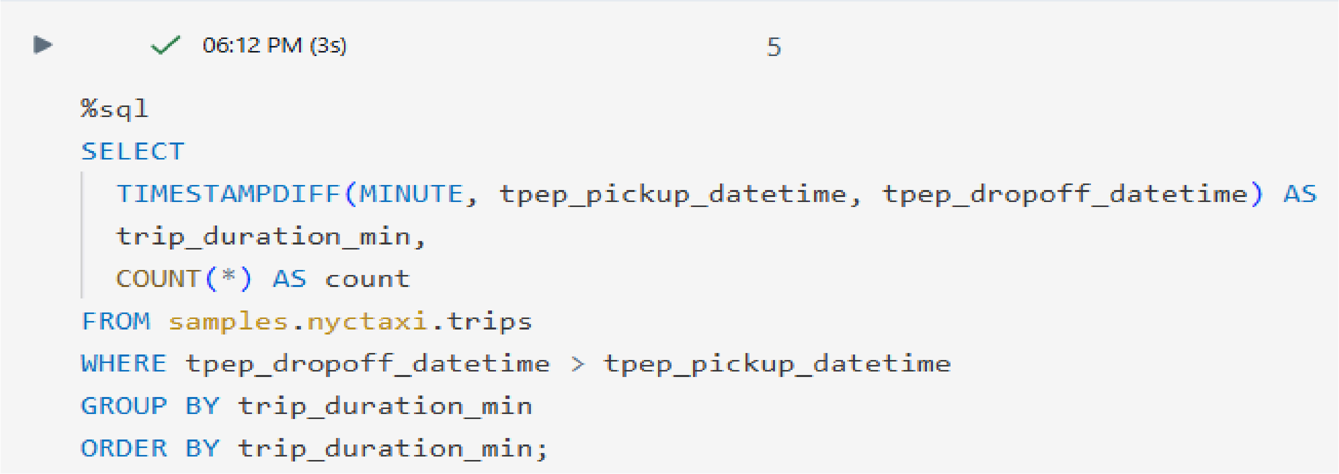


```
%sql
SELECT
  pickup_zip,
  COUNT(*) AS trip_count
FROM samples.nyctaxi.trips
WHERE pickup_zip IS NOT NULL
GROUP BY pickup_zip
ORDER BY trip_count DESC
LIMIT 10;
```

OUTPUT :

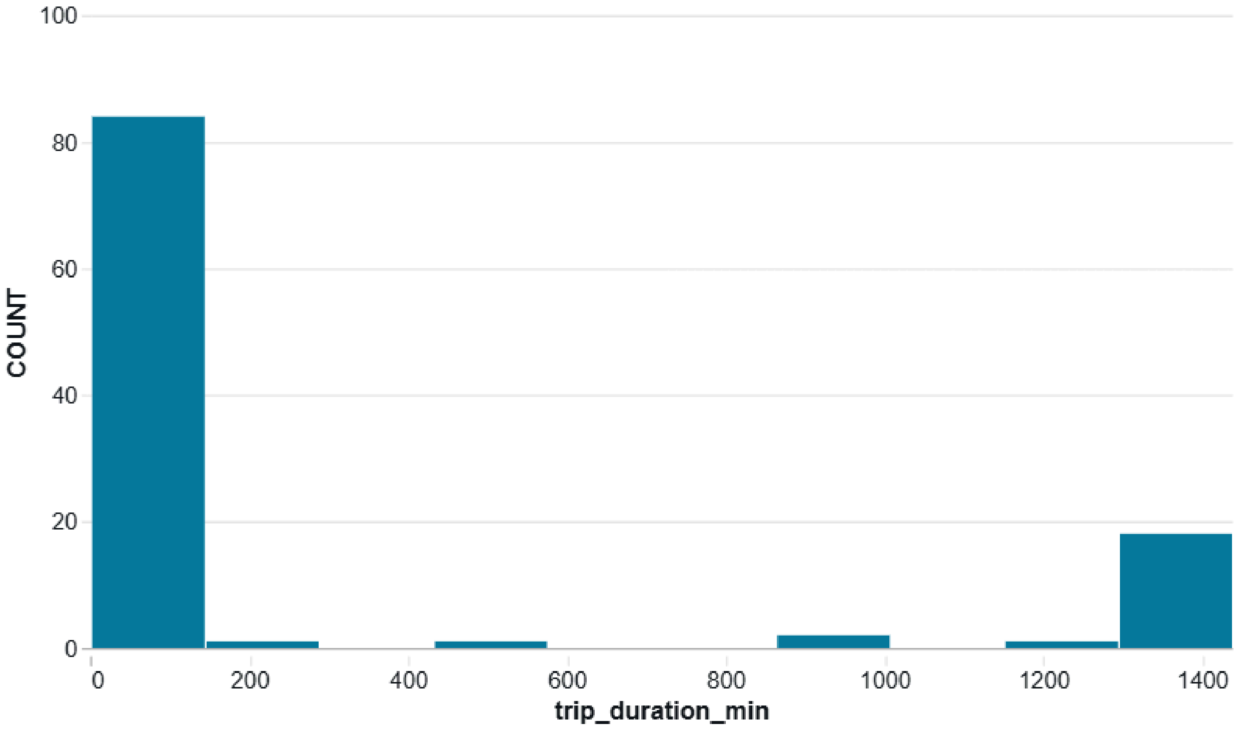


2. Top 10 Drop-off Zip Codes by Trip Count

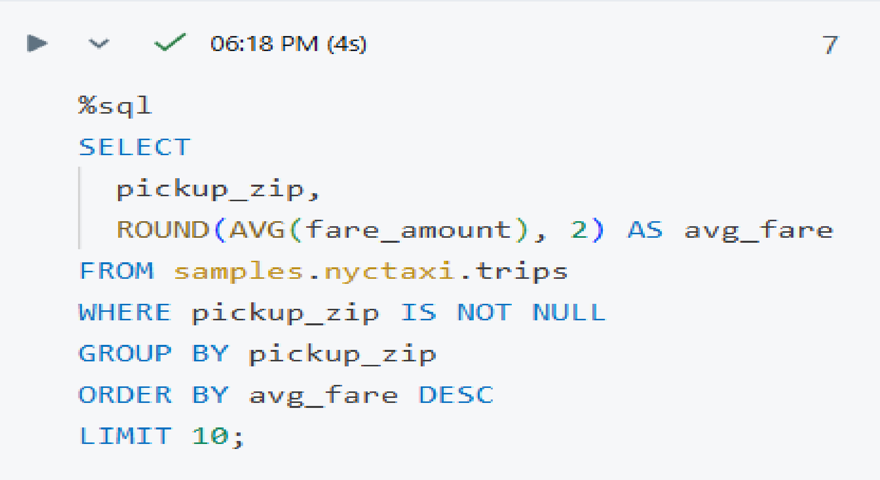


```
%sql
SELECT
  TIMESTAMPDIFF(MINUTE, tpep_pickup_datetime, tpep_dropoff_datetime) AS
  trip_duration_min,
  COUNT(*) AS count
FROM samples.nyctaxi.trips
WHERE tpep_dropoff_datetime > tpep_pickup_datetime
GROUP BY trip_duration_min
ORDER BY trip_duration_min;
```

OUTPUT :

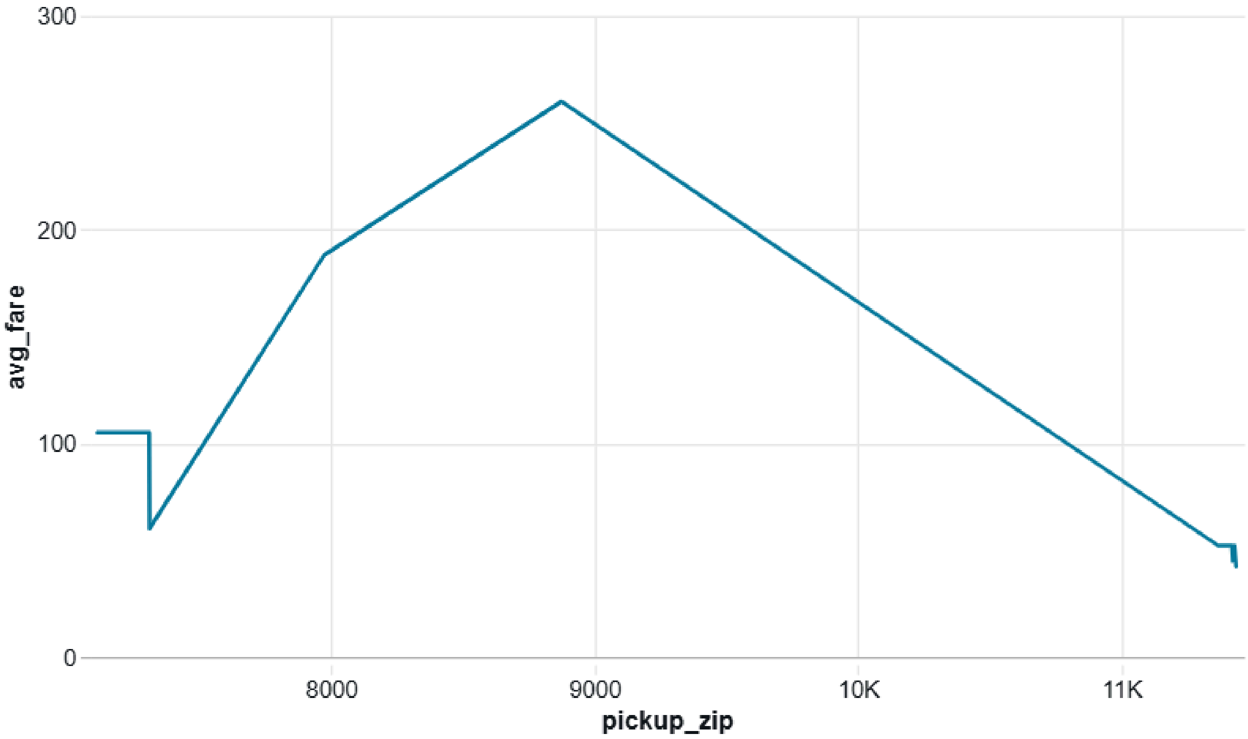


3. Average Fare by Pickup Zip Code

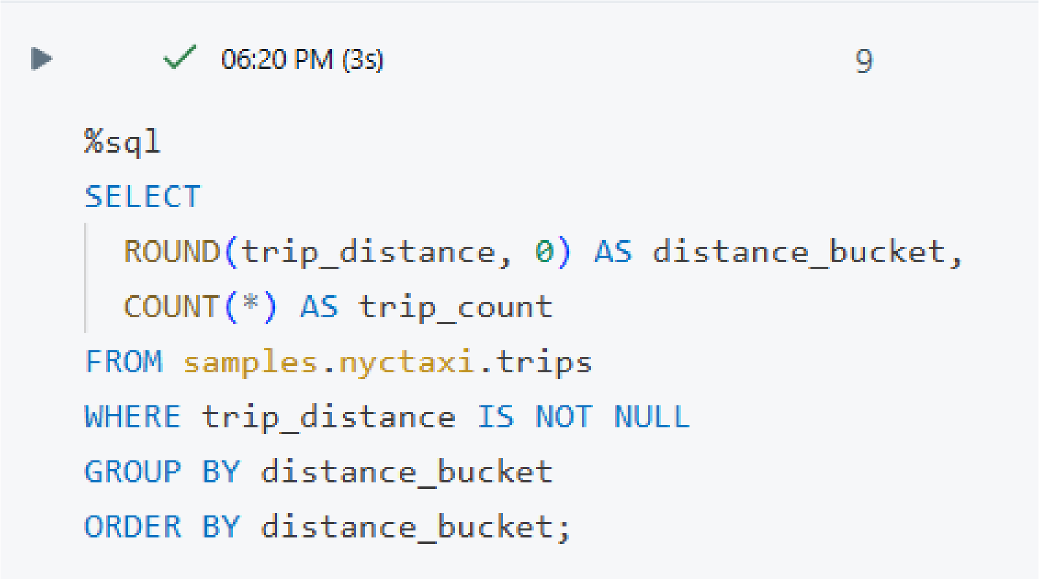


```
%sql
SELECT
  pickup_zip,
  ROUND(AVG(fare_amount), 2) AS avg_fare
FROM samples.nyctaxi.trips
WHERE pickup_zip IS NOT NULL
GROUP BY pickup_zip
ORDER BY avg_fare DESC
LIMIT 10;
```

OUTPUT :

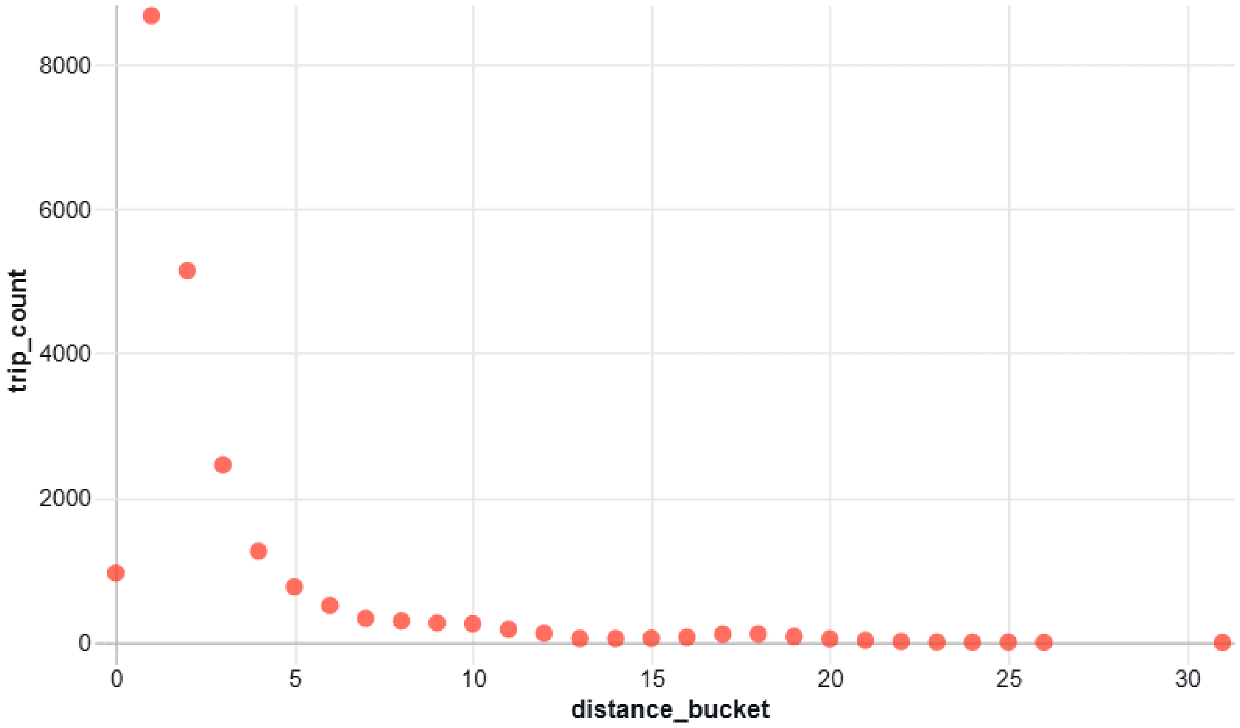


4. Trip Distance Distribution

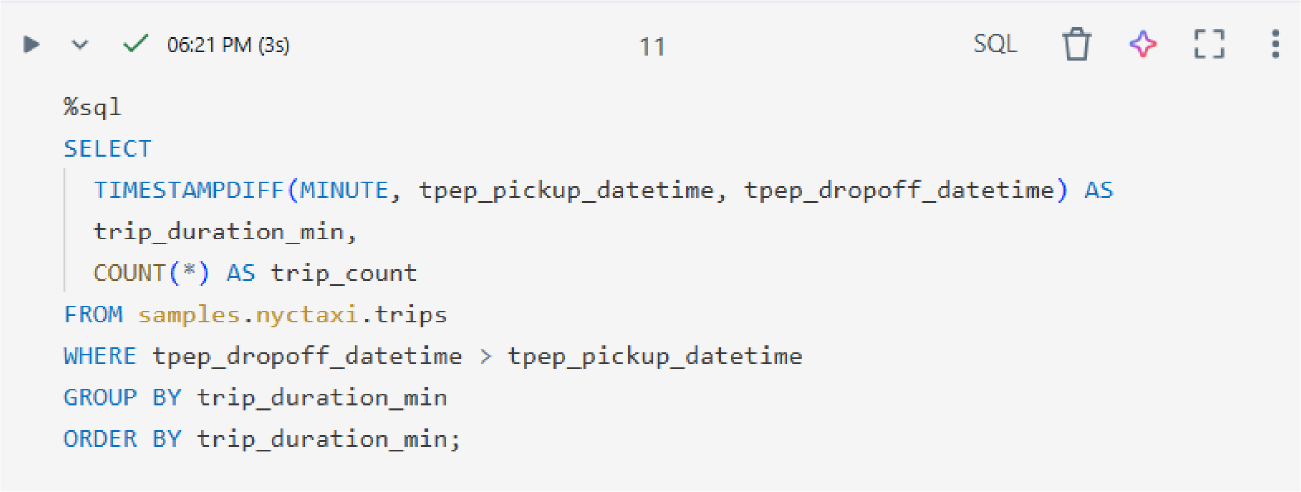


```
%sql
SELECT
  ROUND(trip_distance, 0) AS distance_bucket,
  COUNT(*) AS trip_count
FROM samples.nyctaxi.trips
WHERE trip_distance IS NOT NULL
GROUP BY distance_bucket
ORDER BY distance_bucket;
```

OUTPUT :

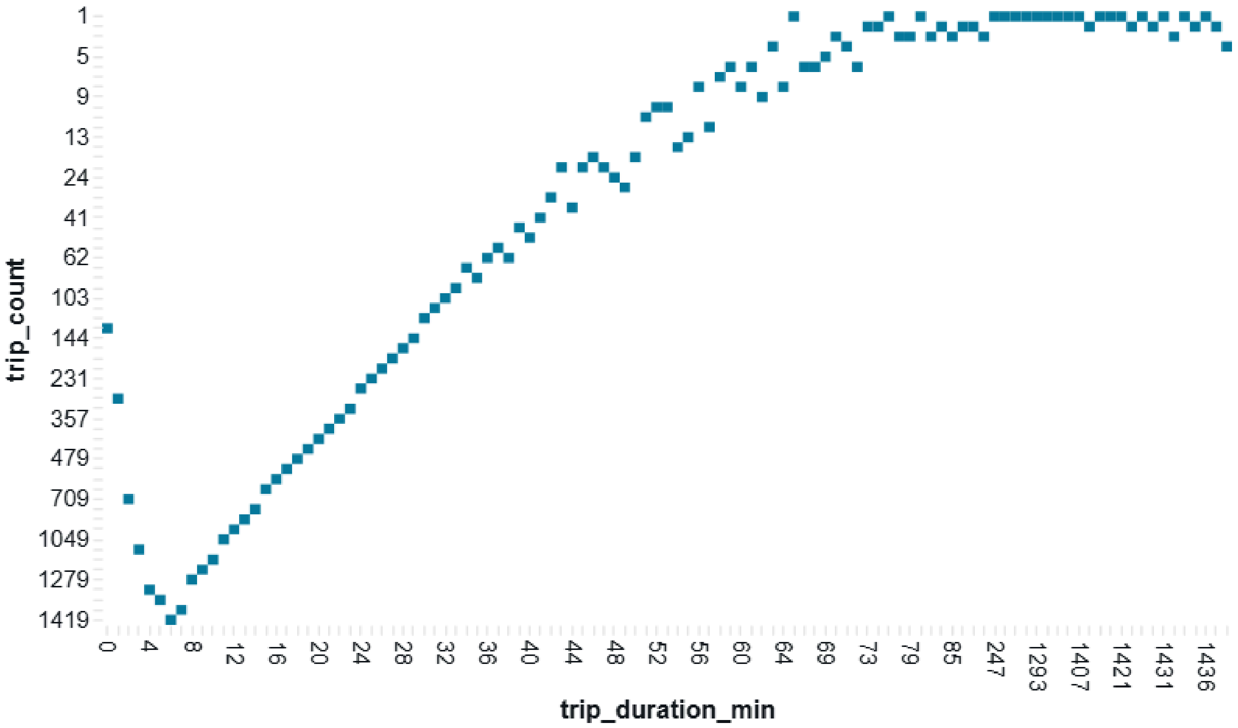


5. Trip Duration in Minutes

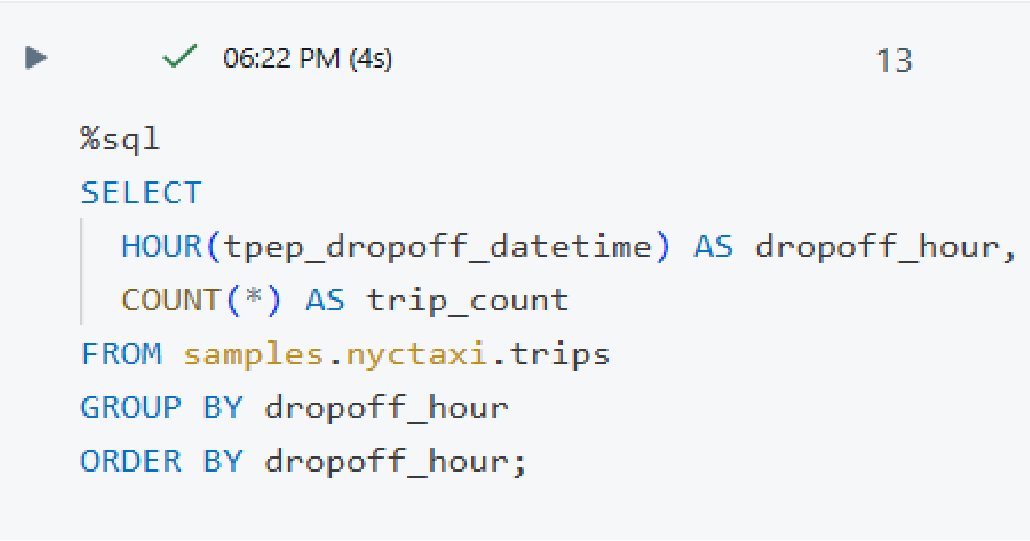


```
%sql
SELECT
  TIMESTAMPDIFF(MINUTE, tpep_pickup_datetime, tpep_dropoff_datetime) AS
  trip_duration_min,
  COUNT(*) AS trip_count
FROM samples.nyctaxi.trips
WHERE tpep_dropoff_datetime > tpep_pickup_datetime
GROUP BY trip_duration_min
ORDER BY trip_duration_min;
```

OUTPUT :

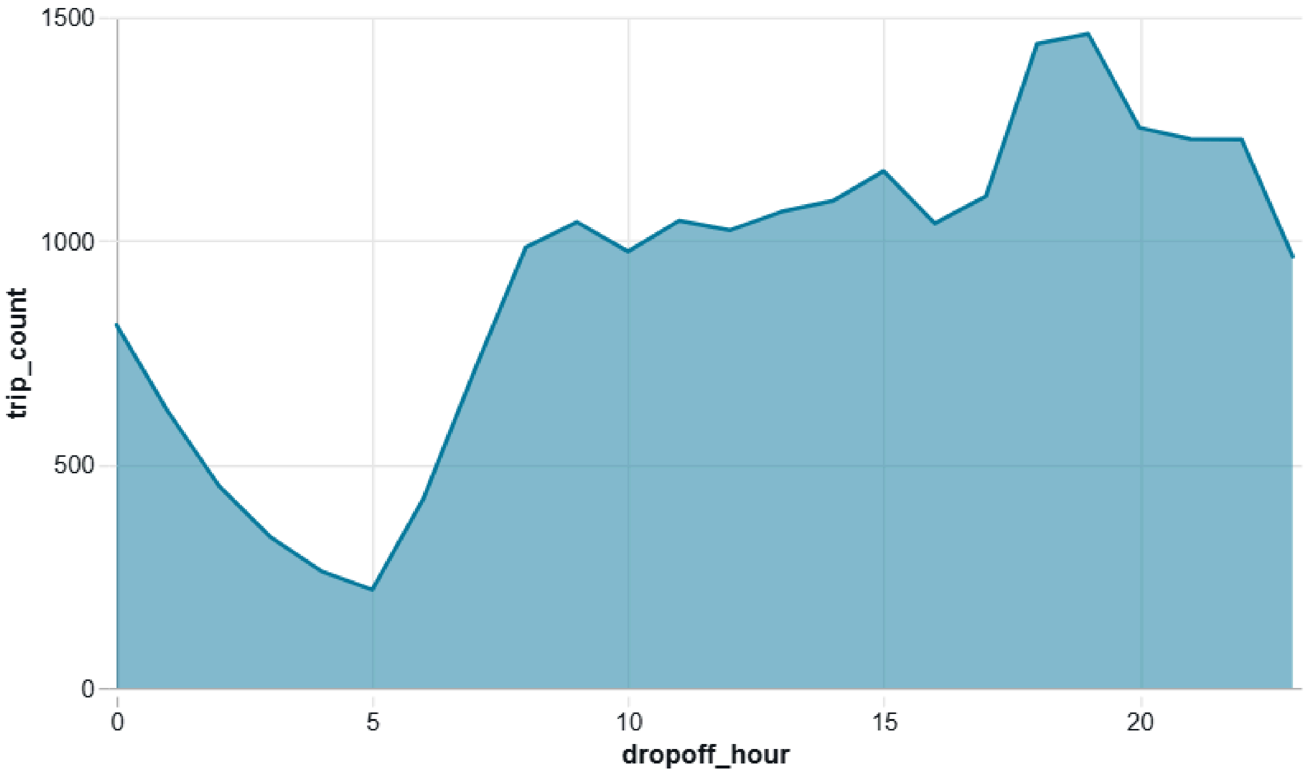


6. Number of Trips by Drop-off Hour

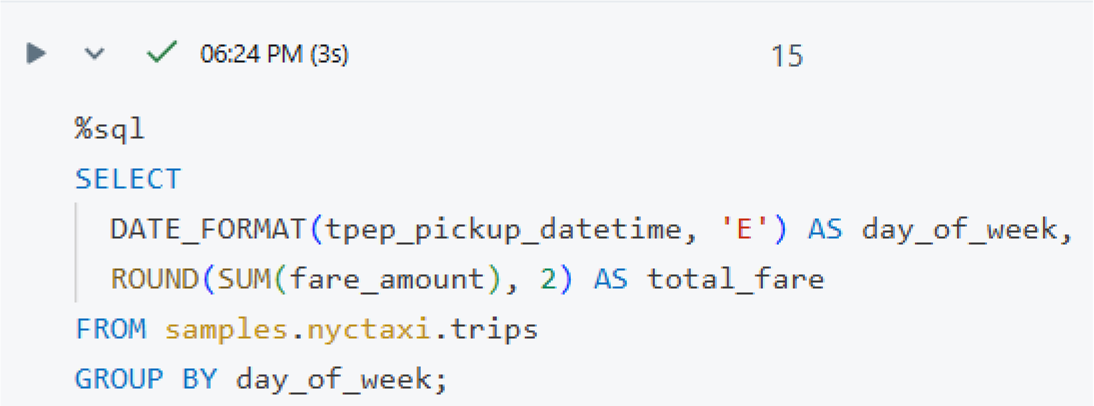


```
%sql
SELECT
  HOUR(tpep_dropoff_datetime) AS dropoff_hour,
  COUNT(*) AS trip_count
FROM samples.nyctaxi.trips
GROUP BY dropoff_hour
ORDER BY dropoff_hour;
```

OUTPUT :

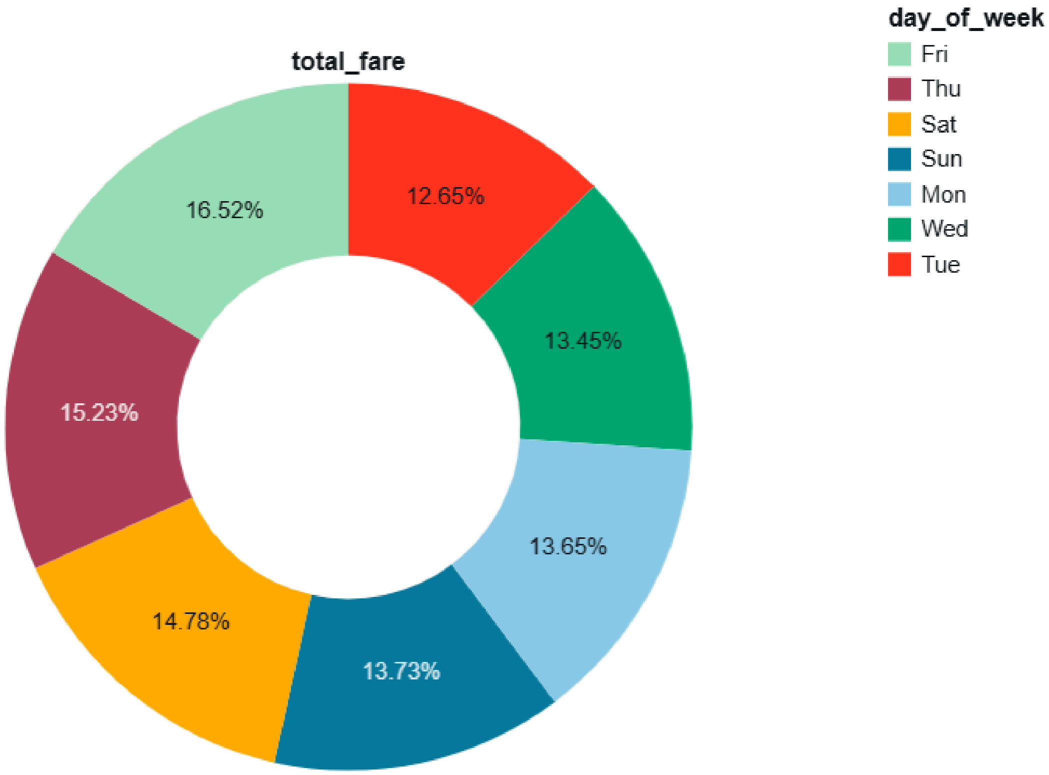


7. Total Fare Amount by Day of Week



```
%sql
SELECT
  DATE_FORMAT(tpep_pickup_datetime, 'E') AS day_of_week,
  ROUND(SUM(fare_amount), 2) AS total_fare
FROM samples.nyctaxi.trips
GROUP BY day_of_week;
```

OUTPUT :


total_fare
day_of_week
Fri
Thu
Sat
Sun
Mon
Wed
Tue