Trip Analysis Report – PySpark on Google Colab

Problem Statement

Analyze the trip dataset using PySpark and present: 1) Top 10 busiest pickup locations 2) Top 10 busiest drop locations 3) City-wise distribution of trips 4) Payment type distribution (Cash vs Online) 5) Trips with highest fare amounts

Step 1: Load & Schema

We load the trip CSV data into Spark with explicit schema to avoid type issues.

Step 2: Clean & Normalize

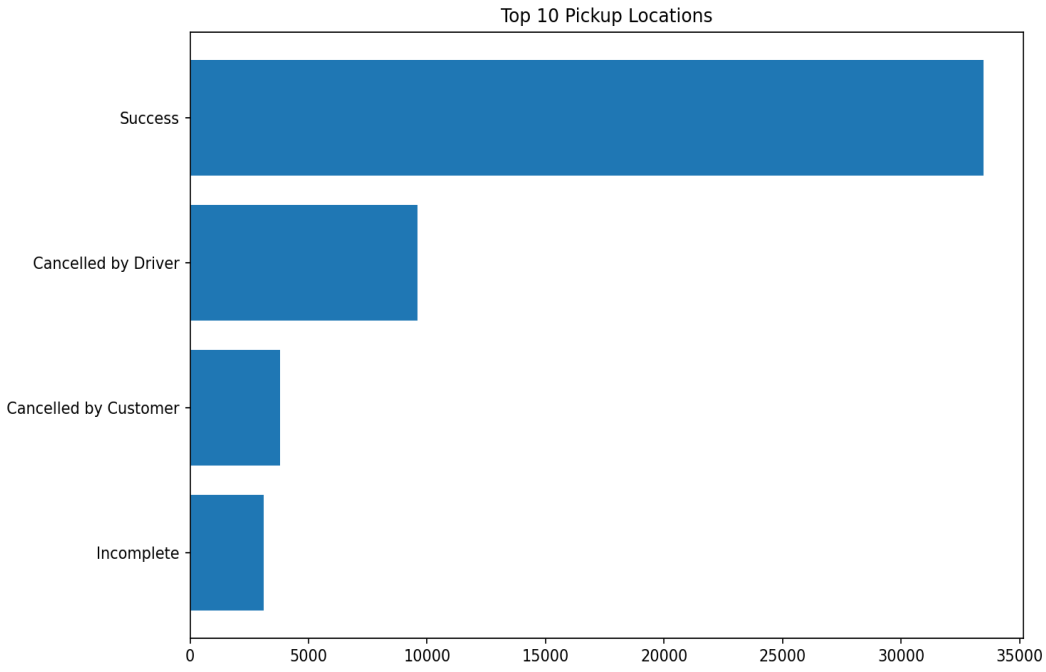
We filter out trips with null pickup/drop times or fares, normalize fare amounts.

Step 3: Compute Aggregations

We calculate top locations, city-wise summaries, payment type distribution, and high fare trips.

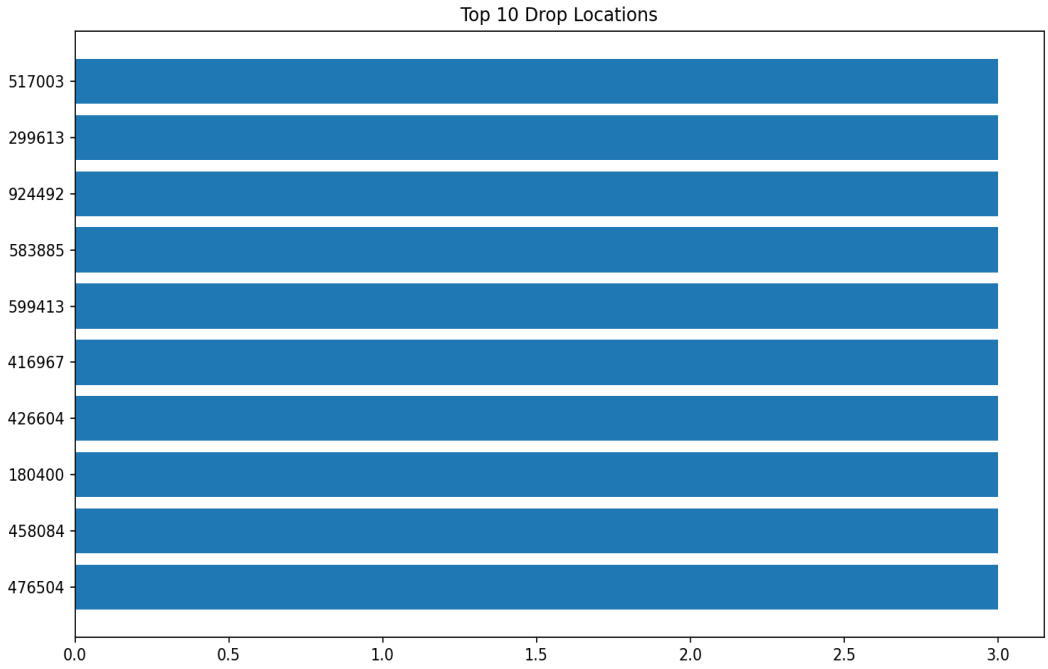
Top 10 Pickup Locations

Pickup_Location	trip_count
Success	33484
Cancelled by Driver	9610
Cancelled by Customer	3799
Incomplete	3106



Top 10 Drop Locations

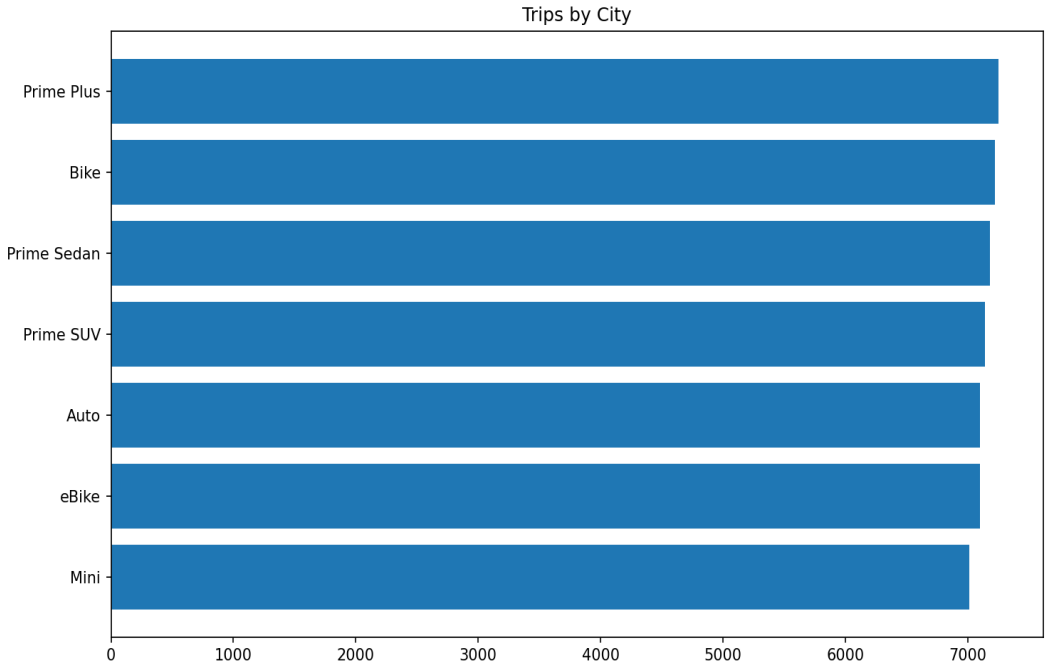
Drop_Location	trip_count
517003	3
299613	3
924492	3
583885	3
599413	3
416967	3
426604	3
180400	3
458084	3
476504	3



Trips by City

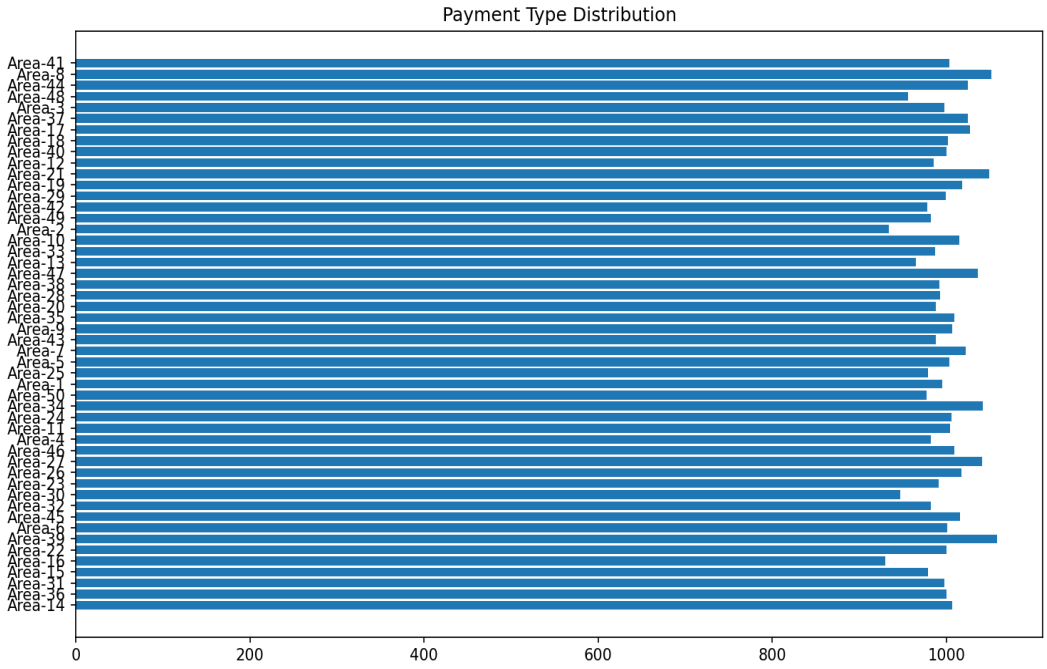
City	num_trips	total_fare
Prime Plus	7252	
Bike	7223	
Prime Sedan	7179	
Prime SUV	7140	
Auto	7098	
eBike	7097	

City	num_trips	total_fare
Mini	7010	



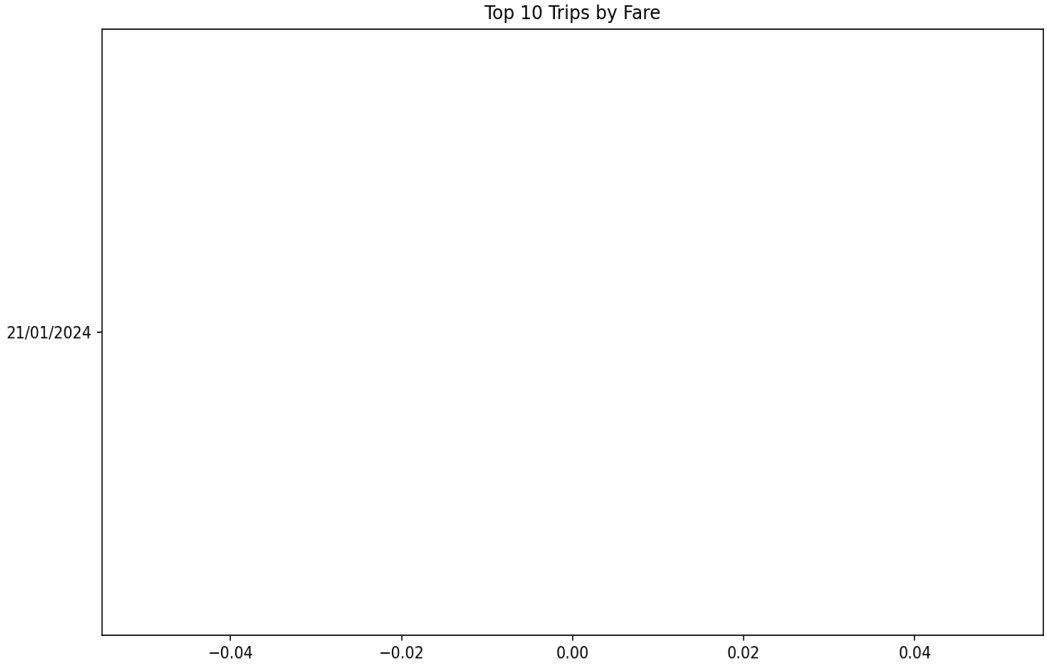
Payment Type Distribution

Payment_Type	num_trips
Area-41	1003
Area-8	1052
Area-44	1025
Area-48	956
Area-3	998
Area-37	1025
Area-17	1027
Area-18	1002
Area-40	1000
Area-12	985



Top 10 Trips by Fare

Trip_ID	Pickup_DateTime	Drop_DateTime	Pickup_Location	Drop_Location	City	Fare_Amount	Payment_Type	fare_num
21/01/2024	2025-09-24 15:00:00	NaT	Success	768113	Mini		Area-25	
28/01/2024	2025-09-24 06:00:00	NaT	Success	329258	Auto		Area-2	
27/01/2024	2025-09-24 19:00:00	NaT	Success	724658	Prime Plus		Area-9	
26/01/2024	2025-09-24 03:00:00	NaT	Cancelled by Driver	201414	Mini		Area-6	
19/01/2024	2025-09-24 20:00:00	NaT	Success	535163	Prime Plus		Area-19	
15/01/2024	2025-09-24 16:00:00	NaT	Cancelled by Driver	301629	Bike		Area-24	
18/01/2024	2025-09-24 15:00:00	NaT	Success	998107	Auto		Area-28	
02/01/2024	2025-09-24 22:00:00	NaT	Cancelled by Driver	319684	Prime Sedan		Area-24	
22/01/2024	2025-09-24 21:00:00	NaT	Success	260604	Prime Sedan		Area-43	
30/01/2024	2025-09-24 22:00:00	NaT	Incomplete	330283	Bike		Area-45	



Conclusions & Insights

Total trips analyzed: **49999**.

Total fare amount (sum): **0.0**.

Cash trips share: **n/a**; Online trips share: **n/a**.