

MIS 6309 – Business Data Warehousing

Final Project -Yellow Taxi Dataset

Group 13:

Krishna Kanth Bandi –KXB220036

Alaparthi Charan Sai-AXX220013

Sree Varsha Rayavarapu-VXR210033

Sai Madhan Muthyam-SXM22016

Sai Srinivas Penumaka -SXP210153

Contents

About Yellow Taxi Data----- 3

Logical Design ----- 4

ER Diagram ----- 5

Insights using SQL-----6-11

Tableau Insights-----11-13

About Yellow Taxi data:

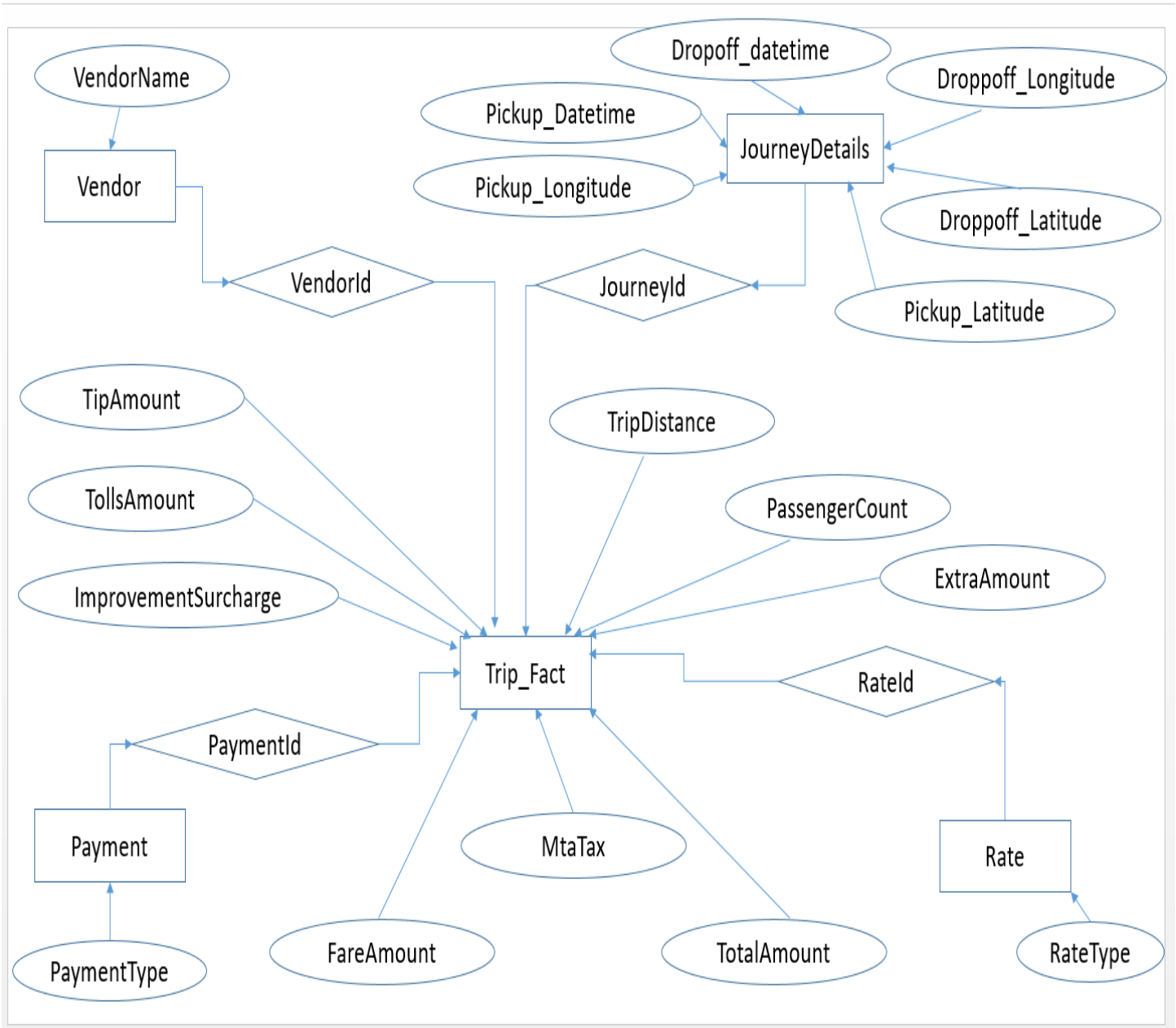
This paper will present a complete examination of the NYC Yellow Taxi Trip Data, a rich dataset that captures the minute details of taxi rides conducted by New York City's distinctive yellow cabs. We research and reveal significant insights into the city's taxi industry's mobility patterns, passenger behavior, and general dynamics by this huge dataset.

Transportation networks are critical to the operation and efficiency of a thriving city such as New York City. Taxis, notably the distinctive yellow taxis, are a popular method of transportation for both locals and visitors. The NYC Yellow Taxi Trip Data provides a one-of-a-kind chance to look into the massive quantity of data collected from these journeys, shining light on numerous facets of the city's transportation scene.

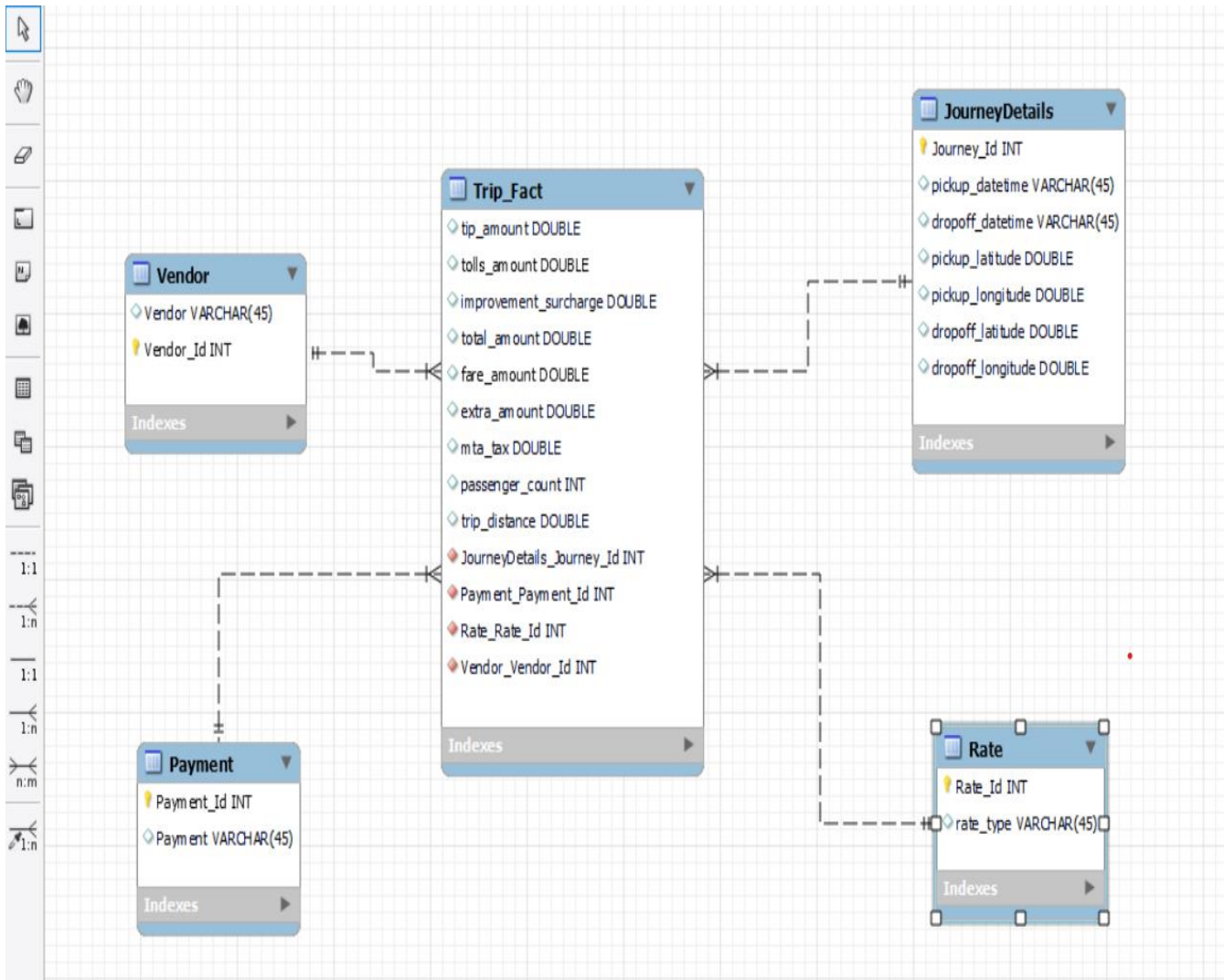
Furthermore, we investigate the fee structure and pricing dynamics of yellow cabs, looking at parameters such as distance traveled, time of day, and traffic congestion. Through this investigation, we want to acquire insights into fee optimization tactics and potential changes to the tariff computation system.

To extract useful insights from the NYC Yellow Taxi Trip Data, we use a variety of data visualization approaches, statistical analysis, and machine learning algorithms throughout this study. By combining data-driven analysis with domain expertise, we strive to provide a comprehensive and actionable assessment of New York City's yellow taxi industry.

Logical Design



ER Diagram:



Insights using SQL

Insight 1:

At what hour customers are traveling farthest and what the average fare amount is during that hour?

```
1 • SELECT DATE_FORMAT(tpep_pickup_datetime, '%H') AS pickup_hour,  
2 round(AVG(trip_distance),2) AS avg_distance,  
3 round(AVG(fare_amount),2) AS avg_fare  
4 FROM TAXI WHERE trip_distance < 1000  
5 GROUP BY pickup_hour  
6 order by avg_distance desc;  
7
```

Result Grid	Filter Rows:	Export:	Wrap Cell Content:
	pickup_hour	avg_distance	avg_fare
▶	05	4.52	15.64
	04	4.02	14.62
	06	3.48	12.87
	03	3.42	13.02
	00	3.25	12.71
	23	3.21	12.64
	01	3.21	12.57
	02	3.21	12.53
	22	3.04	12.21
	21	2.93	11.89
	07	2.83	11.76
	16	2.79	12.12
	15	2.76	12.27

From the query, the average trip distance and average fare amount for each pickup hour was found out. Analyzing taxi data by pickup hour, distance, and fare amount can provide valuable insights into customer behavior, pricing strategy, and operational efficiency for taxi companies. By understanding the busiest hours for pickups, companies can allocate their resources effectively. Analyzing the average fare amount by pickup hour can help companies set their pricing strategy, while analyzing the average trip distance can provide insights into customer behavior and travel patterns. Overall, this analysis can help taxi companies make more informed decisions about their operations, pricing, and marketing strategies, which can lead to better customer service and higher profits.

Insight 2:

What payment type is most preferred?

```
1 • SELECT payment_type, COUNT(*) AS num_trips  
2 FROM TAXI  
3 GROUP BY payment_type;
```

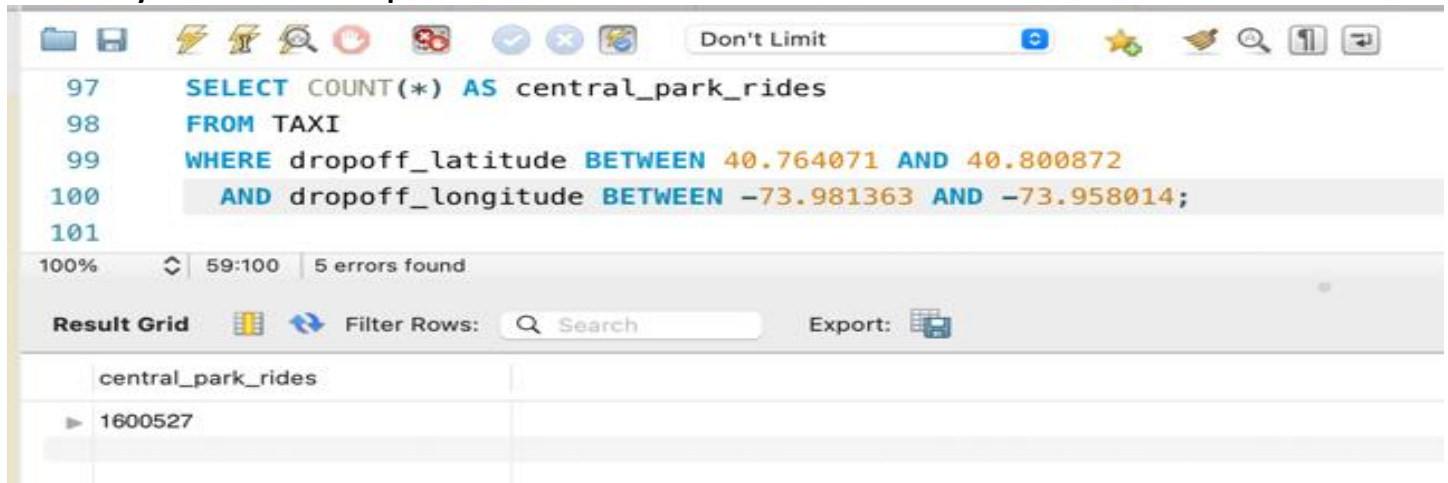
- 1 – Credit
- 2 – Cash
- 3 – No charge
- 4 – Dispute
- 5 -- Unknown

Result Grid	Filter Rows:	Export:	Wrap Cell Content:
	payment_type	num_trips	
▶	1	7881388	
	2	4816992	
	3	38632	
	4	11972	
	5	2	

From the query, number of trips for each payment method was found out. And from the results, we can observe credit card is most preferred payment followed by others. Analyzing payment behavior can help taxi companies optimize their payment processing systems and marketing strategies. They can ensure they're accepting the most popular payment forms and develop targeted campaigns that encourage the use of certain payment methods. This analysis can also provide data for financial reporting, such as the percentage of trips paid for with different methods. Understanding payment behavior can help companies identify trends in customer preferences, which can inform business decisions.

Insight 3:

How many rides had central park as destination?



```

97  SELECT COUNT(*) AS central_park_rides
98  FROM TAXI
99  WHERE dropoff_latitude BETWEEN 40.764071 AND 40.800872
100     AND dropoff_longitude BETWEEN -73.981363 AND -73.958014;
101

```

100% 59:100 5 errors found

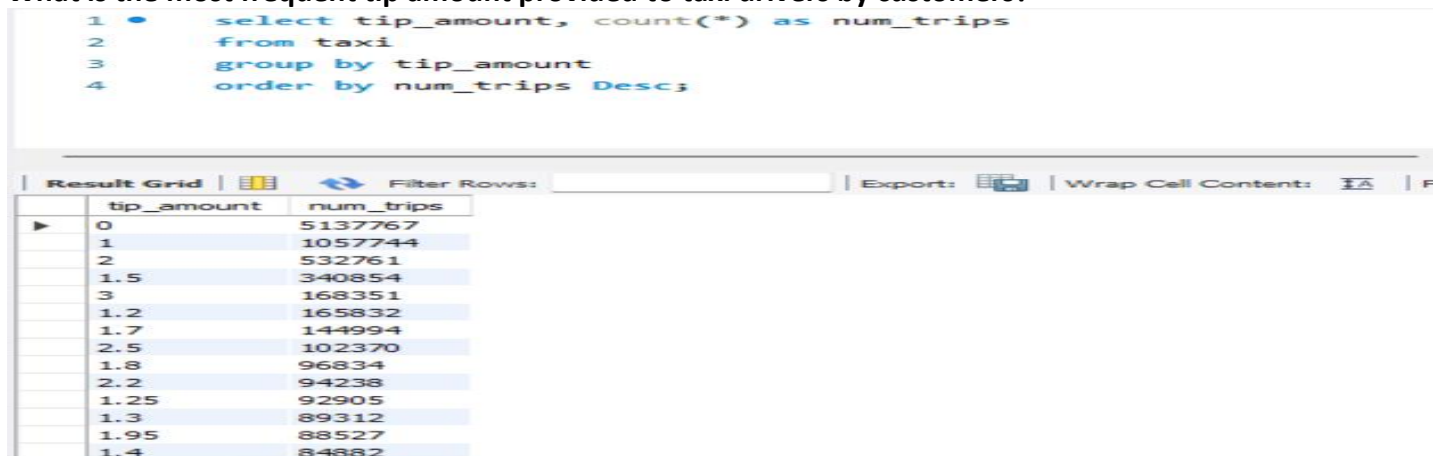
Result Grid Filter Rows: Search Export:

central_park_rides
1600527

The above SQL query counts the number of taxi rides that only include taxi rides where the drop-off location is within the boundaries of Central Park. The analysis can provide insights into the popularity of Central Park as a destination for taxi rides and help taxi companies allocate their resources accordingly. It can also inform marketing and promotional strategies, such as offering discounts or packages for customers traveling to Central Park. Additionally, this analysis can help city planners and tourism organizations understand the level of demand for transportation to Central Park and inform decisions about infrastructure and resource allocation in the area.

Insight 4:

What is the most frequent tip amount provided to taxi drivers by customers?



```

1  select tip_amount, count(*) as num_trips
2  from taxi
3  group by tip_amount
4  order by num_trips Desc;

```

Result Grid Filter Rows: Search Export: Wrap Cell Content:

tip_amount	num_trips
0	5137767
1	1057744
2	532761
1.5	340854
3	168351
1.2	165832
1.7	144994
2.5	102370
1.8	96834
2.2	94238
1.25	92905
1.3	89312
1.95	88527
1.4	84882

This SQL query groups taxi trips by the tip amount and counts the number of trips that fall into each tip amount category. This analysis can provide insights into customer tipping behavior and help taxi companies understand how much customers are willing to tip. It can also inform marketing and promotional strategies, such as offering incentives for customers to tip more or encouraging drivers to provide excellent service to increase tips. Additionally, this analysis can help taxi companies optimize their pricing strategies by considering how tip amounts affect the total fare paid by customers. Overall, analyzing taxi data by tip amount can provide valuable insights into customer behavior and help taxi companies optimize their operations and pricing strategies.

Insight 5

Determine the most common passenger counts for taxi rides and distribution of passenger counts in the dataset?

```
1 • select passenger_count, count(*) as num_trips
2   from taxi
3   group by passenger_count
4   order by num_trips Desc;
```

Result Grid			Filter Rows:	Export:	Wrap Cell Content:
	passenger_count	num_trips			
▶	1	8993870			
	2	1814594			
	5	697645			
	3	528486			
	6	454568			
	4	253228			
	0	6565			
	9	11			
	8	10			
	7	9			

By analyzing the passenger counts in the dataset, we can gain insights into the typical number of passengers for taxi rides. For example, we may find that most taxi rides have one or two passengers, or that there is a high frequency of rides with a particular number of passengers (e.g. airport trips with four passengers). These insights can be used to inform decisions around vehicle selection and capacity planning for taxi companies.

Insight 6

What is the average fare per person for various taxi passenger groups

```
1 • select *, round(avg_total_amt/passenger_count,2) as average_per_person
2   from ((select passenger_count, count(*) as num_trips, round(avg(total_amount),2) as avg_total_amt
3   from taxi
4   where passenger_count in (1,2,3,4,5,6)
5   group by passenger_count)) as a order by average_per_person;
```

Result Grid	Filter Rows:	Export:	Wrap Cell Content:
passenger_count	num_trips	avg_total_amt	average_per_person
6	454568	14.63	2.44
5	697645	14.87	2.97
4	253228	14.93	3.73
3	528486	14.95	4.98
2	1814594	15.38	7.69
1	8993870	15.11	15.11

we can determine the average fare per person for various taxi passenger groups. For example, we may find that the average fare per person is higher for rides with only one passenger compared to rides with multiple passengers, or that rides to/from airports have higher fares per person compared to standard trips. These insights can be used to inform pricing strategies and marketing efforts for taxi companies.

Insights 7

Compare the performance of taxi vendors in terms of trips made, average fare amount, and total fare amount earned

```
1 select vendorid, count(*) as num_trips, round(avg(fare_amount),2) as average,
2   round(sum(fare_amount),2) as total
3   from taxi
4   group by vendorid;
5
```

Result Grid	Filter Rows:	Export:	Wrap Cell Content:
vendorid	num_trips	average	total
2	6647797	12.01	79817100.46
1	6101189	11.8	71967984.88

Analytic methods can be used to compare taxi vendor performance. Here are a few actions we could take:

- Calculate the total number of trips taken by each taxi vendor after grouping the data by vendor.
- Divide the total fare collected by the total number of trips to determine the average fare for each vendor.

- By adding up the prices for all the vendors' trips, determine the total amount of fare that was earned by each one.
- For each vendor, compare the outcomes across the three metrics.

We can determine which vendors have the most trips, the highest average fare, and the greatest total fare earned by conducting this type of data analysis. Additionally, we may spot any patterns or trends in the data, such as whether some suppliers do better at particular times of the day or in particular places.

It's important to keep in mind that this analysis may have missed some elements that affect how well taxi vendors perform, such customer happiness or driver ratings. Therefore, when assessing the performance of taxi providers, this research should be seen as just one component of the picture.

Insights 8

Examine the distribution of various fare categories in the dataset

```
1 • select round(sum(fare_amount)/sum(total_amount)*100,2) as fare_amount_percent,
2     round(sum(tip_amount)/sum(total_amount)*100,2) as tip_amount_percent,
3     round(sum(tolls_amount)/sum(total_amount)*100,2) as tolls_amount_percent,
4     round(sum(mta_tax)/sum(total_amount)*100,2) as mta_tax_percent,
5     round(sum(extra)/sum(total_amount)*100,2) as extra_percent,
6     round(sum(improvement_surcharge)/sum(total_amount)*100,2) as improvement_surcharge_percent
7 from taxi;
8
```

Result Grid Filter Rows: <input type="text"/> Export: Wrap Cell Content:						
	fare_amount_percent	tip_amount_percent	tolls_amount_percent	mta_tax_percent	extra_percent	improvement_surcharge_percent
▶	78.8	12.27	1.61	3.29	2.04	1.87

By examining the distribution of various fare categories in the dataset, we can gain insights into the factors that influence taxi fares. For example, we may find that fares are higher during peak hours or in certain locations, or that longer trips have higher fares. These insights can be used to inform pricing strategies and operational decisions for taxi companies.

Insight 9

Analyzing the fare amount and ride count distribution in the taxi dataset for different RatecodeIDs

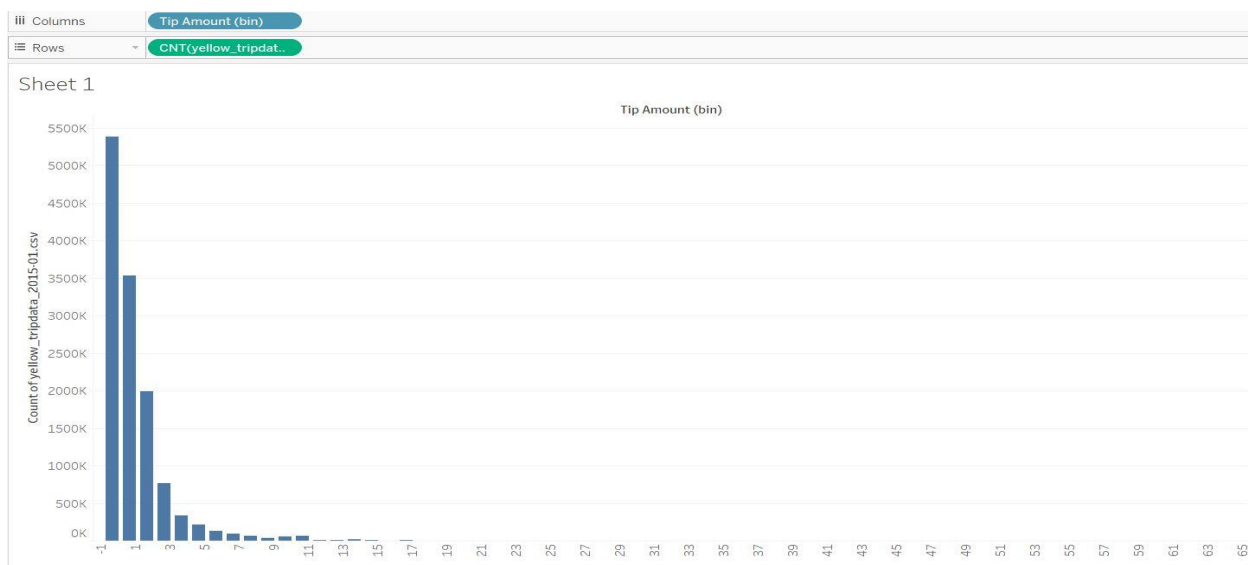
```
1 • select RatecodeID, round(avg(fare_amount),2) as average_fare_amount,  
2     count(*) as number_of_trips  
3     from taxi  
4     group by RatecodeID;
```

	RatecodeID	average_fare_amount	number_of_trips
▶	1	10.97	12464898
	2	51.82	224723
	5	55.19	36896
	3	62.54	17700
	4	59.28	4128
	99	18.84	507
	6	8.6	134

We may learn more about how fare amounts and ride count fluctuate across various types of taxi rides by examining the fare amount and ride count distribution in the taxi dataset for various RatecodeIDs. RatecodeIDs are codes that identify several taxi fare categories, including regular journeys, excursions to and from airports, and trips outside of the city limits.

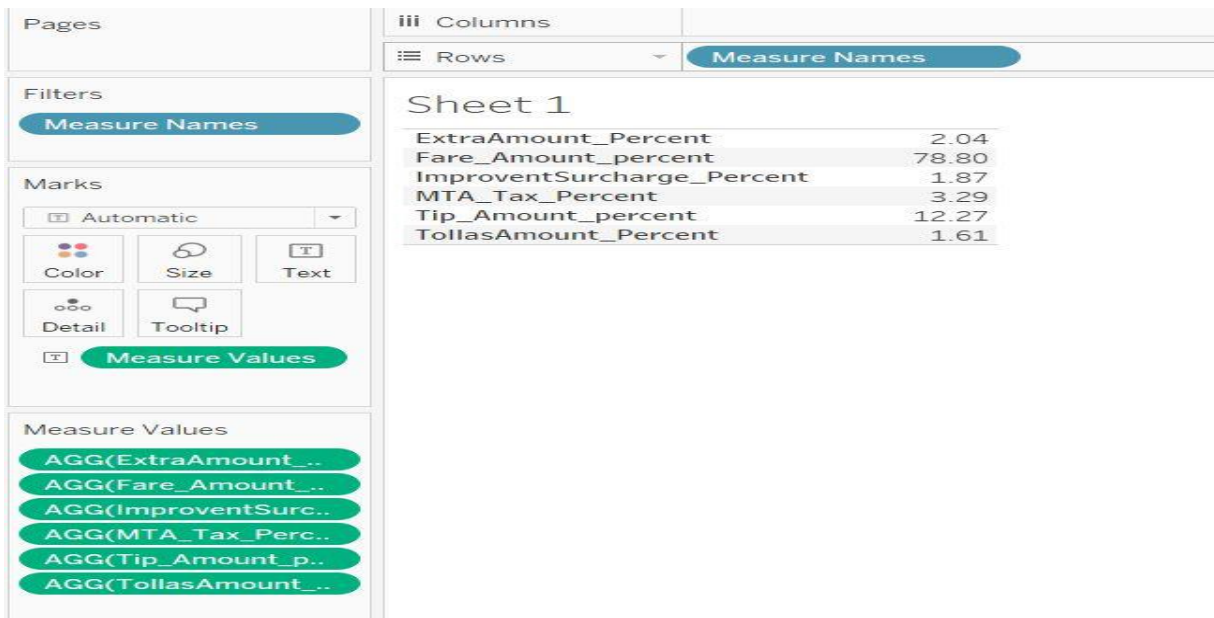
By analyzing the fare amount and ride count distribution for different RatecodeIDs, we can gain insights into the pricing and demand patterns for different types of taxi rides. For example, we may find that trips to/from airports have higher fares, but lower ride counts compared to standard trips, or that trips outside the city limits have lower fares but higher ride counts. These insights can be used to inform pricing strategies and operational decisions for taxi companies.

Insight 10:



The amount of tips that most people chose to leave is our tenth insight. Out of 1 million trips recorded in dataset, only 450k trips were awarded with tip amount, which means 55% of people preferred not to give any tip. These may be due to several factors like not arrival on time, poor quality of service etc.

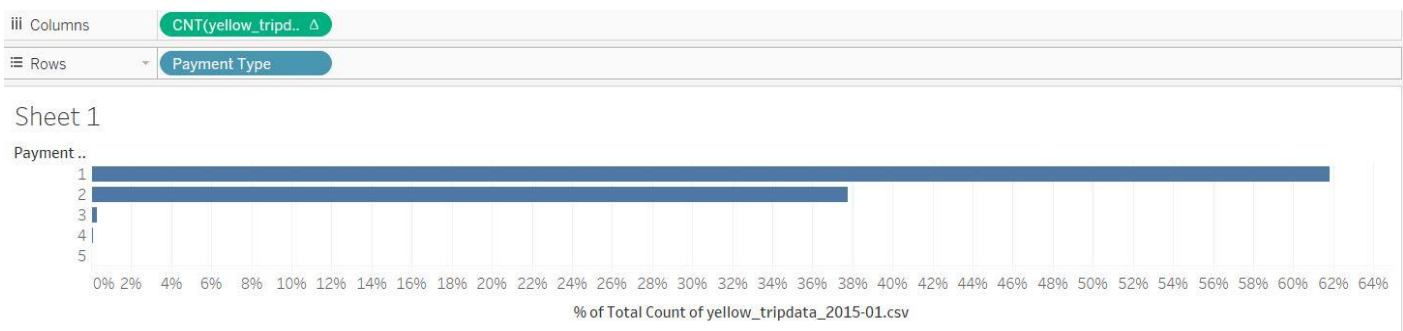
Insight 11:



Our 11th insight is about how total amount is distributed among different fares. From the visualization it is clear that 78% is fare amount, 12 % is tip amount and remaining 10% is distributed among tolls, improvement surcharge, MTA tax.

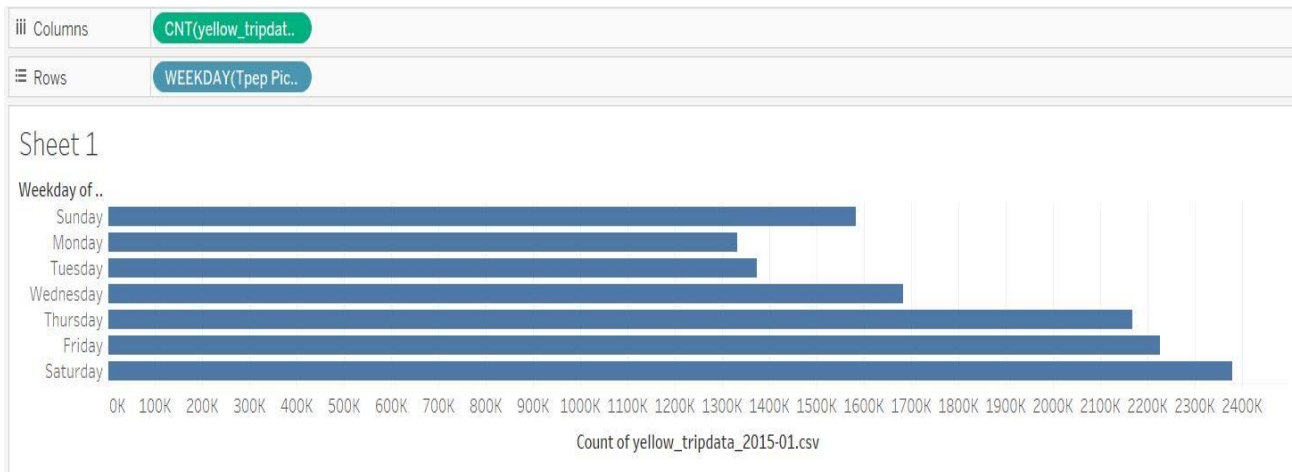
So, it is evident that 78% of total amount goes to Yellow tax company after excluding taxes etc.

Insight 12:



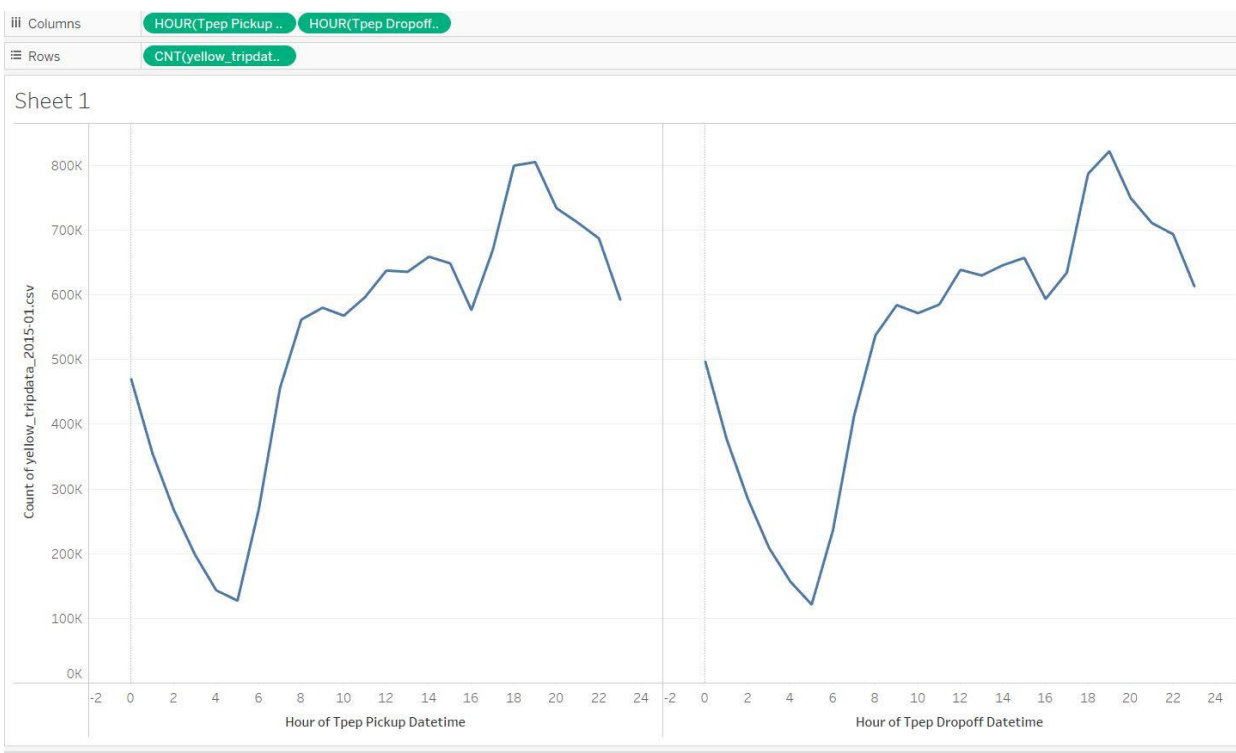
Our 12th insight is about Mode of payments. 62 % of people are preferred to pay through credit And 36% preferred to pay through cash. We may say that from 2015 onwards credit card penetration is increasing slowly as the dataset belongs to 2015.

Insight 13:



Our 13th insight is to figure out on what days of a week does the volume of traffic is high. From visulization it is clear that yellow tax business is getting increased from Monday to Saturday and slightly decreases on Sunday.

Insight 14:



Our 14th insight is about high peak hours in a day. Visualization shows that traffic is high at 10 PM In the night and then decrease gradually till 6 AM in the morning and starts increasing from there till night 10 PM night.