Yellow Taxi Trip Record for Dec'20

Introduction

- Last 10 years have brought about a profound change in the way that people get around.
- One of the greatest change agents has been the rise of rideshare apps like Uber and Lyft. These apps have made life increasingly easy for those looking to catch a ride across town.
- The following is an exploratory data analysis that looks at the state of the New York City taxi industry, specifically yellow taxi and some key metrics measuring their cost structure.

About The Dataset

- The dataset used in this analysis contains approximately 1.4 million rows of observations where each row contains 18 variables collected from yellow taxi rides in New York City in the month of December 2020.
- The dataset was imported to data bricks to figure the solutions if the real- world problems.
- All observations containing Null values were excluded from the dataset within most of the queries for the purposes of this analysis.

Methodology utilized to analyse

- Chosen data bricks platform to perform analysis on the dataset and runtime version was 8.3 (includes Apache Spark 3.1.1, Scala 2.12) and once the cluster gets started then created a notebook named Final_Assignment and used SQL language to do further analysis.

Business Questions to be Answered

1

Which taxi vendor performed best in December 2020?

2

What is the composition of most taxi rides in New York City?

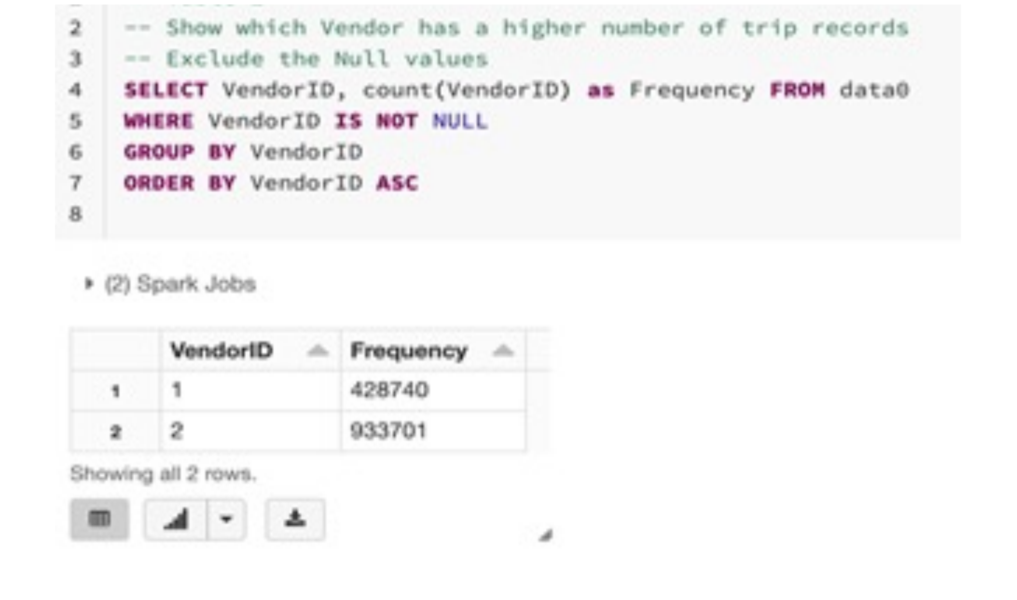
3

How does the data stored in vehicle memory differ from the entire data?

Q1. Which taxi vendor performed best in December 2020?

Showing which vendor has the higher number of trip records.

- Vendor ID 1 [Creative Mobile Technologies]: 428,740
- Vendor ID 2 [VeriFone Inc.]: 933,701

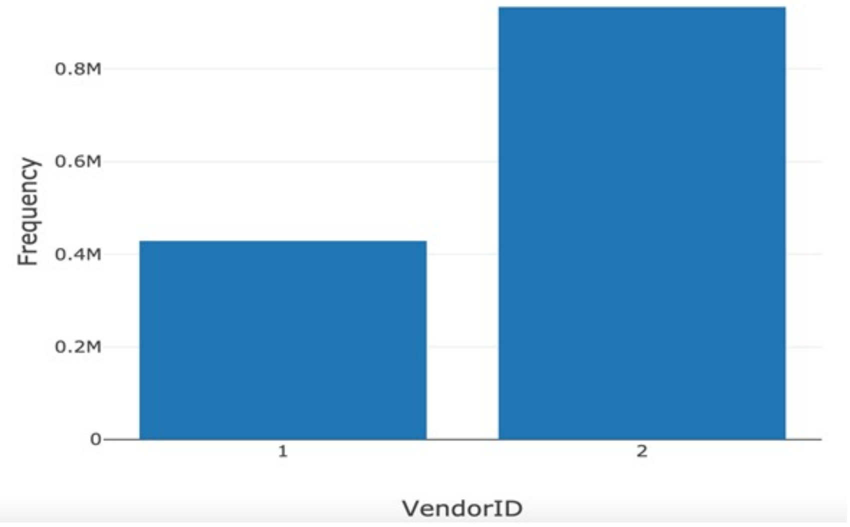


```sql
-- Show which Vendor has a higher number of trip records
-- Exclude the Null values
SELECT VendorID, count(VendorID) as Frequency FROM data0
WHERE VendorID IS NOT NULL
GROUP BY VendorID
ORDER BY VendorID ASC
```

▸ (2) Spark Jobs

	VendorID	Frequency
1	1	428740
2	2	933701

Showing all 2 rows.



Q1. Which taxi vendor performed best in December 2020?

Showing each vendor’s fare amount, tip amount and total amount in ascending order

Fare Amount

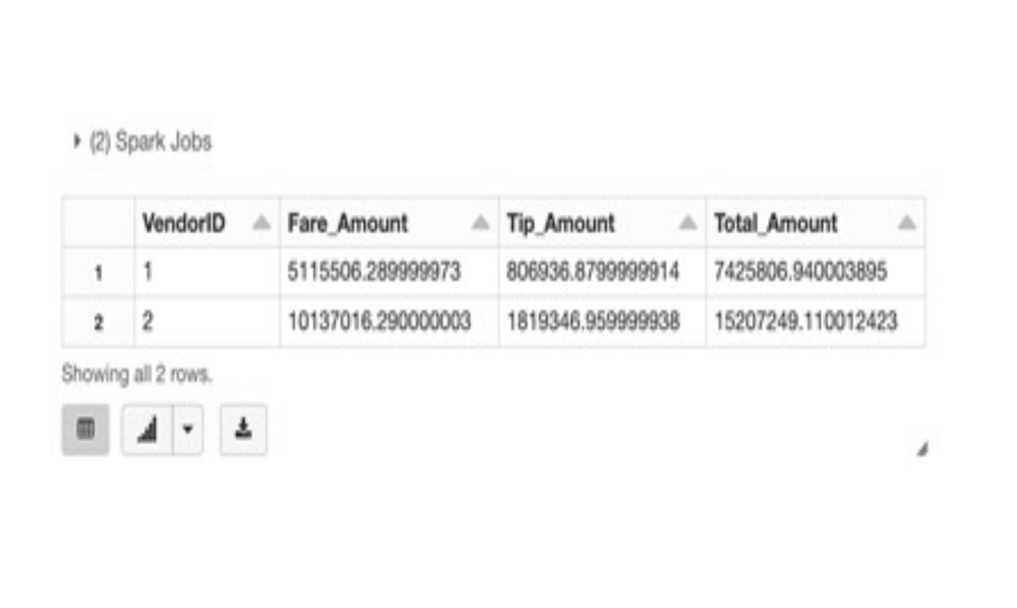
- Vendor ID 1 [Creative Mobile Technologies]: 5,115,506
- Vendor ID 2 [VeriFone Inc.]: 10,137,016

Tip Amount

- Vendor ID 1 [Creative Mobile Technologies]: 806,936.90
- Vendor ID 2 [VeriFone Inc.]: 1,819,347.00

Total Amount

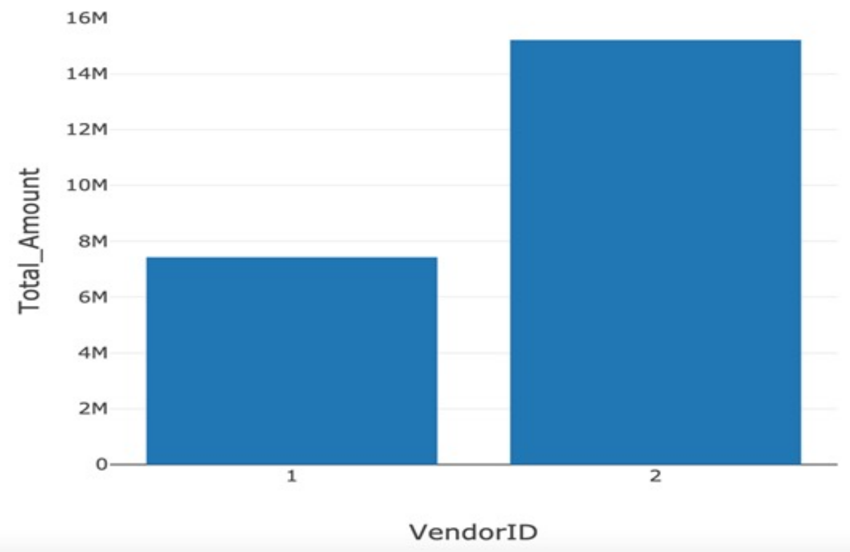
- Vendor ID 1 [Creative Mobile Technologies]: 7,425,807
- Vendor ID 2 [VeriFone Inc.]: 15,207,249



▸ (2) Spark Jobs

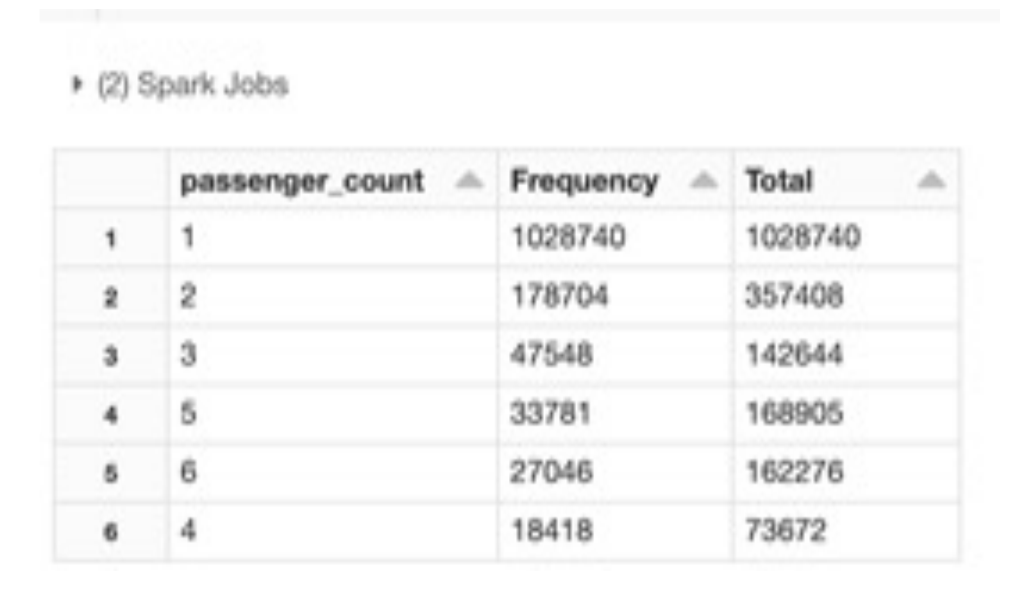
	VendorID	Fare_Amount	Tip_Amount	Total_Amount
1	1	5115506.289999973	806936.8799999914	7425806.940003895
2	2	10137016.290000003	1819346.959999938	15207249.110012423

Showing all 2 rows.



Q2. What is the composition of most taxi rides in New York City?

Checking how many passengers are in most rides and total number of passengers in descending order

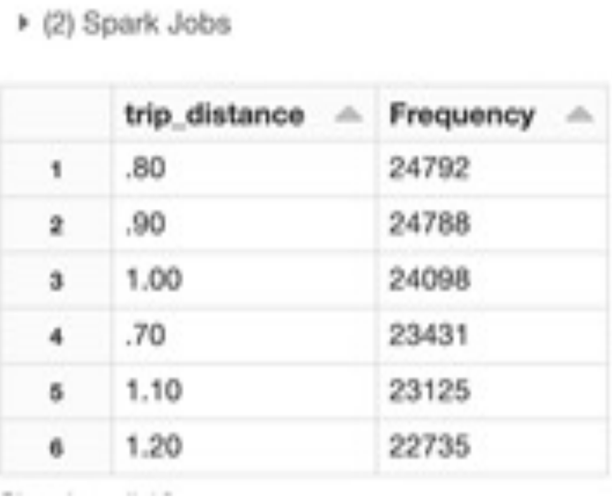


▸ (2) Spark Jobs

	passenger_count	Frequency	Total
1	1	1028740	1028740
2	2	178704	357408
3	3	47548	142644
4	5	33781	168905
5	6	27046	162276
6	4	18418	73672

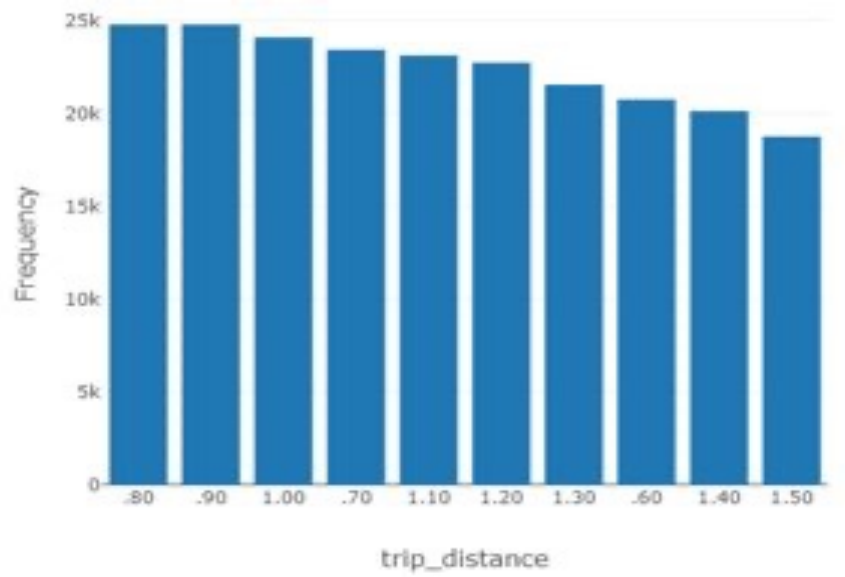
Finding the distance of most rides

ANS: From top 10 trip distance it can be inferred 224,184 rides are between 0.6 miles to 1.5 miles



▸ (2) Spark Jobs

	trip_distance	Frequency
1	.80	24792
2	.90	24788
3	1.00	24098
4	.70	23431
5	1.10	23125
6	1.20	22735

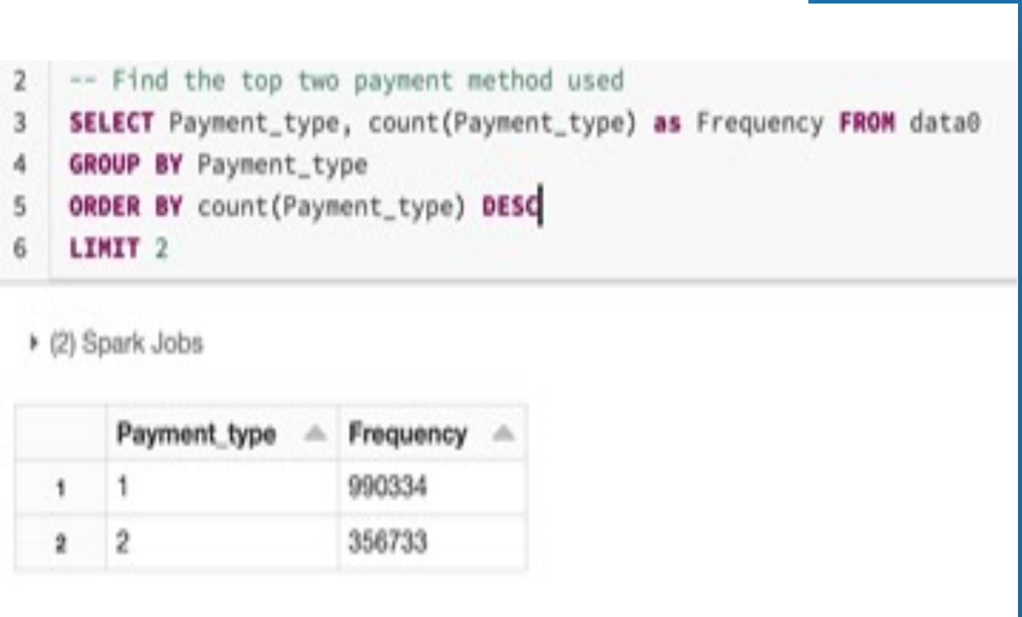


Q2. What is the composition of most taxi rides in New York City?

Finding top two payment methods used.

ANS:

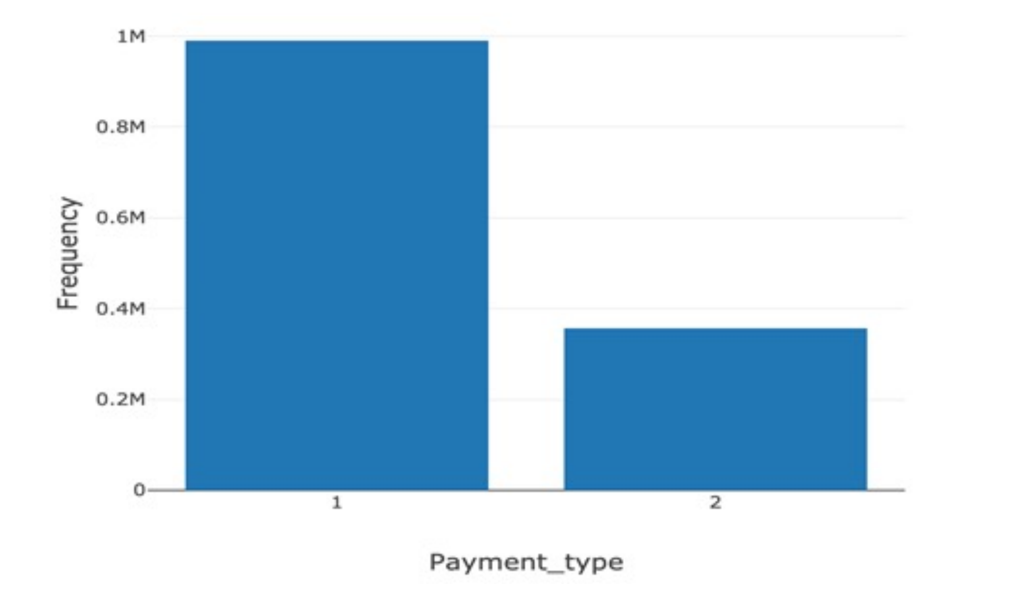
- 1[Credit Card] was used 990,334 times whereas
- 2 [Cash] was used 356,733 times.



```
-- Find the top two payment method used
SELECT Payment_type, count(Payment_type) as Frequency FROM data0
GROUP BY Payment_type
ORDER BY count(Payment_type) DESC
LIMIT 2
```

▸ (2) Spark Jobs

	Payment_type	Frequency
1	1	990334
2	2	356733

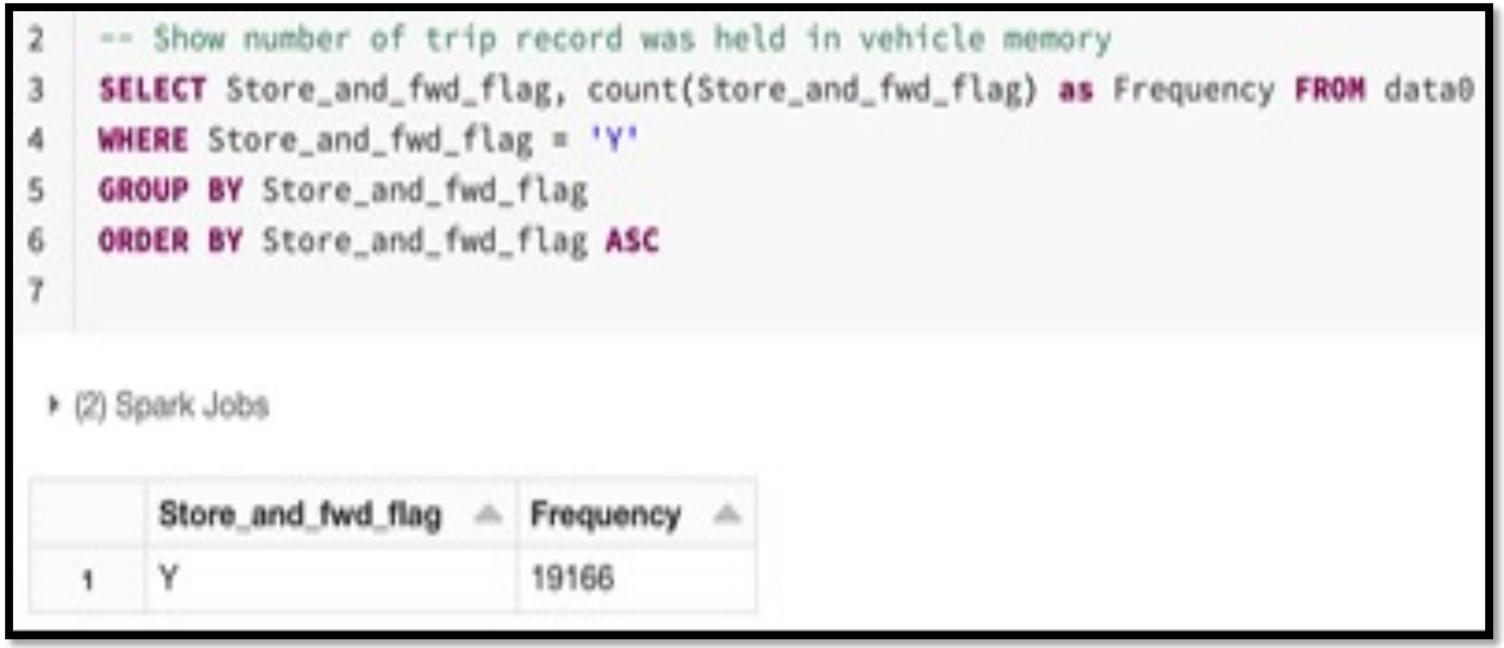


Q3. How does the data stored in vehicle memory differ from the entire data?

Showing number of trip records held in vehicle memory

ANS:

- There are 19,166 trips recorded in the vehicle memory



```
-- Show number of trip record was held in vehicle memory
SELECT Store_and_fwd_flag, count(Store_and_fwd_flag) as Frequency FROM data0
WHERE Store_and_fwd_flag = 'Y'
GROUP BY Store_and_fwd_flag
ORDER BY Store_and_fwd_flag ASC

```

▸ (2) Spark Jobs

	Store_and_fwd_flag	Frequency
1	Y	19166

Q3. How does the data stored in vehicle memory differ from the entire data?

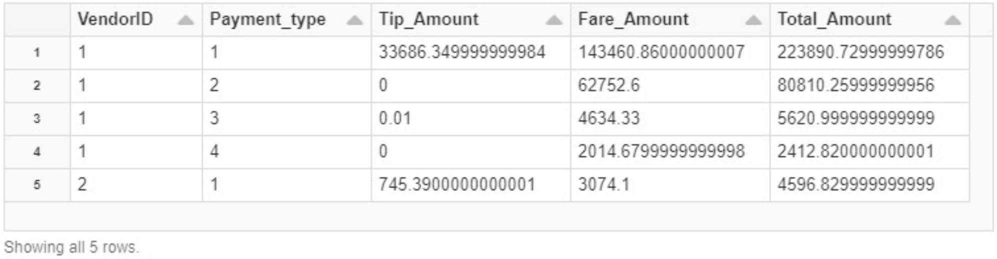
Each vendor’s fare amount and tip when the data is stored in vehicle memory

- Tips are only provided when passenger pays by credit card
- It seems that Vendor2 only accept Credit Card Payments

From the Full Data, we notice

- Tip amount is the highest when passenger pays by credit card
- Vendor 2 accepts all the same payment types as per Vendor 1

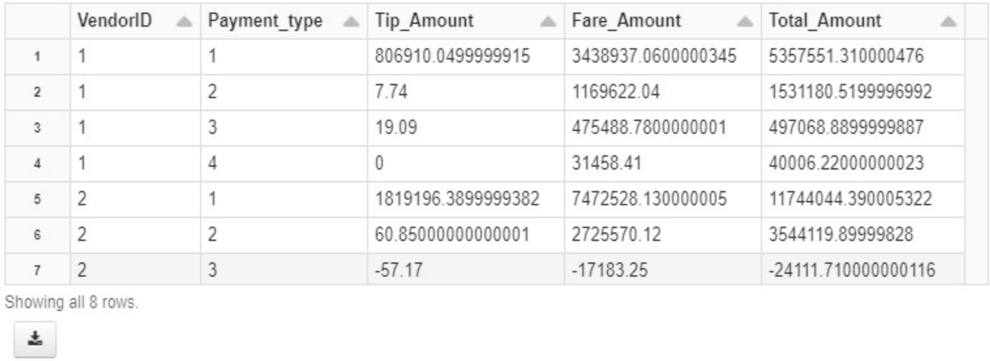
Data from Vehicle Memory Only



	VendorID	Payment_type	Tip_Amount	Fare_Amount	Total_Amount
1	1	1	33686.349999999984	143460.86000000007	223890.72999999786
2	1	2	0	62752.6	80810.25999999956
3	1	3	0.01	4634.33	5620.999999999999
4	1	4	0	2014.6799999999998	2412.820000000001
5	2	1	745.3900000000001	3074.1	4596.829999999999

Showing all 5 rows.

Full Data



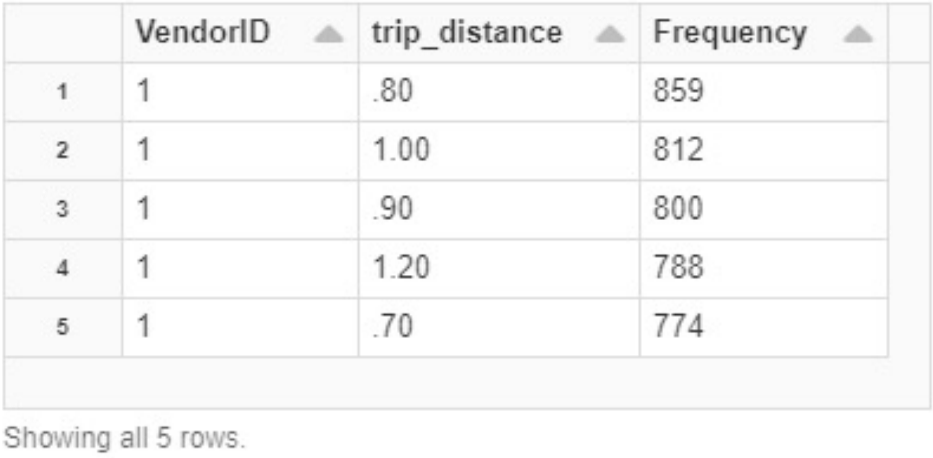
	VendorID	Payment_type	Tip_Amount	Fare_Amount	Total_Amount
1	1	1	806910.0499999915	3438937.0600000345	5357551.310000476
2	1	2	7.74	1169622.04	1531180.5199996992
3	1	3	19.09	475488.7800000001	497068.8899999887
4	1	4	0	31458.41	40006.2200000023
5	2	1	1819196.3899999382	7472528.130000005	11744044.390005322
6	2	2	60.85000000000001	2725570.12	3544119.89999828
7	2	3	-57.17	-17183.25	-24111.710000000116

Showing all 8 rows.

Q3. How does the data stored in vehicle memory differ from the entire data?

Showing the differences for Vendor 1 trip distance and frequency when the data is stored vehicle memory versus the full dataset

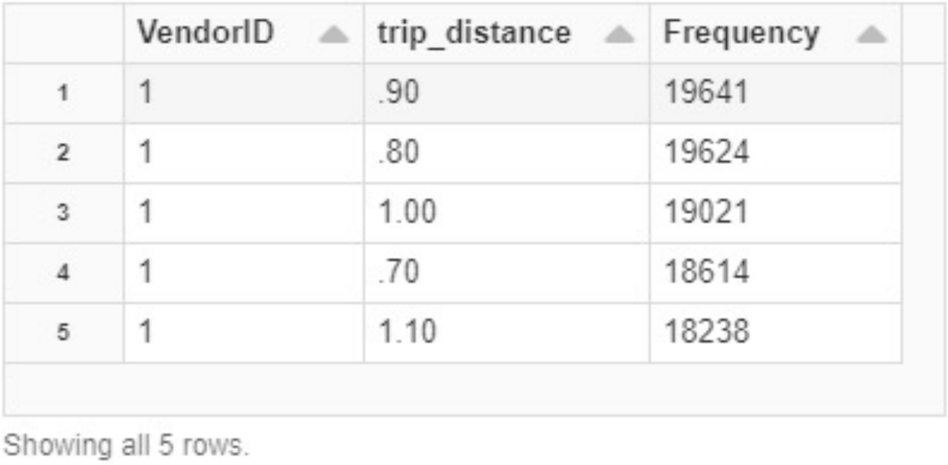
Vendor1 Data from Vehicle Memory Only



	VendorID	trip_distance	Frequency
1	1	.80	859
2	1	1.00	812
3	1	.90	800
4	1	1.20	788
5	1	.70	774

Showing all 5 rows.

Vendor1 from Full Data



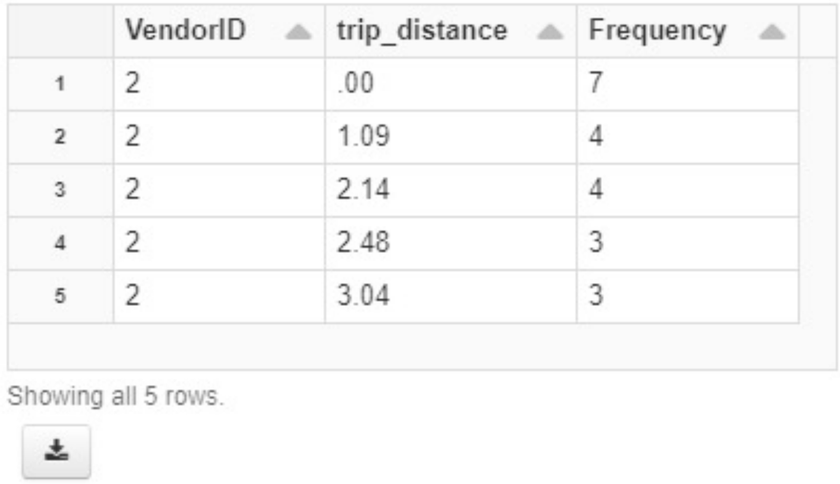
	VendorID	trip_distance	Frequency
1	1	.90	19641
2	1	.80	19624
3	1	1.00	19021
4	1	.70	18614
5	1	1.10	18238

Showing all 5 rows.

Q3. How does the data stored in vehicle memory differ from the entire data?

Showing the differences for Vendor 2 trip distance and frequency when the data is stored vehicle memory versus the full dataset

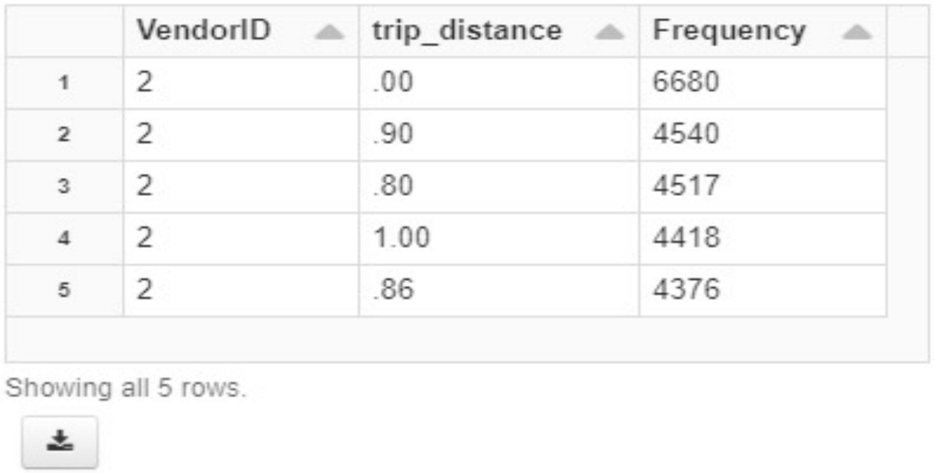
Vendor2 Data from Vehicle Memory Only



	VendorID	trip_distance	Frequency
1	2	.00	7
2	2	1.09	4
3	2	2.14	4
4	2	2.48	3
5	2	3.04	3

Showing all 5 rows.

Vendor2 from Full Data



	VendorID	trip_distance	Frequency
1	2	.00	6680
2	2	.90	4540
3	2	.80	4517
4	2	1.00	4418
5	2	.86	4376

Showing all 5 rows.

Takeaway

Benefits from the project

- Able to prioritize the importance of each task
- Find and reveal answers to real world problems
- How to code in SQL
- Figure out more features within Data Bricks

Drawbacks

- Data Bricks has only 2 hours run time limit

Conclusion

The total revenue for Verifone Inc. in December 2020 amounted to $15 Million. Creative Mobile Technologies were falling behind by a large margin.

- Further analysis will be helpful to determine what makes Verifone Inc much stronger than its competitor.

Majority of the rides are singular passenger and were paid in credit card. Most of the trips in the dataset were one mile or less.

- Benchmarking the yellow taxi trip records with other taxi services like Green Taxi, Uber or Lyft to view their performance in the market.

Less than 1% of the data was recorded in the vehicle memory.

- Standardizing data collection methods will create a higher accuracy on tracking the company's revenue.

References

TLC TRIP RECORD DATA - TLC. (2021). RETRIEVED 4 AUGUST 2021, FROM HTTPS://WWW1.NYC.GOV/SITE/TLC/ABOUT/TLC-TRIP-RECORD-DATA.PAGE

DATA DICTIONARY- YELLOW TAXI TRIP RECORDS (2021). RETRIEVED 4 AUGUST 2021, FROM HTTPS://WWW1.NYC.GOV/ASSETS/TLC/DOWNLOADS/PDF/DATA_DICTIONARY_TRIP_RECORDS_YELLOW.PDF

DOWNLOAD YELLOW TAXI IN THE NEW YORK CITY STREET FOR FREE. (2021). RETRIEVED 18 AUGUST 2021, FROM HTTPS://WWW.VECTEEZY.COM/VECTOR-ART/273204-YELLOW-TAXI-IN-THE-NEW-YORK-CITY-STREET

Thank You