Project Phase-1

Group-3

1. **Link to the dataset** - https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

   The data set provided by the NYC Taxi & Limousine Commission (TLC) contains detailed records of taxi and for-hire vehicle (FHV) trips within New York City. This data is crucial for understanding transportation patterns, demand in specific areas, and the operational efficiency of taxi services in the city. Each trip record includes information on the pickup and drop-off locations, fare details, time of the trip, passenger count, and payment methods. This dataset is invaluable for transportation planning, economic analysis, and urban mobility research.

   Note: This project focuses solely on Yellow Taxi trip records, as they provide enough data to extract meaningful insights.

2. Data Description:

Trip		
Represents the core table, storing detailed trip information, including passenger count, trip distance, fare breakdown, and related foreign keys to other tables		
S.No.	**Name**	**Description**
1	ID	Unique Trip Identifier
2	Vendor	A code indicating the TPEP provider that provided the record
3	PaymentType	A numeric code signifying how the passenger paid for the trip
4	Ratecode	The final rate code in effect at the end of the trip
5	PickUpLocation	TLC Taxi Zone in which the taximeter was engaged
6	DropOffLocation	TLC Taxi Zone in which the taximeter was disengaged
7	PassengerCount	The number of passengers in the vehicle. This is a driver-entered value.
8	TripDistance	The elapsed trip distance in miles reported by the taximeter.
9	StoreAndFwdFlag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server.
10	FareAmount	The time-and-distance fare calculated by the meter.
11	Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
12	MTATax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
13	ImprovementSurcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
14	TipAmount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
15	TollsAmount	Total amount of all tolls paid in trip.
16	TotalAmount	The total amount charged to passengers. Does not include cash tips.
17	CongestionSurcharge	Total amount collected in trip for NYS congestion surcharge.
18	AirportFee	$1.25 for pick up only at LaGuardia and John F. Kennedy Airports

Vendor		
Stores information about the taxi or FHV service providers, typically associated with specific vendor codes.		
S.No.	**Name**	**Description**
1	ID	A unique identifier for each vendor
2	Description	The name of the vendor or service provider, eg. Creative Mobile Technologies, LLC and VeriFone Inc.

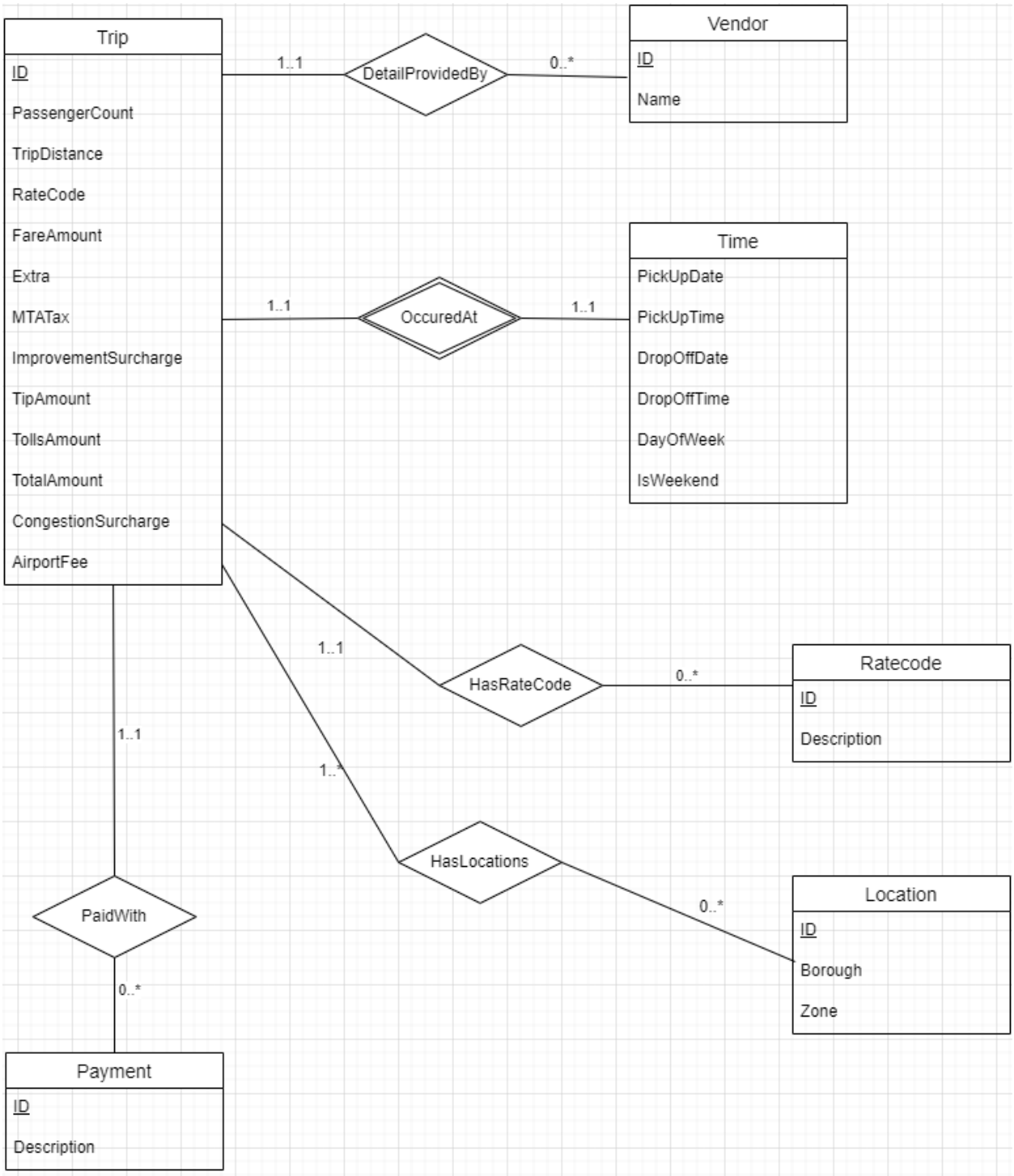
Payment		
Records the different payment methods used by passengers to settle fares.		
S.No.	**Name**	**Description**
1	ID	Unique identifier for each payment type
2	Description	Specifies the type of payment used (e.g., "Credit Card", "Cash", "No Charge")

Time		
Contains information about the start and end times of each trip, and additional time-related details.		
S.No.	**Name**	**Description**
1	TripID	Unique Trip Identifier
2	PickUpDate	The date when the meter was engaged.
3	PickUpTime	The time when the meter was engaged.
4	DropOffDate	The date when the meter was disengaged.
5	DropOffTime	The time when the meter was disengaged.
6	DayOfWeek	Day of the week on the trip day.
7	IsWeekend	Whether the trip day lies on a weekend.

Location		
Captures information about the geographic locations where trips start and end		
S.No.	**Name**	**Description**
1	ID	A unique identifier for each location
2	Borough	The borough in which the location is situated (e.g., Manhattan, Brooklyn)
3	Zone	The specific zone or neighborhood within the borough where the trip originated or ended.

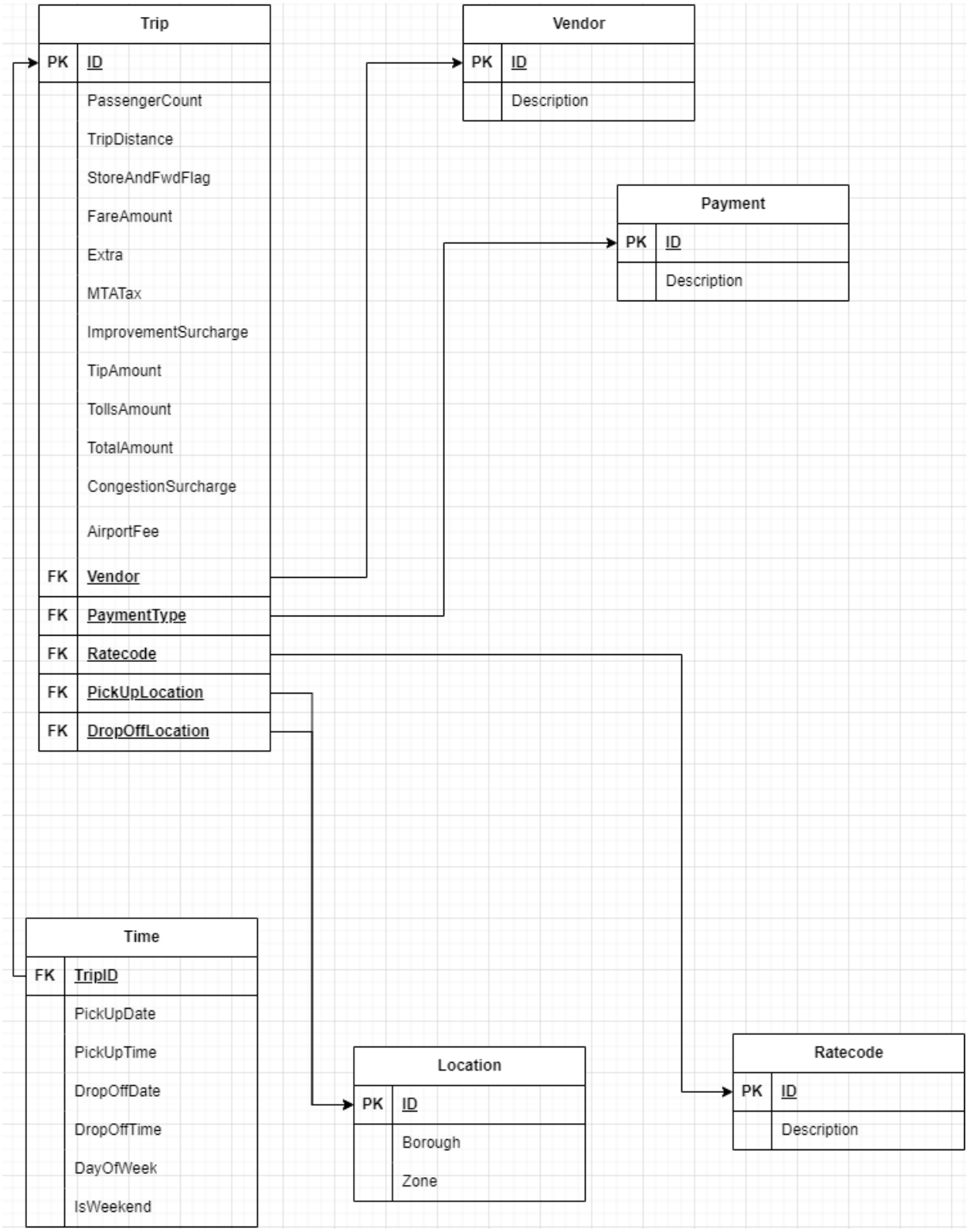
Ratecode		
Describes the various rate codes applied to trips, which may include standard rates, negotiated fares, or special rates for specific destinations.		
S.No.	**Name**	**Description**
1	ID	Unique identifier for each rate code
2	Description	A brief description of the rate type (e.g., "Standard Rate", "JFK Flat Rate").

3. **ER Model:**



Note – Time is a weak entity set.

4. Relational Model:



Relational Model description:

a. **Trips** are the core of the dataset, capturing details about each journey such as distance, fares, and passenger count. These records are linked to other aspects of the dataset through foreign keys.
b. **Vendors** provide the service for each trip, and their information is stored in the **Vendor** table, which links to the **Trip** table via the **Vendor** foreign key.
c. **Payment methods** used to pay for trips are recorded in the **Payment** table, connected to the **Trip** table via the **PaymentType** foreign key.
d. **Rate codes** determine how fares are calculated for each trip, and these are detailed in the **RateCode** table, with a foreign key in the **Trip** table.
e. **Locations** (both pickup and drop-off) are essential geographic data points that are captured in the **Location** table, with foreign keys in the **Trip** table for both the pickup and drop-off locations.
f. **Time details** such as pickup and drop-off dates and times are stored separately in the **Time** table, which is linked to each trip via the **TripID** foreign key.

5. **Programs to load the data and create tables are included in the Data_Import directory of the submitted code.**
   - **a.** Steps to load the data**:**
     i. Store the data files downloaded from the given reference link in 'raw_data' folder.

```
borough_info = "raw_data\\taxi_zone_lookup.csv"
data_2024_01 = "raw_data\\yellow_tripdata_2024-01.parquet"
data_2024_02 = "raw_data\\yellow_tripdata_2024-02.parquet"
data_2024_03 = "raw_data\\yellow_tripdata_2024-03.parquet"
data_2024_04 = "raw_data\\yellow_tripdata_2024-04.parquet"
data_2024_05 = "raw_data\\yellow_tripdata_2024-05.parquet"
data_2024_06 = "raw_data\\yellow_tripdata_2024-06.parquet"
data_2024_07 = "raw_data\\yellow_tripdata_2024-07.parquet"
```

     ii. Run load_from_kaggle.py. Once the execution is complete.
     iii. Run table_creation.sql. This will create all the tables and load the data.

Query | Query History

```
1  SELECT * from trip LIMIT(10)
```

Scratch Pad ×

Data Output | Messages | Notifications

	id [PK] integer	passengercount integer	tripdistance numeric	storeandfwdflag character varying	fareamount numeric	extra numeric	mtatax numeric	improvementsurcharge integer	tipamount numeric	tollsamount numeric	totalamount numeric	congestionsurcharge numeric
1	0	1	2	N	18	1	0	1	0	0	23	2
2	1	1	2	N	10	4	0	1	4	0	19	2
3	2	1	5	N	23	4	0	1	3	0	31	2
4	3	1	1	N	10	4	0	1	2	0	17	2
5	4	1	1	N	8	4	0	1	3	0	16	2
6	5	1	5	N	30	4	0	1	7	0	42	2
7	6	2	11	N	46	6	0	1	10	0	65	0
8	7	0	3	N	25	4	0	1	0	0	30	2
9	8	1	5	N	31	1	0	1	0	0	36	2
10	9	1	0	N	3	1	0	1	0	0	8	2

airportfee numeric	vendor integer	paymenttype integer	ratecode integer	pickuplocation integer	dropofflocation integer
0	2	2	1	186	79
0	1	1	1	140	236
0	1	1	1	236	79
0	1	1	1	79	211
0	1	1	1	211	148
0	1	1	1	148	141
2	2	1	1	138	181
0	1	2	1	246	231
0	2	2	1	161	261
0	2	2	1	113	113

Query | Query History

```
SELECT * from time limit(10)
```

Data Output | Messages | Notifications

	tripid integer	pickupdate date	pickuptime time without time zone	dropoffdate date	dropofftime time without time zone	dayofweek integer	isweekend boolean
1	1340	2024-06-01	00:06:07	2024-06-01	00:11:48	5	true
2	1341	2024-06-01	00:57:38	2024-06-01	01:40:52	5	true
3	1342	2024-06-01	00:06:46	2024-06-01	00:30:52	5	true
4	1343	2024-06-01	00:21:52	2024-06-01	00:31:39	5	true
5	1344	2024-06-01	00:39:15	2024-06-01	00:49:46	5	true
6	1345	2024-06-01	00:36:47	2024-06-01	00:53:21	5	true
7	1346	2024-06-01	00:50:35	2024-06-01	00:50:39	5	true
8	1347	2024-06-01	00:50:56	2024-06-01	01:36:20	5	true
9	1348	2024-06-01	00:05:27	2024-06-01	00:24:00	5	true
10	1349	2024-06-01	00:09:03	2024-06-01	00:20:06	5	true

```
SELECT * from location limit(10)
```

Data Output | Messages | Notifications

	id [PK] integer	borough character varying	zone character varying
1	1	EWR	Newark Airport
2	2	Queens	Jamaica Bay
3	3	Bronx	Allerton/Pelham Gardens
4	4	Manhattan	Alphabet City
5	5	Staten Island	Arden Heights
6	6	Staten Island	Arrochar/Fort Wadsworth
7	7	Queens	Astoria
8	8	Queens	Astoria Park
9	9	Queens	Auburndale
10	10	Queens	Baisley Park

6. a. The inserted data pertains to the year 2024.
b. A total of 44,563,283 records (tuples) have been inserted.
c. Note: Additional records can be added using data from previous years, if necessary.