Phase 1 of NYC Taxi dataset Analysis

Group 35
Abhishek Gawali
Akhil Shetty
Deepika Kini
Shounak Desai

1. Dataset:

The dataset we have chosen for the project is the NYC TLC trip record data.
Link to dataset:
https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Scope:
We have taken 3 taxi types: green, yellow and High volume For Hire (HVFH) since we are able to find some correlation to conduct analysis.
The common columns between the 3 taxis are present in our dataset and a few columns that show some significant information are kept in the separate taxi files.
6 mapping / dimension files are added to provide more information about the taxi data.

Considering the data files in this are very huge, we are still determining the number of years we will be including. As of now, 2022 data will be brought in.

Some static files we have made into tables:

- Ratecode file:
  1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride 99=Unknown

- Payment file:
  1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip

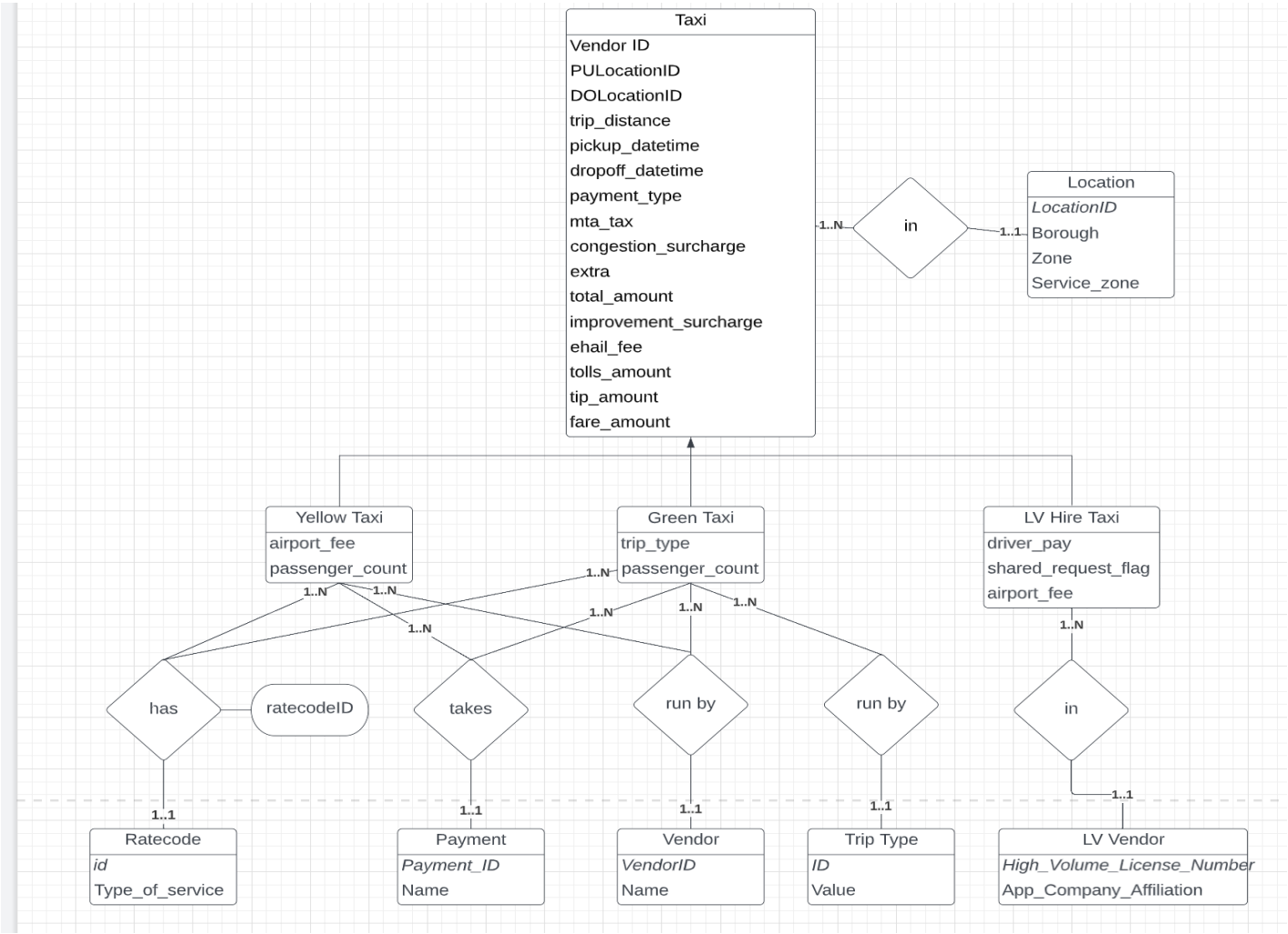
- Trip type file:
  1= Street-hail 2= Dispatch

- Vendor
  1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.

The other mapping files are:

- Location file (tells the location using the locationID in the fact tables)
- HL Vendor and Base file (provides details of vendor and bases the taxis are kept for the HVFH dataset)

2. ER Diagram:



Note:

The ER doesn't exactly conform to the Relational model. We add specialization to show the commonality between the taxi datasets and to consolidate the attributes.

Primary keys are italicised

3. Explaining datasets:

Below is the metadata dictionary for columns , common and ones associated to a certain table.

Common columns:

Vendor ID - A code indicating the LPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.

PULocationID - TLC Taxi Zone in which the taximeter was engaged

DOLocationID - TLC Taxi Zone in which the taximeter was disengaged

Trip_distance - The elapsed trip distance in miles reported by the taximeter

Pep_pickup_datetime - The date and time when the meter was engaged.

Pep_dropoff_datetime - The date and time when the meter was disengaged.

Payment_type - A numeric code signifying how the passenger paid for the trip.
1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip

Total_amount - The total amount charged to passengers. Does not include cash tips

Mta_tax - $0.50 MTA tax that is automatically triggered based on the metered rate in use.

Congestion_surcharge - Total amount collected in trip for NYS congestion surcharge

Extra - Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.

Total_amount - The total amount charged to passengers. Does not include cash tips

Improvement_surcharge - $0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015

Tolls_amount - Total amount of all tolls paid in trip

Tip_amount - Total amount collected from tips

Fare_amount - The time-and-distance fare calculated by the meter

Yellow Taxi Columns:

Passenger_count - The number of passengers in the vehicle
Airport_fee - $1.25 for pick up only at LaGuardia and John F. Kennedy Airports

Green Taxi Columns:

Passenger_count - The number of passengers in the vehicle
trip_type - A code indicating whether the trip was a street-hail or a dispatch that is automatically assigned based on the metered rate in use but can be altered by the driver.
1= Street-hail, 2= Dispatch

LV Hire Taxi Columns:

driver_pay - total driver pay (not including tolls or tips and net of commission, surcharges, or taxes)
Shared_request_flag- Y/N for a shared/pooled ride
Airport_fee- $2.50 for both drop off and pick up at LaGuardia, Newark, and John F. Kennedy airports

4. Relational model in Postgres:

We have 9 tables similar to what is shown in ER, except the Taxi entity.
We use PostgreSQL as database, DBeaver as IDE and load data into this using copy command. Then we create primary (which ever table has it) and foreign key mapping.
We run a python script to convert parquet to csv

5. Programs to create relational tables and load data:

- Python parquet to csv: ParquetToCSV.py
- create tables: create_table_scripts.sql
- foreign key mapping: constrainsts.sql
- SQL file to insert rows in certain tables and Copy command script for loading from csv: dataloadingscript.sql

File size for Jan 2022 data which is loaded:

LVHire file : 12 million rows approx.
Yellow taxi file: 1 million rows approx.
Green taxi file: 62K rows approx.

- Loaded data snapshots:

Below is snapshot of yellow taxi data

```
5 select  * from  yellowcab y limit 100;
```

cab 1 ×

* from yellowcab y limit 100 | Enter a SQL expression to filter results (use Ctrl+Space)

vendorid	tpep_pickup_datetime	tpep_dropoff_datetime	pass	trip_distance	ratecodeid	pulo	dolocationid	paym	fare_amount	extra
1	01-01-2022 00:35	01-01-2022 00:53	2	3.8	1	142	236	1	14.5	3
1	01-01-2022 00:33	01-01-2022 00:42	1	2.1	1	236	42	1	8	0.5
2	01-01-2022 00:53	01-01-2022 01:02	1	0.97	1	166	166	1	7.5	0.5
2	01-01-2022 00:25	01-01-2022 00:35	1	1.09	1	114	68	2	8	0.5
2	01-01-2022 00:36	01-01-2022 01:14	1	4.3	1	68	163	1	23.5	0.5
1	01-01-2022 00:40	01-01-2022 01:09	1	10.3	1	138	161	1	33	3
2	01-01-2022 00:20	01-01-2022 00:34	1	5.07	1	233	87	1	17	0.5
2	01-01-2022 00:13	01-01-2022 00:22	1	2.02	1	238	152	2	9	0.5
2	01-01-2022 00:30	01-01-2022 00:44	1	2.71	1	166	236	1	12	0.5
2	01-01-2022 00:48	01-01-2022 00:53	1	0.78	1	236	141	2	5	0.5
2	01-01-2022 00:55	01-01-2022 01:04	1	1.91	1	141	229	2	8.5	0.5
2	01-01-2022 00:31	01-01-2022 00:34	3	0.82	1	114	90	2	4.5	0.5
2	01-01-2022 00:41	01-01-2022 00:44	3	0.73	1	234	113	2	4.5	0.5
2	01-01-2022 00:56	01-01-2022 01:12	2	2.16	1	246	79	1	11.5	0.5
2	01-01-2022 00:39	01-01-2022 00:47	4	1.43	1	43	140	1	7.5	0.5
2	01-01-2022 00:58	01-01-2022 01:05	1	1.58	1	239	151	2	8	0.5
1	01-01-2022 00:33	01-01-2022 00:47	3	4.2	1	148	141	1	14	2.5
1	01-01-2022 00:53	01-01-2022 01:05	2	2.2	1	237	107	1	9.5	2.5
1	01-01-2022 00:00	01-01-2022 00:05	1	0.2	1	7	7	2	5	0.5
1	01-01-2022 00:35	01-01-2022 00:48	2	3.9	1	107	263	1	13	3
1	01-01-2022 00:49	01-01-2022 00:58	2	3.2	1	263	107	1	11	3
1	01-01-2022 00:06	01-01-2022 00:08	1	0	1	161	161	4	2.5	3
1	01-01-2022 00:09	01-01-2022 00:17	1	1.2	1	161	43	1	7	3
1	01-01-2022 00:30	01-01-2022 00:38	1	1.7	1	239	24	1	7	3
1	01-01-2022 00:46	01-01-2022 00:57	1	1.6	1	239	263	1	8	3
2	01-01-2022 00:14	01-01-2022 00:35	5	2.48	1	246	233	1	14	0.5
2	01-01-2022 00:46	01-01-2022 01:08	6	5.95	1	79	238	1	20	0.5

Snapshot for green taxi:

```
105 select  * from  greencab g  limit 100;
```

greencab 1 ×

select * from greencab g limit 100 | Enter a SQL expression to filter results (use Ctrl+Space)

	vendorid	lpep_pickup_datetime	lpep_dropoff_datetime	ratecodeid	pulocationid	dolocationid	passenger_count	trip_distance	fare_amount	extra
1	2	01-01-2022 00:14	01-01-2022 00:15	1	42	42	1	0.44	3.5	0.5
2	1	01-01-2022 00:20	01-01-2022 00:29	1	116	41	1	2.1	9.5	0.5
3	1	01-01-2022 00:57	01-01-2022 01:13	1	41	140	1	3.7	14.5	3.25
4	2	01-01-2022 00:07	01-01-2022 00:15	1	181	181	1	1.69	8	0.5
5	2	01-01-2022 00:07	01-01-2022 00:28	1	33	170	1	6.26	22	0.5
6	1	01-01-2022 00:47	01-01-2022 00:54	1	150	210	1	1.3	7	0.5
7	2	01-01-2022 00:13	01-01-2022 00:33	1	66	67	1	6.47	22.5	0.5
8	2	01-01-2022 00:43	01-01-2022 00:49	1	40	195	1	1.15	6	0.5
9	2	01-01-2022 00:41	01-01-2022 00:47	1	112	80	1	1.3	6	0.5
10	2	01-01-2022 00:51	01-01-2022 01:09	1	256	186	1	4.75	17	0.5
11	2	31-12-2021 23:44	31-12-2021 23:57	1	75	4	1	6.03	18.5	0.5
12	2	01-01-2022 00:06	01-01-2022 00:21	1	41	116	1	2.82	12.5	0.5
13	2	01-01-2022 00:47	01-01-2022 00:50	1	75	74	1	0.7	4.5	0.5
14	2	01-01-2022 00:10	01-01-2022 23:10	1	74	260	1	5.48	17.5	0.5
15	2	01-01-2022 00:52	01-01-2022 01:25	1	74	231	1	7.54	26.5	0.5
16	1	01-01-2022 00:47	01-01-2022 00:49	1	74	42	1	0.9	4.5	0.5
17	1	01-01-2022 00:59	01-01-2022 01:21	1	41	94	1	6.2	21.5	0.5
18	2	01-01-2022 00:12	01-01-2022 00:26	5	213	174	1	5.57	20	0
19	2	01-01-2022 00:47	01-01-2022 00:53	1	51	185	1	1.87	7	0.5
20	2	01-01-2022 00:54	01-01-2022 01:17	5	185	119	1	6.6	25	0
21	2	01-01-2022 00:37	01-01-2022 00:38	1	129	260	1	0.26	3	0.5
22	2	01-01-2022 01:02	01-01-2022 01:30	1	82	260	1	1.79	17.5	0.5
23	2	01-01-2022 00:56	01-01-2022 01:04	1	235	42	1	2.34	9	0.5
24	2	01-01-2022 00:35	01-01-2022 00:52	1	255	225	1	3.66	14.5	0.5
25	2	01-01-2022 00:02	01-01-2022 00:18	5	66	234	4	3.96	25	0
26	2	01-01-2022 00:45	01-01-2022 00:59	5	33	233	4	6.2	60	0
27	2	01-01-2022 00:24	01-01-2022 00:32	1	95	135	1	2.19	9.5	0.5

Snapshot for HLFH taxi:

```
105 select * from lvhiretaxi l limit 100;
```

retaxi 1 ×

ct * from lvhiretaxi l limit 100 | Enter a SQL expression to filter results (use Ctrl+Space)

hvfhs_license_num	dispatching_base_num	originating_base_num	request_datetime	on_scene_datetime	pickup_datetime	dropoff_datetime	pulocationid	dolocatic
HV0003	B03404	B03404	2022-01-01 00:05:31	2022-01-01 00:05:40	2022-01-01 00:07:24	2022-01-01 00:18:28	170	
HV0003	B03404	B03404	2022-01-01 00:19:27	2022-01-01 00:22:08	2022-01-01 00:22:32	2022-01-01 00:30:12	237	
HV0003	B03404	B03404	2022-01-01 00:43:53	2022-01-01 00:57:37	2022-01-01 00:57:37	2022-01-01 01:07:32	237	
HV0003	B03404	B03404	2022-01-01 00:15:36	2022-01-01 00:17:08	2022-01-01 00:18:02	2022-01-01 00:23:05	262	
HV0003	B03404	B03404	2022-01-01 00:25:45	2022-01-01 00:26:01	2022-01-01 00:28:01	2022-01-01 00:35:42	229	
HV0003	B03404	B03404	2022-01-01 00:34:44	2022-01-01 00:36:52	2022-01-01 00:38:50	2022-01-01 00:51:32	263	
HV0003	B03404	B03404	2022-01-01 00:47:51	2022-01-01 00:52:00	2022-01-01 00:53:25	2022-01-01 01:08:56	113	
HV0003	B03404	B03404	2022-01-01 00:06:21	2022-01-01 00:06:58	2022-01-01 00:08:58	2022-01-01 00:23:01	151	
HV0003	B03404	B03404	2022-01-01 00:27:54	2022-01-01 00:30:26	2022-01-01 00:32:25	2022-01-01 00:44:15	263	
HV0003	B03404	B03404	2022-01-01 00:44:59	2022-01-01 00:48:23	2022-01-01 00:50:23	2022-01-01 01:15:30	237	
HV0003	B03404	B03404	2022-01-01 00:13:49	2022-01-01 00:16:15	2022-01-01 00:17:02	2022-01-01 00:40:09	261	
HV0003	B03404	B03404	2022-01-01 00:39:10	2022-01-01 00:42:59	2022-01-01 00:43:20	2022-01-01 00:47:31	223	
HV0003	B03404	B03404	2022-01-01 00:45:50	2022-01-01 00:52:15	2022-01-01 00:52:29	2022-01-01 01:01:48	223	
HV0005	B03406	[NULL]	2022-01-01 00:36:54	[NULL]	2022-01-01 00:45:34	2022-01-01 00:54:11	88	
HV0003	B03404	B03404	2022-01-01 00:07:13	2022-01-01 00:12:03	2022-01-01 00:12:03	2022-01-01 00:35:07	246	
HV0003	B03404	B03404	2022-01-01 00:53:32	2022-01-01 00:58:14	2022-01-01 00:58:26	2022-01-01 01:07:23	243	
HV0005	B03406	[NULL]	2022-01-01 00:28:07	[NULL]	2022-01-01 00:34:59	2022-01-01 00:50:15	239	
HV0005	B03406	[NULL]	2022-01-01 00:45:44	[NULL]	2022-01-01 00:58:09	2022-01-01 01:28:23	170	
HV0003	B03404	B03404	2022-01-01 00:21:54	2022-01-01 00:26:15	2022-01-01 00:28:15	2022-01-01 00:38:52	223	
HV0003	B03404	B03404	2022-01-01 00:35:08	2022-01-01 00:43:55	2022-01-01 00:45:55	2022-01-01 00:53:47	179	
HV0003	B03404	B03404	2022-01-01 00:48:14	2022-01-01 00:55:03	2022-01-01 00:55:30	2022-01-01 01:03:26	7	
HV0003	B03404	B03404	2022-01-01 00:25:44	2022-01-01 00:29:54	2022-01-01 00:30:45	2022-01-01 00:43:22	165	
HV0003	B03404	B03404	2022-01-01 00:37:55	2022-01-01 00:45:30	2022-01-01 00:47:30	2022-01-01 01:06:49	165	
HV0005	B03406	[NULL]	2022-01-01 00:14:34	[NULL]	2022-01-01 00:19:47	2022-01-01 00:40:46	238	
HV0005	B03406	[NULL]	2022-01-01 00:41:46	[NULL]	2022-01-01 00:52:25	2022-01-01 01:10:00	138	
HV0003	B03404	B03404	2022-01-01 00:23:53	2022-01-01 00:25:22	2022-01-01 00:25:33	2022-01-01 00:31:35	234	
HV0003	B03404	B03404	2022-01-01 00:33:53	2022-01-01 00:40:18	2022-01-01 00:41:46	2022-01-01 00:51:25	158	

Refresh | Save | Cancel | Export data | 200 | 100 | 100 row(s) fetched - 2ms, on 2023-03-03 at 18:34:41

Snapshot for LVFH Vendor information:

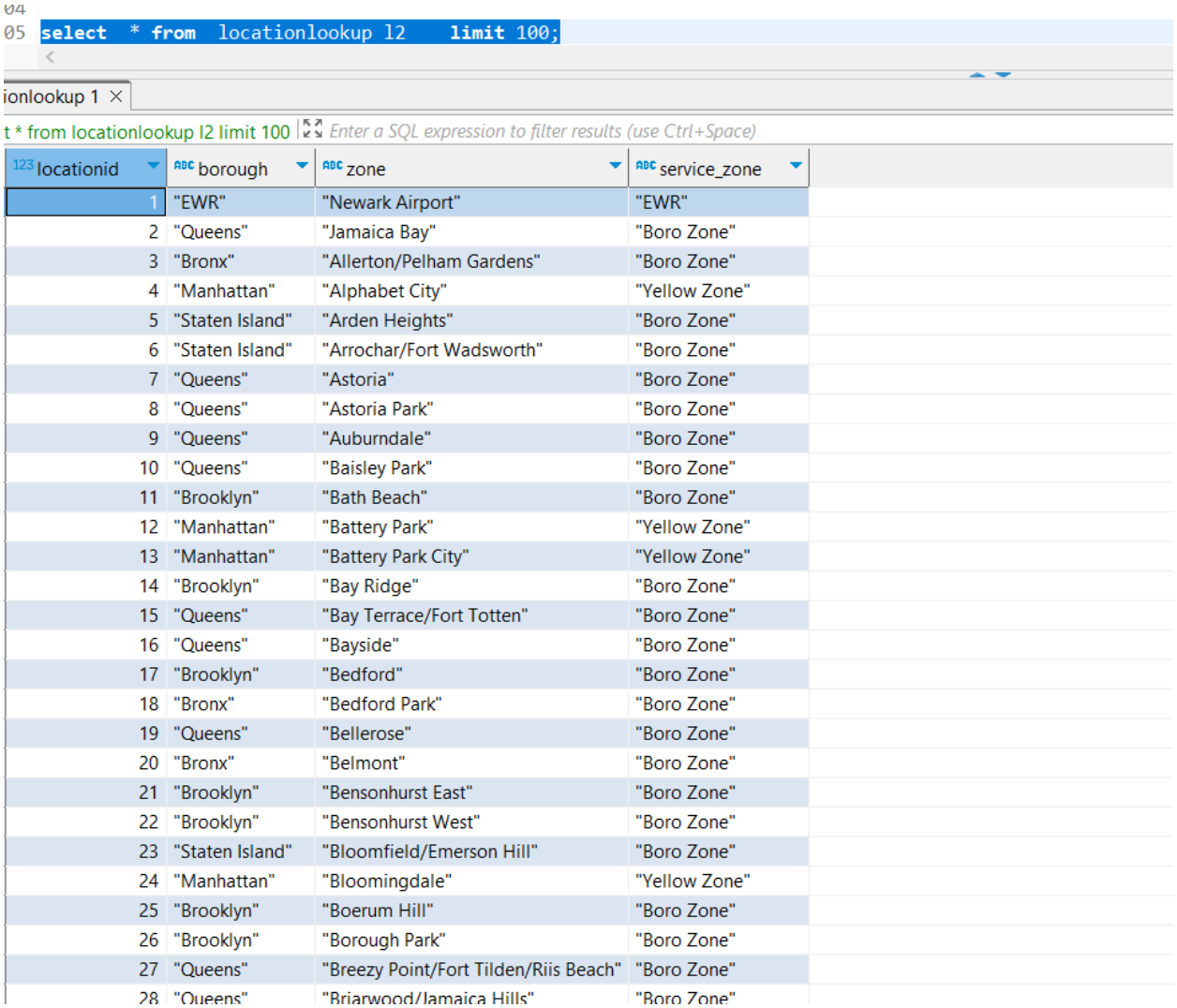
```
105 select * from lvhiretaxivendor l limit 100;
```

lvhiretaxivendor 1 ×

select * from lvhiretaxivendor l limit 100 | Enter a SQL expression to filter results (use Ctrl+Space)

	High_Volume_License_Number	License_Number	Base_Name	App_Company_Affiliation
1	HV0002	B02914	VULCAN CARS LLC	Juno
2	HV0002	B02907	SABO ONE LLC	Juno
3	HV0002	B02908	SABO TWO LLC	Juno
4	HV0002	B03035	OMAHA LLC	Juno
5	HV0005	B02510	TRI-CITY,LLC	Lyft
6	HV0005	B02844	ENDOR CAR & DRIVER LLC	Lyft
7	HV0003	B02877	ZWOLF-NY, LLC	Uber
8	HV0003	B02866	ZWEI-NY,LLC	Uber
9	HV0003	B02882	ZWANZIG-NY,LLC	Uber
10	HV0003	B02869	ZEHN-NY,LLC	Uber
11	HV0003	B02617	WEITER LLC	Uber
12	HV0003	B02876	VIERZEHN-NY, LLC	Uber
13	HV0003	B02865	VIER-NY,LLC	Uber
14	HV0003	B02512	UNTER LLC	Uber
15	HV0003	B02888	SIEBZEHN-NY,LLC	Uber
16	HV0003	B02864	SIEBEN-NY,LLC	Uber
17	HV0003	B02883	SECHZEHN-NY, LLC	Uber
18	HV0003	B02875	SECHS-NY, LLC	Uber
19	HV0003	B02682	SCHMECKEN LLC	Uber
20	HV0003	B02880	NEUNZEHN-NY, LLC	Uber
21	HV0003	B02870	NEUN-NY,LLC	Uber
22	HV0003	B02404	KUCHEN,LLC	Uber
23	HV0003	B02598	HINTER LLC	Uber
24	HV0003	B02765	GRUN LLC	Uber
25	HV0003	B02879	FUNFZEHN-NY, LLC	Uber
26	HV0003	B02867	FUNF-NY, LLC	Uber
27	HV0003	B02878	ELF-NY,LLC	Uber
28	HV0003	B02887	EINUNDZWANZIG-NY, LLC	Uber

Snapshot for location data:



```
04
05 select * from locationlookup l2 limit 100;
```

ionlookup 1 ×

t * from locationlookup l2 limit 100 | Enter a SQL expression to filter results (use Ctrl+Space)

locationid	borough	zone	service_zone
1	"EWR"	"Newark Airport"	"EWR"
2	"Queens"	"Jamaica Bay"	"Boro Zone"
3	"Bronx"	"Allerton/Pelham Gardens"	"Boro Zone"
4	"Manhattan"	"Alphabet City"	"Yellow Zone"
5	"Staten Island"	"Arden Heights"	"Boro Zone"
6	"Staten Island"	"Arrochar/Fort Wadsworth"	"Boro Zone"
7	"Queens"	"Astoria"	"Boro Zone"
8	"Queens"	"Astoria Park"	"Boro Zone"
9	"Queens"	"Auburndale"	"Boro Zone"
10	"Queens"	"Baisley Park"	"Boro Zone"
11	"Brooklyn"	"Bath Beach"	"Boro Zone"
12	"Manhattan"	"Battery Park"	"Yellow Zone"
13	"Manhattan"	"Battery Park City"	"Yellow Zone"
14	"Brooklyn"	"Bay Ridge"	"Boro Zone"
15	"Queens"	"Bay Terrace/Fort Totten"	"Boro Zone"
16	"Queens"	"Bayside"	"Boro Zone"
17	"Brooklyn"	"Bedford"	"Boro Zone"
18	"Bronx"	"Bedford Park"	"Boro Zone"
19	"Queens"	"Bellerose"	"Boro Zone"
20	"Bronx"	"Belmont"	"Boro Zone"
21	"Brooklyn"	"Bensonhurst East"	"Boro Zone"
22	"Brooklyn"	"Bensonhurst West"	"Boro Zone"
23	"Staten Island"	"Bloomfield/Emerson Hill"	"Boro Zone"
24	"Manhattan"	"Bloomingdale"	"Yellow Zone"
25	"Brooklyn"	"Boerum Hill"	"Boro Zone"
26	"Brooklyn"	"Borough Park"	"Boro Zone"
27	"Queens"	"Breezy Point/Fort Tilden/Riis Beach"	"Boro Zone"
28	"Queens"	"Briarwood/Jamaica Hills"	"Boro Zone"

Snapshot for ratecode data:

```
select * from ratecodelookup r limit 100;
```

delookup 1

```
* from ratecodelookup r limit 100
```

Enter a SQL expression to filter results (use C

id	typeofservice
1	Standard rate
2	JFK
3	Newark
4	Nassau or Westchester
5	Negotiated fare
6	Group ride/Shared Ride

Snapshot for payment data:

```
select * from paymentlookup p limit 100;
```

ntlookup 1

```
* from paymentlookup p limit 100
```

Enter a SQL expression to filter results (us

id	typeofpayment
1	Credit card
2	Cash
3	No charge
4	Dispute
5	Unknown
6	Voided trip

Snapshot for Vendor for green and yellow taxi data:

```
select  * from  vendortaxilookup v   limit 100;
```

axilookup 1 ×

from vendortaxilookup v limit 100 | Enter a SQL expression to filter re

id	vendortype
1	Creative Mobile Technologies,
2	VeriFone Inc.

Snapshot for trip type data:

```
05 select  * from   triptype t   limit 100;
```

pe 1 ×

t * from triptype t limit 100 | Enter a SQL expression to filter res

123 id	value
1	Street-Hail
2	Dispatch

Data model after creating and loading data: