In [22]:
```
import io
import pandas as pd
import requests
```

In [23]:
```
url = 'https://storage.googleapis.com/ubder-data-project-rohan/uber_data.csv'
response = requests.get(url)
```

In [24]:
```
df = pd.read_csv(io.StringIO(response.text), sep=',')
```

In [25]:
```
df['tpep_pickup_datetime'] = pd.to_datetime(df['tpep_pickup_datetime'])
df['tpep_dropoff_datetime'] = pd.to_datetime(df['tpep_dropoff_datetime'])
```

In [26]:
```
df = df.drop_duplicates().reset_index(drop=True)
df['trip_id'] = df.index
```

In [28]:
```
df.head()
```

Out[28]:

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	pickup_longit
0	1	2016-03-01	2016-03-01 00:07:55	1	2.50	-73.976
1	1	2016-03-01	2016-03-01 00:11:06	1	2.90	-73.983
2	2	2016-03-01	2016-03-01 00:31:06	2	19.98	-73.782
3	2	2016-03-01	2016-03-01 00:00:00	3	10.78	-73.863
4	2	2016-03-01	2016-03-01 00:00:00	5	30.43	-73.971

In [29]:
```
datetime_dim = df[['tpep_pickup_datetime','tpep_dropoff_datetime']].reset_index(drop=Tr
datetime_dim['tpep_pickup_datetime'] = datetime_dim['tpep_pickup_datetime']
datetime_dim['pick_hour'] = datetime_dim['tpep_pickup_datetime'].dt.hour
datetime_dim['pick_day'] = datetime_dim['tpep_pickup_datetime'].dt.day
datetime_dim['pick_month'] = datetime_dim['tpep_pickup_datetime'].dt.month
datetime_dim['pick_year'] = datetime_dim['tpep_pickup_datetime'].dt.year
datetime_dim['pick_weekday'] = datetime_dim['tpep_pickup_datetime'].dt.weekday

datetime_dim['tpep_dropoff_datetime'] = datetime_dim['tpep_dropoff_datetime']
datetime_dim['drop_hour'] = datetime_dim['tpep_dropoff_datetime'].dt.hour
datetime_dim['drop_day'] = datetime_dim['tpep_dropoff_datetime'].dt.day
datetime_dim['drop_month'] = datetime_dim['tpep_dropoff_datetime'].dt.month
datetime_dim['drop_year'] = datetime_dim['tpep_dropoff_datetime'].dt.year
datetime_dim['drop_weekday'] = datetime_dim['tpep_dropoff_datetime'].dt.weekday


datetime_dim['datetime_id'] = datetime_dim.index

# datetime_dim = datetime_dim.rename(columns={'tpep_pickup_datetime': 'datetime_id'}).r
datetime_dim = datetime_dim[['datetime_id', 'tpep_pickup_datetime', 'pick_hour', 'pick_
                             'tpep_dropoff_datetime', 'drop_hour', 'drop_day', 'drop_mo
```

```
#
datetime_dim.head()
```

Out[29]:

	datetime_id	tpep_pickup_datetime	pick_hour	pick_day	pick_month	pick_year	pick_weekday	tpep_c
0	0	2016-03-01	0	1	3	2016	1	20
1	1	2016-03-01	0	1	3	2016	1	20
2	2	2016-03-01	0	1	3	2016	1	20
3	3	2016-03-01	0	1	3	2016	1	20
4	4	2016-03-01	0	1	3	2016	1	20

In [30]:

```
passenger_count_dim = df[['passenger_count']].reset_index(drop=True)
passenger_count_dim['passenger_count_id'] = passenger_count_dim.index
passenger_count_dim = passenger_count_dim[['passenger_count_id','passenger_count']]

trip_distance_dim = df[['trip_distance']].reset_index(drop=True)
trip_distance_dim['trip_distance_id'] = trip_distance_dim.index
trip_distance_dim = trip_distance_dim[['trip_distance_id','trip_distance']]
```

In [31]:

```
rate_code_type = {
    1:"Standard rate",
    2:"JFK",
    3:"Newark",
    4:"Nassau or Westchester",
    5:"Negotiated fare",
    6:"Group ride"
}

rate_code_dim = df[['RatecodeID']].reset_index(drop=True)
rate_code_dim['rate_code_id'] = rate_code_dim.index
rate_code_dim['rate_code_name'] = rate_code_dim['RatecodeID'].map(rate_code_type)
rate_code_dim = rate_code_dim[['rate_code_id','RatecodeID','rate_code_name']]

# rate_code_dim.head()
```

In [32]:

```
rate_code_dim.head()
```

Out[32]:

	rate_code_id	RatecodeID	rate_code_name
0	0	1	Standard rate
1	1	1	Standard rate
2	2	1	Standard rate
3	3	1	Standard rate
4	4	3	Newark

In [33]:

```
pickup_location_dim = df[['pickup_longitude', 'pickup_latitude']].reset_index(drop=True
pickup_location_dim['pickup_location_id'] = pickup_location_dim.index
```

```
pickup_location_dim = pickup_location_dim[['pickup_location_id','pickup_latitude','pick


dropoff_location_dim = df[['dropoff_longitude', 'dropoff_latitude']].reset_index(drop=T
dropoff_location_dim['dropoff_location_id'] = dropoff_location_dim.index
dropoff_location_dim = dropoff_location_dim[['dropoff_location_id','dropoff_latitude','
```

In [34]:

```
payment_type_name = {
    1:"Credit card",
    2:"Cash",
    3:"No charge",
    4:"Dispute",
    5:"Unknown",
    6:"Voided trip"
}
payment_type_dim = df[['payment_type']].reset_index(drop=True)
payment_type_dim['payment_type_id'] = payment_type_dim.index
payment_type_dim['payment_type_name'] = payment_type_dim['payment_type'].map(payment_ty
payment_type_dim = payment_type_dim[['payment_type_id','payment_type','payment_type_nam
```

In [35]:

```
fact_table = df.merge(passenger_count_dim, left_on='trip_id', right_on='passenger_count
             .merge(trip_distance_dim, left_on='trip_id', right_on='trip_distance_id')
             .merge(rate_code_dim, left_on='trip_id', right_on='rate_code_id') \
             .merge(pickup_location_dim, left_on='trip_id', right_on='pickup_location_i
             .merge(dropoff_location_dim, left_on='trip_id', right_on='dropoff_location
             .merge(datetime_dim, left_on='trip_id', right_on='datetime_id') \
             .merge(payment_type_dim, left_on='trip_id', right_on='payment_type_id') \
             [['trip_id','VendorID', 'datetime_id', 'passenger_count_id',
               'trip_distance_id', 'rate_code_id', 'store_and_fwd_flag', 'pickup_locati
               'payment_type_id', 'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tol
               'improvement_surcharge', 'total_amount']]
```

In [36]:

```
payment_type_dim.columns
```

Out[36]:

```
Index(['payment_type_id', 'payment_type', 'payment_type_name'], dtype='object')
```

In [37]:

```
fact_table.columns
```

Out[37]:

```
Index(['trip_id', 'VendorID', 'datetime_id', 'passenger_count_id',
       'trip_distance_id', 'rate_code_id', 'store_and_fwd_flag',
       'pickup_location_id', 'dropoff_location_id', 'payment_type_id',
       'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tolls_amount',
       'improvement_surcharge', 'total_amount'],
      dtype='object')
```

In [38]:

```
fact_table
```

Out[38]:

	trip_id	VendorID	datetime_id	passenger_count_id	trip_distance_id	rate_code_id	store_and_fwd_
0	0	1	0	0	0	0	
1	1	1	1	1	1	1	
2	2	2	2	2	2	2	

	trip_id	VendorID	datetime_id	passenger_count_id	trip_distance_id	rate_code_id	store_and_fwd_
3	3	2	3	3	3	3	
4	4	2	4	4	4	4	
...	...	...	...	...	...	...	
99995	99995	1	99995	99995	99995	99995	
99996	99996	1	99996	99996	99996	99996	
99997	99997	1	99997	99997	99997	99997	
99998	99998	2	99998	99998	99998	99998	
99999	99999	1	99999	99999	99999	99999	

100000 rows × 17 columns

In [ ]: