In [1]:

import pandas as pd

In [2]:

df = pd.read\_csv("https://raw.githubusercontent.com/darshilparmar/uber-etl-pipeline-data-engineering-project/main/data/uber\_data.csv")

In [3]:

df.head()

Out[3]:

|  | **VendorID** | **tpep\_pickup\_datetime** | **tpep\_dropoff\_datetime** | **passenger\_count** | **trip\_distance** | **pickup\_longitude** | **pickup\_latitude** | **RatecodeID** | **store\_and\_fwd\_flag** | **dropoff\_longitude** | **dropoff\_latitude** | **payment\_type** | **fare\_amount** | **extra** | **mta\_tax** | **tip\_amount** | **tolls\_amount** | **improvement\_surcharge** | **total\_amount** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 2016-03-01 00:00:00 | 2016-03-01 00:07:55 | 1 | 2.50 | -73.976746 | 40.765152 | 1 | N | -74.004265 | 40.746128 | 1 | 9.0 | 0.5 | 0.5 | 2.05 | 0.00 | 0.3 | 12.35 |
| **1** | 1 | 2016-03-01 00:00:00 | 2016-03-01 00:11:06 | 1 | 2.90 | -73.983482 | 40.767925 | 1 | N | -74.005943 | 40.733166 | 1 | 11.0 | 0.5 | 0.5 | 3.05 | 0.00 | 0.3 | 15.35 |
| **2** | 2 | 2016-03-01 00:00:00 | 2016-03-01 00:31:06 | 2 | 19.98 | -73.782021 | 40.644810 | 1 | N | -73.974541 | 40.675770 | 1 | 54.5 | 0.5 | 0.5 | 8.00 | 0.00 | 0.3 | 63.80 |
| **3** | 2 | 2016-03-01 00:00:00 | 2016-03-01 00:00:00 | 3 | 10.78 | -73.863419 | 40.769814 | 1 | N | -73.969650 | 40.757767 | 1 | 31.5 | 0.0 | 0.5 | 3.78 | 5.54 | 0.3 | 41.62 |
| **4** | 2 | 2016-03-01 00:00:00 | 2016-03-01 00:00:00 | 5 | 30.43 | -73.971741 | 40.792183 | 3 | N | -74.177170 | 40.695053 | 1 | 98.0 | 0.0 | 0.0 | 0.00 | 15.50 | 0.3 | 113.80 |

In [4]:

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 100000 entries, 0 to 99999

Data columns (total 19 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 VendorID 100000 non-null int64

1 tpep\_pickup\_datetime 100000 non-null object

2 tpep\_dropoff\_datetime 100000 non-null object

3 passenger\_count 100000 non-null int64

4 trip\_distance 100000 non-null float64

5 pickup\_longitude 100000 non-null float64

6 pickup\_latitude 100000 non-null float64

7 RatecodeID 100000 non-null int64

8 store\_and\_fwd\_flag 100000 non-null object

9 dropoff\_longitude 100000 non-null float64

10 dropoff\_latitude 100000 non-null float64

11 payment\_type 100000 non-null int64

12 fare\_amount 100000 non-null float64

13 extra 100000 non-null float64

14 mta\_tax 100000 non-null float64

15 tip\_amount 100000 non-null float64

16 tolls\_amount 100000 non-null float64

17 improvement\_surcharge 100000 non-null float64

18 total\_amount 100000 non-null float64

dtypes: float64(12), int64(4), object(3)

memory usage: 14.5+ MB

In [5]:

#converted pickup & dropoff datetime details in DATETIME format

df["tpep\_pickup\_datetime"] = pd.to\_datetime(df["tpep\_pickup\_datetime"])

df["tpep\_dropoff\_datetime"] = pd.to\_datetime(df["tpep\_dropoff\_datetime"])

In [6]:

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 100000 entries, 0 to 99999

Data columns (total 19 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 VendorID 100000 non-null int64

1 tpep\_pickup\_datetime 100000 non-null datetime64[ns]

2 tpep\_dropoff\_datetime 100000 non-null datetime64[ns]

3 passenger\_count 100000 non-null int64

4 trip\_distance 100000 non-null float64

5 pickup\_longitude 100000 non-null float64

6 pickup\_latitude 100000 non-null float64

7 RatecodeID 100000 non-null int64

8 store\_and\_fwd\_flag 100000 non-null object

9 dropoff\_longitude 100000 non-null float64

10 dropoff\_latitude 100000 non-null float64

11 payment\_type 100000 non-null int64

12 fare\_amount 100000 non-null float64

13 extra 100000 non-null float64

14 mta\_tax 100000 non-null float64

15 tip\_amount 100000 non-null float64

16 tolls\_amount 100000 non-null float64

17 improvement\_surcharge 100000 non-null float64

18 total\_amount 100000 non-null float64

dtypes: datetime64[ns](2), float64(12), int64(4), object(1)

memory usage: 14.5+ MB

In [7]:

#dropping duplicate values

df = df.drop\_duplicates().reset\_index(drop=True)

df["trip\_id"] = df.index

In [8]:

df.head()

Out[8]:

|  | **VendorID** | **tpep\_pickup\_datetime** | **tpep\_dropoff\_datetime** | **passenger\_count** | **trip\_distance** | **pickup\_longitude** | **pickup\_latitude** | **RatecodeID** | **store\_and\_fwd\_flag** | **dropoff\_longitude** | **dropoff\_latitude** | **payment\_type** | **fare\_amount** | **extra** | **mta\_tax** | **tip\_amount** | **tolls\_amount** | **improvement\_surcharge** | **total\_amount** | **trip\_id** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 2016-03-01 | 2016-03-01 00:07:55 | 1 | 2.50 | -73.976746 | 40.765152 | 1 | N | -74.004265 | 40.746128 | 1 | 9.0 | 0.5 | 0.5 | 2.05 | 0.00 | 0.3 | 12.35 | 0 |
| **1** | 1 | 2016-03-01 | 2016-03-01 00:11:06 | 1 | 2.90 | -73.983482 | 40.767925 | 1 | N | -74.005943 | 40.733166 | 1 | 11.0 | 0.5 | 0.5 | 3.05 | 0.00 | 0.3 | 15.35 | 1 |
| **2** | 2 | 2016-03-01 | 2016-03-01 00:31:06 | 2 | 19.98 | -73.782021 | 40.644810 | 1 | N | -73.974541 | 40.675770 | 1 | 54.5 | 0.5 | 0.5 | 8.00 | 0.00 | 0.3 | 63.80 | 2 |
| **3** | 2 | 2016-03-01 | 2016-03-01 00:00:00 | 3 | 10.78 | -73.863419 | 40.769814 | 1 | N | -73.969650 | 40.757767 | 1 | 31.5 | 0.0 | 0.5 | 3.78 | 5.54 | 0.3 | 41.62 | 3 |
| **4** | 2 | 2016-03-01 | 2016-03-01 00:00:00 | 5 | 30.43 | -73.971741 | 40.792183 | 3 | N | -74.177170 | 40.695053 | 1 | 98.0 | 0.0 | 0.0 | 0.00 | 15.50 | 0.3 | 113.80 | 4 |

In [9]:

datetime\_dim = df[['tpep\_pickup\_datetime','tpep\_dropoff\_datetime']].reset\_index(drop=True)

datetime\_dim['tpep\_pickup\_datetime'] = datetime\_dim['tpep\_pickup\_datetime']

datetime\_dim['pick\_hour'] = datetime\_dim['tpep\_pickup\_datetime'].dt.hour

datetime\_dim['pick\_month'] = datetime\_dim['tpep\_pickup\_datetime'].dt.month

datetime\_dim['pick\_day'] = datetime\_dim['tpep\_pickup\_datetime'].dt.day

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\_pickup\_datetime'].dt.hour

datetime\_dim['drop\_month'] = datetime\_dim['tpep\_pickup\_datetime'].dt.month

datetime\_dim['drop\_day'] = datetime\_dim['tpep\_pickup\_datetime'].dt.day

datetime\_dim['drop\_year'] = datetime\_dim['tpep\_pickup\_datetime'].dt.year

datetime\_dim['drop\_weekday'] = datetime\_dim['tpep\_pickup\_datetime'].dt.weekday

datetime\_dim['datetime\_id'] = datetime\_dim.index

datetime\_dim = datetime\_dim[['datetime\_id', 'tpep\_pickup\_datetime','pick\_hour','pick\_month','pick\_day',

'pick\_year','pick\_weekday' , 'tpep\_dropoff\_datetime','drop\_hour','drop\_month',

'drop\_day','drop\_year','drop\_weekday']]

datetime\_dim

Out[9]:

|  | **datetime\_id** | **tpep\_pickup\_datetime** | **pick\_hour** | **pick\_month** | **pick\_day** | **pick\_year** | **pick\_weekday** | **tpep\_dropoff\_datetime** | **drop\_hour** | **drop\_month** | **drop\_day** | **drop\_year** | **drop\_weekday** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 | 2016-03-01 00:07:55 | 0 | 3 | 1 | 2016 | 1 |
| **1** | 1 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 | 2016-03-01 00:11:06 | 0 | 3 | 1 | 2016 | 1 |
| **2** | 2 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 | 2016-03-01 00:31:06 | 0 | 3 | 1 | 2016 | 1 |
| **3** | 3 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 |
| **4** | 4 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 | 2016-03-01 00:00:00 | 0 | 3 | 1 | 2016 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **99995** | 99995 | 2016-03-01 06:17:10 | 6 | 3 | 1 | 2016 | 1 | 2016-03-01 06:22:15 | 6 | 3 | 1 | 2016 | 1 |
| **99996** | 99996 | 2016-03-01 06:17:10 | 6 | 3 | 1 | 2016 | 1 | 2016-03-01 06:32:41 | 6 | 3 | 1 | 2016 | 1 |
| **99997** | 99997 | 2016-03-01 06:17:10 | 6 | 3 | 1 | 2016 | 1 | 2016-03-01 06:37:23 | 6 | 3 | 1 | 2016 | 1 |
| **99998** | 99998 | 2016-03-01 06:17:10 | 6 | 3 | 1 | 2016 | 1 | 2016-03-01 06:22:09 | 6 | 3 | 1 | 2016 | 1 |
| **99999** | 99999 | 2016-03-01 06:17:11 | 6 | 3 | 1 | 2016 | 1 | 2016-03-01 06:22:00 | 6 | 3 | 1 | 2016 | 1 |

100000 rows × 13 columns

In [10]:

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 [11]:

passenger\_count\_dim.head()

Out[11]:

|  | **passenger\_count\_id** | **passenger\_count** |
| --- | --- | --- |
| **0** | 0 | 1 |
| **1** | 1 | 1 |
| **2** | 2 | 2 |
| **3** | 3 | 3 |
| **4** | 4 | 5 |

In [12]:

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

Out[12]:

|  | **RatecodeID** | **Rate\_code\_id** | **Rate\_code\_name** |
| --- | --- | --- | --- |
| **0** | 1 | 0 | Standard rate |
| **1** | 1 | 1 | Standard rate |
| **2** | 1 | 2 | Standard rate |
| **3** | 1 | 3 | Standard rate |
| **4** | 3 | 4 | Newark |
| **...** | ... | ... | ... |
| **99995** | 1 | 99995 | Standard rate |
| **99996** | 1 | 99996 | Standard rate |
| **99997** | 1 | 99997 | Standard rate |
| **99998** | 1 | 99998 | Standard rate |
| **99999** | 1 | 99999 | Standard rate |

100000 rows × 3 columns

In [13]:

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\_longitude','pickup\_latitude']]

In [14]:

drop\_location\_dim = df[['dropoff\_longitude','dropoff\_latitude']].reset\_index(drop=True)

drop\_location\_dim['drop\_location\_id'] = drop\_location\_dim.index

drop\_location\_dim = drop\_location\_dim[['drop\_location\_id','dropoff\_longitude','dropoff\_latitude']]

drop\_location\_dim

Out[14]:

|  | **drop\_location\_id** | **dropoff\_longitude** | **dropoff\_latitude** |
| --- | --- | --- | --- |
| **0** | 0 | -74.004265 | 40.746128 |
| **1** | 1 | -74.005943 | 40.733166 |
| **2** | 2 | -73.974541 | 40.675770 |
| **3** | 3 | -73.969650 | 40.757767 |
| **4** | 4 | -74.177170 | 40.695053 |
| **...** | ... | ... | ... |
| **99995** | 99995 | -73.998245 | 40.750462 |
| **99996** | 99996 | -73.982361 | 40.752529 |
| **99997** | 99997 | -73.865028 | 40.770512 |
| **99998** | 99998 | -73.970695 | 40.759148 |
| **99999** | 99999 | -73.980354 | 40.754910 |

100000 rows × 3 columns

In [15]:

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\_type\_name)

payment\_type\_dim = payment\_type\_dim[['payment\_type\_id', 'payment\_type' , 'payment\_type\_name']]

payment\_type\_dim.head()

Out[15]:

|  | **payment\_type\_id** | **payment\_type** | **payment\_type\_name** |
| --- | --- | --- | --- |
| **0** | 0 | 1 | Credit card |
| **1** | 1 | 1 | Credit card |
| **2** | 2 | 1 | Credit card |
| **3** | 3 | 1 | Credit card |
| **4** | 4 | 1 | Credit card |

In [18]:

fact\_table = df.merge(passenger\_count\_dim, left\_on='trip\_id', right\_on='passenger\_count\_id') \

.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\_id') \

.merge(drop\_location\_dim, left\_on='trip\_id', right\_on='drop\_location\_id')\

.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\_location\_id', 'drop\_location\_id',

'payment\_type\_id', 'fare\_amount', 'extra', 'mta\_tax', 'tip\_amount', 'tolls\_amount',

'improvement\_surcharge', 'total\_amount']]

In [19]:

fact\_table

Out[19]:

|  | **trip\_id** | **VendorID** | **datetime\_id** | **passenger\_count\_id** | **trip\_distance\_id** | **Rate\_code\_id** | **store\_and\_fwd\_flag** | **pickup\_location\_id** | **drop\_location\_id** | **payment\_type\_id** | **fare\_amount** | **extra** | **mta\_tax** | **tip\_amount** | **tolls\_amount** | **improvement\_surcharge** | **total\_amount** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 1 | 0 | 0 | 0 | 0 | N | 0 | 0 | 0 | 9.0 | 0.5 | 0.5 | 2.05 | 0.00 | 0.3 | 12.35 |
| **1** | 1 | 1 | 1 | 1 | 1 | 1 | N | 1 | 1 | 1 | 11.0 | 0.5 | 0.5 | 3.05 | 0.00 | 0.3 | 15.35 |
| **2** | 2 | 2 | 2 | 2 | 2 | 2 | N | 2 | 2 | 2 | 54.5 | 0.5 | 0.5 | 8.00 | 0.00 | 0.3 | 63.80 |
| **3** | 3 | 2 | 3 | 3 | 3 | 3 | N | 3 | 3 | 3 | 31.5 | 0.0 | 0.5 | 3.78 | 5.54 | 0.3 | 41.62 |
| **4** | 4 | 2 | 4 | 4 | 4 | 4 | N | 4 | 4 | 4 | 98.0 | 0.0 | 0.0 | 0.00 | 15.50 | 0.3 | 113.80 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **99995** | 99995 | 1 | 99995 | 99995 | 99995 | 99995 | N | 99995 | 99995 | 99995 | 5.0 | 0.0 | 0.5 | 0.00 | 0.00 | 0.3 | 5.80 |
| **99996** | 99996 | 1 | 99996 | 99996 | 99996 | 99996 | N | 99996 | 99996 | 99996 | 14.0 | 0.0 | 0.5 | 2.00 | 0.00 | 0.3 | 16.80 |
| **99997** | 99997 | 1 | 99997 | 99997 | 99997 | 99997 | N | 99997 | 99997 | 99997 | 29.0 | 0.0 | 0.5 | 8.80 | 5.54 | 0.3 | 44.14 |
| **99998** | 99998 | 2 | 99998 | 99998 | 99998 | 99998 | N | 99998 | 99998 | 99998 | 5.5 | 0.5 | 0.5 | 1.36 | 0.00 | 0.3 | 8.16 |
| **99999** | 99999 | 1 | 99999 | 99999 | 99999 | 99999 | N | 99999 | 99999 | 99999 | 6.0 | 0.0 | 0.5 | 0.00 | 0.00 | 0.3 | 6.80 |

100000 rows × 17 columns

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: