In [1]:

```
!pwd
```

```
/home/ubuntu/notebooks/final exam
```

In [2]:

```
# import command.
import os
os.listdir()
```

Out[2]:

```
['2017Q3-capitalbikeshare-tripdata.csv',
 '2017q1-4.csv',
 'zomato.csv.zip',
 'practice-DM_BI_v2.ipynb',
 '09_FInalProject.ipynb',
 '2017Q2-capitalbikeshare-tripdata.csv',
 'Draft_FInalProject (1).ipynb',
 '2017-Q1-trips.zip',
 '2017Q4-capitalbikeshare-tripdata.csv',
 '2017q1.csv',
 '.ipynb_checkpoints']
```

In [4]:

```
!unzip zomato.csv
```

```
Archive:  zomato.csv.zip
  inflating: zomato.csv.csv
```

Use xsv command to find the headings of csv file in order to remove the column we do not need for better analysis.

In [5]:

```
!xsv headers zomato.csv
```

```
1   url
2   address
3   name
4   online_order
5   book_table
6   rate
7   votes
8   phone
9   location
10  rest_type
11  dish_liked
12  cuisines
13  approx_cost(for two people)
14  reviews_list
15  menu_item
16  listed_in(type)
17  listed_in(city)
```

There are 17 columns in this csv files. The following columns are not needed in the further analysis.

1 url

2 name

8 phone

14 reviews_list

We will remove these 4 columns and take the rest and name it a new csv file using csvcut command. -z is to exapnd the maximum length of characters.

In [6]:

```
!csvcut -z 2500000 -c 3,4,5,6,7,9,10,11,12,13,15,16,17 zomato.csv > zomato2.csv
```

Chekcing the existing columns for new csv file.

In [7]:

```
!csvcut -n zomato2.csv
```

```
  1: name
  2: online_order
  3: book_table
  4: rate
  5: votes
  6: location
  7: rest_type
  8: dish_liked
  9: cuisines
 10: approx_cost(for two people)
 11: menu_item
 12: listed_in(type)
 13: listed_in(city)
```

In [8]:

```
!csvstat zomato2.csv
```

```
  1. "name"

        Type of data:          Text
        Contains null values:  False
        Unique values:         8792
        Longest value:         159 characters
        Most common values:    Cafe Coffee Day (96x)
                               Onesta (85x)
                               Just Bake (73x)
                               Empire Restaurant (71x)
                               Five Star Chicken (70x)

  2. "online_order"

        Type of data:          Boolean
        Contains null values:  False
        Unique values:         2
        Most common values:    True (30444x)
                               False (21273x)
```

Check if the new csv file has common syntax errors

In [9]:

```
!csvclean zomato2.csv
```

```
No errors.
```

Our data has done the simple cleaning and we can create table now.

In [10]:

```
!pip freeze | grep -E 'ipython-sql|psycopg2'
```

```
ipython-sql==0.4.1
psycopg2==2.9.5
psycopg2-binary==2.9.5
```

In [11]:

```
 %load_ext sql
```

In [12]:

```
!dropdb -U student GP9
```

In [13]:

```
!createdb -U student GP9
```

In [14]:

```
%sql postgresql://student@/GP9
```

In [15]:

```
!psql --version
```

```
psql (PostgreSQL) 12.12 (Ubuntu 12.12-0ubuntu0.20.04.1)
```

## Creating ZOMATO table

In [18]:

```
%%sql
DROP TABLE IF EXISTS ZOMATO Cascade;
CREATE TABLE ZOMATO (
    name VARCHAR(100),
    online_order VARCHAR(100),
    book_table VARCHAR(100),
    rate VARCHAR(10),
    votes INTEGER,
    location VARCHAR(100),
    rest_type VARCHAR(100),
    dish_liked VARCHAR(100),
    cuisines VARCHAR(100),
    approx_cost_two_people INTEGER,
    menu_item VARCHAR(100),
    listed_in_type VARCHAR(100),
    listed_in_city VARCHAR(100)
);
```

```
 * postgresql://student@/GP9
Done.
Done.
```

Out[18]:

```
[]
```

In [19]:

```
%%sql
select * from ZOMATO;
```

```
 * postgresql://student@/GP9
0 rows affected.
```

Out[19]:

| name | online_order | book_table | rate | votes | location | rest_type | dish_liked | cuisines | approx_c |
|---|---|---|---|---|---|---|---|---|---|

In [31]:

```
%%sql
COPY ZOMATO FROM '/home/ubuntu/notebooks/final exam/zomato2.csv'
CSV
HEADER;
```

```
 * postgresql://student@/GP9
(psycopg2.errors.BadCopyFileFormat) missing data for column "location_key"
CONTEXT:  COPY zomato, line 2: "Jalsa,Yes,Yes,4.1/5,775,Banashankari,Casual
Dining,"Pasta, Lunch Buffet, Masala Papad, Paneer Lajawa..."

[SQL: COPY ZOMATO FROM '/home/ubuntu/notebooks/final exam/zomato2.csv'
CSV
HEADER;]
(Background on this error at: https://sqlalche.me/e/14/9h9h) (https://sqlalc
he.me/e/14/9h9h))
```

## star schema

In [ ]:

```

```

## Create location table as a dimension table

In [21]:

```
%%sql
DROP TABLE IF EXISTS location;
CREATE TABLE location(
         Key SERIAL PRIMARY KEY,
         location VARCHAR(100),
         cuisines VARCHAR(100),
         menu_item VARCHAR(100)
         );
```

```
 * postgresql://student@/GP9
Done.
Done.
```

Out[21]:

```
[]
```

### Populate the location table with data from table ZOMATO

In [22]:

```
%%sql
INSERT INTO location(location ,cuisines,menu_item)
SELECT DISTINCT location , cuisines ,menu_item
FROM ZOMATO;
```

```
 * postgresql://student@/GP9
0 rows affected.
```

Out[22]:

```
[]
```

In [23]:

```
%%sql
select * from location limit 10
```

```
 * postgresql://student@/GP9
0 rows affected.
```

Out[23]:

| key | location | cuisines | menu_item |
|---|---|---|---|

In [24]:

```
%%sql
ALTER TABLE ZOMATO
ADD COLUMN location_key INTEGER,
ADD CONSTRAINT fk_location
    FOREIGN KEY (location_key)
    REFERENCES location (key);
```

```
 * postgresql://student@/GP9
Done.
```

Out[24]:

```
[]
```

In [25]:

```
%%sql
UPDATE ZOMATO
SET location_key = location.key
FROM location
```

```
 * postgresql://student@/GP9
0 rows affected.
```

Out[25]:

```
[]
```

## Create cuisines table as a dimension table

In [26]:

```
%%sql
DROP TABLE IF EXISTS cuisines;
CREATE TABLE cuisines(
         Key SERIAL PRIMARY KEY,
         cuisines VARCHAR(100),
         menu_item VARCHAR(100),
         rate VARCHAR(10)
         );
```

```
 * postgresql://student@/GP9
Done.
Done.
```

Out[26]:

```
[]
```

## Populate the cuisines table with data from table ZOMATO

In [27]:

```
%%sql
INSERT INTO cuisines(cuisines ,menu_item,rate)
SELECT DISTINCT cuisines , menu_item , rate
FROM ZOMATO;
```

```
 * postgresql://student@/GP9
0 rows affected.
```

Out[27]:

```
[]
```

In [28]:

```
%%sql
select * from cuisines limit 10
```

```
 * postgresql://student@/GP9
0 rows affected.
```

Out[28]:

| key | cuisines | menu_item | rate |
|---|---|---|---|

In [ ]:

```

```