MySQL EDA

MySQL EDA had to load df.csv 100MM records from the MySQL command line with the following:

Note: do the following first before proceeding to connect to MySQL
MySQL EDA had to load df.csv 100MM records from the MySQL command line with the following:

sb@DESKTOP-P48C40B c:\xampp\mysql\bin

#> mysql.exe -u root --password
#> Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g. Your MariaDB connection id is 432 Server version: 10.4.6-MariaDB mariadb.org binary distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> use netflixstudy
Database changed

MariaDB [netflixstudy]> load data local infile 'df_test.csv' into table ratings;
Query OK, 108 rows affected, 432 warnings (0.016 sec) Records: 108 Deleted: 0 Skipped: 0 Warnings: 432

MariaDB [netflixstudy]> load data local infile 'df_test.csv' into table ratings fields terminated by ',';
Query OK, 108 rows affected, 111 warnings (0.004 sec) Records: 108 Deleted: 0 Skipped: 0 Warnings: 111

MariaDB [netflixstudy]> load data local infile 'df.csv' into table ratings fields terminated by ',';
Query OK, 71833510 rows affected, 65535 warnings (6 min 47.817 sec) Records: 71833510 Deleted: 0 Skipped: 0 Warnings: 71833513

MariaDB [netflixstudy]>

In [3]:
```
# Import necessary modules
import pandas as pd
import mysql.connector
from mysql.connector import errorcode
```

In [4]:
```
# from https://dev.mysql.com/doc/connector-python/en/connector-python-example-connecting.html
try:
    db = mysql.connector.connect(
        host="localhost",
        user="root",
        password="",
        database="netflixstudy"
        )
except mysql.connector.Error as err:
  if err.errno == errorcode.ER_ACCESS_DENIED_ERROR:
    print("Something is wrong with your user name or password")
  elif err.errno == errorcode.ER_BAD_DB_ERROR:
    print("Database does not exist")
  else:
    print(err)
# else:
#   cnx.close()
# db.close()   # at some point need to close the connection with this instruction

cursor = db.cursor(buffered=True)  # to avoid [error](https://stackoverflow.com/questions/29772337/python-mysql-connector-unread-result-found-when-using-fetchone)
```

In [2]:
```
# Run this when done querying, otherwise, don't use
# cursor.close()
# db.close()
```

In [10]:
```
# Build select statement for ratings table:
query = (
        "SELECT Cust_Id,Movie_Id,Rating "
            "FROM ratings "
            "where Cust_Id = 97"
        )

# Execute the statement and fetch the results: results
cursor.execute(query)

result = cursor.fetchall()

data = {}
df = pd.DataFrame(data)

print('First read of netflixstudy database :\n')

# Both print result to screen and append to dataframe
for r in result:
    data = {
        'Cust_Id': r[0],
        'Movie_Id': r[1],
        'Rating': r[2]
    }
    df = df.append(data, ignore_index=True)

df['Cust_Id'] = df['Cust_Id'].astype(int)
df['Movie_Id'] = df['Movie_Id'].astype(int)

df
```

```
Members and guests that paid more than $30 for a facility :
```

Out[10]:

	Cust_Id	Movie_Id	Rating
0	97	83	4.0
1	97	167	4.0
2	97	175	3.0
3	97	270	3.0
4	97	275	4.0
...	...	...	...
479	97	17441	5.0
480	97	17479	2.0
481	97	17621	4.0
482	97	17627	5.0
483	97	17692	2.0

484 rows × 3 columns

In [15]:
```
# Build select statement for ratings table, average rating for a customer ID:
query = (
        "SELECT Cust_Id, avg(Rating) "
            "FROM ratings "
            "where Cust_Id = 97"
        )

# Execute the statement and fetch the results: results
cursor.execute(query)

result2 = cursor.fetchall()

data = {}
df2 = pd.DataFrame(data)

print('Average Rating for a Customer ID :\n')

# Both print result to screen and append to dataframe
for r in result2:
    data = {
        'Cust_Id': r[0],
        'Avg_Rating': r[1]
    }
    df2 = df2.append(data, ignore_index=True)

df2 = df2[['Cust_Id','Avg_Rating']]          # rearrange columns from alphabet
ic default order
df2['Cust_Id'] = df2['Cust_Id'].astype(int)  # remove .0 from customer ID

df2
```

```
Average Rating for a Customer ID :
```

Out[15]:

	Cust_Id	Avg_Rating
0	97	3.225207

In [16]:
```
# Build select statement for ratings table, average rating for a Movie ID:
query = (
        "SELECT Movie_Id, avg(Rating) "
            "FROM ratings "
            "where Movie_Id = 1001"
        )

# Execute the statement and fetch the results: results
cursor.execute(query)

result3 = cursor.fetchall()

data = {}
df3 = pd.DataFrame(data)

print('Average Rating for Movie ID :\n')

# Both print result to screen and append to dataframe
for r in result3:
    data = {
        'Movie_Id': r[0],
        'Avg_Rating': r[1]
    }
    df3 = df3.append(data, ignore_index=True)

df3 = df3[['Movie_Id','Avg_Rating']]
df3['Movie_Id'] = df3['Movie_Id'].astype(int)


df3
```

```
Average Rating for Movie ID :
```

Out[16]:

	Movie_Id	Avg_Rating
0	1001	3.292859