**CIS 593 – BIG DATA**

**LAB ASSIGNMENT – 2**

**Name: Pooja A. Khatri**

**CSU ID: 2783752**

1. **Platforms**

Language used: Python

Python IDE: PyCharm

Database design tool: MySQL Workbench

Interface for connecting to a MySQL database: pymysql

1. **Files included**

CSV files: json\_businessmaster.csv , json\_categories.csv, json\_attributes.csv, json\_hours.csv

Python files: businessmaster.py , businesscategories.py , businessattributes.py , businesshours.py

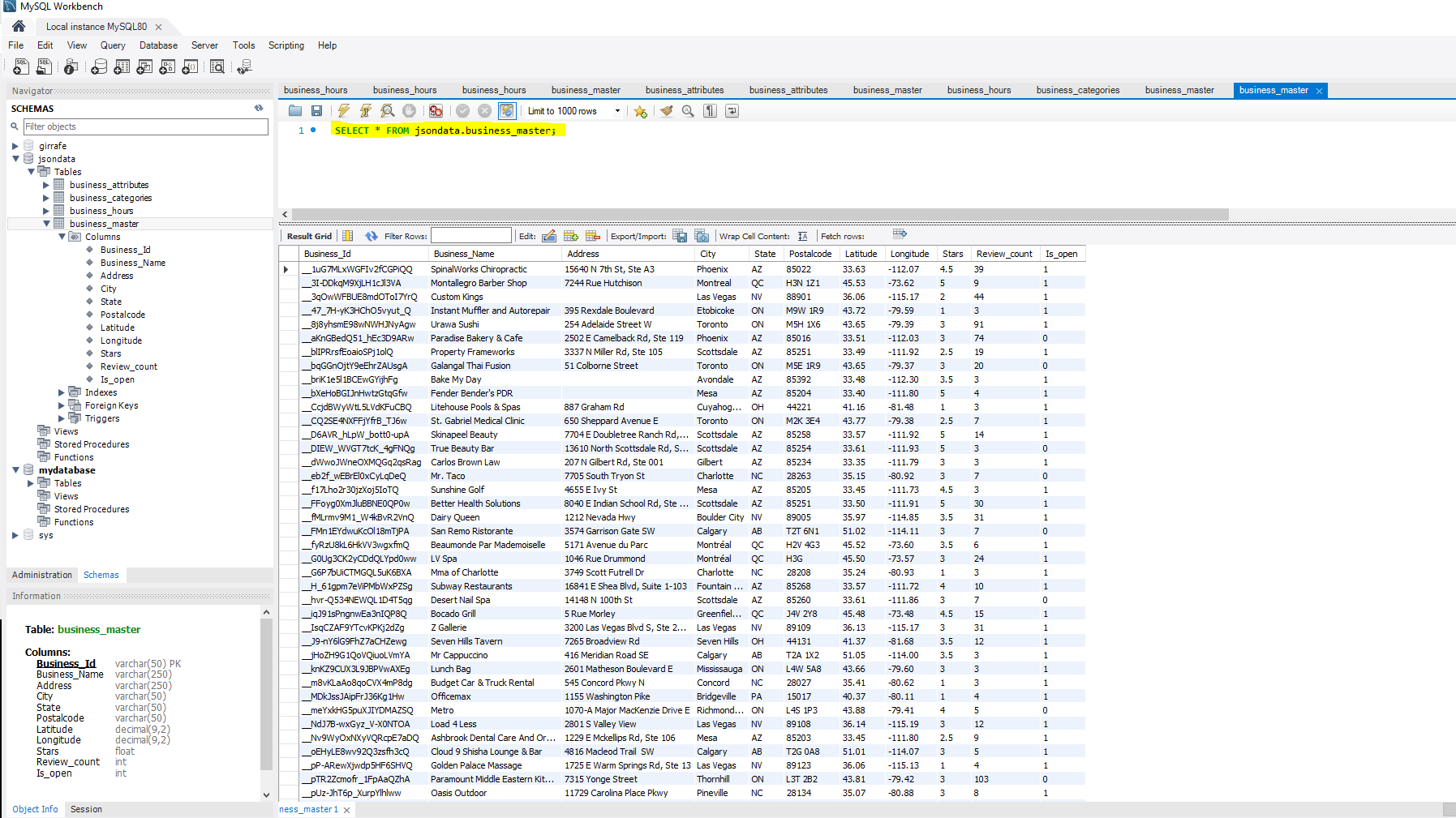
Json file: yelp\_academic\_dataset\_business.json

1. **Steps to get the result**

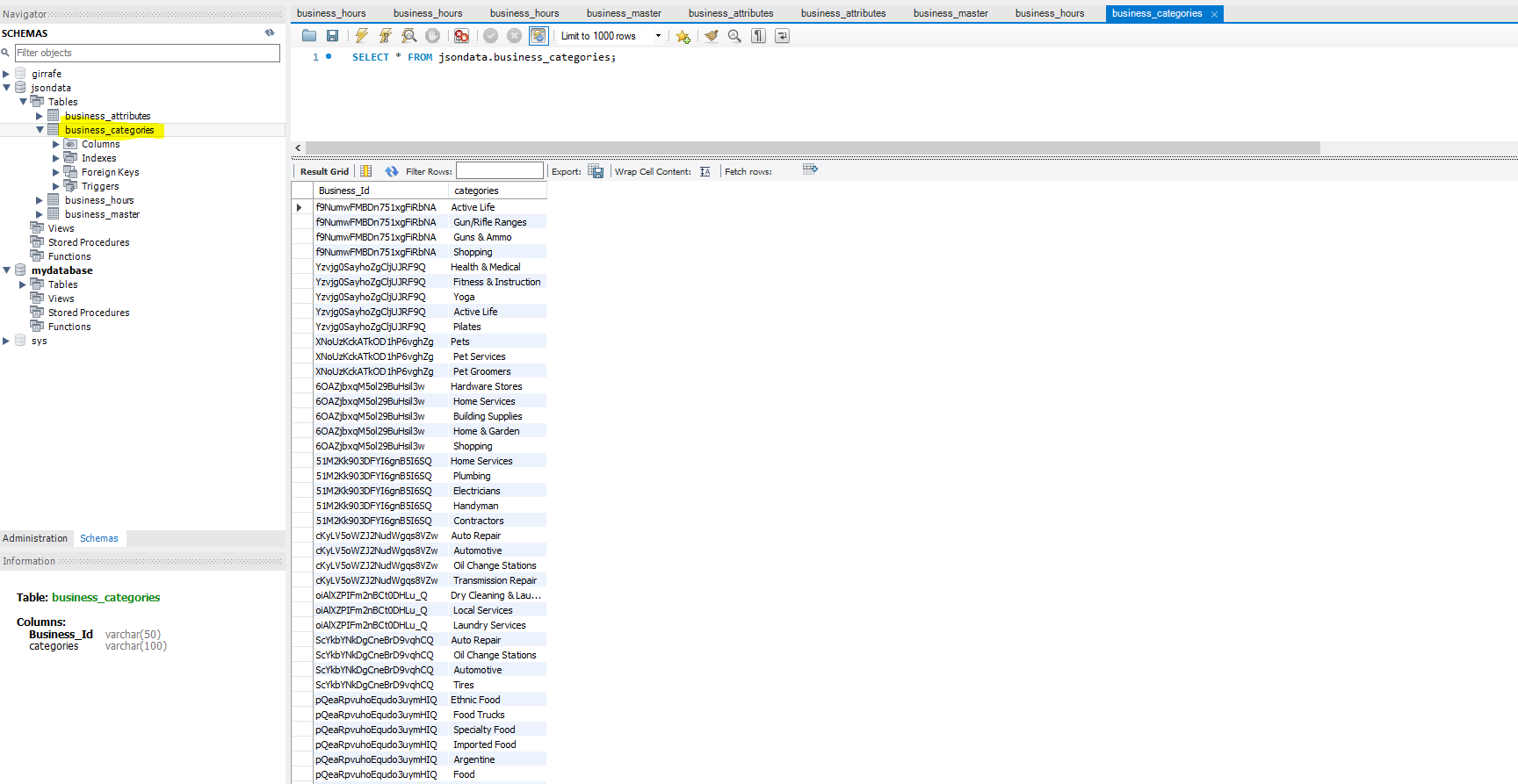
* Generate one master table which has the keys and values of each records of json file which are not dictionaries
* Generate one table for hours dictionary records inside the main dictionary. Store it in table in 1NF form so that separate row is created in table for each days for each record of file
* Generate one table to store the attributes dictionary records inside all records. Again, separate rows created for all attributes of single business ID
* Generate one table to store the categories dictionary records inside all records. Same way, segregated for all multiple values in multiple rows with business Id
* Fetch records from multiple tables linking with business ID as business ID is primary key in business master table and foreign key in all other tables.

1. **Screenshots**

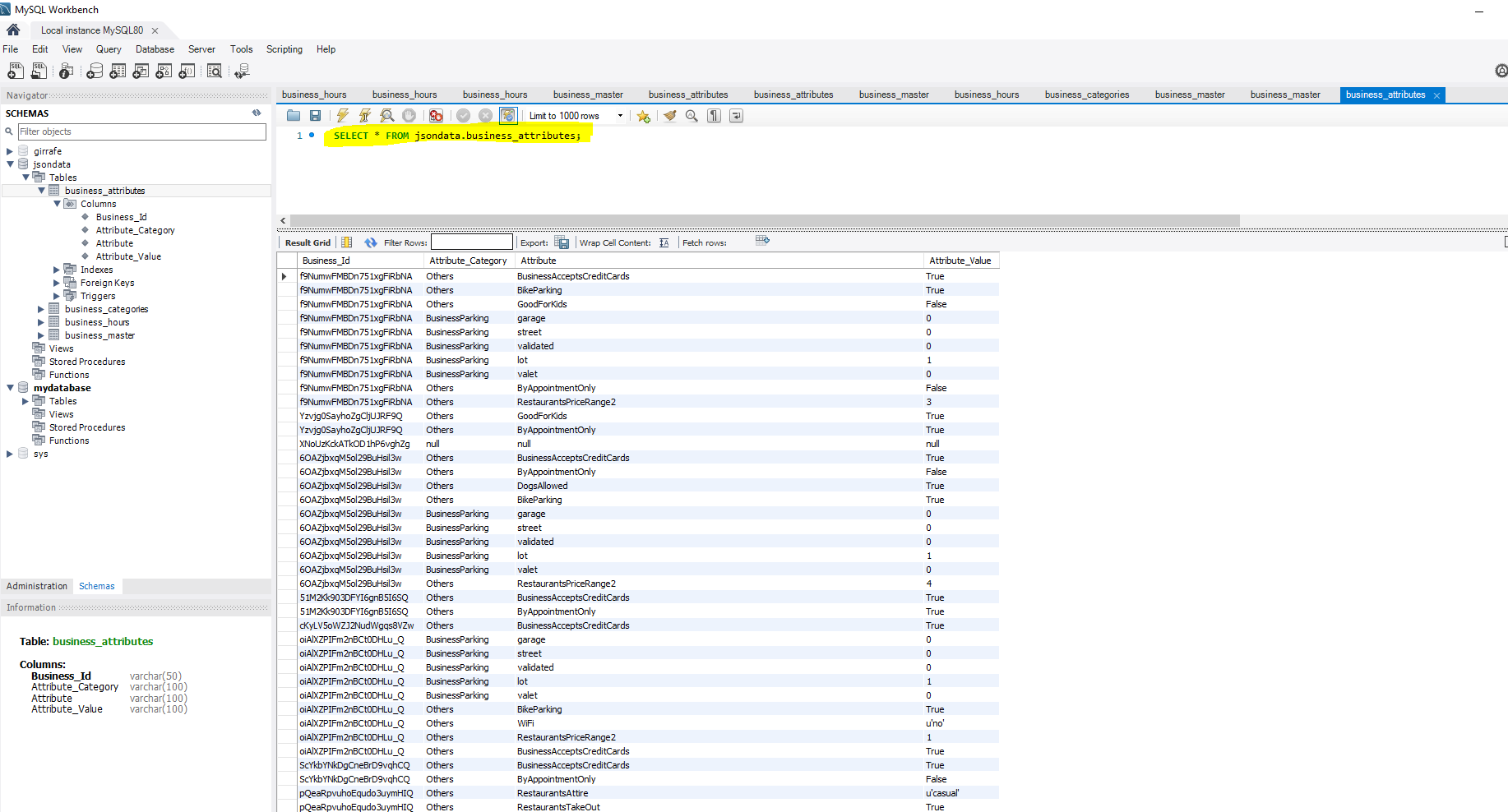
Business\_master table showing all other records except dictionary records inside main dictionary.[Business\_Id is Primary key here]



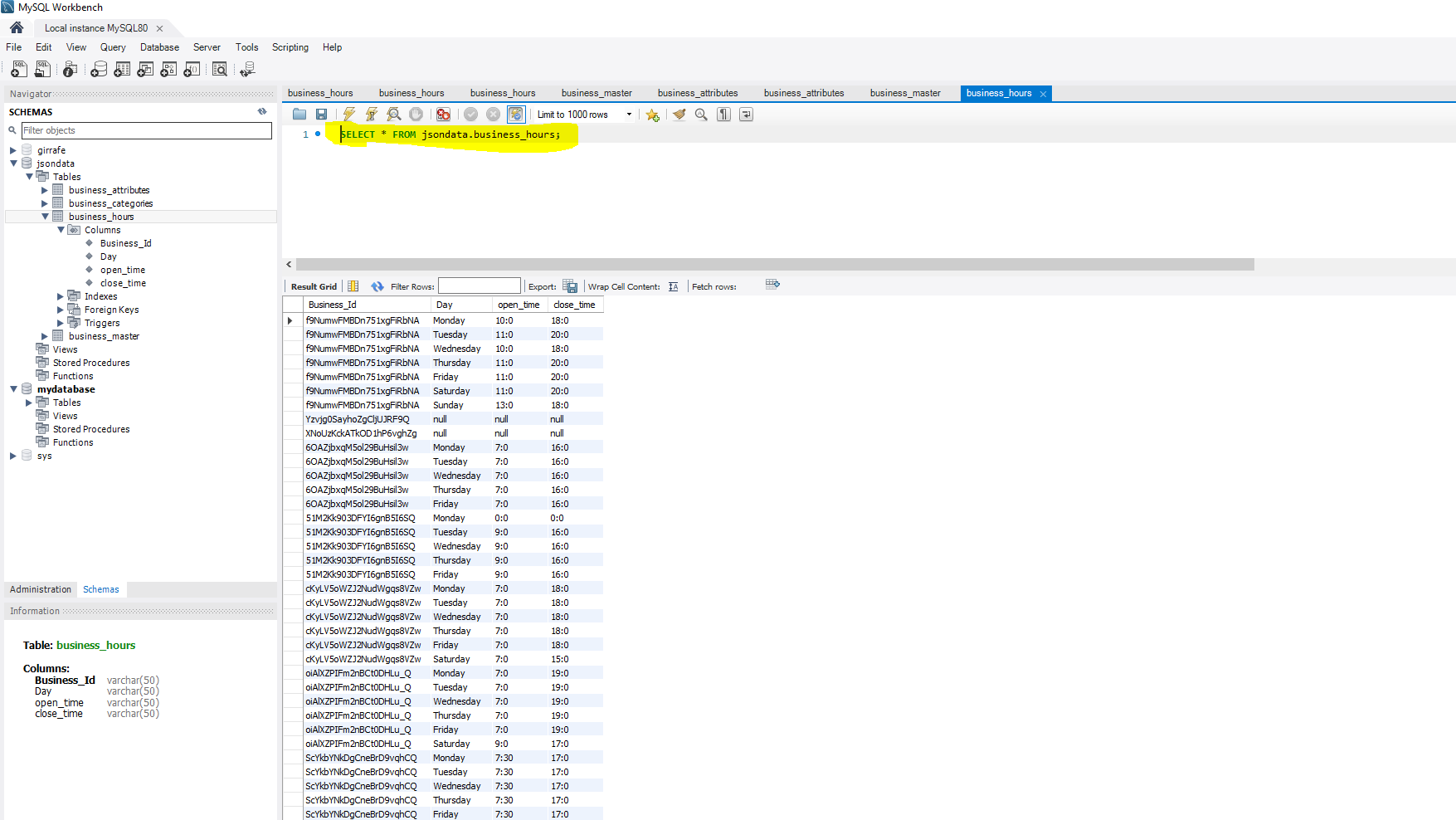
Business\_categories table showing all category dictionary records with Respective Business ID(Foreign Key)



Business\_attributes table showing all attributes dictionary records with Respective Business ID(Foreign Key)

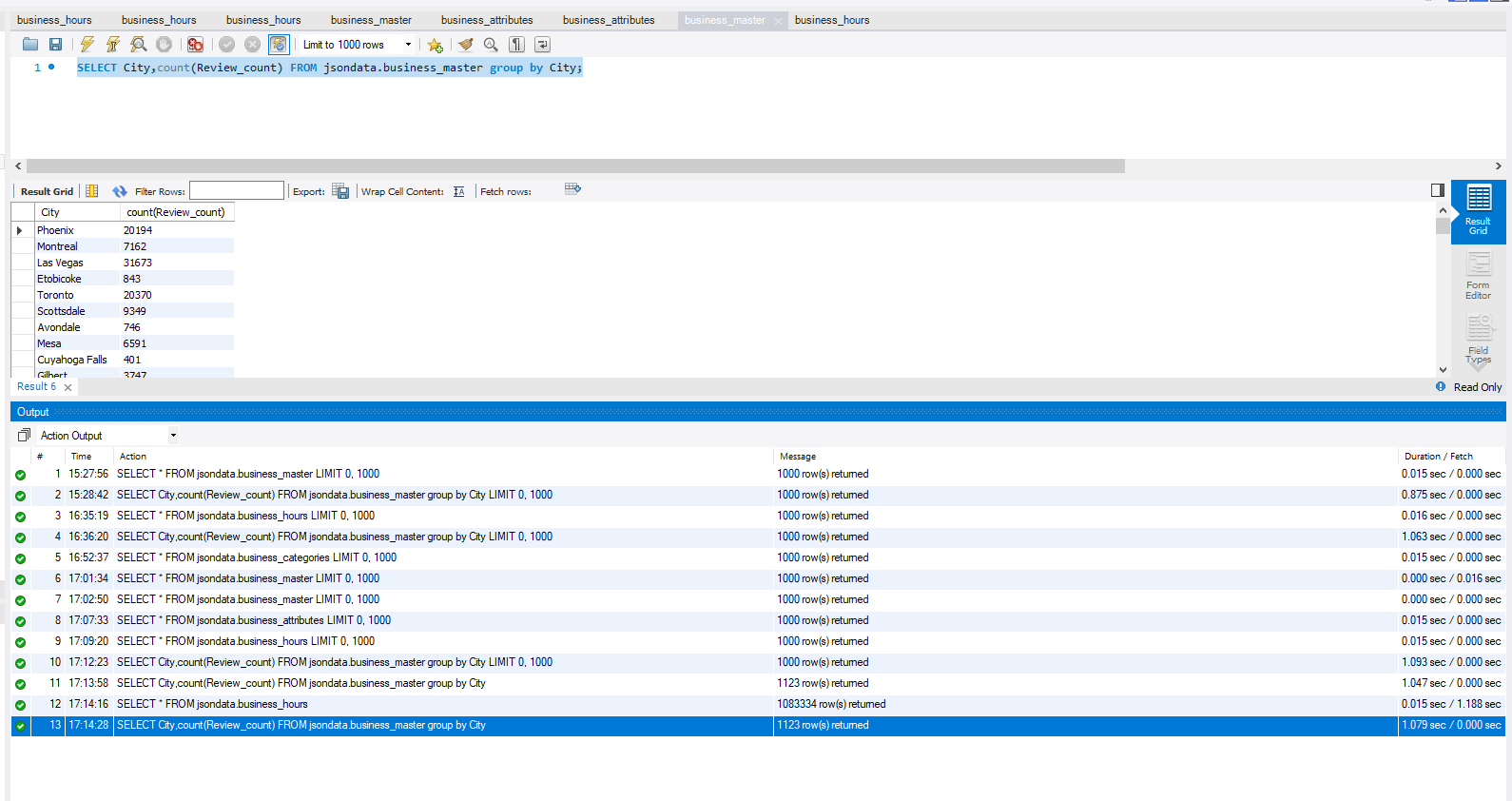


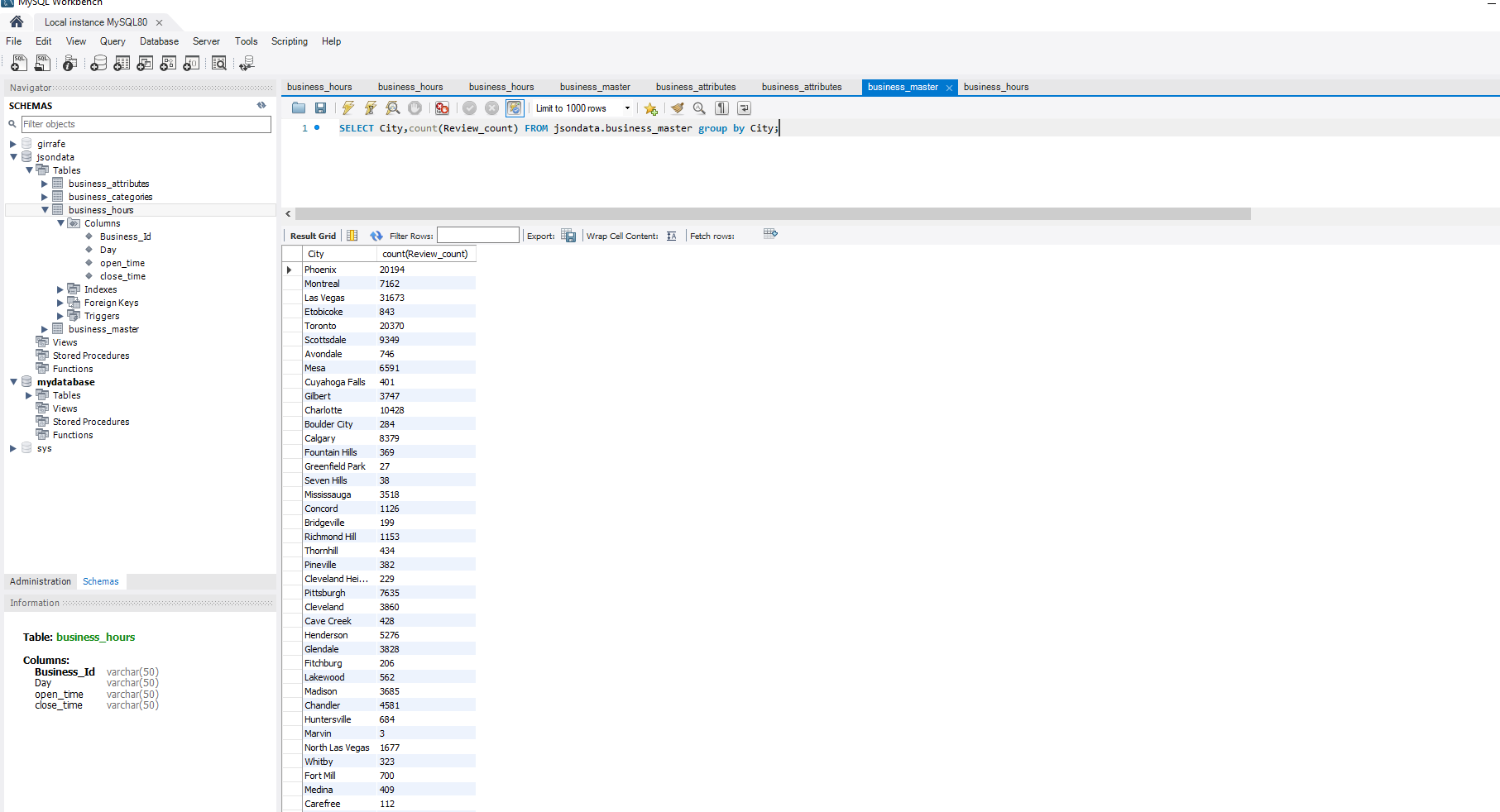
business\_hours table showing all hours dictionary records with Respective Business ID(Foreign Key)



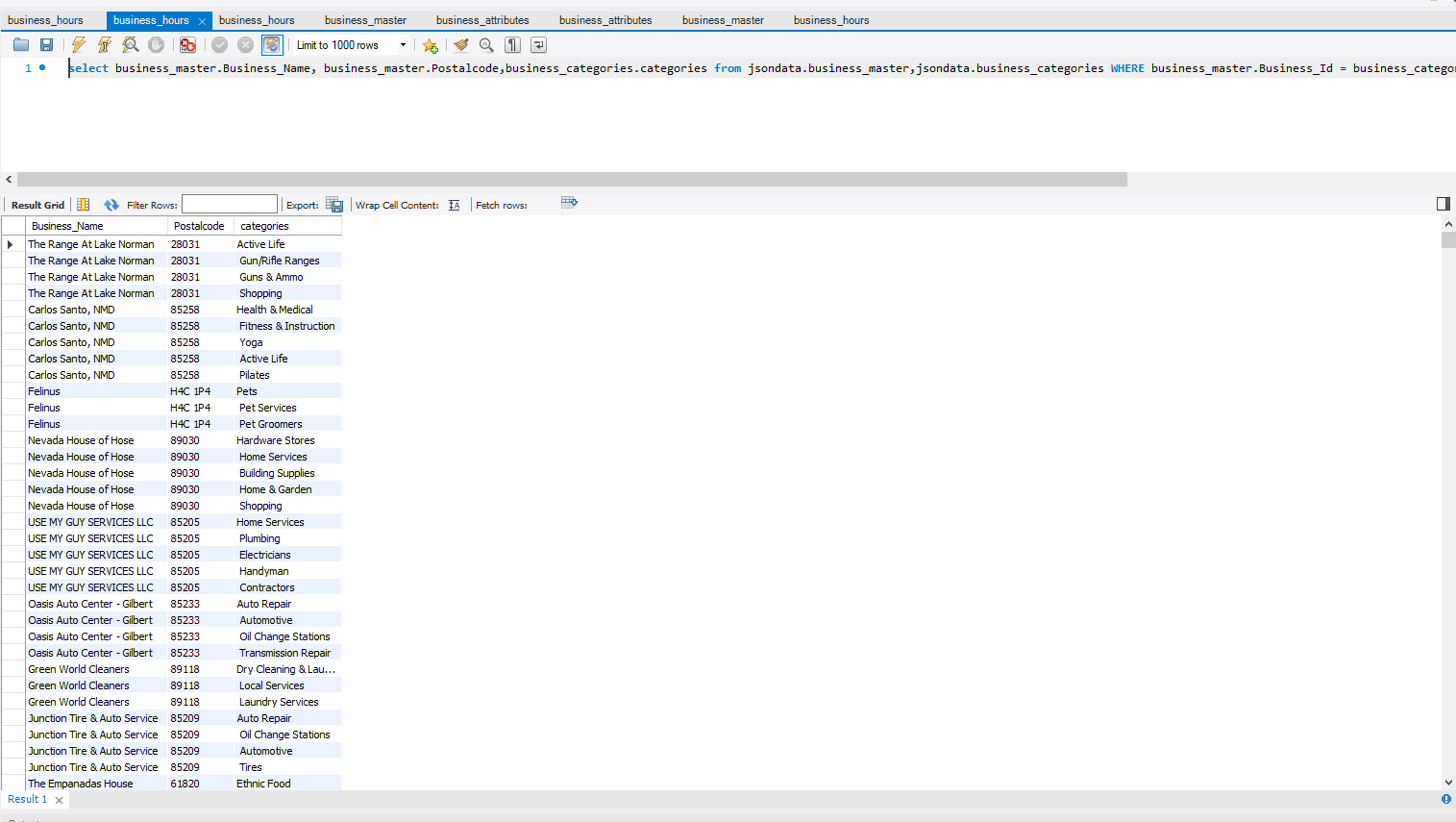
Submit the following SQL result and show how long the query takes to return the query result

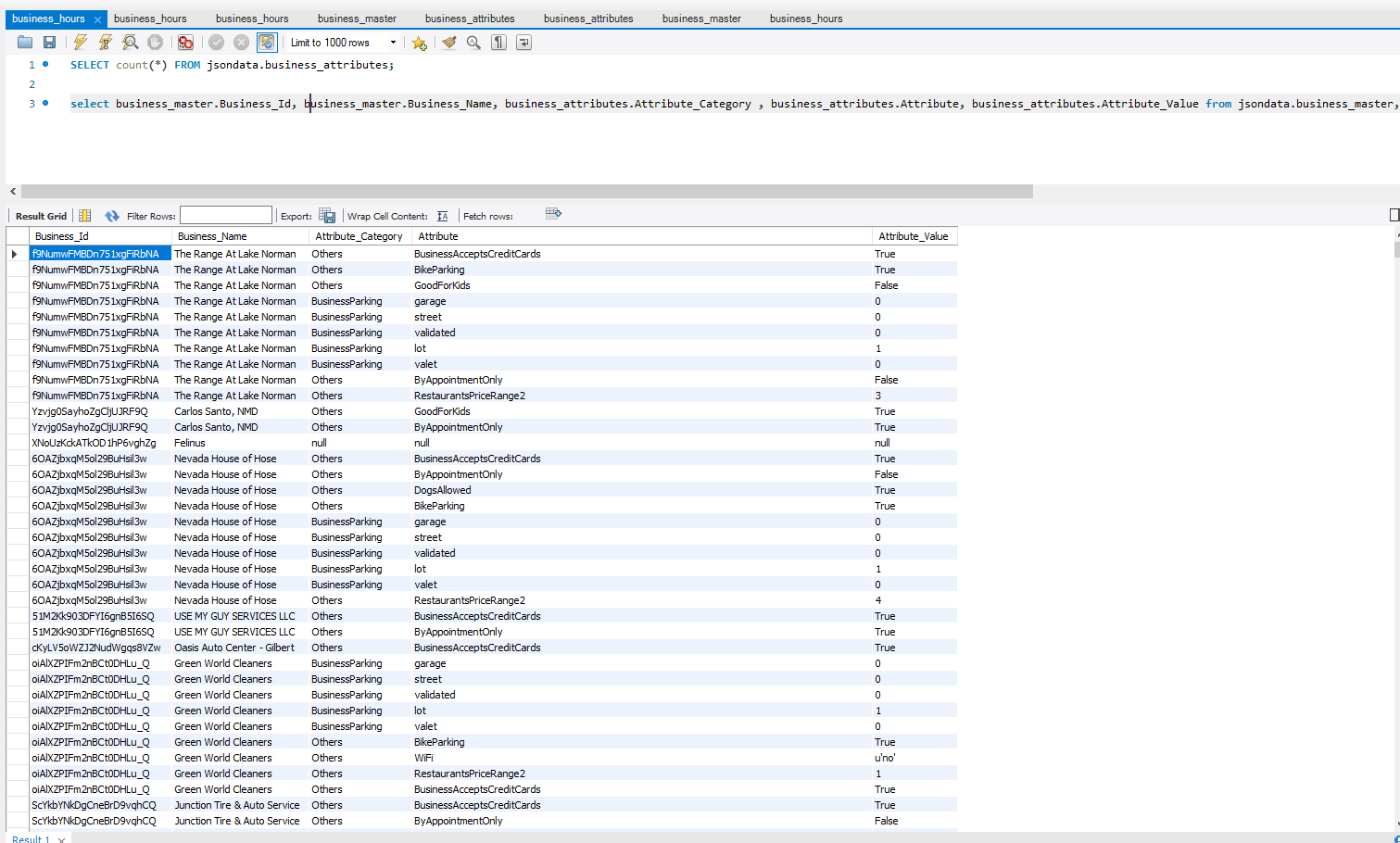
Select city, COUNT(review\_count) From Business Group by city;





Fetching records from multiple tables

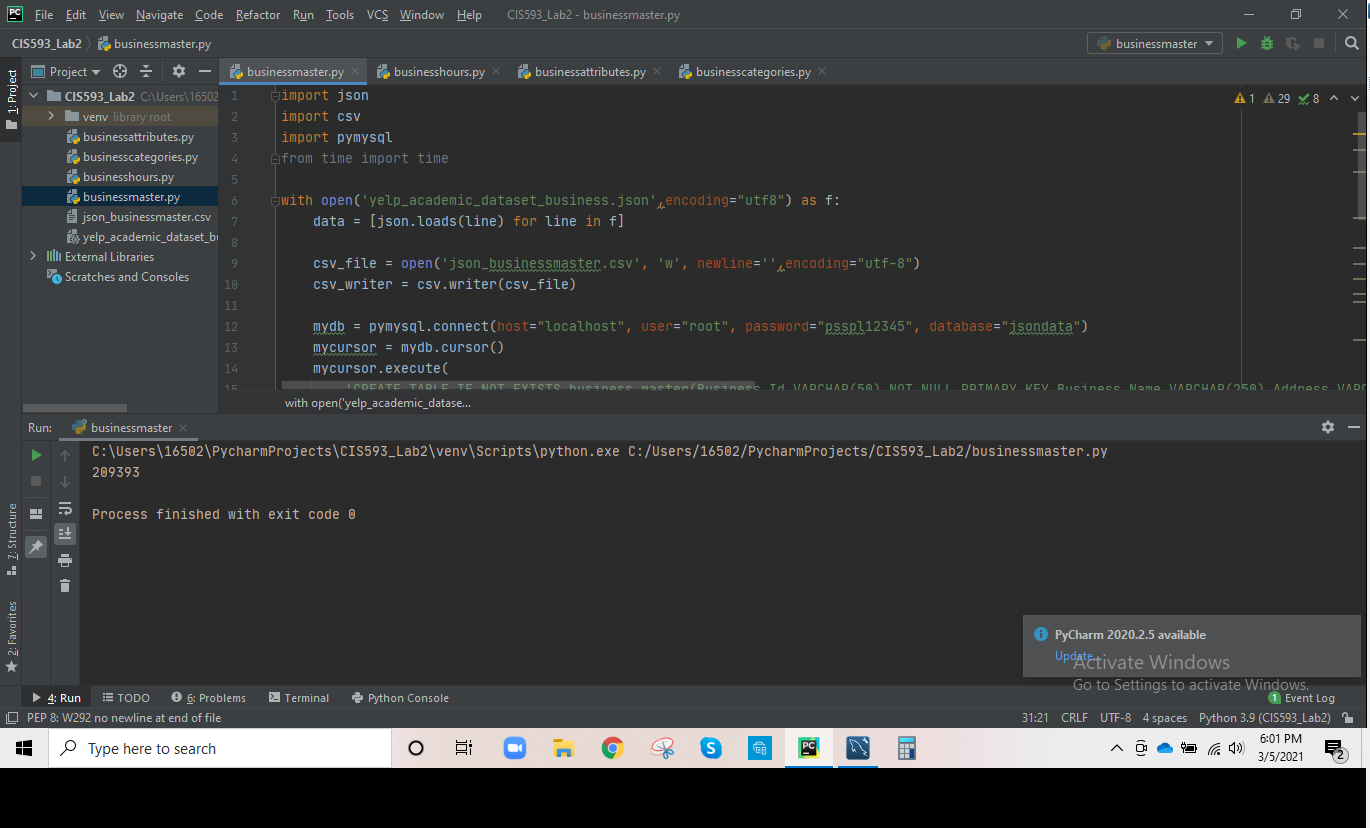


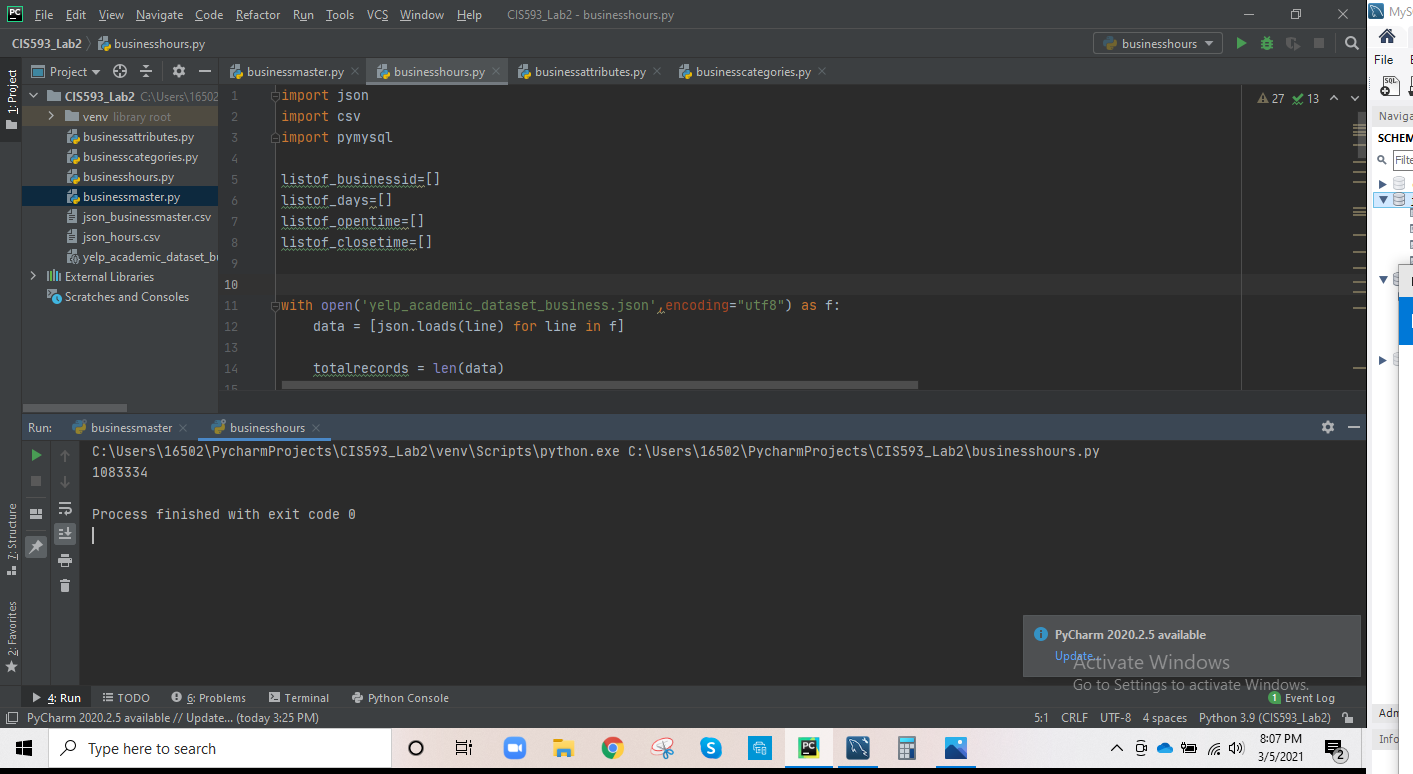


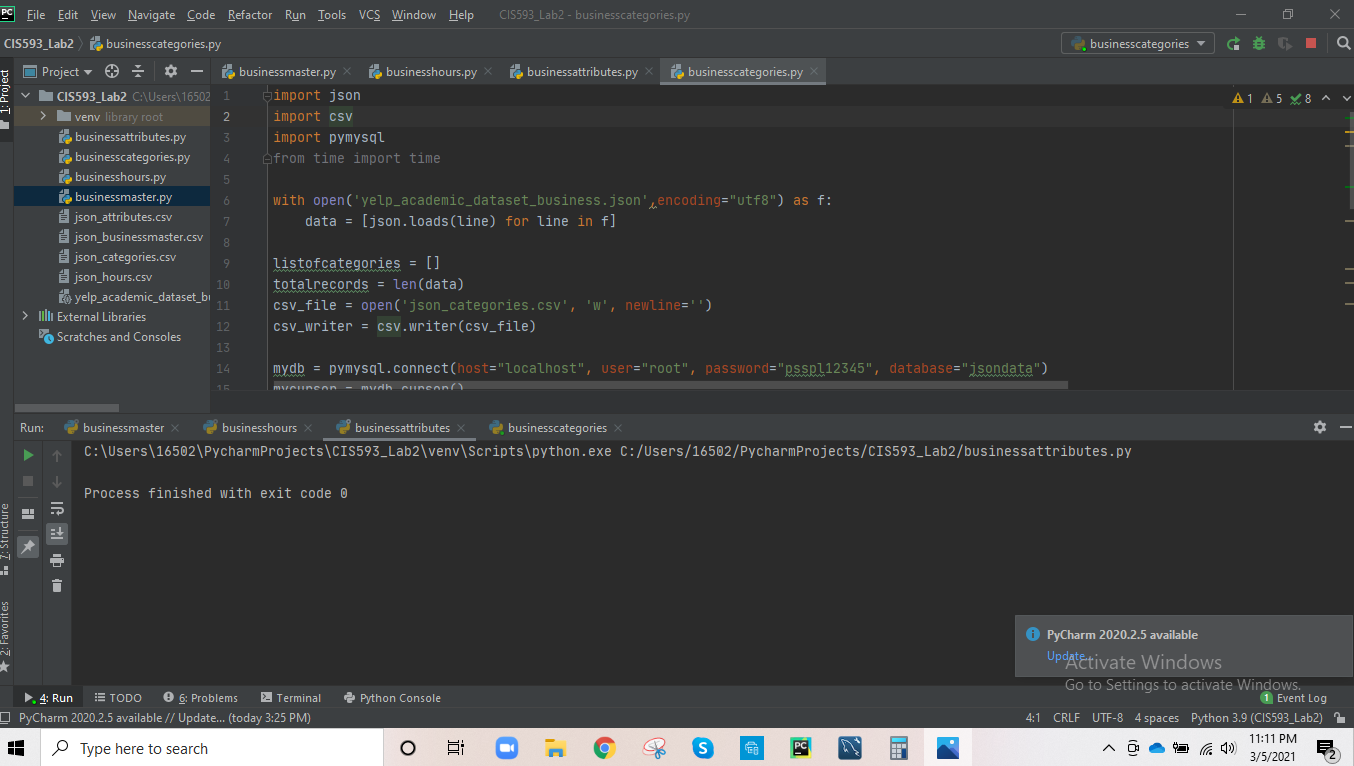
Time taken and Total records returned by above two queries:

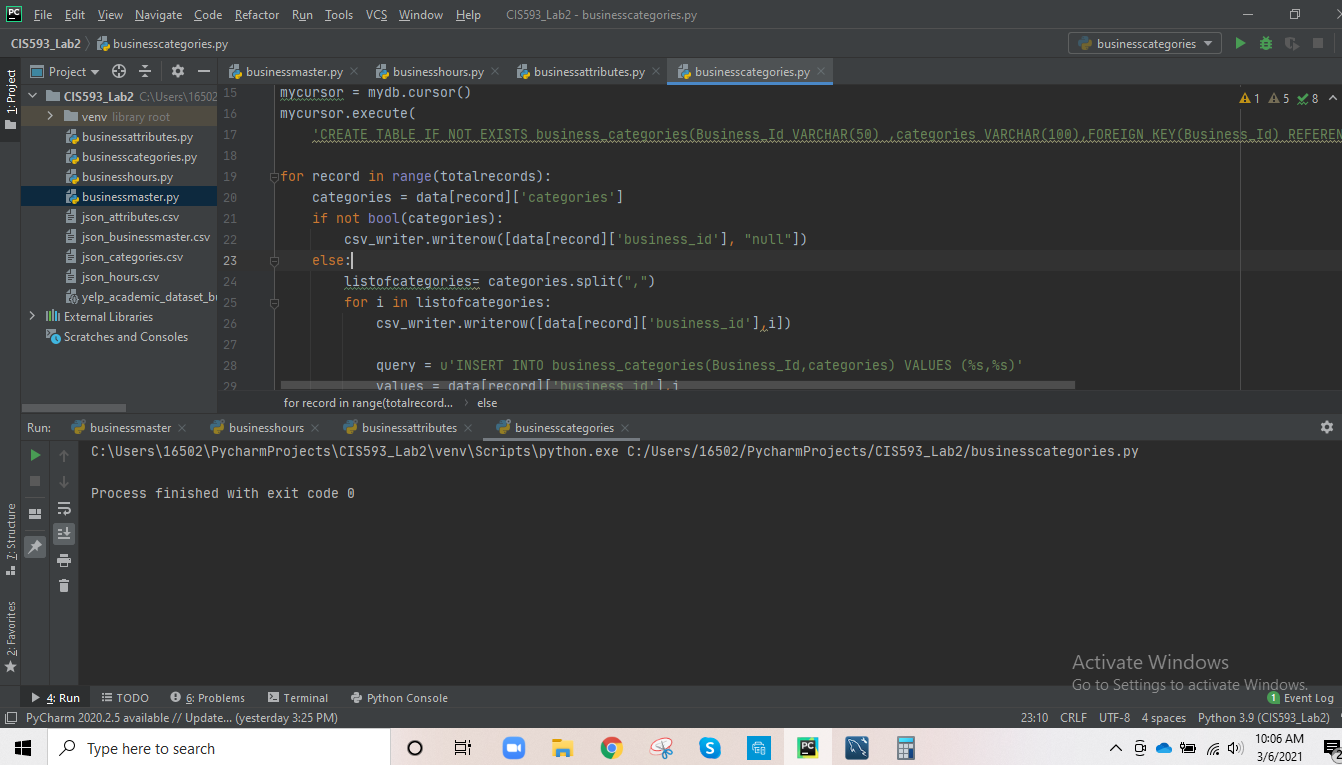


Pycharm showing successful execution of all four files:









1. **Source Code:**

Businessmaster.py

import json  
import csv  
import pymysql  
  
with open('yelp\_academic\_dataset\_business.json',encoding="utf8") as f:  
 data = [json.loads(line) for line in f]  
  
 csv\_file = open('json\_businessmaster.csv', 'w', newline='',encoding="utf-8")  
 csv\_writer = csv.writer(csv\_file)  
  
 mydb = pymysql.connect(host="localhost", user="root", password="psspl12345", database="jsondata")  
 mycursor = mydb.cursor()  
 mycursor.execute(  
 'CREATE TABLE IF NOT EXISTS business\_master(Business\_Id VARCHAR(50) NOT NULL PRIMARY KEY,Business\_Name VARCHAR(250),Address VARCHAR(250),City VARCHAR(50),State VARCHAR(50),Postalcode VARCHAR(50),Latitude DECIMAL(9,2),Longitude DECIMAL(9,2), Stars FLOAT,Review\_count INT ,Is\_open INT)')  
  
  
  
 totalrecords = len(data)  
 print(totalrecords)  
 #Fetch data hours data and store it in lists  
 for record in range(totalrecords):  
 csv\_writer.writerow([data[record]['business\_id'],data[record]['name'],data[record]['address'],data[record]['city'],data[record]['state'],data[record]['postal\_code'],data[record]['latitude'],data[record]['longitude'],data[record]['stars'],data[record]['review\_count'],data[record]['is\_open']])  
  
 query = u'INSERT INTO business\_master(Business\_Id,Business\_Name,Address,City,State,Postalcode,Latitude,Longitude,Stars,Review\_count,Is\_open) VALUES (%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s)'  
 values = data[record]['business\_id'],data[record]['name'],data[record]['address'],data[record]['city'],data[record]['state'],data[record]['postal\_code'],data[record]['latitude'],data[record]['longitude'],data[record]['stars'],data[record]['review\_count'],data[record]['is\_open']  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 csv\_file.close()  
 mycursor.close()

businesscategories.py

import json  
import csv  
import pymysql  
  
with open('yelp\_academic\_dataset\_business.json',encoding="utf8") as f:  
 data = [json.loads(line) for line in f]  
  
listofcategories = []  
totalrecords = len(data)  
csv\_file = open('json\_categories.csv', 'w', newline='')  
csv\_writer = csv.writer(csv\_file)  
  
mydb = pymysql.connect(host="localhost", user="root", password="psspl12345", database="jsondata")  
mycursor = mydb.cursor()  
mycursor.execute(  
 'CREATE TABLE IF NOT EXISTS business\_categories(Business\_Id VARCHAR(50) ,categories VARCHAR(100),FOREIGN KEY(Business\_Id) REFERENCES jsondata.business\_master(Business\_Id))')  
  
for record in range(totalrecords):  
 categories = data[record]['categories']  
 if not bool(categories):  
 csv\_writer.writerow([data[record]['business\_id'], "null"])  
 else:  
 listofcategories= categories.split(",")  
 for i in listofcategories:  
 csv\_writer.writerow([data[record]['business\_id'],i])  
  
 query = u'INSERT INTO business\_categories(Business\_Id,categories) VALUES (%s,%s)'  
 values = data[record]['business\_id'],i  
 mycursor.execute(query, values)  
 mydb.commit()  
  
csv\_file.close()  
mycursor.close()

businessattributes.py

import json  
import csv  
import pymysql  
  
  
with open('yelp\_academic\_dataset\_business.json',encoding="utf8") as f:  
 data = [json.loads(line) for line in f]  
  
  
  
totalrecords = len(data)  
csv\_file = open('json\_attributes.csv', 'w', newline='')  
csv\_writer = csv.writer(csv\_file)  
  
mydb = pymysql.connect(host="localhost", user="root", password="psspl12345", database="jsondata")  
mycursor = mydb.cursor()  
mycursor.execute(  
 'CREATE TABLE IF NOT EXISTS business\_attributes(Business\_Id VARCHAR(50) ,Attribute\_Category VARCHAR(100),Attribute VARCHAR(100) NULL,Attribute\_Value VARCHAR(100) NULL,FOREIGN KEY(Business\_Id) REFERENCES jsondata.business\_master(Business\_Id))')  
  
for record in range(totalrecords):  
 has\_items = bool(data[record]['attributes'])  
 if has\_items == True:  
 for k,v in data[record]['attributes'].items():  
 if k == "Ambience":  
 ambiencestring = data[record]['attributes']['Ambience']  
 ambiencedict = eval(ambiencestring)  
  
 if not bool(ambiencedict):  
 csv\_writer.writerow([data[record]['business\_id'], 'Ambience', "null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'Ambience', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 else:  
 for key, val in ambiencedict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'Ambience', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'Ambience', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 elif k == "GoodForMeal":  
 GoodForMealstring = data[record]['attributes']['GoodForMeal']  
 GoodForMealdict = eval(GoodForMealstring)  
 if not bool(GoodForMealdict):  
 csv\_writer.writerow([data[record]['business\_id'], 'GoodForMeal',"null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'GoodForMeal', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
 else:  
 for key, val in GoodForMealdict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'GoodForMeal', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'GoodForMeal', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 elif k == "BusinessParking":  
 businessparkingstring = data[record]['attributes']['BusinessParking']  
 businessparkingdict = eval(businessparkingstring)  
 if not bool(businessparkingdict):  
 csv\_writer.writerow([data[record]['business\_id'], 'BusinessParking',"null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'BusinessParking', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
 else:  
 for key, val in businessparkingdict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'BusinessParking', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'BusinessParking', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 elif k == "Music":  
 Musicstring = data[record]['attributes']['Music']  
 Musicdict = eval(Musicstring)  
 if not bool(Musicdict):  
 csv\_writer.writerow([data[record]['business\_id'], 'Music',"null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'Music', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
 else:  
 for key, val in Musicdict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'Music', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'Music', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 elif k == "BestNights":  
 BestNightsstring = data[record]['attributes']['BestNights']  
 BestNightsdict = eval(BestNightsstring)  
 if not bool(BestNightsdict):  
 csv\_writer.writerow([data[record]['business\_id'], 'BestNights',"null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'BestNights', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
 else:  
 for key, val in BestNightsdict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'BestNights', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'BestNights', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 elif k == "DietaryRestrictions":  
 DietaryRestrictionsstring = data[record]['attributes']['DietaryRestrictions']  
 DietaryRestrictionsdict = eval(DietaryRestrictionsstring)  
 if not bool(DietaryRestrictionsdict):  
 csv\_writer.writerow([data[record]['business\_id'], 'DietaryRestrictions',"null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'DietaryRestrictions', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
 else:  
 for key, val in DietaryRestrictionsdict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'DietaryRestrictions', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'DietaryRestrictions', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 elif k == "HairSpecializesIn":  
 HairSpecializesInstring = data[record]['attributes']['HairSpecializesIn']  
 HairSpecializesIndict = eval(HairSpecializesInstring)  
 if not bool(HairSpecializesIndict):  
 csv\_writer.writerow([data[record]['business\_id'], 'HairSpecializesIn',"null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'HairSpecializesIn', "null", "null"  
 mycursor.execute(query, values)  
 mydb.commit()  
 else:  
 for key, val in HairSpecializesIndict.items():  
 csv\_writer.writerow([data[record]['business\_id'], 'HairSpecializesIn', key, val])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'HairSpecializesIn', key, val  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 else:  
 csv\_writer.writerow([data[record]['business\_id'], "Others" ,k, v])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'Others', k, v  
 mycursor.execute(query, values)  
 mydb.commit()  
  
 else:  
 csv\_writer.writerow([data[record]['business\_id'], "null"])  
  
 query = u'INSERT INTO business\_attributes(Business\_Id,Attribute\_Category,Attribute,Attribute\_Value) VALUES (%s,%s,%s,%s)'  
 values = data[record]['business\_id'], 'null', 'null','null'  
 mycursor.execute(query, values)  
 mydb.commit()  
  
csv\_file.close()  
mycursor.close()

businesshours.py

import json  
import csv  
import pymysql  
  
listof\_businessid=[]  
listof\_days=[]  
listof\_opentime=[]  
listof\_closetime=[]  
  
  
with open('yelp\_academic\_dataset\_business.json',encoding="utf8") as f:  
 data = [json.loads(line) for line in f]  
  
 totalrecords = len(data)  
  
 #Fetch data hours data and store it in lists  
 for record in range(totalrecords):  
 has\_items = bool(data[record]['hours'])  
 if has\_items == True:  
 for i in data[record]['hours']:  
 day = i  
 open\_time = data[record]['hours'][i].split("-")[0]  
 close\_time = data[record]['hours'][i].split("-")[1]  
 listof\_businessid.append(data[record]['business\_id'])  
 listof\_days.append(day)  
 listof\_opentime.append(open\_time)  
 listof\_closetime.append(close\_time)  
  
 else:  
 day= "null"  
 open\_time= "null"  
 close\_time= "null"  
 listof\_businessid.append(data[record]['business\_id'])  
 listof\_days.append(day)  
 listof\_opentime.append(open\_time)  
 listof\_closetime.append(close\_time)  
  
  
  
  
  
 #open csv  
  
 csv\_file = open('json\_hours.csv', 'w', newline='')  
 csv\_writer = csv.writer(csv\_file)  
  
 #open connection to write to dbms  
  
 mydb = pymysql.connect(host="localhost", user="root", password="psspl12345", database="jsondata")  
 mycursor = mydb.cursor()  
 mycursor.execute(  
 'CREATE TABLE IF NOT EXISTS business\_hours(Business\_Id VARCHAR(50) ,Day VARCHAR(50),open\_time VARCHAR(50),close\_time VARCHAR(50), FOREIGN KEY(Business\_Id) REFERENCES jsondata.business\_master(Business\_Id))')  
  
 print(len(listof\_businessid))  
 for i in range(len(listof\_businessid)):  
 csv\_writer.writerow([listof\_businessid[i],listof\_days[i],listof\_opentime[i],listof\_closetime[i]])  
  
 query = u'INSERT INTO business\_hours(Business\_Id,Day,open\_time,close\_time) VALUES (%s,%s,%s,%s)'  
 values = listof\_businessid[i],listof\_days[i],listof\_opentime[i],listof\_closetime[i]  
 mycursor.execute(query, values)  
 mydb.commit()  
  
  
 csv\_file.close()  
 mycursor.close()