Analysis of Yelp Business Intelligence Data

Part I: Installation and Initial Setup

Begin by installing the necessary libraries that you may need to conduct your analysis. At the very least, you must install pandas and matplotlib

```
%%info
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
sc.install_pypi_package("matplotlib==3.2.1")
sc.install_pypi_package("pandas==1.0.3")
sc.install_pypi_package("scipy==1.7.1")
sc.install_pypi_package("seaborn==0.11.2")
```

```
{"version_major":2,"version_minor":0,"model_id":"b0f9da27c02f4ecf99eda
433c9996d29"}
```

```
Starting Spark application
```

```
<IPython.core.display.HTML object>
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```

```
SparkSession available as 'spark'.
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```

```
Collecting matplotlib==3.2.1
  Using cached
https://files.pythonhosted.org/packages/b2/c2/71fcf957710f3ba1f09088b3
5776a799ba7dd95f7c2b195ec800933b276b/matplotlib-3.2.1-cp37-cp37m-
manylinux1_x86_64.whl
Collecting python-dateutil>=2.1 (from matplotlib==3.2.1)
  Using cached
https://files.pythonhosted.org/packages/36/7a/87837f39d0296e723bb9b62b
bb257d0355c7f6128853c78955f57342a56d/python_dateutil-2.8.2-py2.py3-
none-any.whl
Collecting pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 (from
matplotlib==3.2.1)
  Using cached
https://files.pythonhosted.org/packages/d9/41/d9cfb4410589805cd787f8a8
2cddd13142d9bf7449d12adf2d05a4a7d633/pyparsing-3.0.8-py3-none-any.whl
Collecting cycler>=0.10 (from matplotlib==3.2.1)
  Using cached
https://files.pythonhosted.org/packages/5c/f9/695d6bedebd747e5eb0fe8fa
d57b72fdf25411273a39791cde838d5a8f51/cycler-0.11.0-py3-none-any.whl
Requirement already satisfied: numpy>=1.11 in
/usr/local/lib64/python3.7/site-packages (from matplotlib==3.2.1)
```

```
Collecting kiwisolver>=1.0.1 (from matplotlib==3.2.1)
  Using cached
https://files.pythonhosted.org/packages/51/50/9a9a94afa26c50fc5d912727
2737806990aa698c7a1c220b8e5075e70304/kiwisolver-1.4.2-cp37-cp37m-
manylinux_2_5_x86_64.manylinux1_x86_64.whl
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.7/site-packages (from python-dateutil>=2.1-
>matplotlib==3.2.1)
Collecting typing-extensions; python_version < "3.8" (from
kiwisolver>=1.0.1->matplotlib==3.2.1)
  Using cached
https://files.pythonhosted.org/packages/75/e1/932e06004039dd670c9d5e1d
f0cd606bf46e29a28e65d5bb28e894ea29c9/typing_extensions-4.2.0-py3-none-
any.whl
Installing collected packages: python-dateutil, pyparsing, cycler,
typing-extensions, kiwisolver, matplotlib
Successfully installed cycler-0.11.0 kiwisolver-1.4.2 matplotlib-3.2.1
pyparsing-3.0.8 python-dateutil-2.8.2 typing-extensions-4.2.0

Collecting pandas==1.0.3
  Using cached
https://files.pythonhosted.org/packages/4a/6a/94b219b8ea0f2d580169e85e
d1edc0163743f55aaeca8a44c2e8fc1e344e/pandas-1.0.3-cp37-cp37m-
manylinux1_x86_64.whl
Requirement already satisfied: pytz>=2017.2 in
/usr/local/lib/python3.7/site-packages (from pandas==1.0.3)
Requirement already satisfied: numpy>=1.13.3 in
/usr/local/lib64/python3.7/site-packages (from pandas==1.0.3)
Requirement already satisfied: python-dateutil>=2.6.1 in
/mnt/tmp/1651336224280-0/lib/python3.7/site-packages (from
pandas==1.0.3)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.7/site-packages (from python-dateutil>=2.6.1-
>pandas==1.0.3)
Installing collected packages: pandas
Successfully installed pandas-1.0.3

Collecting scipy==1.7.1
  Using cached
https://files.pythonhosted.org/packages/b5/6b/8bc0b61ebf824f8c3979a313
68bbe38dd247590049a994ab0ed077cb56dc/scipy-1.7.1-cp37-cp37m-
manylinux_2_5_x86_64.manylinux1_x86_64.whl
Requirement already satisfied: numpy<1.23.0,>=1.16.5 in
/usr/local/lib64/python3.7/site-packages (from scipy==1.7.1)
Installing collected packages: scipy
Successfully installed scipy-1.7.1

Collecting seaborn==0.11.2
  Using cached
https://files.pythonhosted.org/packages/10/5b/0479d7d845b5ba410ca702ff
```

```
cd7f2cd95a14a4dfff1fde2637802b258b9b/seaborn-0.11.2-py3-none-any.whl
Requirement already satisfied: numpy>=1.15 in
/usr/local/lib64/python3.7/site-packages (from seaborn==0.11.2)
Requirement already satisfied: scipy>=1.0 in /mnt/tmp/1651336224280-
0/lib/python3.7/site-packages (from seaborn==0.11.2)
Requirement already satisfied: matplotlib>=2.2 in
/mnt/tmp/1651336224280-0/lib/python3.7/site-packages (from
seaborn==0.11.2)
Requirement already satisfied: pandas>=0.23 in /mnt/tmp/1651336224280-
0/lib/python3.7/site-packages (from seaborn==0.11.2)
Requirement already satisfied: python-dateutil>=2.1 in
/mnt/tmp/1651336224280-0/lib/python3.7/site-packages (from
matplotlib>=2.2->seaborn==0.11.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!
=2.1.6,>=2.0.1 in /mnt/tmp/1651336224280-0/lib/python3.7/site-packages
(from matplotlib>=2.2->seaborn==0.11.2)
Requirement already satisfied: cycler>=0.10 in /mnt/tmp/1651336224280-
0/lib/python3.7/site-packages (from matplotlib>=2.2->seaborn==0.11.2)
Requirement already satisfied: kiwisolver>=1.0.1 in
/mnt/tmp/1651336224280-0/lib/python3.7/site-packages (from
matplotlib>=2.2->seaborn==0.11.2)
Requirement already satisfied: pytz>=2017.2 in
/usr/local/lib/python3.7/site-packages (from pandas>=0.23-
>seaborn==0.11.2)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.7/site-packages (from python-dateutil>=2.1-
>matplotlib>=2.2->seaborn==0.11.2)
Requirement already satisfied: typing-extensions; python_version <
"3.8" in /mnt/tmp/1651336224280-0/lib/python3.7/site-packages (from
kiwisolver>=1.0.1->matplotlib>=2.2->seaborn==0.11.2)
Installing collected packages: seaborn
Successfully installed seaborn-0.11.2
```

Importing

Now, import the installed packages from the previous block below.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy
from pyspark.sql.functions import col
```


```
{"version_major":2,"version_minor":0,"model_id":"902a9879b1ea4f4a9d8f6
feaf8edb224"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```

Part II: Analyzing Categories

Loading Data

We are finally ready to load data. Using spark load the data from S3 into a dataframe object that we can manipulate further down in our analysis.

```
#Loading the dataset
df_business =
spark.read.json('s3://cis9760-yelpdata/yelp_academic_dataset_business.
json')
```


```
{"version_major":2,"version_minor":0,"model_id":"b30b638d0d494b09928b2
e29ecc0dca5"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
df_business.printSchema()
df_business.show(10)
df_business.count()
```


```
{"version_major":2,"version_minor":0,"model_id":"6324edb73dec4da9a2356
d2cca21b612"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
root
 |-- address: string (nullable = true)
 |-- attributes: struct (nullable = true)
 |    |-- AcceptsInsurance: string (nullable = true)
 |    |-- AgesAllowed: string (nullable = true)
 |    |-- Alcohol: string (nullable = true)
 |    |-- Ambience: string (nullable = true)
 |    |-- BYOB: string (nullable = true)
 |    |-- BYOBCorkage: string (nullable = true)
 |    |-- BestNights: string (nullable = true)
 |    |-- BikeParking: string (nullable = true)
 |    |-- BusinessAcceptsBitcoin: string (nullable = true)
 |    |-- BusinessAcceptsCreditCards: string (nullable = true)
 |    |-- BusinessParking: string (nullable = true)
 |    |-- ByAppointmentOnly: string (nullable = true)
 |    |-- Caters: string (nullable = true)
 |    |-- CoatCheck: string (nullable = true)
 |    |-- Corkage: string (nullable = true)
 |    |-- DietaryRestrictions: string (nullable = true)
 |    |-- DogsAllowed: string (nullable = true)
 |    |-- DriveThru: string (nullable = true)
 |    |-- GoodForDancing: string (nullable = true)
 |    |-- GoodForKids: string (nullable = true)
 |    |-- GoodForMeal: string (nullable = true)
 |    |-- HairSpecializesIn: string (nullable = true)
 |    |-- HappyHour: string (nullable = true)
```

```
 |    |-- HasTV: string (nullable = true)
 |    |-- Music: string (nullable = true)
 |    |-- NoiseLevel: string (nullable = true)
 |    |-- Open24Hours: string (nullable = true)
 |    |-- OutdoorSeating: string (nullable = true)
 |    |-- RestaurantsAttire: string (nullable = true)
 |    |-- RestaurantsCounterService: string (nullable = true)
 |    |-- RestaurantsDelivery: string (nullable = true)
 |    |-- RestaurantsGoodForGroups: string (nullable = true)
 |    |-- RestaurantsPriceRange2: string (nullable = true)
 |    |-- RestaurantsReservations: string (nullable = true)
 |    |-- RestaurantsTableService: string (nullable = true)
 |    |-- RestaurantsTakeOut: string (nullable = true)
 |    |-- Smoking: string (nullable = true)
 |    |-- WheelchairAccessible: string (nullable = true)
 |    |-- WiFi: string (nullable = true)
 |-- business_id: string (nullable = true)
 |-- categories: string (nullable = true)
 |-- city: string (nullable = true)
 |-- hours: struct (nullable = true)
 |    |-- Friday: string (nullable = true)
 |    |-- Monday: string (nullable = true)
 |    |-- Saturday: string (nullable = true)
 |    |-- Sunday: string (nullable = true)
 |    |-- Thursday: string (nullable = true)
 |    |-- Tuesday: string (nullable = true)
 |    |-- Wednesday: string (nullable = true)
 |-- is_open: long (nullable = true)
 |-- latitude: double (nullable = true)
 |-- longitude: double (nullable = true)
 |-- name: string (nullable = true)
 |-- postal_code: string (nullable = true)
 |-- review_count: long (nullable = true)
 |-- stars: double (nullable = true)
 |-- state: string (nullable = true)

+--------------------+--------------------+--------------------
+--------------------+-------------+--------------------+-------
+----------+------------+--------------------+-----------+------------
+-----+-----+
|             address|          attributes|         business_id|
categories|          city|               hours|is_open|  latitude|
longitude|                name|postal_code|review_count|stars|state|
+--------------------+--------------------+--------------------
+--------------------+-------------+--------------------+-------
+----------+------------+--------------------+-----------+------------
+-----+-----+
|1616 Chapala St, ...|[,,,,,,,,,,, True...|Pns2l4eNsfO8kk83d...|
Doctors, Traditio...| Santa Barbara|                null|      0|
34.4266787|-119.7111968|Abby Rappoport, L...|     93101|           7|
```

```
5.0|   CA|
|87 Grasso Plaza S...|[,,,,,,,,, True,,...|mpf3x-BjTdTEA3yCZ...|
Shipping Centers,...|        Affton|[8:0-18:30, 0:0-0...|      1|
38.551126|  -90.335695|       The UPS Store|      63123|          15|
3.0|   MO|
|5255 E Broadway Blvd|[,,,,,,, True,, T...|tUFrWirKiKi_TAnsV...|
Department Stores...|        Tucson|[8:0-23:0, 8:0-22...|      0|
32.223236| -110.880452|              Target|      85711|          22|
3.5|   AZ|
|         935 Race St|[,, u'none',,,,, ...|MTSW4McQd7CbVtyjq...|
Restaurants, Food...|  Philadelphia|[7:0-21:0, 7:0-20...|      1|
39.9555052| -75.1555641|  St Honore Pastries|      19107|          80|
4.0|   PA|
|       101 Walnut St|[,,,,,,, True,, T...|mWMc6_wTdE0EUBKIG...|
Brewpubs, Breweri...|    Green Lane|[12:0-22:0,, 12:0...|      1|
40.3381827| -75.4716585|Perkiomen Valley ...|      18054|          13|
4.5|   PA|
|       615 S Main St|[,, u'none', None...|CF33F8-E6oudUQ46H...|
Burgers, Fast Foo...|  Ashland City|[9:0-0:0, 0:0-0:0...|      1|
36.269593|  -87.058943|      Sonic Drive-In|      37015|           6|
2.0|   TN|
|8522 Eager Road, ...|[,,,,,,, True,, T...|n_0UpQx1hsNbnPUSl...|
Sporting Goods, F...|     Brentwood|[10:0-18:0, 0:0-0...|      1|
38.627695|  -90.340465|     Famous Footwear|      63144|          13|
2.5|   MO|
|  400 Pasadena Ave S|               null|qkRM_2X51Yqxk3btl...|
Synagogues, Relig...|St. Petersburg|[9:0-17:0, 9:0-17...|      1|
27.76659|  -82.732983|      Temple Beth-El|      33707|           5|
3.5|   FL|
|   8025 Mackenzie Rd|[,, u'full_bar', ...|k0hlBqXX-Bt0vf1op...|Pubs,
Restaurants...|        Affton|                null|      0|38.5651648|
-90.3210868|Tsevi's Pub And G...|      63123|          19|  3.0|   MO|
| 2312 Dickerson Pike|[,, u'none',,,,,,,...|bBDDEgkFA1Otx9Lfe...|Ice
Cream & Froze...|     Nashville|[6:0-16:0, 0:0-0:...|      1|
36.2081024| -86.7681696|      Sonic Drive-In|      37207|          10|
1.5|   TN|
+--------------------+--------------------+--------------------
+--------------------+-------------+--------------------+-------
+----------+------------+--------------------+-----------+------------
+-----+-----+
only showing top 10 rows

150346
```

Overview of Data

Display the number of rows and columns in our dataset.

```
print("Number of columns in Business table: ",
str(len(df_business.columns)))
print("Number of rows in Business table: ",str(df_business.count()))
```

{"version_major":2,"version_minor":0,"model_id":"13342eb5e01f48f5bc9fc639b5f78ca9"}

{"version_major":2,"version_minor":0,"model_id":""}

```
Number of columns in Business table:  14
Number of rows in Business table:  150346
```

Display the DataFrame schema below.

```
df_business.printSchema()
```

{"version_major":2,"version_minor":0,"model_id":"3e7d352bc6964e7fad80f4d77ff55c00"}

{"version_major":2,"version_minor":0,"model_id":""}

```
root
 |-- address: string (nullable = true)
 |-- attributes: struct (nullable = true)
 |    |-- AcceptsInsurance: string (nullable = true)
 |    |-- AgesAllowed: string (nullable = true)
 |    |-- Alcohol: string (nullable = true)
 |    |-- Ambience: string (nullable = true)
 |    |-- BYOB: string (nullable = true)
 |    |-- BYOBCorkage: string (nullable = true)
 |    |-- BestNights: string (nullable = true)
 |    |-- BikeParking: string (nullable = true)
 |    |-- BusinessAcceptsBitcoin: string (nullable = true)
 |    |-- BusinessAcceptsCreditCards: string (nullable = true)
 |    |-- BusinessParking: string (nullable = true)
 |    |-- ByAppointmentOnly: string (nullable = true)
 |    |-- Caters: string (nullable = true)
 |    |-- CoatCheck: string (nullable = true)
 |    |-- Corkage: string (nullable = true)
 |    |-- DietaryRestrictions: string (nullable = true)
 |    |-- DogsAllowed: string (nullable = true)
 |    |-- DriveThru: string (nullable = true)
 |    |-- GoodForDancing: string (nullable = true)
 |    |-- GoodForKids: string (nullable = true)
 |    |-- GoodForMeal: string (nullable = true)
 |    |-- HairSpecializesIn: string (nullable = true)
 |    |-- HappyHour: string (nullable = true)
 |    |-- HasTV: string (nullable = true)
 |    |-- Music: string (nullable = true)
 |    |-- NoiseLevel: string (nullable = true)
 |    |-- Open24Hours: string (nullable = true)
 |    |-- OutdoorSeating: string (nullable = true)
 |    |-- RestaurantsAttire: string (nullable = true)
 |    |-- RestaurantsCounterService: string (nullable = true)
 |    |-- RestaurantsDelivery: string (nullable = true)
 |    |-- RestaurantsGoodForGroups: string (nullable = true)
```

```
 |    |-- RestaurantsPriceRange2: string (nullable = true)
 |    |-- RestaurantsReservations: string (nullable = true)
 |    |-- RestaurantsTableService: string (nullable = true)
 |    |-- RestaurantsTakeOut: string (nullable = true)
 |    |-- Smoking: string (nullable = true)
 |    |-- WheelchairAccessible: string (nullable = true)
 |    |-- WiFi: string (nullable = true)
 |-- business_id: string (nullable = true)
 |-- categories: string (nullable = true)
 |-- city: string (nullable = true)
 |-- hours: struct (nullable = true)
 |    |-- Friday: string (nullable = true)
 |    |-- Monday: string (nullable = true)
 |    |-- Saturday: string (nullable = true)
 |    |-- Sunday: string (nullable = true)
 |    |-- Thursday: string (nullable = true)
 |    |-- Tuesday: string (nullable = true)
 |    |-- Wednesday: string (nullable = true)
 |-- is_open: long (nullable = true)
 |-- latitude: double (nullable = true)
 |-- longitude: double (nullable = true)
 |-- name: string (nullable = true)
 |-- postal_code: string (nullable = true)
 |-- review_count: long (nullable = true)
 |-- stars: double (nullable = true)
 |-- state: string (nullable = true)
```

Display the first 5 rows with the following columns:

- business_id
- name
- city
- state
- categories

```
#print(df_business.show(5))
df_business.select("business_id","name","city","state","categories").s
how(5)
```


```
{"version_major":2,"version_minor":0,"model_id":"8fb5d0e816bb40cab1798
7ed702adef1"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+--------------------+------------+-----
+--------------------+
|         business_id|                name|        city|state|
categories|
```

```
+--------------------+--------------------+------------+-----
+--------------------+
|Pns2l4eNsfO8kk83d...|Abby Rappoport, L...|Santa Barbara|   CA|
Doctors, Traditio...|
|mpf3x-BjTdTEA3yCZ...|       The UPS Store|       Affton|   MO|
Shipping Centers,...|
|tUFrWirKiKi_TAnsV...|              Target|       Tucson|   AZ|
Department Stores...|
|MTSW4McQd7CbVtyjq...|  St Honore Pastries| Philadelphia|   PA|
Restaurants, Food...|
|mWMc6_wTdE0EUBKIG...|Perkiomen Valley ...|   Green Lane|   PA|
Brewpubs, Breweri...|
+--------------------+--------------------+------------+-----
+--------------------+
only showing top 5 rows
```

Analyzing Categories

```
df_business.select("business_id", "categories").show(5)
```


```
{"version_major":2,"version_minor":0,"model_id":"df1c291d12524bc5bf097
143f43d9258"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+--------------------+
|         business_id|          categories|
+--------------------+--------------------+
|Pns2l4eNsfO8kk83d...|Doctors, Traditio...|
|mpf3x-BjTdTEA3yCZ...|Shipping Centers,...|
|tUFrWirKiKi_TAnsV...|Department Stores...|
|MTSW4McQd7CbVtyjq...|Restaurants, Food...|
|mWMc6_wTdE0EUBKIG...|Brewpubs, Breweri...|
+--------------------+--------------------+
only showing top 5 rows
```

Display the first 5 rows of your association table below.

```
from pyspark.sql.functions import explode, split
df_business_cat =
df_business.withColumn('category',explode(split('categories',", ")))
df_business_cat.select("business_id", "category").show(5)
```


```
{"version_major":2,"version_minor":0,"model_id":"2f00b93e6d774b2fbe1ba
9bd1cb3528f"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+--------------------+
|         business_id|            category|
+--------------------+--------------------+
|Pns2l4eNsfO8kk83d...|             Doctors|
|Pns2l4eNsfO8kk83d...|Traditional Chine...|
```

```
|Pns2l4eNsfO8kk83d...|Naturopathic/Holi...|
|Pns2l4eNsfO8kk83d...|         Acupuncture|
|Pns2l4eNsfO8kk83d...|    Health & Medical|
+--------------------+--------------------+
only showing top 5 rows
```

Total Unique Categories

what is the total number of unique categories available?

Below, implement the code necessary to calculate this figure.

```
from pyspark.sql.functions import countDistinct
df_business_cat.select(countDistinct("category")).show()
```


```
{"version_major":2,"version_minor":0,"model_id":"a14ed0f903d1479883b02
d53a744b940"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+------------------------+
|count(DISTINCT category)|
+------------------------+
|                    1311|
+------------------------+
```

Top Categories By Business

Counts of Businesses / Category

```
df_business_cat \
    .groupBy(df_business_cat.category) \
    .count().show(20)
```


```
{"version_major":2,"version_minor":0,"model_id":"4e729646b66141a98ed6e
5be10314983"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+-----+
|            category|count|
+--------------------+-----+
|      Dermatologists|  336|
|      Paddleboarding|   98|
|        Aerial Tours|   12|
|Faith-based Crisi...|    1|
|         Hobby Shops|  552|
|          Bubble Tea|  477|
|            Handyman|  356|
|             Tanning|  667|
|      Aerial Fitness|   19|
|             Falafel|  103|
|        Summer Camps|  232|
```

```
|       Outlet Stores|  182|
|     Clothing Rental|   37|
|      Sporting Goods| 1662|
|     Cooking Schools|   76|
|  Lactation Services|   27|
|Ski & Snowboard S...|   40|
|             Museums|  413|
|              Doulas|   31|
|     Baseball Fields|   14|
+--------------------+-----+
only showing top 20 rows
```

Bar Chart of Top Categories

```
df_topcat = df_business_cat \
            .groupBy(df_business_cat.category) \
            .count() \
            .sort(col("count").desc()).limit(20)
df_topcat.show()
```


```
{"version_major":2,"version_minor":0,"model_id":"fd8bd13211da4b90b217a
e8111c416d7"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+-----+
|            category|count|
+--------------------+-----+
|         Restaurants|52268|
|                Food|27781|
|            Shopping|24395|
|       Home Services|14356|
|       Beauty & Spas|14292|
|           Nightlife|12281|
|    Health & Medical|11890|
|      Local Services|11198|
|                Bars|11065|
|          Automotive|10773|
|Event Planning & ...| 9895|
|          Sandwiches| 8366|
|American (Traditi...| 8139|
|         Active Life| 7687|
|               Pizza| 7093|
|        Coffee & Tea| 6703|
|           Fast Food| 6472|
|  Breakfast & Brunch| 6239|
|      American (New)| 6097|
|     Hotels & Travel| 5857|
+--------------------+-----+
```

```
df_topcat = df_topcat.toPandas()
df_topcat
```

{"version_major":2,"version_minor":0,"model_id":"756df7594b1c4e9b97ae7cf37ce8a9ec"}

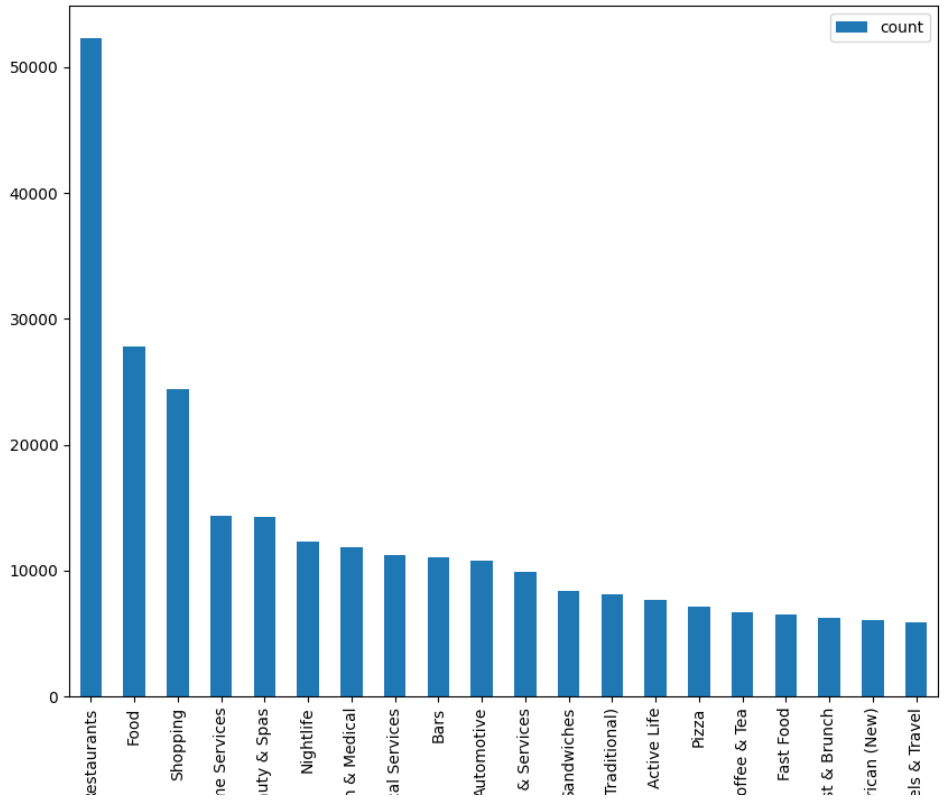
{"version_major":2,"version_minor":0,"model_id":""}

	category	count
0	Restaurants	52268
1	Food	27781
2	Shopping	24395
3	Home Services	14356
4	Beauty & Spas	14292
5	Nightlife	12281
6	Health & Medical	11890
7	Local Services	11198
8	Bars	11065
9	Automotive	10773
10	Event Planning & Services	9895
11	Sandwiches	8366
12	American (Traditional)	8139
13	Active Life	7687
14	Pizza	7093
15	Coffee & Tea	6703
16	Fast Food	6472
17	Breakfast & Brunch	6239
18	American (New)	6097
19	Hotels & Travel	5857

```
plt.figure(figsize =(10,8))
df_topcat.plot(kind='bar', x='category', figsize=(10,8), rot=90)
%matplot plt
```

{"version_major":2,"version_minor":0,"model_id":"19fe217b7e4f45c8b73a517c13ef89c0"}

{"version_major":2,"version_minor":0,"model_id":""}



Part III: Do Yelp Reviews Skew Negative?

Loading User Data

```
df_review =
spark.read.json('s3://cis9760-yelpdata/yelp_academic_dataset_review.json')
```

```
{"version_major":2,"version_minor":0,"model_id":"2edf47545be14206b68c9fb1900679c8"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```

```
df_review.printSchema()
df_review.show(10)
df_review.count()
```

```
{"version_major":2,"version_minor":0,"model_id":"bdeedc7c845846e59209540630027f71"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```

```
root
 |-- business_id: string (nullable = true)
 |-- cool: long (nullable = true)
```

```
 |-- date: string (nullable = true)
 |-- funny: long (nullable = true)
 |-- review_id: string (nullable = true)
 |-- stars: double (nullable = true)
 |-- text: string (nullable = true)
 |-- useful: long (nullable = true)
 |-- user_id: string (nullable = true)

+--------------------+----+--------------------+-----
+--------------------+-----+--------------------+------
+--------------------+
|         business_id|cool|                date|funny|
review_id|stars|                text|useful|             user_id|
+--------------------+----+--------------------+-----
+--------------------+-----+--------------------+------
+--------------------+
|XQfwVwDr-v0ZS3_Cb...|   0|2018-07-07 22:09:11|    0|KU_O5udG6zpxOg-
Vc...|  3.0|If you decide to ...|     0|mh_-eMZ6K5RLWhZyI...|
|7ATYjTIgM3jUlt4UM...|   1|2012-01-03 15:28:18|    0|
BiTunyQ73aT9WBnpR...|  5.0|I've taken a lot ...|     1|
OyoGAe7OKpv6SyGZT...|
|YjUWPpI6HXG530lwP...|   0|2014-02-05 20:30:30|    0|
saUsX_uimxRlCVr67...|  3.0|Family diner. Had...|     0|
8g_iMtfSiwikVnbP2...|
|kxX2SOes4o-D3ZQBk...|   1|2015-01-04 00:01:03|    0|
AqPFMleE6RsU23_au...|  5.0|Wow!  Yummy, diff...|     1|
_7bHUi9Uuf5__HHc_...|
|e4Vwtrqf-wpJfwesg...|   1|2017-01-14 20:54:15|    0|Sx8TMOWLNuJBWer-
0...|  4.0|Cute interior and...|     1|bcjbaE6dDog4jkNY9...|
|04UD14gamNjLY0IDY...|   1|2015-09-23 23:10:31|    2|JrIxlS1TzJ-
iCu79u...|  1.0|I am a long term ...|     1|eUta8W_HdHMXPzLBB...|
|gmjsEdUsKpj9Xxu6p...|   0|2015-01-03 23:21:18|    2|
6AxgBCNX_PNTOxmbR...|  5.0|Loved this tour! ...|     0|
r3zeYsv1XFBRA4dJp...|
|LHSTtnW3YHCeUkRDG...|   0|2015-08-07 02:29:16|    0|
_ZeMknuYdlQcUqng_...|  5.0|Amazingly amazing...|     2|
yfFzsLmaWF2d4Sr0U...|
|B5XSoSG3SfvQGtKEG...|   0|2016-03-30 22:46:33|    1|
ZKvDG2sBvHVdF5oBN...|  3.0|This easter inste...|     1|wSTuiTk-
sKNdcFypr...|
|gebiRewfieSdtt17P...|   0|2016-07-25 07:31:06|    0|
pUycOfUwM8vqX7KjR...|  3.0|Had a party of 6 ...|     0|
59MxRhNVhU9MYndMk...|
+--------------------+----+--------------------+-----
+--------------------+-----+--------------------+------
+--------------------+
only showing top 10 rows

6990280
```

Let's begin by listing the business_id and stars columns together for the user reviews data.

```
df_review.select('business_id', 'stars').show(5)
```

{"version_major":2,"version_minor":0,"model_id":"8ccb697f44a24c69a922e7949e6e68ef"}

{"version_major":2,"version_minor":0,"model_id":""}

```
+--------------------+-----+
|         business_id|stars|
+--------------------+-----+
|XQfwVwDr-v0ZS3_Cb...|  3.0|
|7ATYjTIgM3jUlt4UM...|  5.0|
|YjUWPpI6HXG530lwP...|  3.0|
|kxX2SOes4o-D3ZQBk...|  5.0|
|e4Vwtrqf-wpJfwesg...|  4.0|
+--------------------+-----+
only showing top 5 rows
```

Now, let's aggregate along the stars column to get a resultant dataframe that displays average stars per business as accumulated by users who took the time to submit a written review.

```
#from pyspark.sql.functions import col
#nll = '\\N'
#df_txt_review = df_review.filter(col("startYear") != nll)
    #.filter(df_review.titleType == "movie") \
df_review_avg = df_review \
    .groupBy('business_id') \
    .avg('stars')
df_review_avg.show(5)
```

{"version_major":2,"version_minor":0,"model_id":"4148b4881b154ea49a7f337a29fe183c"}

{"version_major":2,"version_minor":0,"model_id":""}

```
+--------------------+------------------+
|         business_id|        avg(stars)|
+--------------------+------------------+
|HSzSGdcNaU7heQe0N...|3.3333333333333335|
|skW4boArIApRw9DXK...|2.3947368421052633|
|zJErbOQMKX-MwHs_u...|2.9279279279279278|
|I0053JmJ5DEFUWSJ8...|2.3956043956043955|
|wS-SWAa_yaJAw6fJm...| 3.357142857142857|
+--------------------+------------------+
only showing top 5 rows
```

Now the fun part - let's join our two dataframes (reviews and business data) by business_id.

```
df_bus_review = df_business.join(df_review_avg, ['business_id'])
```

{"version_major":2,"version_minor":0,"model_id":"c3a8bd9789d14293ab16db232fc120cf"}

{"version_major":2,"version_minor":0,"model_id":""}

```
df_bus_review.select('name', 'city', 'state', 'avg(stars)',
'stars').show(5)
```

{"version_major":2,"version_minor":0,"model_id":"ca0e5d8c18ff473a98dd2
949e74eebad"}

{"version_major":2,"version_minor":0,"model_id":""}

```
+--------------------+------------+-----+------------------+-----+
|                name|        city|state|        avg(stars)|stars|
+--------------------+------------+-----+------------------+-----+
|Champps Penn's La...|Philadelphia|   PA|2.3947368421052633|  2.5|
|Golden Corral Buf...|      Tucson|   AZ|2.3956043956043955|  2.5|
|  NJ Weedman's Joint|     Trenton|   NJ| 4.232558139534884|  4.0|
|Philadelphia Marr...|Philadelphia|   PA|2.9279279279279278|  3.0|
|       A Able Movers|      Tucson|   AZ|             1.875|  2.0|
+--------------------+------------+-----+------------------+-----+
only showing top 5 rows
```

```
df_busreview_skew = df_bus_review.withColumn('skew',
((df_bus_review['avg(stars)'] - df_bus_review['stars']) /
df_bus_review['stars']))
df_busreview_skew.show(5)
```

{"version_major":2,"version_minor":0,"model_id":"7709385851094f9f8ad1d
e6f0e83f38a"}

{"version_major":2,"version_minor":0,"model_id":""}

```
+--------------------+--------------------+--------------------
+--------------------+------------+--------------------+-------
+-------------+---------------+--------------------+-----------
+------------+-----+-----+-----------------+--------------------+
|         business_id|             address|          attributes|
categories|        city|               hours|is_open|     latitude|
longitude|                name|postal_code|review_count|stars|state|
avg(stars)|                skew|
+--------------------+--------------------+--------------------
+--------------------+------------+--------------------+-------
+-------------+---------------+--------------------+-----------
+------------+-----+-----+-----------------+--------------------+
|--gJkxbsiSIwsQKbi...| 4545 W Kennedy Blvd|[,,,,,,, True, Fa...|Skin
Care, Hair S...|       Tampa|[8:0-20:0, 8:0-20...|      1|
27.9451223|    -82.5210814|Salon Lofts - Wes...|      33609|
6|  5.0|   FL|4.833333333333333|-0.03333333333333...|
|-02xFuruu85XmDn2x...|7475 E Tanque Ver...|[True,,,,,,, True...|
Shopping, Ophthal...|      Tucson|[8:30-17:30, 0:0-...|      1|
32.2510387|   -110.8331729|Family Vision Center|      85715|
109|  4.5|   AZ| 4.68595041322314|0.041322314049586716|
|-0EdehHjIQc0DtYU8...|7040 Land O Lakes...|[,,, {'touristy':...|
```

```
Restaurants, Chinese|Land O Lakes|                null|      1|
28.2601545|    -82.4748164|New Tung Tung Two...|      34638|
35|  3.0|   FL|3.138888888888889| 0.04629629629629628|
|-0dKgi_Hpcis921nO...|    4983 Glenwood St|[,, u'none', {'ro...|
Restaurants, Barb...| Garden City|[11:30-20:0, 0:0-...|      0|
43.6502742721|-116.2818321878|  Cutter's Grand BBQ|      83714|
28|  4.5|   ID|4.678571428571429|0.039682539682539736|
|-0iIxySkp97WNlwK6...|538 S Virginia St...|[,, 'none', {'tou...|
Caterers, Sandwic...|        Reno|[6:45-15:30, 0:0-...|      1|
39.5202401|    -119.810022|Truckee Bagel Com...|      89501|
219|  3.5|   NV|3.721030042918455| 0.06315144083384425|
+--------------------+--------------------+--------------------
+--------------------+-----------+--------------------+-------
+-------------+--------------+--------------------+----------
+-----------+-----+-----+-----------------+--------------------+
only showing top 5 rows

df_busreview_skew.select('name', 'city', 'state', 'avg(stars)',
'stars', 'skew').show(5)

{"version_major":2,"version_minor":0,"model_id":"a744a8a5b6fd4977a318c
d8c3e9d3149"}

{"version_major":2,"version_minor":0,"model_id":""}

+--------------------+-----------+-----+------------------+-----
+--------------------+
|                name|       city|state|       avg(stars)|stars|
skew|
+--------------------+-----------+-----+------------------+-----
+--------------------+
|Gillane's Bar & G...|    Ardmore|   PA|3.3333333333333335|  3.0|
0.11111111111111116|
|Champps Penn's La...|Philadelphia|   PA|2.3947368421052633|  2.5|-
0.04210526315789469|
|Philadelphia Marr...|Philadelphia|   PA|2.9279279279279278|  3.0|-
0.02402402402402...|
|Golden Corral Buf...|     Tucson|   AZ|2.3956043956043955|  2.5|-
0.04175824175824179|
|  Swiss Watch Center|      Tampa|   FL| 3.357142857142857|  3.5|-
0.04081632653061223|
+--------------------+-----------+-----+------------------+-----
+--------------------+
only showing top 5 rows
```

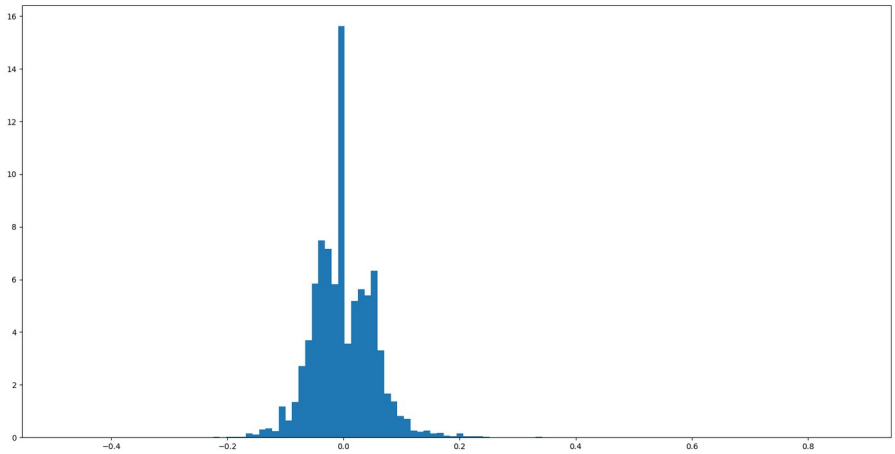
And finally, graph it!

```
df_busreview_skew = df_busreview_skew.toPandas()

{"version_major":2,"version_minor":0,"model_id":"ed9bb4a2d5404436be5e9
23ba1a09f5b"}

{"version_major":2,"version_minor":0,"model_id":""}
```

```
plt.figure(figsize=(20,10))
plt.hist(df_busreview_skew['skew'],bins=120,density=True)
plt.show()
%matplot plt
```

{"version_major":2,"version_minor":0,"model_id":"eed42646a1ca4e528d95eec6382a8c0f"}

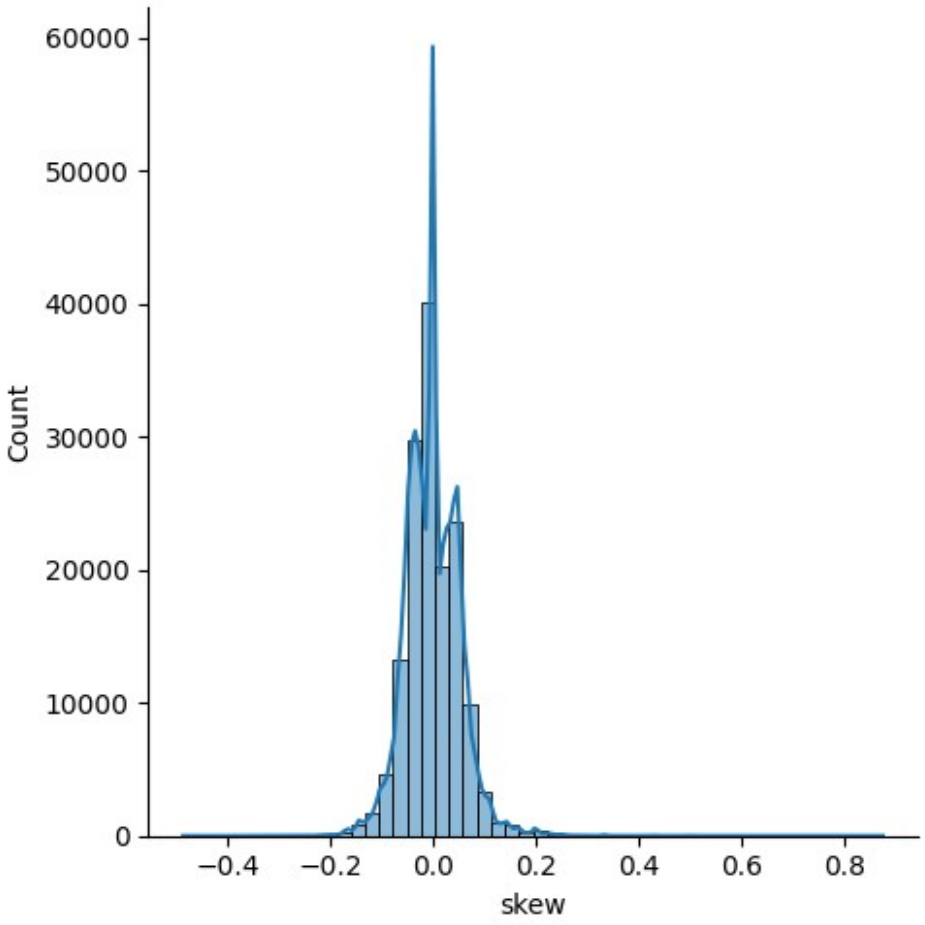
{"version_major":2,"version_minor":0,"model_id":""}




```
X = df_busreview_skew['skew']
sns.displot(X, kde=True, bins=50)
%matplot plt
```

{"version_major":2,"version_minor":0,"model_id":"44f71c4684214d1e920080b8a5ae2d14"}

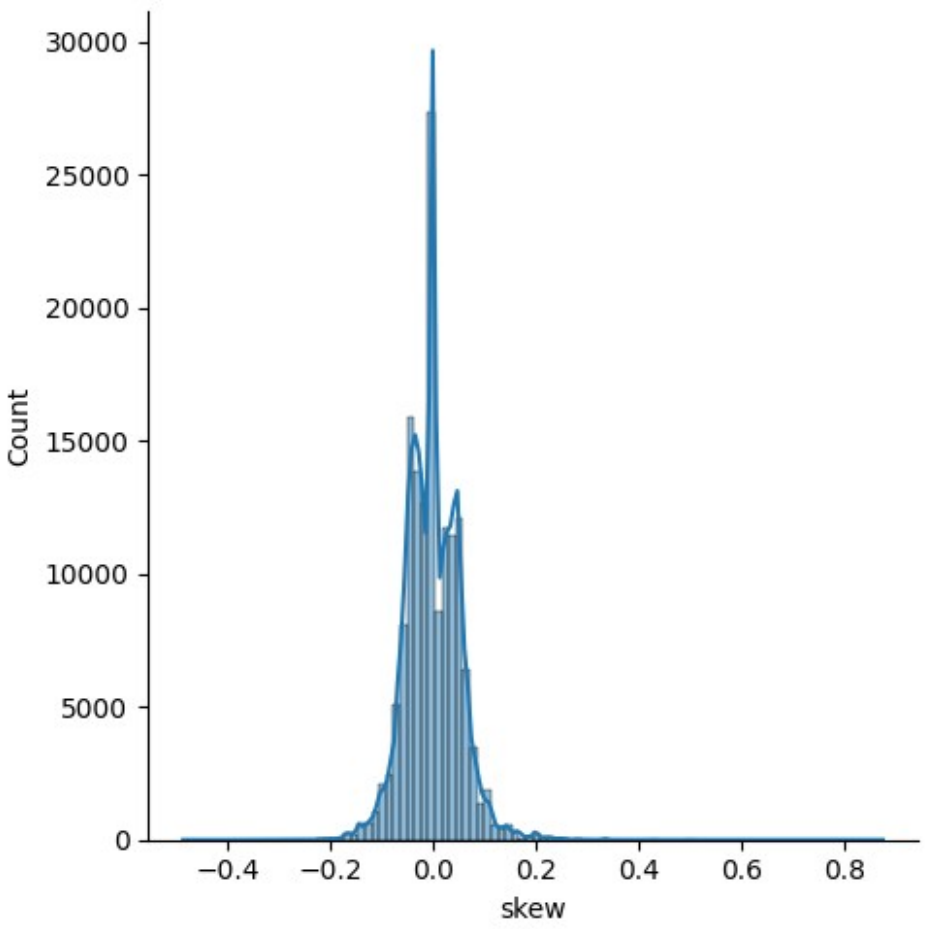
{"version_major":2,"version_minor":0,"model_id":""}




```
X = df_busreview_skew['skew']
sns.displot(X, kde=True, bins=100)
%matplot plt
```


```
{"version_major":2,"version_minor":0,"model_id":"2a13145f3c7c484f839c4
6f6e04c84ca"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```




```
#sns.distplot(X)
#%matplot plt

df_busreview_skew.describe()['skew']
```


```
{"version_major":2,"version_minor":0,"model_id":"f30cfe2e6ed14cc09d6bd
b0238dbc19e"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
count    150346.000000
mean         -0.000075
std           0.053311
min          -0.485294
25%          -0.035714
50%           0.000000
75%           0.034483
max           0.875000
Name: skew, dtype: float64
```

IMPLICATIONS

The density plot obtained of the skew column seems to be evenly distributed to the naked eye. To check if there is actually any skweness in the data, I changed the bin values to 70, 90, 100, 150 etc. From the graph we cannot conclude anything which way it is skewed.

As we know that - If a density curve is left skewed, then the mean is less than the median. If a density curve is right skewed, then the mean is greater than the median. If a density curve has no skew, then the mean is equal to the median.

The descriptive stats of skew column are as below - mean is -0.000075 median is 0.000000 As per these values we can say that the density curve is left skewed by a very small margin.

We can conlcude that there is not much significant difference in the level of satisfaction between users who wrote a written review vs the ones who just gave the star rating.

Should the Elite be Trusted?

```
df_user =
spark.read.json('s3://cis9760-yelpdata/yelp_academic_dataset_user.json
')
df_user.printSchema()
df_user.show(10)
df_user.count()
```


```
{"version_major":2,"version_minor":0,"model_id":"13543ab266ae4165b06fd
2854ce1d47a"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
root
 |-- average_stars: double (nullable = true)
 |-- compliment_cool: long (nullable = true)
 |-- compliment_cute: long (nullable = true)
 |-- compliment_funny: long (nullable = true)
 |-- compliment_hot: long (nullable = true)
 |-- compliment_list: long (nullable = true)
 |-- compliment_more: long (nullable = true)
 |-- compliment_note: long (nullable = true)
 |-- compliment_photos: long (nullable = true)
 |-- compliment_plain: long (nullable = true)
 |-- compliment_profile: long (nullable = true)
 |-- compliment_writer: long (nullable = true)
 |-- cool: long (nullable = true)
 |-- elite: string (nullable = true)
 |-- fans: long (nullable = true)
 |-- friends: string (nullable = true)
 |-- funny: long (nullable = true)
 |-- name: string (nullable = true)
 |-- review_count: long (nullable = true)
 |-- useful: long (nullable = true)
 |-- user_id: string (nullable = true)
```

```
 |-- yelping_since: string (nullable = true)

+------------+--------------+--------------+---------------
+--------------+--------------+--------------+--------------
+----------------+----------------+-----------------
+----------------+-----+--------------------+----
+--------------------+-----+--------+------------+------
+--------------------+-------------------+
|average_stars|compliment_cool|compliment_cute|compliment_funny|
compliment_hot|compliment_list|compliment_more|compliment_note|
compliment_photos|compliment_plain|compliment_profile|
compliment_writer| cool|               elite|fans|
friends|funny|    name|review_count|useful|             user_id|
yelping_since|
+------------+--------------+--------------+---------------
+--------------+--------------+--------------+--------------
+----------------+----------------+-----------------
+----------------+-----+--------------------+----
+--------------------+-----+--------+------------+------
+--------------------+-------------------+
|         3.91|            467|             56|             467|
250|             18|             65|            232|              180|
844|                55|              239| 5994|                2007|
267|NSCy54eWehBJyZdG2...| 1259|  Walker|         585|  7217|
qVc8ODYU5SZjKXVBg...|2007-01-25 16:47:26|
|         3.74|           3131|            157|            3131|
1145|            251|            264|           1847|
1946|            7054|               184|             1521|27281|
2009,2010,2011,20...|3138|ueRPE0CX75ePGMqOF...|13066|  Daniel|
4333| 43091|j14WgRoU_-2ZE1aw1...|2009-01-25 04:35:42|
|         3.32|            119|             17|             119|
89|              3|             13|             66|               18|
96|               10|              35| 1003|2009,2010,2011,20...|
52|LuO3Bn4f3rlhyHIaN...| 1010|   Steph|         665|  2086|
2WnXYQFK0hXEoTxPt...|2008-07-25 10:41:00|
|         4.27|             26|              6|              26|
24|              2|              4|             12|                9|
16|                1|              10|  299|     2009,2010,2011|
28|enx1vVPnfdNUdPho6...|  330|    Gwen|         224|   512|
SZDeASXq7o05mMNLs...|2005-11-29 04:38:33|
|         3.54|              0|              0|               0|
1|              0|             1|             1|                0|
1|                0|              0|    7|                    |
1|PBK4q9KEEBHhFvSXC...|   15|   Karen|          79|    29|hA5lMy-
EnncsH4JoR...|2007-01-05 19:40:59|
|         3.85|           2543|            361|            2543|
1713|            147|            163|           1212|
323|            5696|               191|              815|11211|
2006,2007,2008,20...|1357|xBDpTUbai0DXrvxCe...| 9940|    Jane|
1221| 14953|q_QQ5kBBwlCcbL1s4...|2005-03-14 20:26:35|
```

```
|         2.75|              0|              0|               0|
0|              0|              0|              0|               0|
1|                 0|                0|    0|                 |
1|HDAQ74AEznP-YsMk1...|    1|     Rob|          12|     6|
cxuxXkcihfCbqt5By...|2009-02-24 03:09:06|
|         3.73|             12|              0|              12|
4|              0|              7|              8|               0|
6|                 2|                5|  143|                 |
23|y2GyxJF5VQWohxgw_...|  102|    Mike|         358|   399|
E9kcWJdJUHuTKfQur...|2008-12-11 22:11:56|
|         4.04|              5|              3|               5|
2|              0|              0|              3|               1|
4|                 0|                3|   46|                 |
7|tOQDlz36rI__SOsbL...|   40|Rachelle|          40|   109|lO1iq-
f75hnPNZkTy...|2008-12-29 22:40:56|
|          3.4|              3|              0|               3|
0|              0|              0|              1|               0|
6|                 0|                0|   23|                 |
4|gy5fWeSv3Gamuq9Ox...|   20|    John|         109|   154|
AUi8MPWJ0mLkMfwbu...|2010-01-07 18:32:04|
+-------------+---------------+---------------+----------------
+--------------+---------------+---------------+--------------
+-----------------+----------------+-----------------
+-----------------+-----+--------------------+----
+--------------------+-----+--------+------------+------
+--------------------+------------------+
only showing top 10 rows

1987897

#df_user.select(col('elite')).show(10)

{"version_major":2,"version_minor":0,"model_id":"f6db2ad441da4a9a9d628
8f1c6de98f7"}

{"version_major":2,"version_minor":0,"model_id":""}

bus_only = df_business.select("business_id",
col("stars").alias("bus_stars"))
review_only = df_review.select("business_id", "stars", "user_id")
user_only = df_user.select("elite", "user_id")


bus_review = bus_only.join(review_only, ['business_id'])
bus_review_user = bus_review.join(user_only, ['user_id'])
bus_review_user.show(10)

{"version_major":2,"version_minor":0,"model_id":"addecec6089d47aa88ccc
bc6b945a8bd"}

{"version_major":2,"version_minor":0,"model_id":""}
```

```
+--------------------+--------------------+---------+-----+-----+
|             user_id|         business_id|bus_stars|stars|elite|
+--------------------+--------------------+---------+-----+-----+
|--3Hl2oAvTPlq-f7K...|h0wHeh0hTt6Us4W-1...|      4.0|  5.0|     |
|--3Hl2oAvTPlq-f7K...|BD1FU6xsYPtbQZ8pX...|      4.0|  2.0|     |
|--3Hl2oAvTPlq-f7K...|hodLyDkWXAosXLSTK...|      4.5|  5.0|     |
|--6lqYpHZCBbpW5z2...|mUI4DJagyUyu76qnR...|      3.0|  1.0|     |
|--ChzqcPs4YFWlw1j...|CkENBnSJFfPA1hY0q...|      3.5|  2.0|     |
|--N8bMebkrhJuFYkO...|WO-ofNAvSuJpdRt9A...|      4.5|  5.0|     |
|--QuTu4oQehIyk8VC...|oJxsRgj8Un9VAPXpa...|      3.5|  2.0|     |
|--QuTu4oQehIyk8VC...|Uky0DD3LU4C7eyNDh...|      4.0|  5.0|     |
|--QuTu4oQehIyk8VC...|-KugJyLmuTECAE121...|      4.5|  5.0|     |
|--QuTu4oQehIyk8VC...|6pT7TIyrhpeo-LjTl...|      3.5|  5.0|     |
+--------------------+--------------------+---------+-----+-----+
only showing top 10 rows
```

```
bus_review_user.count()
```


```
{"version_major":2,"version_minor":0,"model_id":"44d85bf68fcd4f95af5d1
ed10ff7d120"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
6990247
```

```
from pyspark.sql.functions import col
nll = ''
bus_review_user_elite = bus_review_user.filter(col("elite") != nll)

bus_review_user_elite.count()
```


```
{"version_major":2,"version_minor":0,"model_id":"f00a7bf979114ffeb0c10
b249fcc70a2"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
1725658

bus_review_user_elite.show(10)
```


```
{"version_major":2,"version_minor":0,"model_id":"0488b8d28ded461b97027
700598e833a"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+--------------------+---------+-----
+--------------------+
|             user_id|         business_id|bus_stars|stars|
elite|
+--------------------+--------------------+---------+-----
+--------------------+
|IeSz60ozr1yAVIH8C...|TV81bpCQ6p6o4Hau5...|      4.5|  4.0|
2017,2018,2019,20...|
```

```
|xW2A0MciHB0pLB4RH...|W4ZEKkva9HpAdZG88...|      4.0|  5.0|
2014,2015,2016,20...|
|SSafXe2aUO0cXgQhE...|E-4t5Hoon6aVFTWDP...|      4.0|  5.0|
2014,2015,2016,20...|
|yiYUEExKfZEv_T8CF...|_pbx96FZ3eHJw-V_R...|      2.5|  3.0|
2015|
|A3EiqW7_k00gvaiQi...|8uF-bhJFgT4Tn6DTb...|      4.5|  5.0|
2019,20,20|
|Zsucq1c-sjuGxs5jZ...|zaC6coZ5Gp8mLjeg7...|      4.5|  4.0|
2011,2012,2013,20...|
|aX3vDE1UmbdrWeOsg...|EqEcDeXqIq1YwnzHg...|      4.5|  5.0|
2018,2019,20,20,2021|
|aHiQYaTXrmQTeG610...|3w7NRntdQ9h0KwDsk...|      2.0|  4.0|
2012,2013,2014,20...|
|g34Qcj06LmCDhKzks...|yE1raqkLX7OZsjmX3...|      4.0|  5.0|
2017,2018,2019,20,20|
|yiYUEExKfZEv_T8CF...|EP2jFD3aGoSBCWb7i...|      4.0|  4.0|
2015|
+--------------------+--------------------+---------+-----
+--------------------+
only showing top 10 rows

bus_review_user_elite_diff =
bus_review_user_elite.withColumn('star_diff',
(bus_review_user_elite['bus_stars'] - bus_review_user_elite['stars']))
bus_review_user_elite_diff.show(5)

{"version_major":2,"version_minor":0,"model_id":"ddb858a7fb714f2b8b9a8
80033ca3d3a"}

{"version_major":2,"version_minor":0,"model_id":""}

+--------------------+--------------------+---------+-----
+--------------------+---------+
|             user_id|         business_id|bus_stars|stars|
elite|star_diff|
+--------------------+--------------------+---------+-----
+--------------------+---------+
|IeSz60ozr1yAVIH8C...|TV81bpCQ6p6o4Hau5...|      4.5|  4.0|
2017,2018,2019,20...|      0.5|
|xW2A0MciHB0pLB4RH...|W4ZEKkva9HpAdZG88...|      4.0|  5.0|
2014,2015,2016,20...|     -1.0|
|SSafXe2aUO0cXgQhE...|E-4t5Hoon6aVFTWDP...|      4.0|  5.0|
2014,2015,2016,20...|     -1.0|
|yiYUEExKfZEv_T8CF...|_pbx96FZ3eHJw-V_R...|      2.5|  3.0|
2015|     -0.5|
|A3EiqW7_k00gvaiQi...|8uF-bhJFgT4Tn6DTb...|      4.5|  5.0|
2019,20,20|     -0.5|
+--------------------+--------------------+---------+-----
+--------------------+---------+
only showing top 5 rows
```

```
bus_review_user_elite_diff = bus_review_user_elite_diff.toPandas()
```

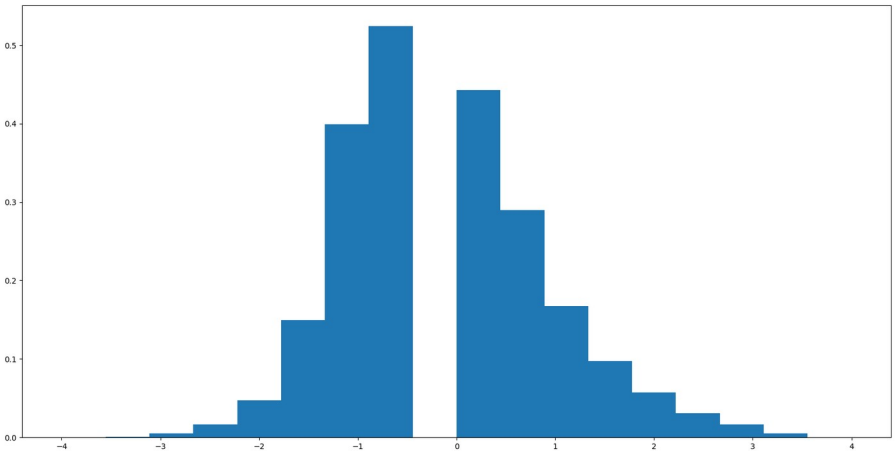

```
{"version_major":2,"version_minor":0,"model_id":"56aa019eeb8d4168a56c5e578ab90839"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
plt.figure(figsize=(20,10))
plt.hist(bus_review_user_elite_diff['star_diff'],bins=18,density=True)
plt.show()
%matplot plt
```


```
{"version_major":2,"version_minor":0,"model_id":"0e807e18f7884ee4bd957d2dd4fe2bc3"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
bus_review_user_notelite = bus_review_user.filter(col("elite") == nll)
bus_review_user_notelite.count()
```


```
{"version_major":2,"version_minor":0,"model_id":"daa67024bd2747edacc812d1484c11d9"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
5264589
```

```
bus_review_user_notelite.show(10)
```


```
{"version_major":2,"version_minor":0,"model_id":"304b24992f3c48e0877339fbdf225fe6"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+--------------------+---------+-----+-----+
|             user_id|         business_id|bus_stars|stars|elite|
```

```
+--------------------+--------------------+---------+-----+-----+
|--3Hl2oAvTPlq-f7K...|BD1FU6xsYPtbQZ8pX...|      4.0|  2.0|     |
|--3Hl2oAvTPlq-f7K...|hodLyDkWXAosXLSTK...|      4.5|  5.0|     |
|--3Hl2oAvTPlq-f7K...|h0wHeh0hTt6Us4W-1...|      4.0|  5.0|     |
|--6lqYpHZCBbpW5z2...|mUI4DJagyUyu76qnR...|      3.0|  1.0|     |
|--ChzqcPs4YFWlw1j...|CkENBnSJFfPA1hY0q...|      3.5|  2.0|     |
|--N8bMebkrhJuFYkO...|WO-ofNAvSuJpdRt9A...|      4.5|  5.0|     |
|--QuTu4oQehIyk8VC...|-KugJyLmuTECAE121...|      4.5|  5.0|     |
|--QuTu4oQehIyk8VC...|oJxsRgj8Un9VAPXpa...|      3.5|  2.0|     |
|--QuTu4oQehIyk8VC...|6pT7TIyrhpeo-LjTl...|      3.5|  5.0|     |
|--QuTu4oQehIyk8VC...|Uky0DD3LU4C7eyNDh...|      4.0|  5.0|     |
+--------------------+--------------------+---------+-----+-----+
only showing top 10 rows

bus_review_user_notelite_diff =
bus_review_user_notelite.withColumn('star_diff',
(bus_review_user_notelite['bus_stars'] -
bus_review_user_notelite['stars']))
bus_review_user_notelite_diff.show(5)

{"version_major":2,"version_minor":0,"model_id":"cb3cdfe136bf4f4787fae
99c6a455d8b"}

{"version_major":2,"version_minor":0,"model_id":""}

+--------------------+--------------------+---------+-----+-----
+---------+
|             user_id|         business_id|bus_stars|stars|elite|
star_diff|
+--------------------+--------------------+---------+-----+-----
+---------+
|--3Hl2oAvTPlq-f7K...|BD1FU6xsYPtbQZ8pX...|      4.0|  2.0|     |
2.0|
|--3Hl2oAvTPlq-f7K...|h0wHeh0hTt6Us4W-1...|      4.0|  5.0|     |
-1.0|
|--3Hl2oAvTPlq-f7K...|hodLyDkWXAosXLSTK...|      4.5|  5.0|     |
-0.5|
|--6lqYpHZCBbpW5z2...|mUI4DJagyUyu76qnR...|      3.0|  1.0|     |
2.0|
|--ChzqcPs4YFWlw1j...|CkENBnSJFfPA1hY0q...|      3.5|  2.0|     |
1.5|
+--------------------+--------------------+---------+-----+-----
+---------+
only showing top 5 rows

bus_review_user_notelite_diff =
bus_review_user_notelite_diff.drop("user_id","business_id","bus_stars"
,"stars","elite")
bus_review_user_notelite_diff =
bus_review_user_notelite_diff.toPandas()
```

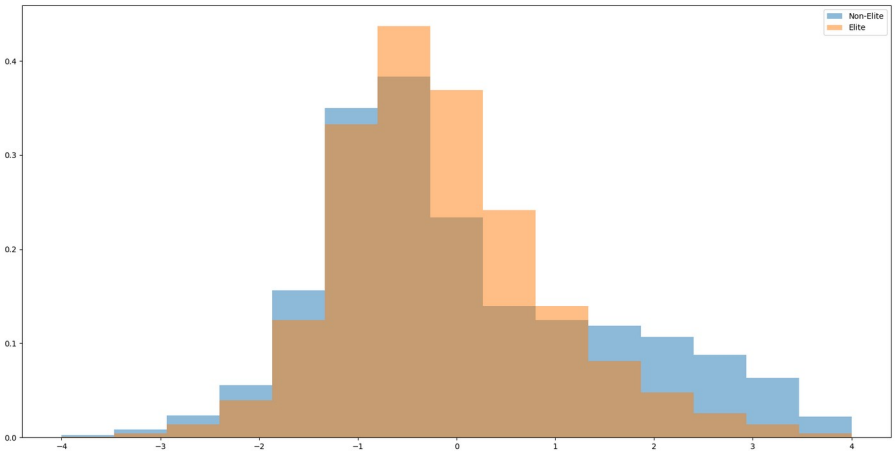
{"version_major":2,"version_minor":0,"model_id":"04119f6dac9e446a8f1e7959402a2217"}

{"version_major":2,"version_minor":0,"model_id":""}

```
plt.figure(figsize=(20,10))
plt.hist(bus_review_user_notelite_diff['star_diff'],bins=15,density=True,alpha=0.5)
plt.hist(bus_review_user_elite_diff['star_diff'],bins=15,density=True,alpha=0.5)
plt.legend(["Non-Elite","Elite"])
plt.show()
%matplot plt
```

{"version_major":2,"version_minor":0,"model_id":"4370ccd230884cc8aea1588d5d256316"}

{"version_major":2,"version_minor":0,"model_id":""}

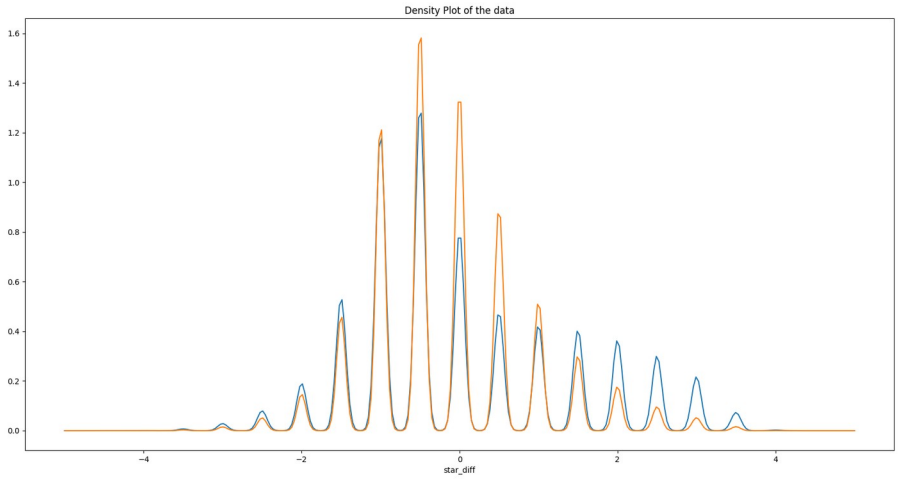



```
from scipy.stats import kde

data1 = bus_review_user_notelite_diff['star_diff']
density1 = kde.gaussian_kde(data1)
x1 = np.linspace(-5,5,300)
y1=density1(x1)

data2 = bus_review_user_elite_diff['star_diff']
density2 = kde.gaussian_kde(data2)
x2 = np.linspace(-5,5,300)
y2=density2(x2)

plt.figure(figsize=(20,10))
plt.plot(x1, y1)
plt.plot(x2, y2)
```

```
plt.title("Density Plot of the data")
plt.xlabel("star_diff")
plt.show()
%matplot plt
```

{"version_major":2,"version_minor":0,"model_id":"84c07573423342edaff65f54956d9099"}

{"version_major":2,"version_minor":0,"model_id":""}




```
bus_review_user_notelite_diff['star_diff'].describe()
bus_review_user_elite_diff['star_diff'].describe()
```

{"version_major":2,"version_minor":0,"model_id":"eee966dd65e0495da2e0e2245b003ed4"}

{"version_major":2,"version_minor":0,"model_id":""}

```
count    1.725658e+06
mean    -1.425587e-01
std      1.003810e+00
min     -4.000000e+00
25%     -1.000000e+00
50%     -5.000000e-01
75%      5.000000e-01
max      4.000000e+00
Name: star_diff, dtype: float64
```

Should the Elite be Trusted? - Analysis

The hist plot of the difference column from both elite and non elite customers seems to be slightly skewed.

As we know that -

If a density curve is left skewed, then the mean is less than the median.

If a density curve is right skewed, then the mean is greater than the median.

If a density curve has no skew, then the mean is equal to the median.

The descriptive stats of difference column of notelite:

mean --> 0.05037288, median --> -0.5000000

As per these values we can say that the density curve is right skewed.

The descriptive stats of difference column of elite:

mean --> -0.1425587, median --> -0.5000000

As per these values we can say that the density curve is right skewed.

However, the if we look at the average rating difference values of elite and notelite groups. The notelite group is closer to zero, suggesting that their star ratings are closer to the actual business rating compared to the elite group.

Additionally, since the values are very small we cannot conclude if we can entirely trust or not trust the elite group. We can say that both the groups have fairly closer ratings to the actual business ratings.

Extra Credit (3 points) - Popular users ratings vs Not Popular users ratings

Joining the (only selected columns) three data frames on common columns

```
bus_only = df_business.select("business_id",
col("stars").alias("bus_stars"), "state", "city")
review_only = df_review.select("business_id", "stars", "user_id")
user_only = df_user.select("average_stars", "fans", "review_count",
"useful", "user_id")


bus_review = bus_only.join(review_only, ['business_id'])
bus_review_user = bus_review.join(user_only, ['user_id'])
bus_review_user.show(10)
```


```
{"version_major":2,"version_minor":0,"model_id":"bcdb592df3694f6bbde64
869d05218c2"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+--------------------+---------+-----
+-------------+-----+-------------+----+-----------+------+
|             user_id|         business_id|bus_stars|state|
city|stars|average_stars|fans|review_count|useful|
+--------------------+--------------------+---------+-----
+-------------+-----+-------------+----+-----------+------+
|--3Hl2oAvTPlq-f7K...|h0wHeh0hTt6Us4W-1...|      4.0|   NV|
```

```
Reno|  5.0|         2.73|   0|          11|    14|
|--3Hl2oAvTPlq-f7K...|hodLyDkWXAosXLSTK...|      4.5|   NV|
Sparks|  5.0|         2.73|   0|          11|    14|
|--3Hl2oAvTPlq-f7K...|BD1FU6xsYPtbQZ8pX...|      4.0|   NV|
Reno|  2.0|         2.73|   0|          11|    14|
|--6lqYpHZCBbpW5z2...|mUI4DJagyUyu76qnR...|      3.0|   FL|
Tampa|  1.0|         3.27|   0|           9|     5|
|--ChzqcPs4YFWlw1j...|CkENBnSJFfPA1hY0q...|      3.5|   CA|Santa
Barbara|  2.0|         3.83|   0|          50|    20|
|--N8bMebkrhJuFYkO...|WO-ofNAvSuJpdRt9A...|      4.5|   LA|  New
Orleans|  5.0|         3.67|   0|           3|     7|
|--QuTu4oQehIyk8VC...|6pT7TIyrhpeo-LjTl...|      3.5|   PA|
Philadelphia|  5.0|         3.1|   0|           7|    22|
|--QuTu4oQehIyk8VC...|oJxsRgj8Un9VAPXpa...|      3.5|   PA|
Philadelphia|  2.0|         3.1|   0|           7|    22|
|--QuTu4oQehIyk8VC...|Uky0DD3LU4C7eyNDh...|      4.0|   PA|
Philadelphia|  5.0|         3.1|   0|           7|    22|
|--QuTu4oQehIyk8VC...|-KugJyLmuTECAE121...|      4.5|   DE|
Claymont|  5.0|         3.1|   0|           7|    22|
+--------------------+--------------------+---------+-----
+------------+-----+------------+----+-----------+------+
only showing top 10 rows
```

Understanding the distribution of fans column

```
bus_review_user_fans =
bus_review_user.drop("user_id","business_id","bus_stars","state","city
","stars","average_stars","review_count")
bus_review_user_fans = bus_review_user_fans.toPandas()
bus_review_user_fans.describe()
#fans : 0 - 12497, 13.20916
#useful: 0 - 206296, 427.9901
```


```
{"version_major":2,"version_minor":0,"model_id":"93392b27a32a40f49e3dc
4d45c3f07b2"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```


```
               fans        useful
count  6.990247e+06  6.990247e+06
mean   1.320916e+01  4.279901e+02
std    8.723713e+01  3.226596e+03
min    0.000000e+00  0.000000e+00
25%    0.000000e+00  3.000000e+00
50%    0.000000e+00  1.900000e+01
75%    4.000000e+00  1.110000e+02
max    1.249700e+04  2.062960e+05
```

splitting the data frame into two groups

bus_review_user_pop -> popular users whose #fans are greater than 100

bus_review_user_notpop -> popular users whose #fans are less than 100

```
bus_review_user_pop = bus_review_user\
                    .filter(bus_review_user.fans > 100)
bus_review_user_notpop = bus_review_user\
                    .filter(bus_review_user.fans <= 100)
```


```
{"version_major":2,"version_minor":0,"model_id":"fbfcee51f98e48ab976d2
2b5f3afcfb9"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
#bus_review_user2.count()
```


```
{"version_major":2,"version_minor":0,"model_id":"7a1ff33b0c0246dc9612d
ea61e9d7a7f"}

{"version_major":2,"version_minor":0,"model_id":""}
```


Calculating the avg star ratings per business under each of the above created groups

```
bus_review_user_pop_avg = bus_review_user_pop\
                          .groupBy(bus_review_user_pop.business_id) \
                          .avg('stars')\
                           .withColumnRenamed("avg(stars)","pop_stars
")
bus_review_user_notpop_avg = bus_review_user_notpop\
                          .groupBy(bus_review_user_notpop.business_id
) \
                          .avg('stars')\
                           .withColumnRenamed("avg(stars)","notpop_st
ars")
```


```
{"version_major":2,"version_minor":0,"model_id":"77a44772fd0846afb9f64
4d7a3b72235"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
bus_review_user_pop_avg.show(5)
bus_review_user_notpop_avg.show(5)
```


```
{"version_major":2,"version_minor":0,"model_id":"f3c95cf112d9487888d79
f7148b427e6"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+------------------+
|         business_id|         pop_stars|
+--------------------+------------------+
|zJErbOQMKX-MwHs_u...|3.1538461538461537|
|aYiAfRcjAQXeGGQu1...|               3.0|
|Yh_KhyVD6ZBwsIQQ1...|3.6666666666666665|
|Ety2Z0CImO6FYDV6L...| 4.428571428571429|
|zk7tV01E9N_VenizN...|               2.0|
```

```
+--------------------+------------------+
only showing top 5 rows

+--------------------+------------------+
|         business_id|      notpop_stars|
+--------------------+------------------+
|yqq1Fvt7WtduIO3Gw...|2.9315068493150687|
|zJErbOQMKX-MwHs_u...|2.9138755980861246|
|XH3mYdTg4ZxWV-8W7...| 4.176470588235294|
|oQ0MQpVVyzGe_JTIL...| 3.782608695652174|
|ZFaG1Q3voENwwZPQA...|2.4285714285714284|
+--------------------+------------------+
only showing top 5 rows
```

Joining the two dataframes, creates a new one that has business id, avg popular user rating, avg nonpopular user rating

```
bus_review_user_avg_all =
bus_review_user_pop_avg.join(bus_review_user_notpop_avg,
['business_id'])
```


```
{"version_major":2,"version_minor":0,"model_id":"bcf362aedfef439a8ffd6
3d9a4301d1c"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
bus_review_user_avg_all.show(5)
```


```
{"version_major":2,"version_minor":0,"model_id":"3ef5bd646b9f4620a03c5
d4ba77597f8"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+---------+------------------+
|         business_id|pop_stars|      notpop_stars|
+--------------------+---------+------------------+
|-0iIxySkp97WNlwK6...|      4.0|3.7161572052401746|
|-2wh7NTLkWEgsrLJv...|      4.0| 4.425531914893617|
|-3e3CP3FFc-rvJj_-...|    3.625|3.7083333333333335|
|-6L_z3ftD1iepJb0F...|      3.5| 4.867924528301887|
|-8562lttAp_PuLWpQ...|      4.5| 4.461538461538462|
+--------------------+---------+------------------+
only showing top 5 rows
```

Top 20 businesses based on popular user star ratings - bus_review_user_avg_pop20

```
bus_review_user_avg_pop20 =
bus_review_user_avg_all.sort(col("pop_stars").desc()).limit(20)
bus_review_user_avg_pop20.show()
```


```
{"version_major":2,"version_minor":0,"model_id":"de74aa13570b4c5a96237
d009f3c894e"}

{"version_major":2,"version_minor":0,"model_id":""}
```

```
+--------------------+---------+------------------+
|         business_id|pop_stars|      notpop_stars|
+--------------------+---------+------------------+
|2Eyu9uCg0xbitNE_d...|      5.0|              4.25|
|4rMUoAD4OylA9e-ED...|      5.0|              3.75|
|2PZdPsjYt2cZ5c0EE...|      5.0| 3.357142857142857|
|0j3QnTbA38xlDFZzP...|      5.0| 4.666666666666667|
|2RxpHCW6eOx2dvuCM...|      5.0| 4.703703703703703|
|0lzdZFAyyiYVdQDqj...|      5.0|             3.475|
|2cEJyupw_tE2P7vPM...|      5.0| 4.052631578947368|
|-a_rcDrFanuh8lDCb...|      5.0|               2.8|
|2pYUBcNkUxD1Bg8FA...|      5.0| 4.431506849315069|
|045a4sFqlTYnmQ9oo...|      5.0| 2.857142857142857|
|3ArvoGCDuPYIj-RII...|      5.0|2.3333333333333335|
|0hXhsAhvwGoMG0kBC...|      5.0|2.6666666666666665|
|3FlUa98PjRhFEfmMk...|      5.0| 4.970588235294118|
|-EEtnHdSUwHEq_oll...|      5.0| 4.512820512820513|
|3Lf3nWp9TcIj7hvw9...|      5.0| 4.717948717948718|
|199cAgEkz82JS1p6a...|      5.0| 4.145161290322581|
|3NObha3nErUqtucmc...|      5.0| 4.927927927927928|
|1SzruLiMEH-gVoQKA...|      5.0|               5.0|
|3c3okC0udBKZD_fAS...|      5.0| 4.666666666666667|
|1iq3rdfVkcmaIIrq9...|      5.0|               4.0|
+--------------------+---------+------------------+
```

Top 20 businesses based on non popular user star ratings - bus_review_user_avg_notpop20

```
bus_review_user_avg_notpop20 =
bus_review_user_avg_all.sort(col("notpop_stars").desc()).limit(20)
bus_review_user_avg_notpop20.show()
```


```
{"version_major":2,"version_minor":0,"model_id":"acb6b5d0d0b74eabac2c0
1c90b94a0c9"}

{"version_major":2,"version_minor":0,"model_id":""}
```


```
+--------------------+-----------------+------------+
|         business_id|        pop_stars|notpop_stars|
+--------------------+-----------------+------------+
|IjRF9ohTpWc8KcVjy...|              5.0|         5.0|
|Tlik-EshZIUhltgxF...|              4.0|         5.0|
|2c7qkJxkNUi8WLp1H...|              5.0|         5.0|
|K38t-YmXMDPShYCNz...|              5.0|         5.0|
|2r3EByiuMsKp7sxSZ...|              4.0|         5.0|
|LblcxMOPjalvJVMzT...|              5.0|         5.0|
|5O6Tg6yWk2ySa9rD7...|              5.0|         5.0|
|NmpFSvxOWQj34rfgs...|              5.0|         5.0|
|NsGDYMsoRFPoBweOW...|              3.0|         5.0|
|HsMr8L-mNuY0ZANik...|4.666666666666667|         5.0|
|OawxYvvdRuHPM0JFi...|              3.0|         5.0|
|6o3MHG1ofTZidJoKQ...|              4.0|         5.0|
|RJGcaGTxCxszS23fo...|              5.0|         5.0|
```

```
|Er2Sa7bYraSXbsA5V...|              5.0|         5.0|
|SJ3iTeHq3Om0SoruO...|              5.0|         5.0|
|Owq1Dovc83YQqC3UY...|              5.0|         5.0|
|PQtxL2Ow_VGd9fr9o...|              4.0|         5.0|
|QENvbR6VwbBL_fZ71...|              4.2|         5.0|
|WpfS-F8Mw38d4SDIc...|              4.0|         5.0|
|XBqrCQbNfRc3SmwTa...|              5.0|         5.0|
+--------------------+-----------------+------------+
```

```
bus_review_user_avg_pop20 = bus_review_user_avg_pop20.toPandas()
bus_review_user_avg_notpop20 = bus_review_user_avg_notpop20.toPandas()
```


```
{"version_major":2,"version_minor":0,"model_id":"97a05f352c654aef94027
55900e9e7d2"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```

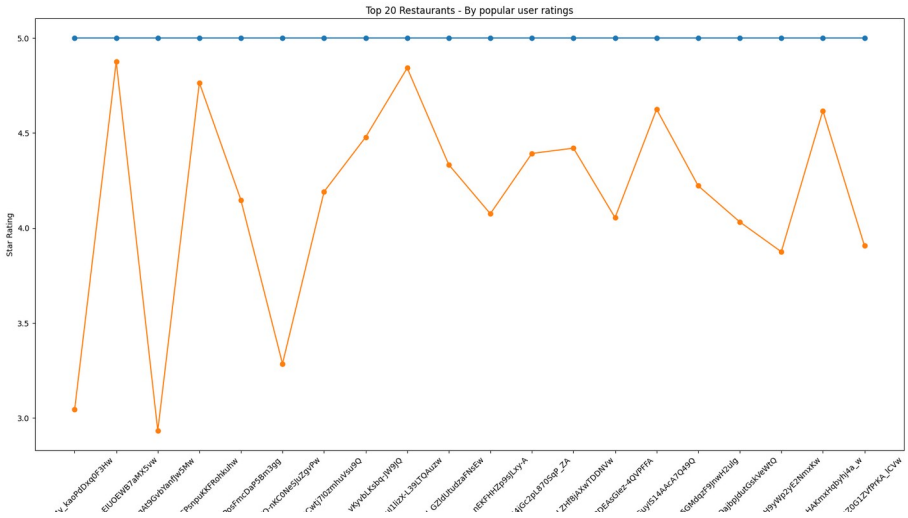

Scatter plot - Top 20 Restaurants - By popular user ratings

Overlaying the popular user ratings and notpopular user ratings of these top20

```
plt.figure(figsize=(20,10))

plt.scatter(bus_review_user_avg_pop20['business_id'],
bus_review_user_avg_pop20['pop_stars'])
plt.plot(bus_review_user_avg_pop20['business_id'],
bus_review_user_avg_pop20['pop_stars'])

plt.scatter(bus_review_user_avg_pop20['business_id'],
bus_review_user_avg_pop20['notpop_stars'])
plt.plot(bus_review_user_avg_pop20['business_id'],
bus_review_user_avg_pop20['notpop_stars'])

plt.title("Top 20 Restaurants - By popular user ratings")
plt.ylabel('Star Rating')
plt.xlabel("Business ID")
plt.xticks(rotation=45)

plt.show()
%matplot plt
```


```
{"version_major":2,"version_minor":0,"model_id":"a007de37826f47cab84bc
8eb46b93372"}
```

```
{"version_major":2,"version_minor":0,"model_id":""}
```



Implication

The top rated restauarants (5.0) (as rated by the popular users) have recieved lesser ratings for the same businesses by the non popular users. We can make a generalised inference that the users who have lesser number of fans tend to be more critical of the restaurant rating.

The reason could be that they wish to be very particular in their critique and hence expect the users to follow them on Yelp.

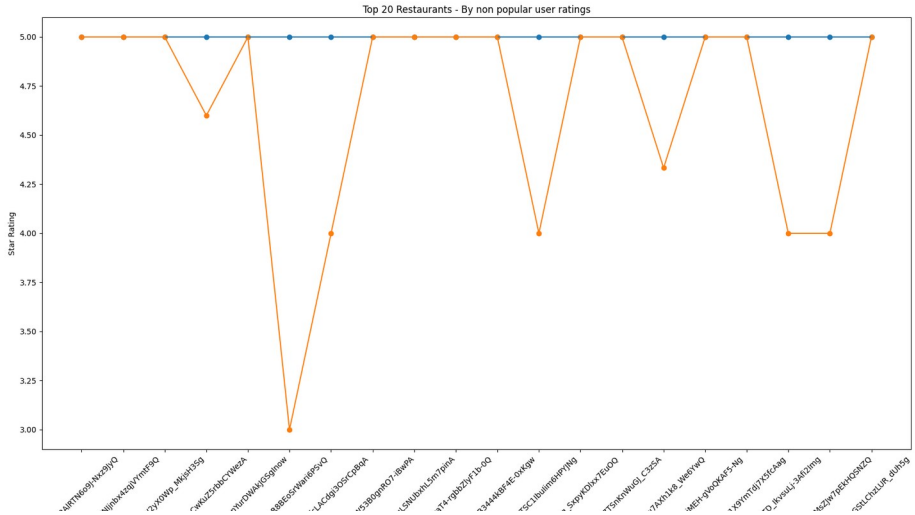
Scatter plot - Top 20 Restaurants - By non popular user ratings

Overlaying the popular user ratings and notpopular user ratings of these top20

```
plt.figure(figsize=(20,10))

plt.scatter(bus_review_user_avg_notpop20['business_id'],
bus_review_user_avg_notpop20['notpop_stars'])
plt.plot(bus_review_user_avg_notpop20['business_id'],
bus_review_user_avg_notpop20['notpop_stars'])

plt.scatter(bus_review_user_avg_notpop20['business_id'],
bus_review_user_avg_notpop20['pop_stars'])
plt.plot(bus_review_user_avg_notpop20['business_id'],
bus_review_user_avg_notpop20['pop_stars'])
plt.title("Top 20 Restaurants - By non popular user ratings")
plt.ylabel('Star Rating')
plt.xlabel("Business ID")
plt.xticks(rotation=45)
plt.show()
%matplot plt
```


```
{"version_major":2,"version_minor":0,"model_id":"f3610f85ecdf423b8b075
c791681d51f"}
```

{"version_major":2,"version_minor":0,"model_id":""}



Implication

The top rated restauarants (5.0) (as rated by the non popular users) are almost inline with the popular user ratings for those businesses.

As we made an inference that not popular users tend to be more critical, we can say that the popular users are also in agreement with their ratings.