Yelp Dataset

EMSE 6586: Database Management for Data Analytics

Presented By : Akash Poddar
Avani Rao
Karthik Iyer
Vidit Sheth

OUTLINE

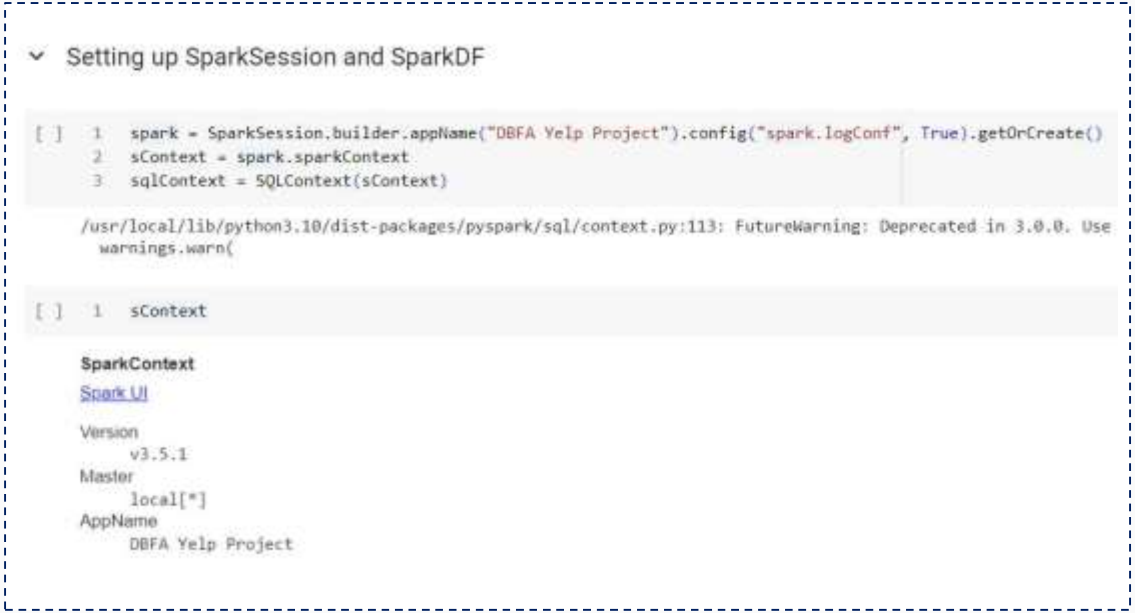
- Introduction
- Setting Up
- Reading
- Data Cleaning
- Connecting to SQL Database
- SQL Queries
- Spark SQL Queries
- Sentiment Analysis
- Conclusion

INTRODUCTION

- The Yelp Dataset contains JSON data about businesses, user reviews, user interactions, and geographical information.
- The dataset encompasses a wide range of businesses, including restaurants, cafes, hotels, salons, retail stores, and more.

SETTING UP

- The Yelp dataset spans 8GB.
- Spark seamlessly scales to manage vast data volumes by distributing processing across multiple nodes.
- With user-friendly APIs supporting various languages like Python, Spark accommodates a wide range of users.
- Its smooth integration with Python and effective big data management render Spark a prime option for our project.



Setting up SparkSession and SparkDF

```
spark = SparkSession.builder.appName("DBFA Yelp Project").config("spark.logConf", True).getOrCreate()
sContext = spark.sparkContext
sqlContext = SQLContext(sContext)
```

```
/usr/local/lib/python3.10/dist-packages/pyspark/sql/context.py:113: FutureWarning: Deprecated in 3.0.0. Use
  warnings.warn(
```

```
sContext
```

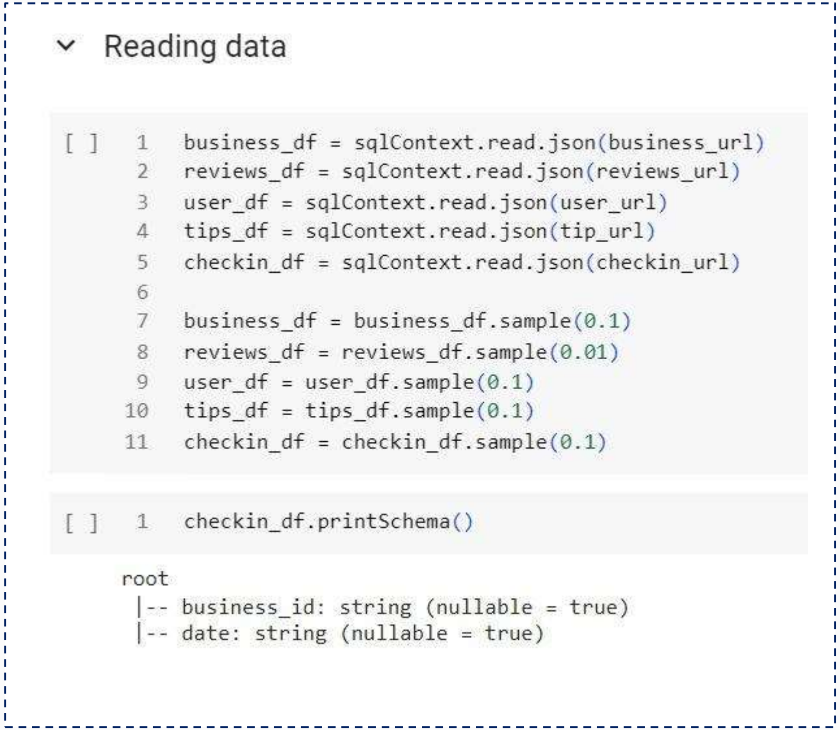
SparkContext

Spark UI

```
Version
        v3.5.1
Master
        local[*]
AppName
        DBFA Yelp Project
```

Reading Data

- **Reading JSON Data:** It reads JSON data from various sources (specified by business_url, reviews_url, user_url, tip_url, and checkin_url) into Spark DataFrames.
- **Sampling Data:** After reading the JSON data into DataFrames, the code samples a portion of each DataFrame using the sample( ) function.
- The argument passed to the sample ( ) function indicates the fraction of data to be sampled.



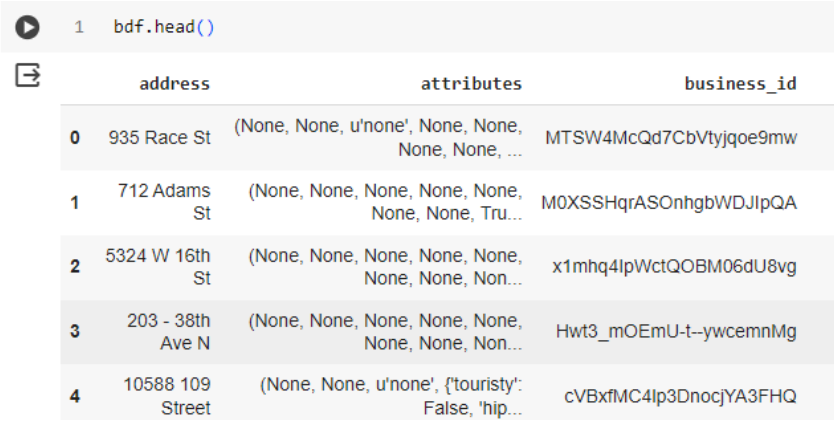
```
business_df = sqlContext.read.json(business_url)
reviews_df = sqlContext.read.json(reviews_url)
user_df = sqlContext.read.json(user_url)
tips_df = sqlContext.read.json(tip_url)
checkin_df = sqlContext.read.json(checkin_url)

business_df = business_df.sample(0.1)
reviews_df = reviews_df.sample(0.01)
user_df = user_df.sample(0.1)
tips_df = tips_df.sample(0.1)
checkin_df = checkin_df.sample(0.1)
```

```
checkin_df.printSchema()
```

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

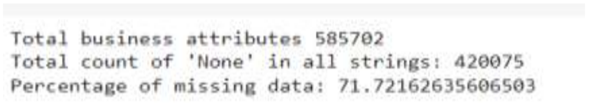
Data Cleaning - Cleaning Business Data

- Dropped null values, especially in 'attributes' (71%), impacting analysis.
- Analyzed 'attributes' for 'None' prevalence to gauge data completeness.



```
bdf.head()
```

	address	attributes	business_id
0	935 Race St	(None, None, u'none', None, None, None, None, ...	MTSW4McQd7CbVtyjqoe9mw
1	712 Adams St	(None, None, None, None, None, None, None, Tru...	M0XSSHqrASOnhgbWDJIpQA
2	5324 W 16th St	(None, None, None, None, None, None, None, Non...	x1mhq4IpWctQOBM06dU8vg
3	203 - 38th Ave N	(None, None, None, None, None, None, None, Non...	Hwt3_mOEmU-t--ywcemnMg
4	10588 109 Street	(None, None, u'none', {'touristy': False, 'hip...	cVBxfMC4Ip3DnocjYA3FHQ

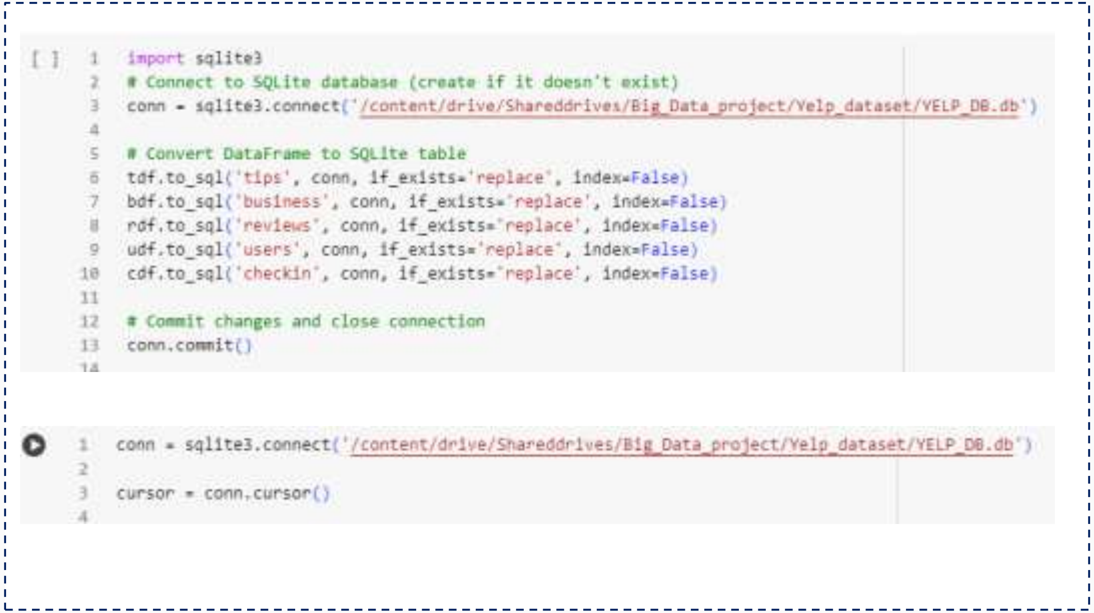


```
Total business attributes 585702
Total count of 'None' in all strings: 420075
Percentage of missing data: 71.72162635606503
```

SQL AND SPARK FOR QUERYING

Connecting to SQL Database

- Connect to SQLite database (create if it doesn't exist)
- Convert DataFrame to SQLite table
- Commit changes and close connection

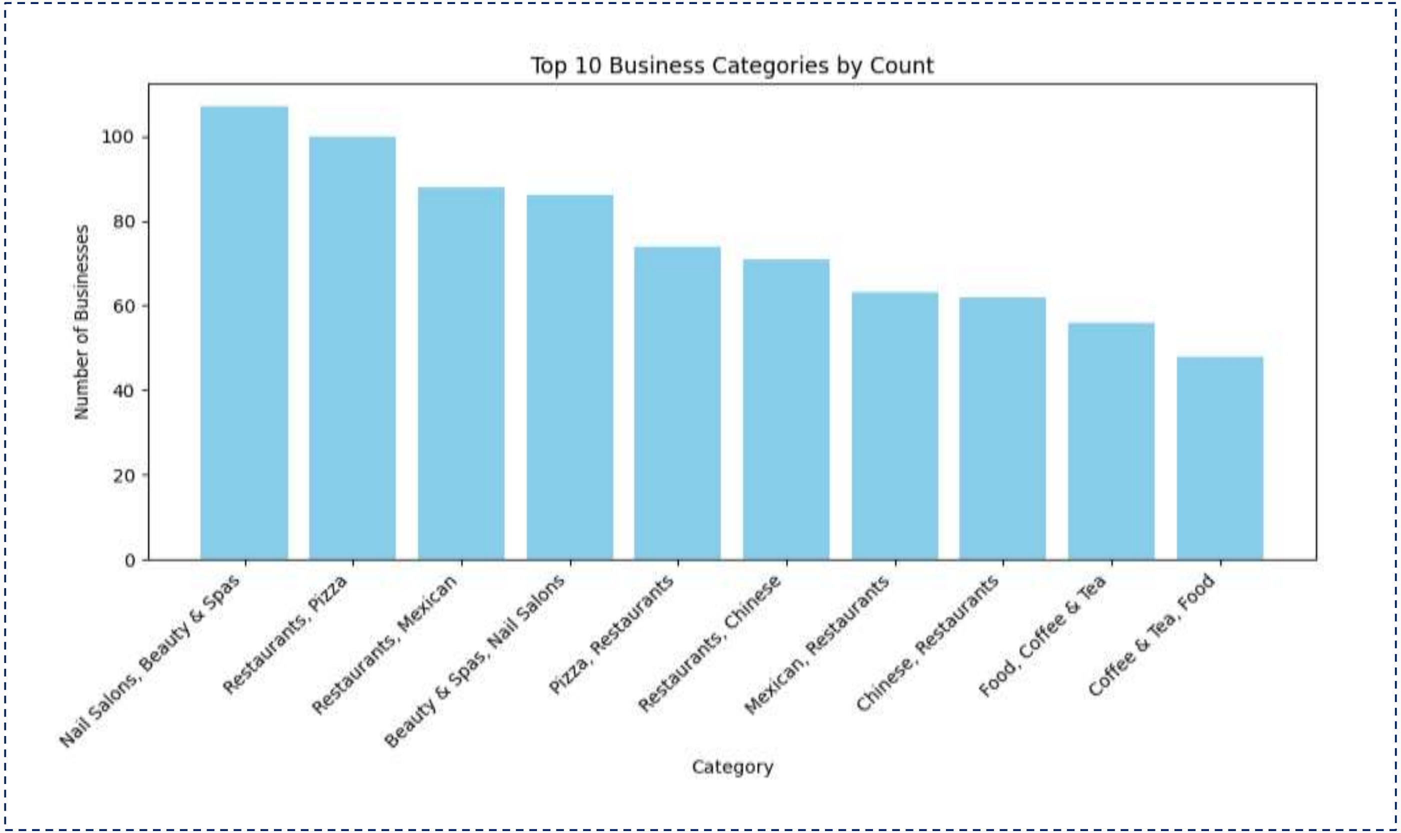


```
import sqlite3
# Connect to SQLite database (create if it doesn't exist)
conn = sqlite3.connect('/content/drive/Shareddrives/Big_Data_project/Yelp_dataset/YELP_DB.db')

# Convert DataFrame to SQLite table
tdf.to_sql('tips', conn, if_exists='replace', index=False)
bdf.to_sql('business', conn, if_exists='replace', index=False)
rdf.to_sql('reviews', conn, if_exists='replace', index=False)
udf.to_sql('users', conn, if_exists='replace', index=False)
cdf.to_sql('checkin', conn, if_exists='replace', index=False)

# Commit changes and close connection
conn.commit()
```

```
conn = sqlite3.connect('/content/drive/Shareddrives/Big_Data_project/Yelp_dataset/YELP_DB.db')

cursor = conn.cursor()
```

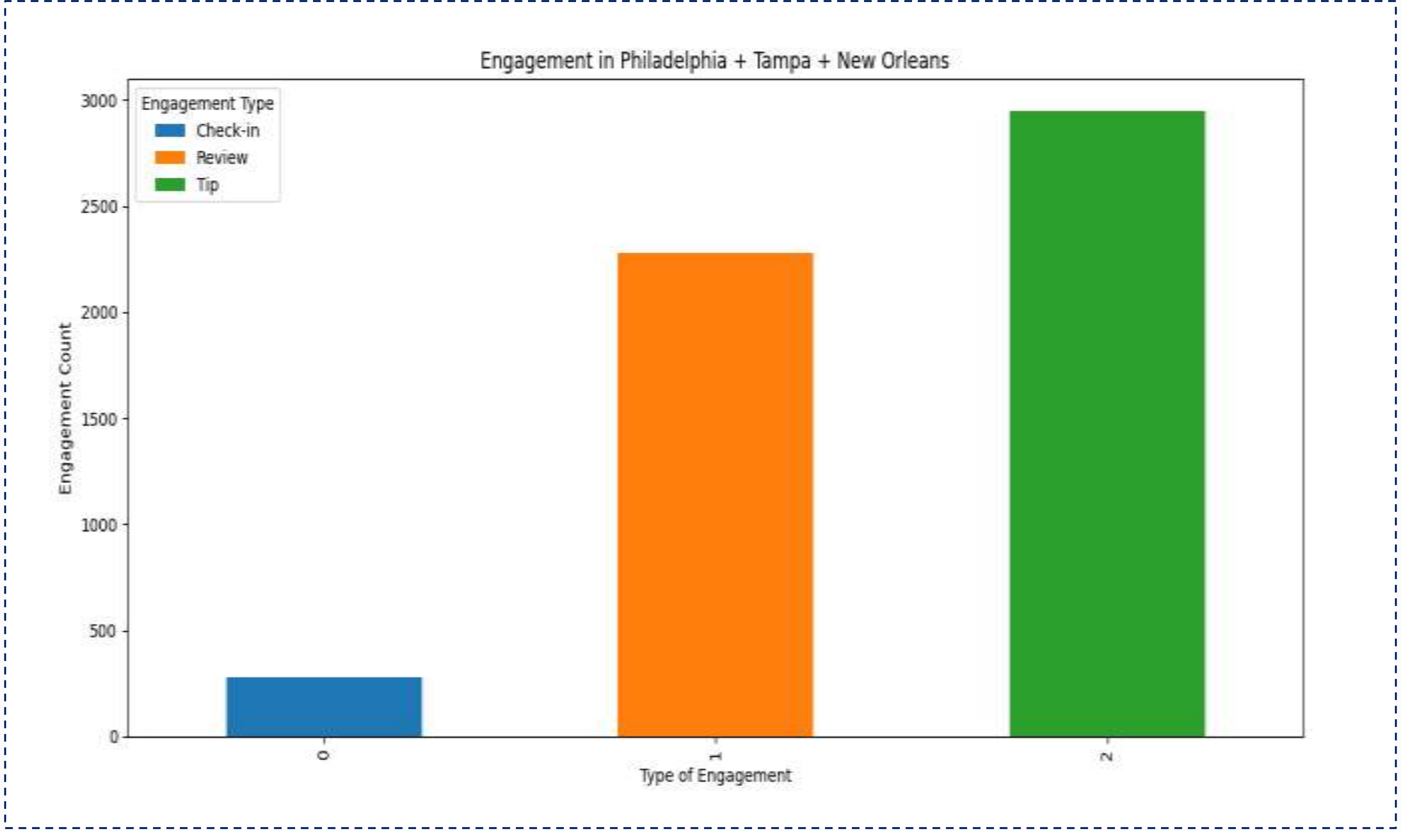
SQL QUERIES:

Query to find out top ten Categories of Businesses reviewed on Yelp



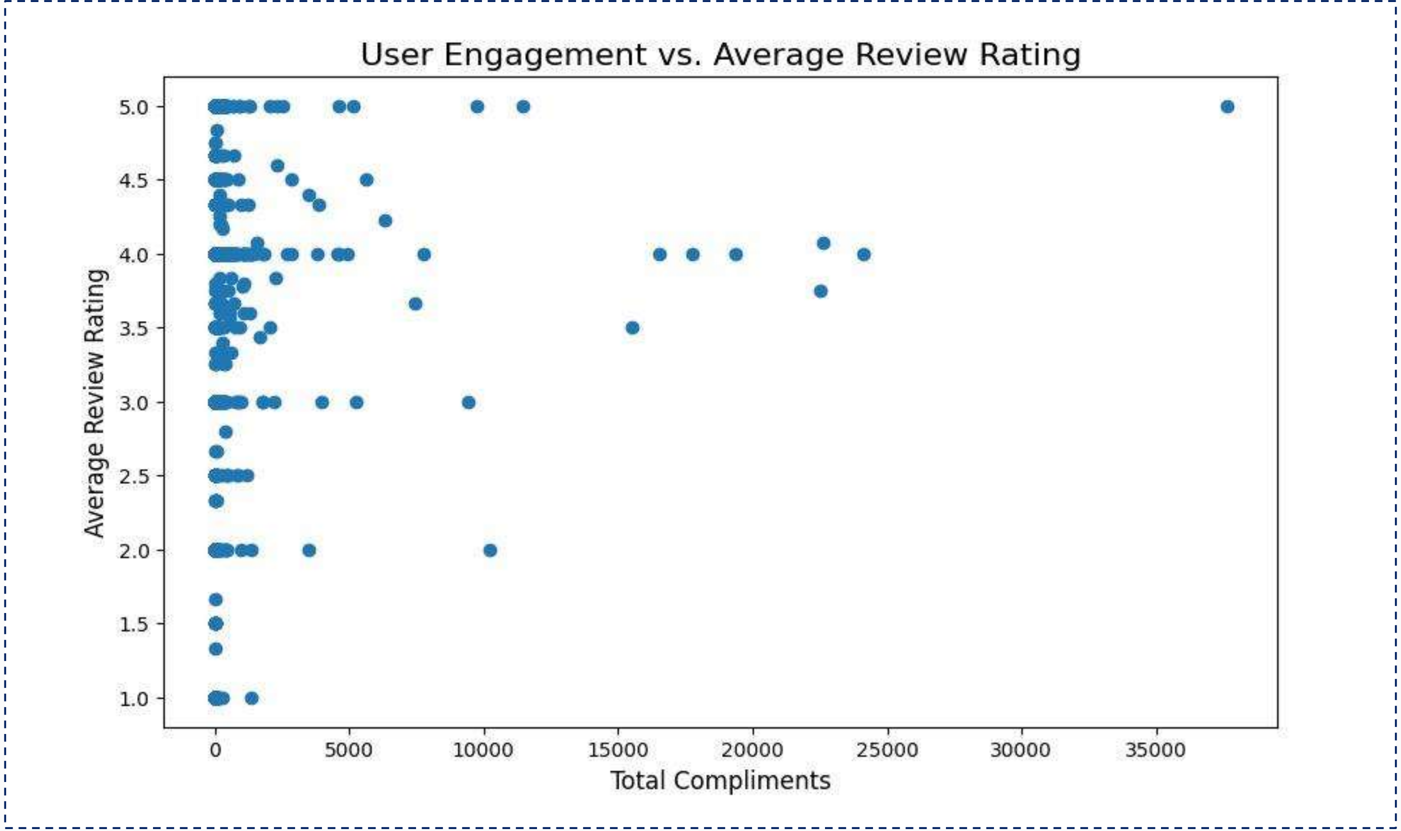
SQL QUERIES:

Query to find out Engagement Count based on Type for Philadelphia, Tampa, New Orleans



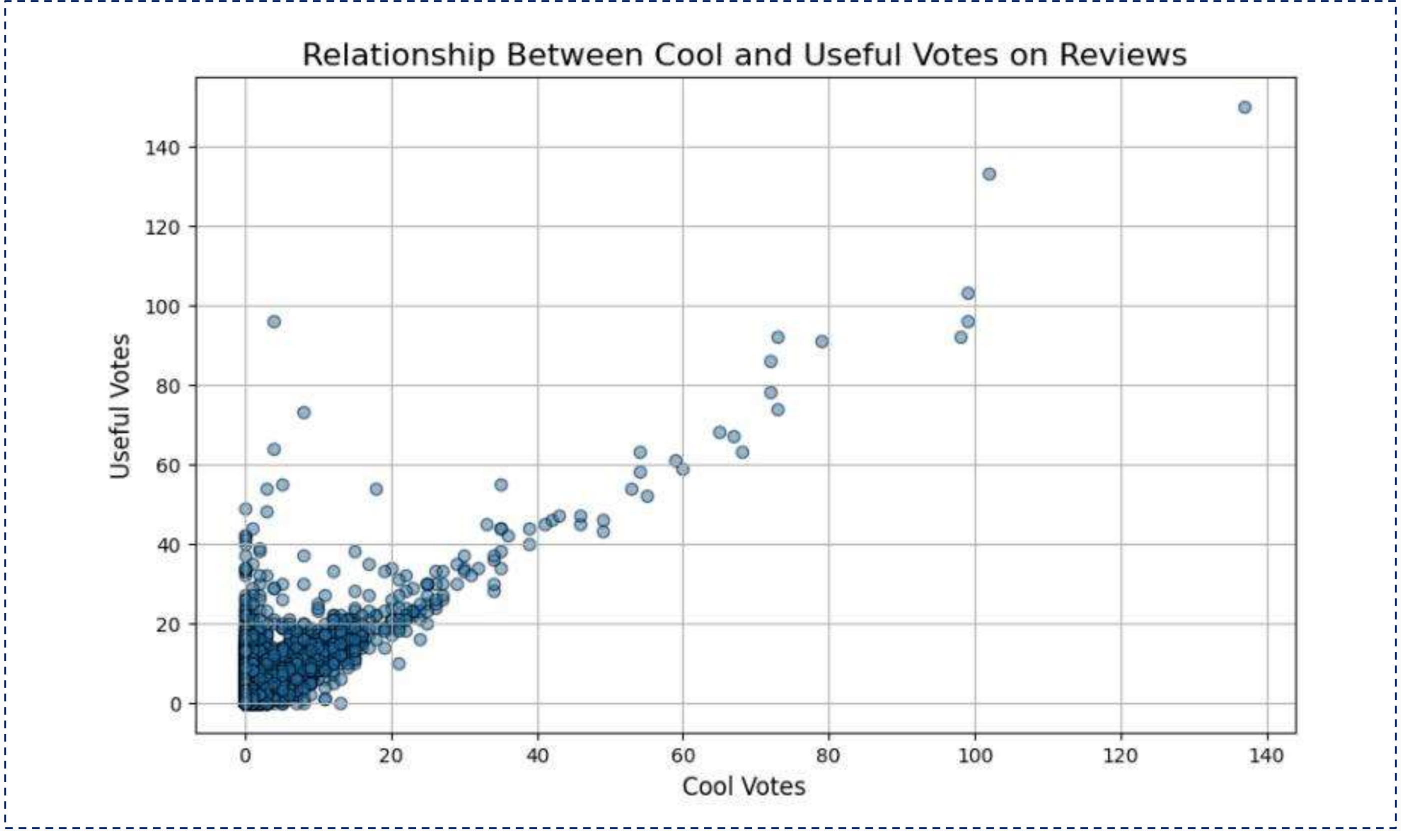
SQL QUERIES:

Query to Investigate if users who are more engaged on Yelp (via compliments and fans)



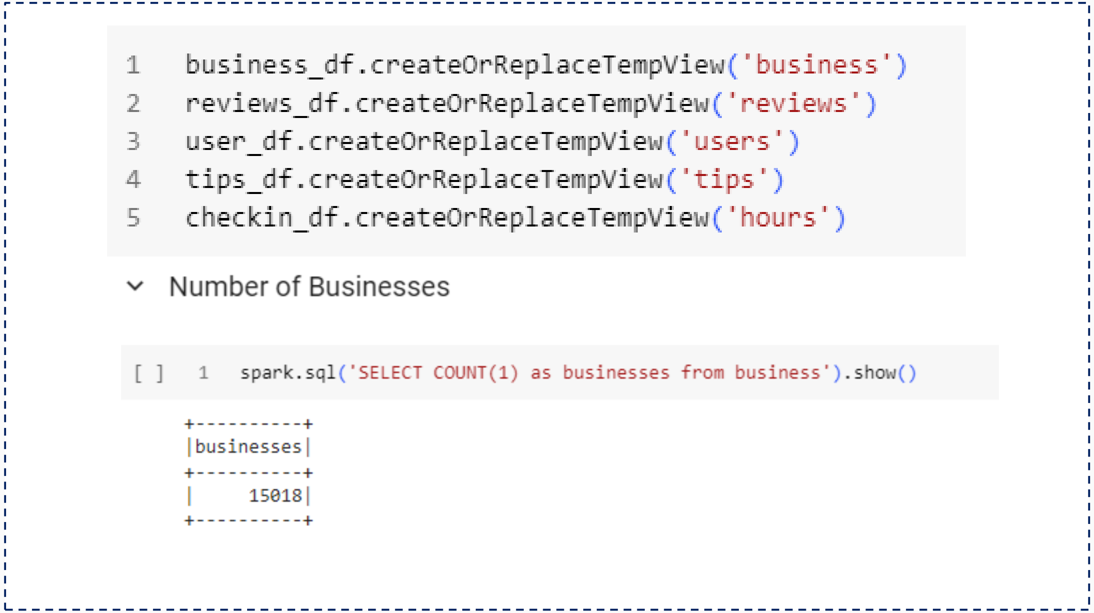
SQL QUERIES:

Query to Investigate Relationship between Cool and Useful Votes on Reviews



Spark Database

- Create a temporary instance of a spark database
- Using spark to query



```
business_df.createOrReplaceTempView('business')
reviews_df.createOrReplaceTempView('reviews')
user_df.createOrReplaceTempView('users')
tips_df.createOrReplaceTempView('tips')
checkin_df.createOrReplaceTempView('hours')
```

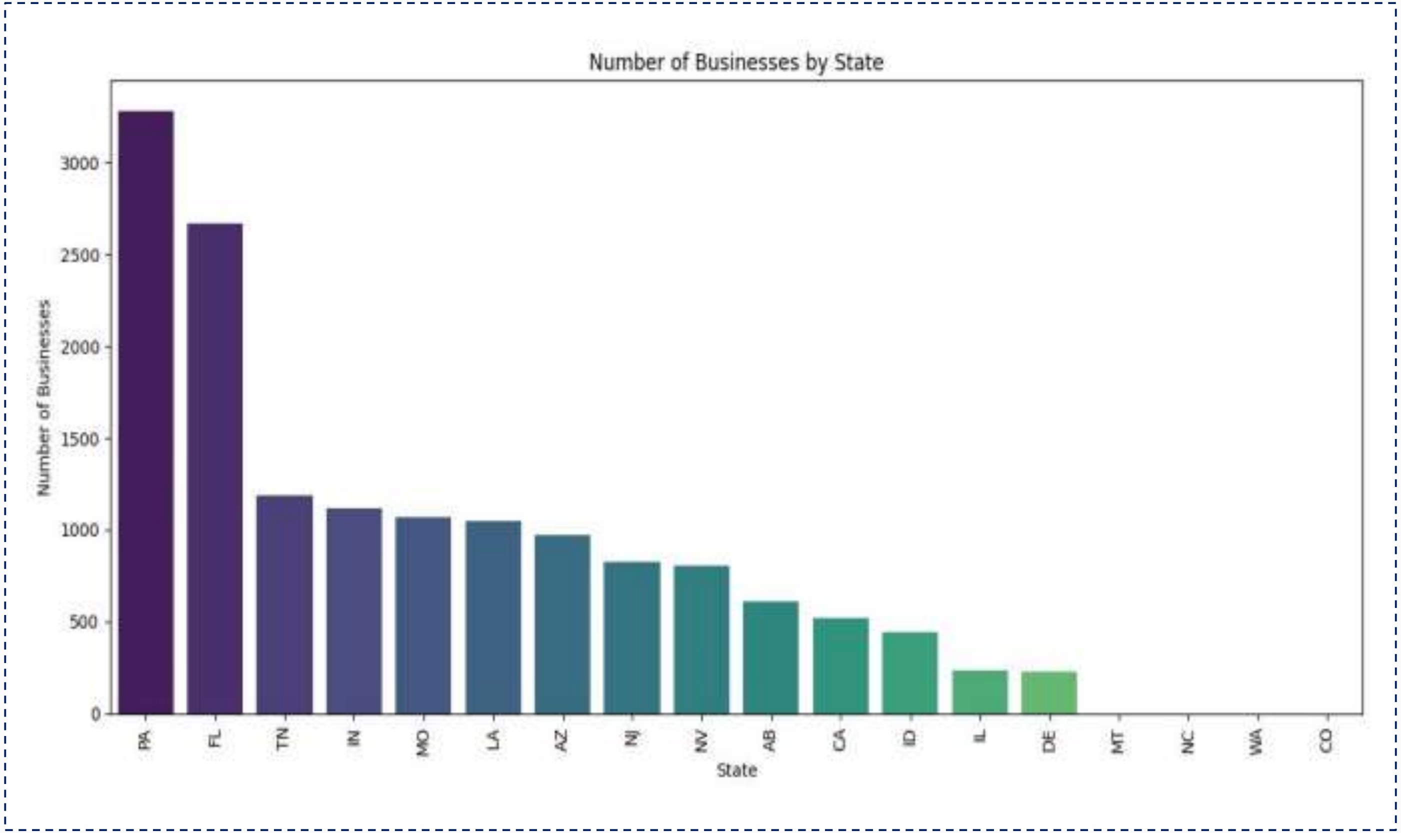
Number of Businesses

```
spark.sql('SELECT COUNT(1) as businesses from business').show()
```

```
+----------+
|businesses|
+----------+
|     15018|
+----------+
```

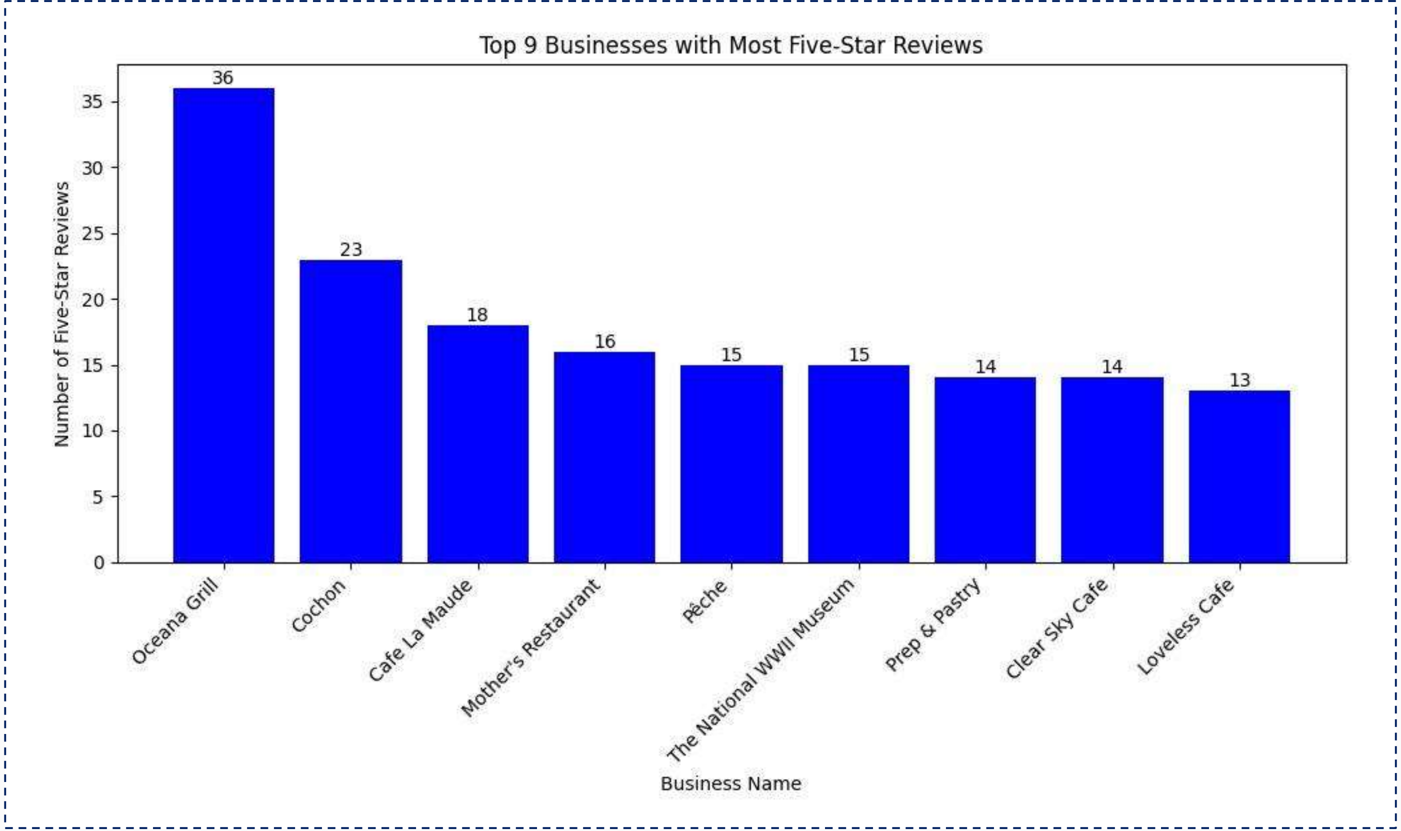
Spark SQL Queries:

Query to find out Number of Businesses by State



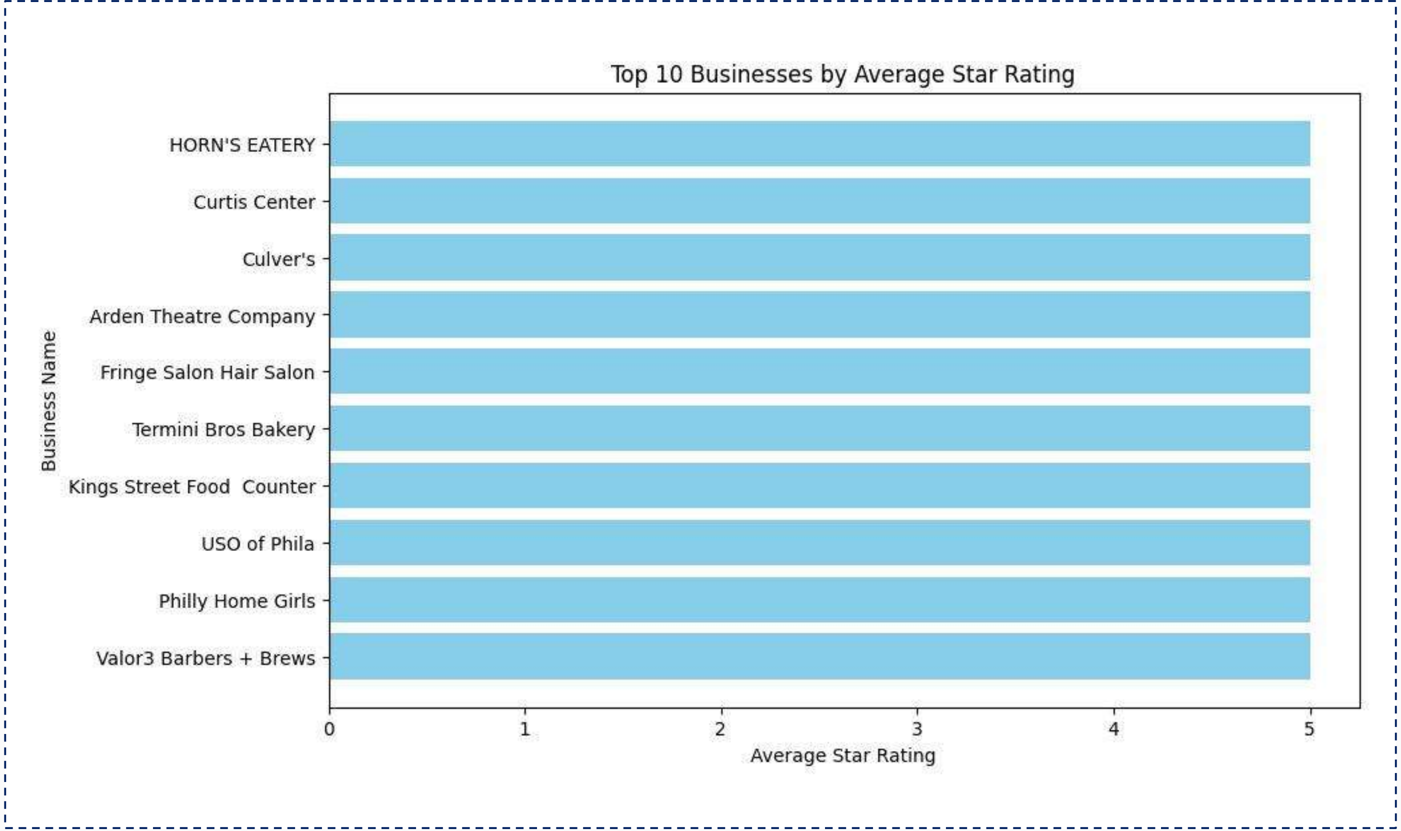
Spark SQL Queries:

Query to find out top ten Business with 5-star Rating (Business)



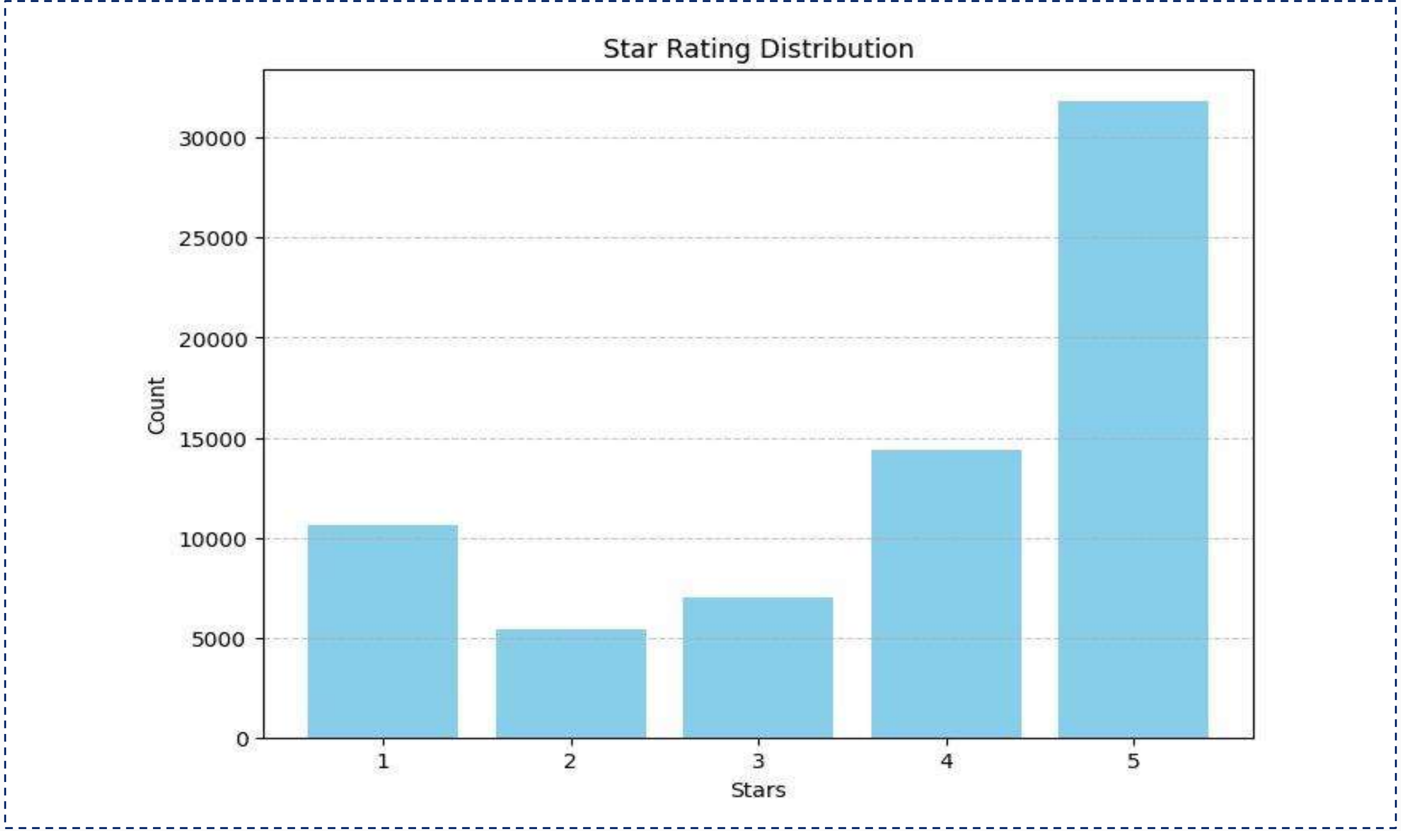
Spark SQL Queries:

Query to find out top ten Business by Average star Rating



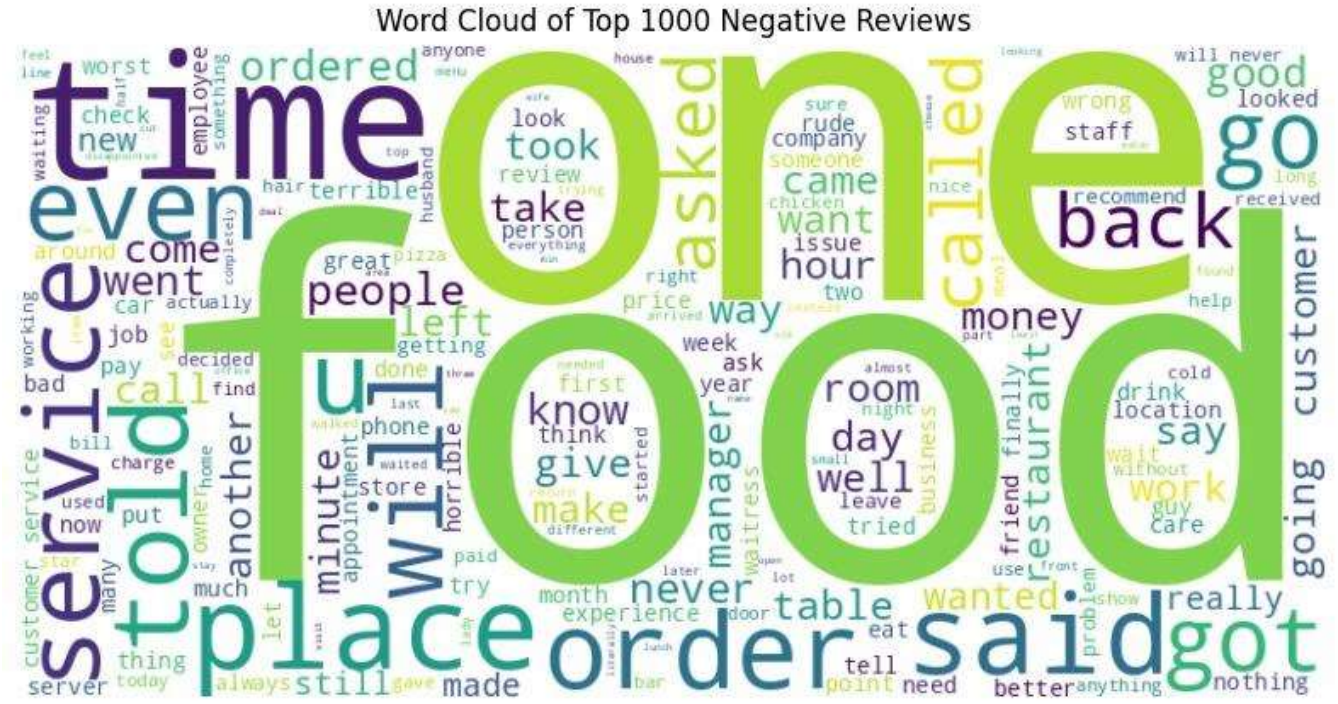
Spark SQL Queries:

Query to find out Star Rating Distribution



Spark SQL Queries:

Query to find out top ten Negative Reviews

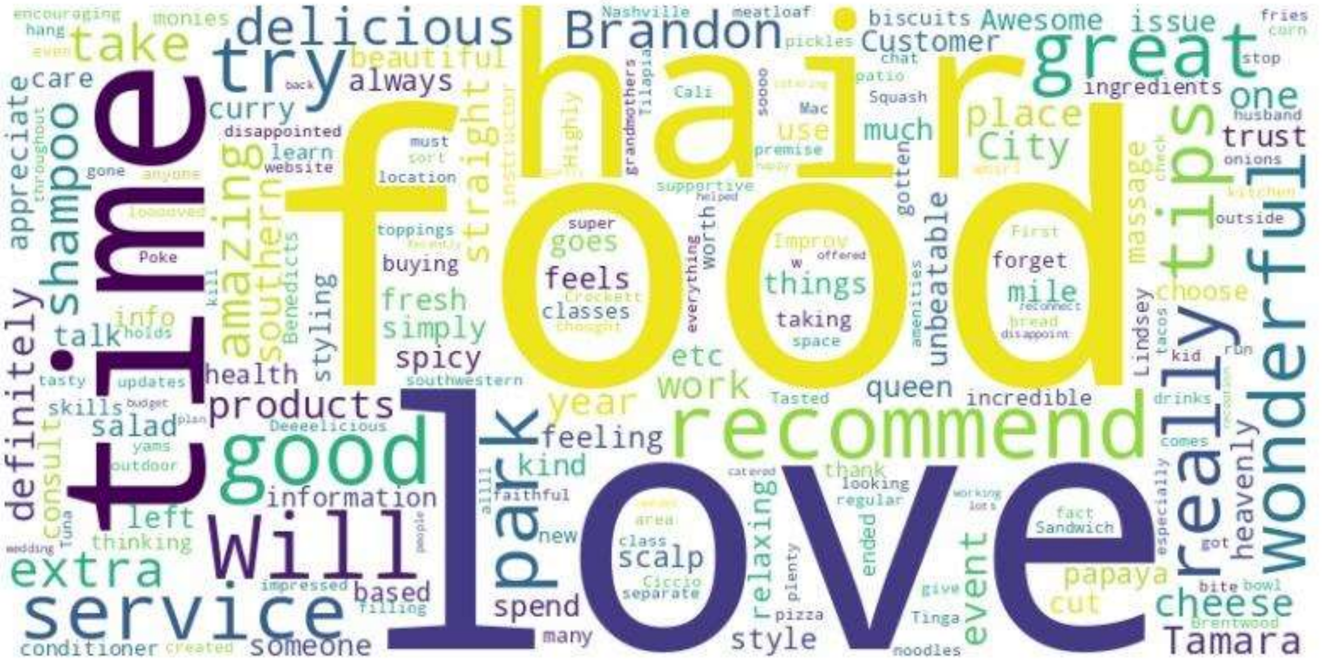


Word Cloud of Top 1000 Negative Reviews

Spark SQL Queries:

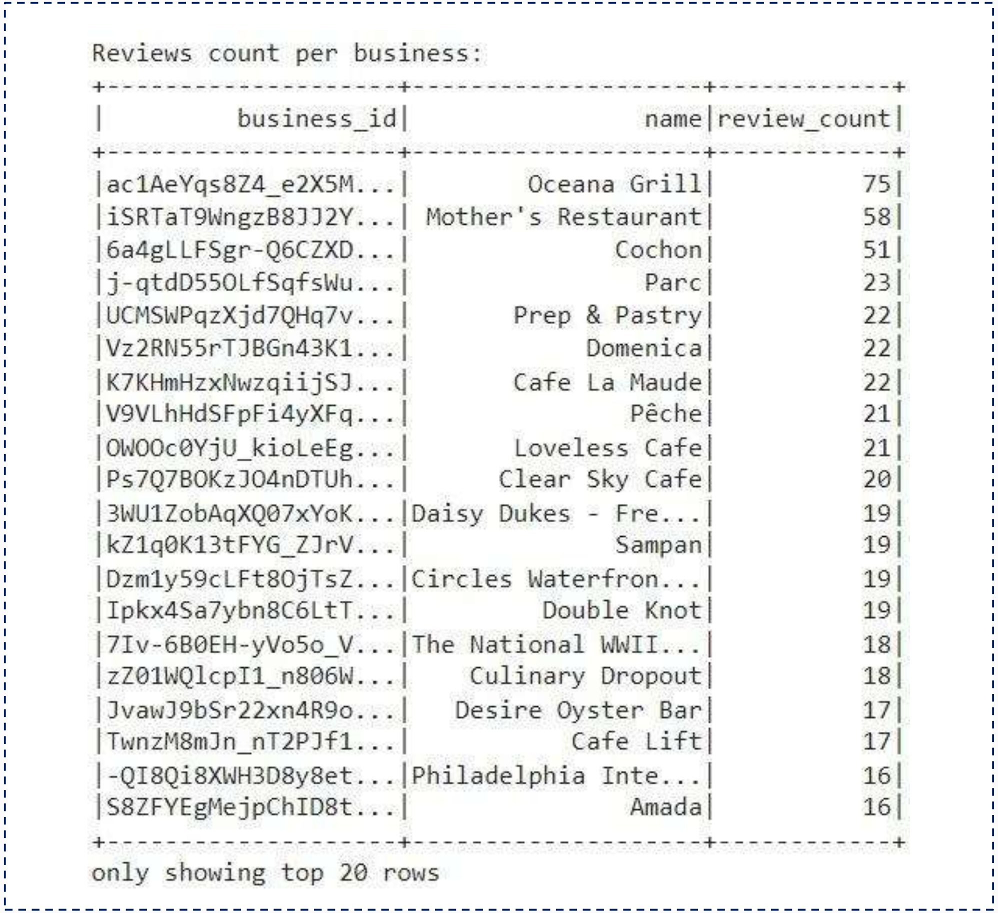
Query to find out top ten Positive Reviews

Word Cloud of Top 1000 Positive Reviews



Sentiment Analysis:

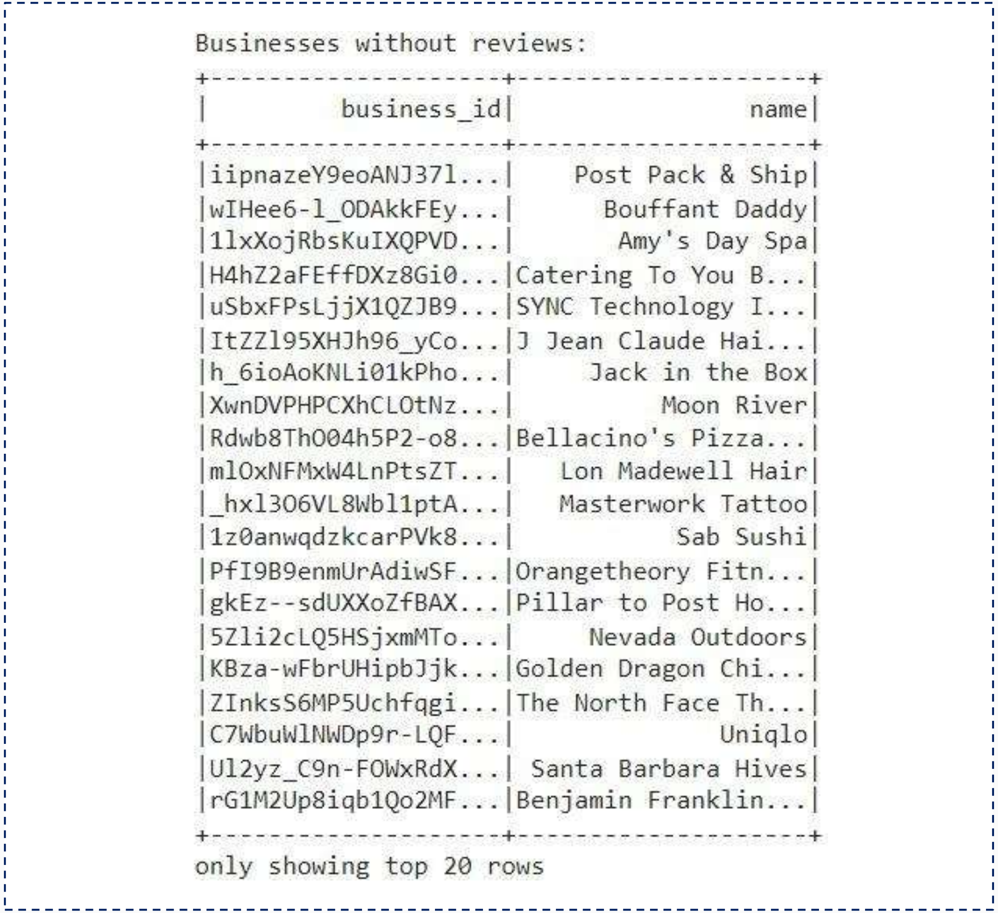
Review Counts per Business



```
Reviews count per business:
+--------------------+--------------------+------------+
|         business_id|                name|review_count|
+--------------------+--------------------+------------+
|ac1AeYqs8Z4_e2X5M...|         Oceana Grill|          75|
|iSRTaT9WngzB8JJ2Y...| Mother's Restaurant|          58|
|6a4gLLFSgr-Q6CZXD...|              Cochon|          51|
|j-qtdD55OLfSqfsWu...|                Parc|          23|
|UCMSWPqzXjd7QHq7v...|       Prep & Pastry|          22|
|Vz2RN55rTJBGn43K1...|            Domenica|          22|
|K7KHmHzxNwzqiijSJ...|        Cafe La Maude|          22|
|V9VLhHdSFpFi4yXFq...|               Pêche|          21|
|OWOOc0YjU_kioLeEg...|        Loveless Cafe|          21|
|Ps7Q7BOKzJO4nDTUh...|      Clear Sky Cafe|          20|
|3WU1ZobAqXQ07xYoK...|Daisy Dukes - Fre...|          19|
|kZ1q0K13tFYG_ZJrV...|              Sampan|          19|
|Dzm1y59cLFt8OjTsZ...|Circles Waterfron...|          19|
|Ipkx4Sa7ybn8C6LtT...|         Double Knot|          19|
|7Iv-6B0EH-yVo5o_V...|The National WWII...|          18|
|zZ01WQlcpI1_n806W...|    Culinary Dropout|          18|
|JvawJ9bSr22xn4R9o...|   Desire Oyster Bar|          17|
|TwnzM8mJn_nT2PJf1...|           Cafe Lift|          17|
|-QI8Qi8XWH3D8y8et...|Philadelphia Inte...|          16|
|S8ZFYEgMejpChID8t...|               Amada|          16|
+--------------------+--------------------+------------+
only showing top 20 rows
```

Sentiment Analysis:

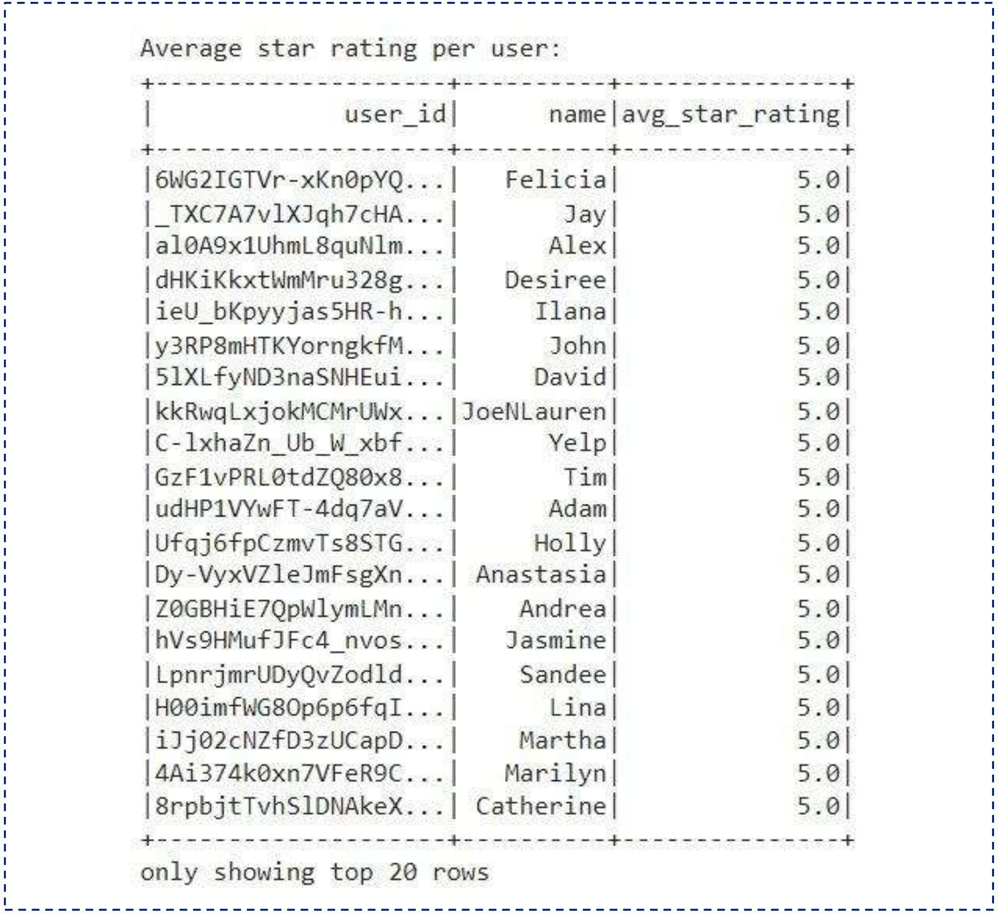
Businesses without Reviews



```
Businesses without reviews:
+--------------------+--------------------+
|         business_id|                name|
+--------------------+--------------------+
|iipnazeY9eoANJ37l...|    Post Pack & Ship|
|wIHee6-l_ODAkkFEy...|      Bouffant Daddy|
|1lxXojRbsKuIXQPVD...|       Amy's Day Spa|
|H4hZ2aFEffDXz8Gi0...|Catering To You B...|
|uSbxFPsLjjX1QZJB9...|SYNC Technology I...|
|ItZZl95XHJh96_yCo...|J Jean Claude Hai...|
|h_6ioAoKNLi01kPho...|     Jack in the Box|
|XwnDVPHPCXhCLOtNz...|          Moon River|
|Rdwb8ThO04h5P2-o8...|Bellacino's Pizza...|
|mlOxNFMxW4LnPtsZT...|   Lon Madewell Hair|
|_hxl3O6VL8Wbl1ptA...|   Masterwork Tattoo|
|1z0anwqdzkcarPVk8...|           Sab Sushi|
|PfI9B9enmUrAdiwSF...|Orangetheory Fitn...|
|gkEz--sdUXXoZfBAX...|Pillar to Post Ho...|
|5Zli2cLQ5HSjxmMTo...|     Nevada Outdoors|
|KBza-wFbrUHipbJjk...|Golden Dragon Chi...|
|ZInksS6MP5Uchfqgi...|The North Face Th...|
|C7WbuWlNWDp9r-LQF...|              Uniqlo|
|Ul2yz_C9n-FOWxRdX...| Santa Barbara Hives|
|rG1M2Up8iqb1Qo2MF...|Benjamin Franklin...|
+--------------------+--------------------+
only showing top 20 rows
```

Sentiment Analysis:

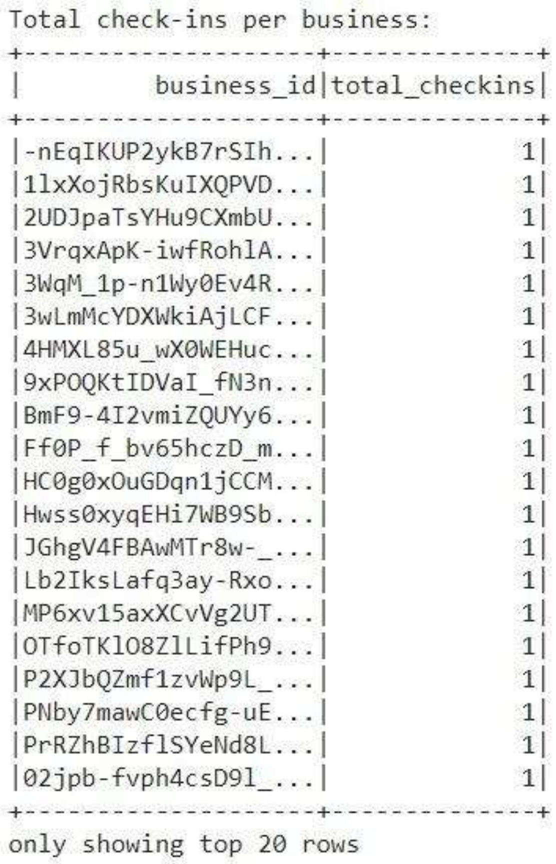
Average Star Rating Per User



```
Average star rating per user:
+--------------------+----------+---------------+
|             user_id|      name|avg_star_rating|
+--------------------+----------+---------------+
|6WG2IGTVr-xKn0pYQ...|   Felicia|            5.0|
|_TXC7A7vlXJqh7cHA...|       Jay|            5.0|
|al0A9x1UhmL8quNlm...|      Alex|            5.0|
|dHKiKkxtWmMru328g...|   Desiree|            5.0|
|ieU_bKpyyjas5HR-h...|     Ilana|            5.0|
|y3RP8mHTKYorngkfM...|      John|            5.0|
|5lXLfyND3naSNHEui...|     David|            5.0|
|kkRwqLxjokMCMrUWx...|JoeNLauren|            5.0|
|C-lxhaZn_Ub_W_xbf...|      Yelp|            5.0|
|GzF1vPRL0tdZQ80x8...|       Tim|            5.0|
|udHP1VYwFT-4dq7aV...|      Adam|            5.0|
|Ufqj6fpCzmvTs8STG...|     Holly|            5.0|
|Dy-VyxVZleJmFsgXn...| Anastasia|            5.0|
|Z0GBHiE7QpWlymLMn...|    Andrea|            5.0|
|hVs9HMufJFc4_nvos...|   Jasmine|            5.0|
|LpnrjmrUDyQvZodld...|    Sandee|            5.0|
|H00imfWG8Op6p6fqI...|      Lina|            5.0|
|iJj02cNZfD3zUCapD...|    Martha|            5.0|
|4Ai374k0xn7VFeR9C...|   Marilyn|            5.0|
|8rpbjtTvhSlDNAkeX...| Catherine|            5.0|
+--------------------+----------+---------------+
only showing top 20 rows
```

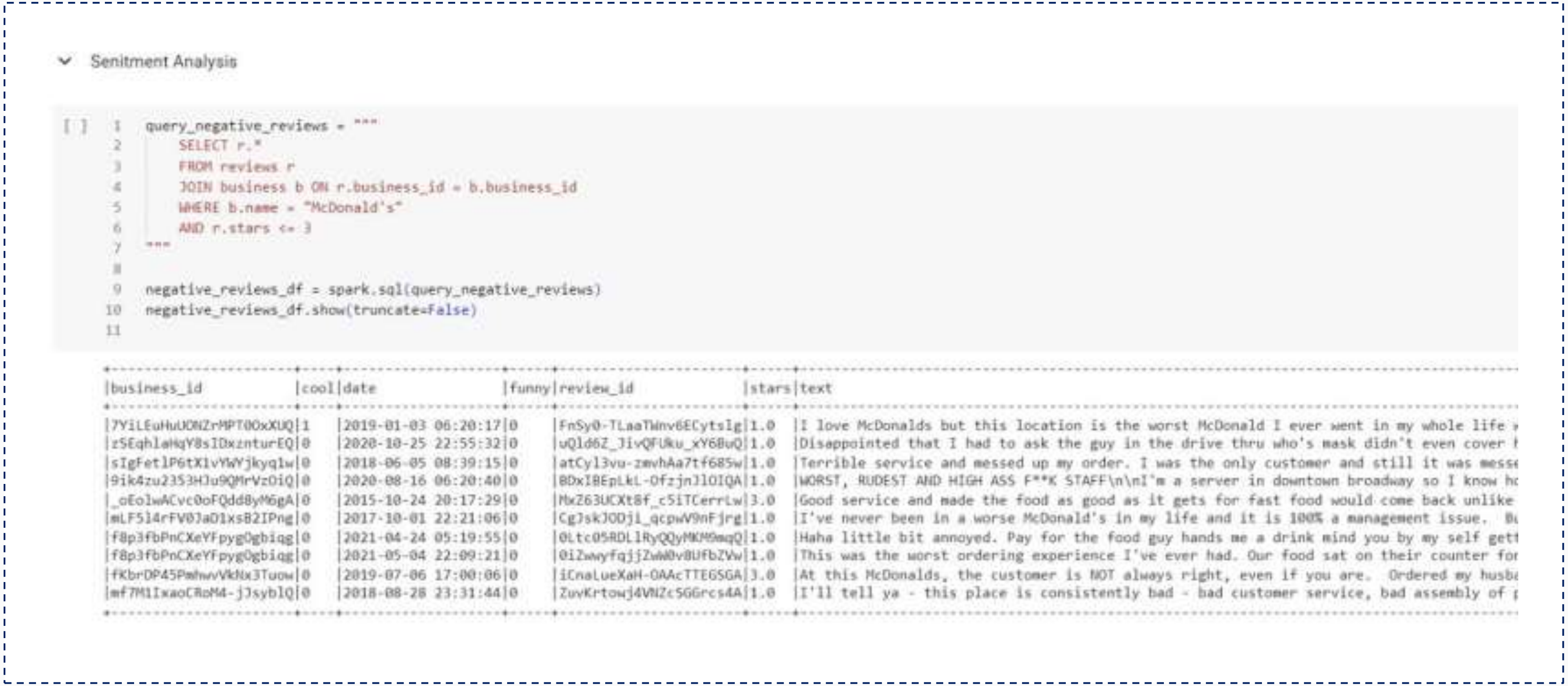
Sentiment Analysis:

Total Check-ins Per Business



```
Total check-ins per business:
+--------------------+--------------+
|         business_id|total_checkins|
+--------------------+--------------+
|-nEqIKUP2ykB7rSIh...|             1|
|1lxXojRbsKuIXQPVD...|             1|
|2UDJpaTsYHu9CXmbU...|             1|
|3VrqxApK-iwfRohlA...|             1|
|3WqM_1p-n1Wy0Ev4R...|             1|
|3wLmMcYDXWkiAjLCF...|             1|
|4HMXL85u_wX0WEHuc...|             1|
|9xPOQKtIDVaI_fN3n...|             1|
|BmF9-4I2vmiZQUYy6...|             1|
|Ff0P_f_bv65hczD_m...|             1|
|HC0g0xOuGDqn1jCCM...|             1|
|Hwss0xyqEHi7WB9Sb...|             1|
|JGhgV4FBAwMTr8w-_...|             1|
|Lb2IksLafq3ay-Rxo...|             1|
|MP6xv15axXCvVg2UT...|             1|
|OTfoTKlO8ZlLifPh9...|             1|
|P2XJbQZmf1zvWp9L_...|             1|
|PNby7mawC0ecfg-uE...|             1|
|PrRZhBIzflSYeNd8L...|             1|
|02jpb-fvph4csD9l_...|             1|
+--------------------+--------------+
only showing top 20 rows
```

Sentiment Analysis

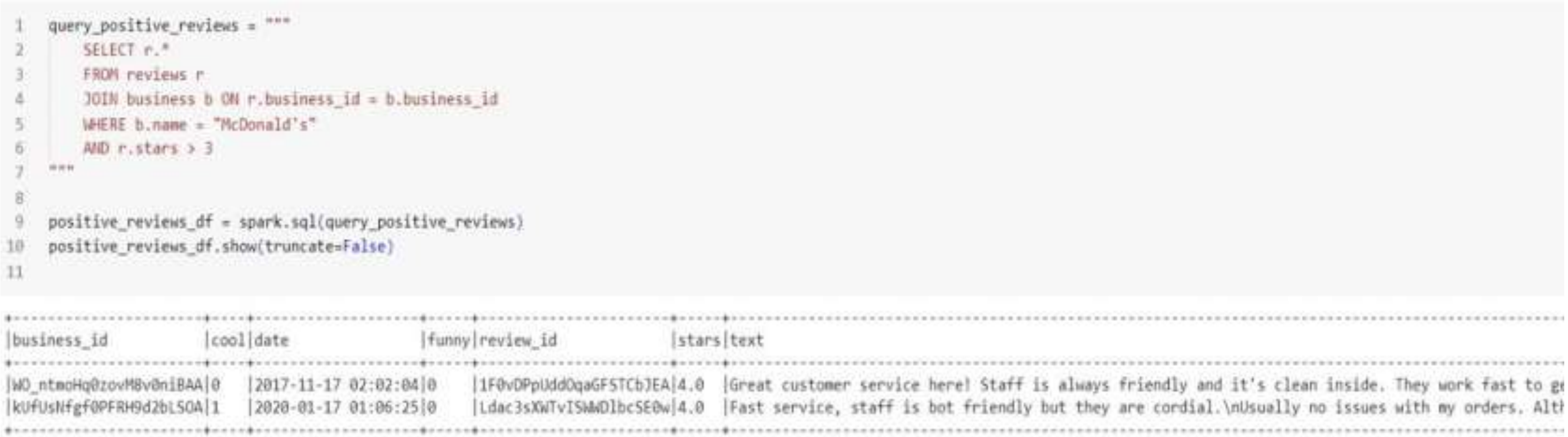


Senitment Analysis

```
[ ]
query_negative_reviews = """
    SELECT r.*
    FROM reviews r
    JOIN business b ON r.business_id = b.business_id
    WHERE b.name = "McDonald's"
    AND r.stars <= 3
"""

negative_reviews_df = spark.sql(query_negative_reviews)
negative_reviews_df.show(truncate=False)

```

```
+----------------------+----+-------------------+-----+----------------------+-----+--------------------------------------------------------------------------------------------
|business_id           |cool|date               |funny|review_id             |stars|text
+----------------------+----+-------------------+-----+----------------------+-----+--------------------------------------------------------------------------------------------
|7YiLEuHuUONZrMPT0OxXUQ|1   |2019-01-03 06:20:17|0    |FnSy0-TLaaTWnv6ECytsIg|1.0  |I love McDonalds but this location is the worst McDonald I ever went in my whole life 
|z5EqhlaHqY8sIDxznturEQ|0   |2020-10-25 22:55:32|0    |uQld6Z_JivQFUku_xY6BuQ|1.0  |Disappointed that I had to ask the guy in the drive thru who's mask didn't even cover 
|sIgFetlP6tX1vYWYjkyq1w|0   |2018-06-05 08:39:15|0    |atCyl3vu-zmvhAa7tf685w|1.0  |Terrible service and messed up my order. I was the only customer and still it was mess
|9ik4zu2353HJu9QMrVzOiQ|0   |2020-08-16 06:20:40|0    |BDxIBEpLkL-OfzjnJlOIQA|1.0  |WORST, RUDEST AND HIGH ASS F**K STAFF\n\nI'm a server in downtown broadway so I know h
|_oEolwACvc0oFQdd8yM6gA|0   |2015-10-24 20:17:29|0    |MxZ63UCXt8f_c5iTCerrLw|3.0  |Good service and made the food as good as it gets for fast food would come back unlike 
|mLF5l4rFV0JaD1xsB2IPng|0   |2017-10-01 22:21:06|0    |CgJskJODji_qcpwV9nFjrg|1.0  |I've never been in a worse McDonald's in my life and it is 100% a management issue.  B
|f8p3fbPnCXeYFpygOgbiqg|0   |2021-04-24 05:19:55|0    |0Ltc05RDLlRyQQyMKM9mqQ|1.0  |Haha little bit annoyed. Pay for the food guy hands me a drink mind you by my self gett
|f8p3fbPnCXeYFpygOgbiqg|0   |2021-05-04 22:09:21|0    |0iZwwyfqjjZwW0v8UfbZVw|1.0  |This was the worst ordering experience I've ever had. Our food sat on their counter for
|fKbrDP45PmhwvVkNx3Tuow|0   |2019-07-06 17:00:06|0    |iCnaLueXaH-OAAcTTEGSGA|3.0  |At this McDonalds, the customer is NOT always right, even if you are.  Ordered my husb
|mf7M1IxaoCRoM4-jJsyblQ|0   |2018-08-28 23:31:44|0    |ZuvKrtowj4VNZc5GGrcs4A|1.0  |I'll tell ya - this place is consistently bad - bad customer service, bad assembly of 
+----------------------+----+-------------------+-----+----------------------+-----+--------------------------------------------------------------------------------------------
```

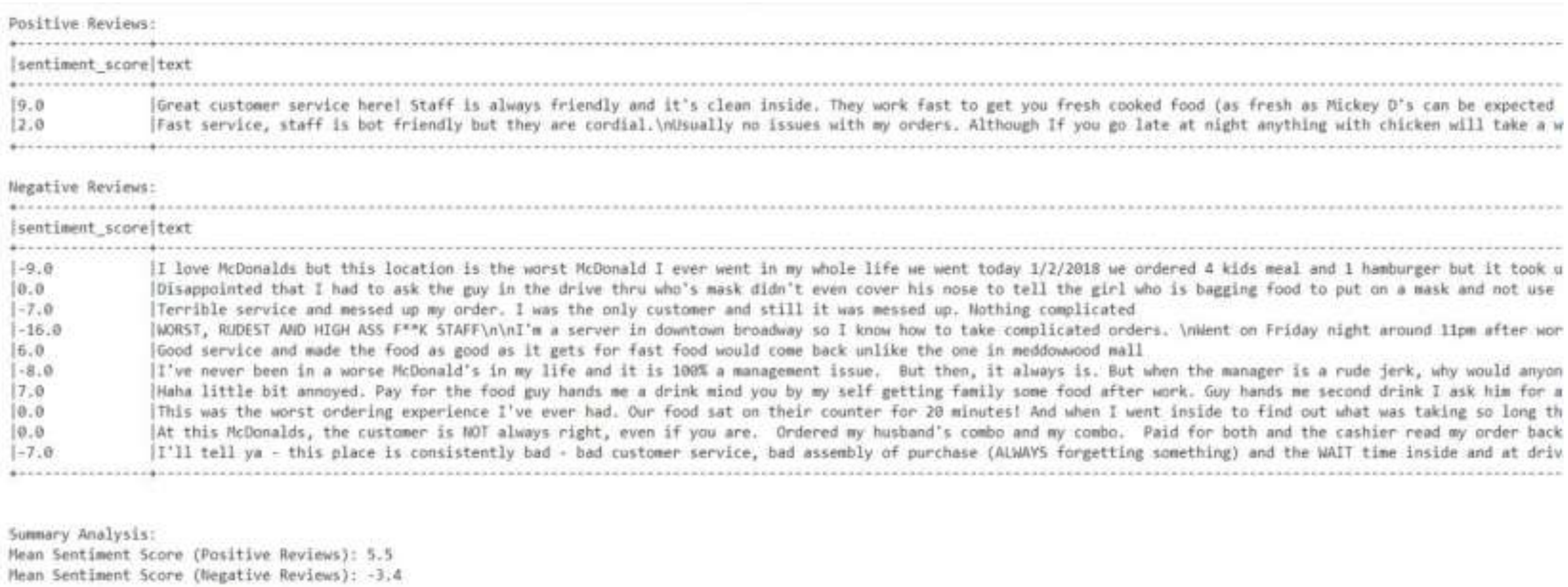
Sentiment Analysis



```
query_positive_reviews = """
    SELECT r.*
    FROM reviews r
    JOIN business b ON r.business_id = b.business_id
    WHERE b.name = "McDonald's"
    AND r.stars > 3
"""

positive_reviews_df = spark.sql(query_positive_reviews)
positive_reviews_df.show(truncate=False)

```

```
+----------------------+----+-------------------+-----+----------------------+-----+-------------------------------------------------------------------------------------------------
|business_id           |cool|date               |funny|review_id             |stars|text
+----------------------+----+-------------------+-----+----------------------+-----+-------------------------------------------------------------------------------------------------
|WO_ntmoHq0zovM8v0niBAA|0   |2017-11-17 02:02:04|0    |1F0vDPpUdd0qaGFSTCbJEA|4.0  |Great customer service here! Staff is always friendly and it's clean inside. They work fast to ge
|kUfUsNfgf0PFRH9d2bLSOA|1   |2020-01-17 01:06:25|0    |Ldac3sXWTvISWWDlbcSE0w|4.0  |Fast service, staff is bot friendly but they are cordial.\nUsually no issues with my orders. Alth
+----------------------+----+-------------------+-----+----------------------+-----+-------------------------------------------------------------------------------------------------
```

Sentiment Analysis



```
Positive Reviews:
+---------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|sentiment_score|text
+---------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|9.0            |Great customer service here! Staff is always friendly and it's clean inside. They work fast to get you fresh cooked food (as fresh as Mickey D's can be expected
|2.0            |Fast service, staff is bot friendly but they are cordial.\nUsually no issues with my orders. Although If you go late at night anything with chicken will take a w
+---------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Negative Reviews:
+---------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|sentiment_score|text
+---------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
|-9.0           |I love McDonalds but this location is the worst McDonald I ever went in my whole life we went today 1/2/2018 we ordered 4 kids meal and 1 hamburger but it took u
|0.0            |Disappointed that I had to ask the guy in the drive thru who's mask didn't even cover his nose to tell the girl who is bagging food to put on a mask and not use
|-7.0           |Terrible service and messed up my order. I was the only customer and still it was messed up. Nothing complicated
|-16.0          |WORST, RUDEST AND HIGH ASS F**K STAFF\n\nI'm a server in downtown broadway so I know how to take complicated orders. \nWent on Friday night around 11pm after wor
|6.0            |Good service and made the food as good as it gets for fast food would come back unlike the one in meddowwood mall
|-8.0           |I've never been in a worse McDonald's in my life and it is 100% a management issue.  But then, it always is. But when the manager is a rude jerk, why would anyon
|7.0            |Haha little bit annoyed. Pay for the food guy hands me a drink mind you by my self getting family some food after work. Guy hands me second drink I ask him for a
|0.0            |This was the worst ordering experience I've ever had. Our food sat on their counter for 20 minutes! And when I went inside to find out what was taking so long th
|0.0            |At this McDonalds, the customer is NOT always right, even if you are.  Ordered my husband's combo and my combo.  Paid for both and the cashier read my order back
|-7.0           |I'll tell ya - this place is consistently bad - bad customer service, bad assembly of purchase (ALWAYS forgetting something) and the WAIT time inside and at driv
+---------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------


Summary Analysis:
Mean Sentiment Score (Positive Reviews): 5.5
Mean Sentiment Score (Negative Reviews): -3.4
```

Conclusion

- The code effectively uses PySpark to analyze Yelp dataset, efficiently managing large datasets with its distributed computing capabilities.
- By employing SQL queries on Spark DataFrames, the code extracts key insights like business distribution across states and top business categories by count, , and correlation between user engagement metrics and review ratings.
- Visualizations such as bar charts, scatter plots, and word clouds present data intuitively, aiding trend identification for deeper analysis.
- Conducting sentiment analysis on reviews provides nuanced understanding of customer opinions.
- Overall, the code illustrates PySpark's versatility and scalability for large-scale data analytics, showcasing its effectiveness in deriving actionable insights from extensive datasets like Yelp.

Thank You