

**Data Technology Solutions**  
**(DSMM Group 2) – 2023**



**Yelp Project Report**

**Submitted To :**

Bhavik Gandhi

**Submitted By :**

Likhitha Jayanthi (C0891732)

Anisha Susan Mathew (C0907393)

Mahmood Hossain (C0896079)

Simarjeet kaur (C0894309)

Arun Kumar Subramaniam (C0891553)

Parminder kaur (C0908143)

Shanmuga Priyan Jeevanandam (C0889053)

## **Contents of Report**

- Introduction
- Objectives
- Problem Definition
- Analysis :
  - To calculate the average review count for businesses in each category using pySpark
  - To calculate the average review count for businesses in each state using pySpark
  - To find the Top 5 Business with most number of 5 Star rating using pySpark
  - To find Is having more varied business gets them more reviews on yelp necessarily using pyspark
  - To calculate the total review count per state for businesses that are marked as open using Mapreduce
- Visualizations :
  - Average Star Rating for Business in each city
  - Popular business categories and the number of businesses in that category
  - States and their average review count
  - complement count of each categories
  - Review count trend over the years as a line chart
  - Highest review counts and observe trends and patterns in the data
  - How review counts vary based on geographical location
- Prescription For Yelp
- Conclusion
- MECE Table
- Git Repo Link

## **Introduction**

Yelp is a user-generated review platform that operates in the local business review and recommendations industry. Its primary focus is on providing a platform for users to discover and review local businesses, while also offering businesses an opportunity to engage with their customers and potentially attract new ones.

## **Objective of the Project**

Gaining insights from Yelp database using Big data technologies and prescribe improvements from the insights gained throughout the project.

## **Problem Definition**

Finding out strategic solutions for leveraging Yelp's community and user engagement for upscaling their business.

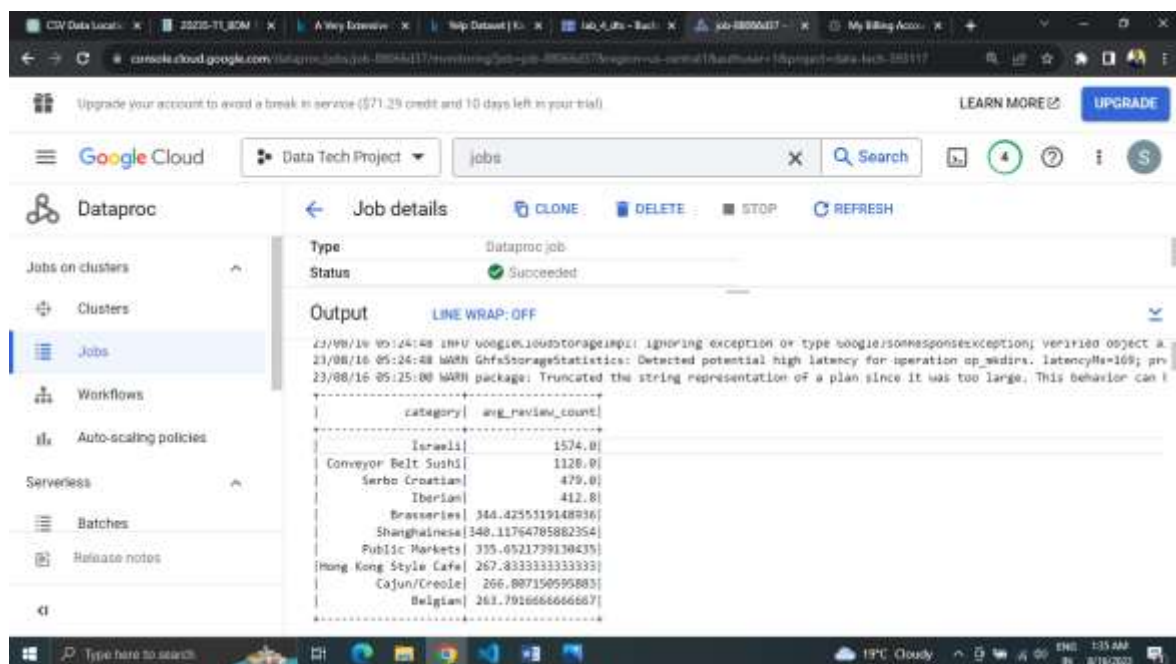
# Analysis

**Spark Job 1 :** To calculate the average review count for businesses in each category

**Insights :** From the output of this PySpark job we can infer the following

1. **Popular Categories:** The top categories with the highest average review counts are likely to be the most popular or well received by customers. These categories are getting more attention and positive reviews from customers.
2. **Customer Satisfaction:** Categories with higher average review counts suggest that businesses in those categories are providing good products or services that customers are satisfied with. It reflects positively on the quality and customer experience.
3. **Demand and Engagement:** Categories with higher average review counts might indicate higher demand and engagement from customers. Businesses in these categories might be attracting more customers and generating more reviews.
4. **Potential Business Focus:** Businesses looking to succeed could consider entering or expanding into categories that have higher average review counts. It could be an indicator of potential success in terms of customer satisfaction and engagement.
5. **Areas of Improvement:** On the flip side, categories with lower average review counts might indicate areas where businesses could improve. Lower review counts might suggest less customer engagement or potential areas for enhancement.
6. **Market Trends:** Analyzing popular categories can provide insights into current market trends and consumer preferences. Businesses can adapt their strategies to align with these trends.

## Output Screenshot :

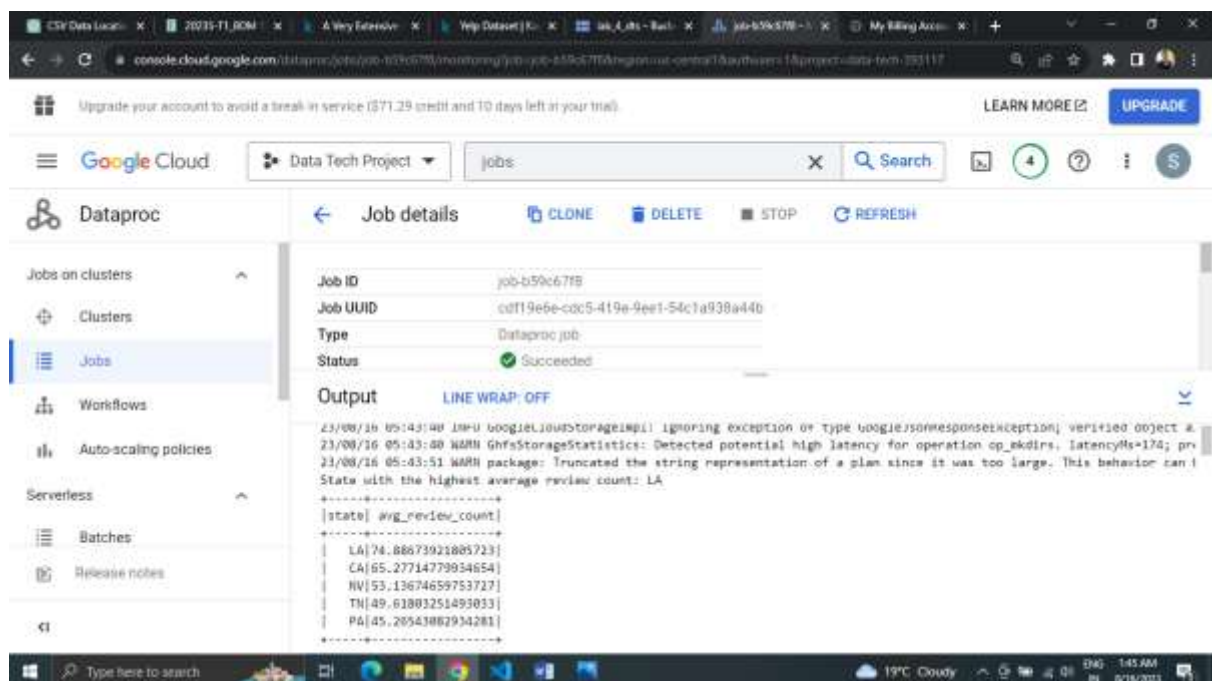


**Spark Job 2 :** To calculate the average review count for businesses in each state

**Insights :** From the output of this PySpark job, we can infer the following

1. **State with the Highest Average Review Count:** The code identifies the state with the highest average review count among businesses. This state is likely to have businesses that are popular and receive a higher number of reviews on average.
2. **Customer Engagement and Satisfaction:** States with higher average review counts suggest that businesses in those states are able to engage customers effectively and provide satisfactory products or services.
3. **Positive Consumer Perception:** States with high average review counts could indicate that businesses in those areas are well regarded by their customers, leading to positive word-of-mouth and higher engagement.
4. **Business Opportunities:** Entrepreneurs or investors might consider exploring business opportunities in the states with high average review counts. These states could have a favorable environment for businesses to thrive.
5. **Regional Preferences:** The analysis offers insights into regional preferences for certain types of businesses. Higher average review counts might suggest that customers in those states are more active in reviewing and engaging with businesses.

### Output Screenshot :



**Spark Job 3 :** To find the Top 5 Business with most number of 5 Star rating

**Insights :** From the output of the PySpark job, we can infer the following

1. **Quality and Popularity:** The displayed restaurants have received 5-star ratings and boast a high number of reviews. This indicates both exceptional quality and strong popularity among customers.
2. **Customer Satisfaction:** The combination of high ratings and numerous reviews suggests that these restaurants consistently deliver positive experiences, making customers more likely to share their satisfaction.
3. **Engagement Matters:** The higher review counts among these top-rated restaurants highlight active customer engagement, demonstrating that they not only attract customers but also encourage them to leave reviews.
4. **Customer-Centric Approach:** The success of these restaurants is rooted in prioritizing customer preferences and needs, leading to both high ratings and a substantial customer base.
5. **Market Leadership:** These top-rated restaurants stand out in the competitive market due to their quality and customer engagement, making them potential market leaders that others in the industry can look up to.

**Output Screenshot :**

The screenshot shows the Google Cloud Dataproc console. The job details for 'job-ca417f30' are displayed, showing it has succeeded. The output section shows a table of restaurant data with columns for business\_id, name, and review\_count.

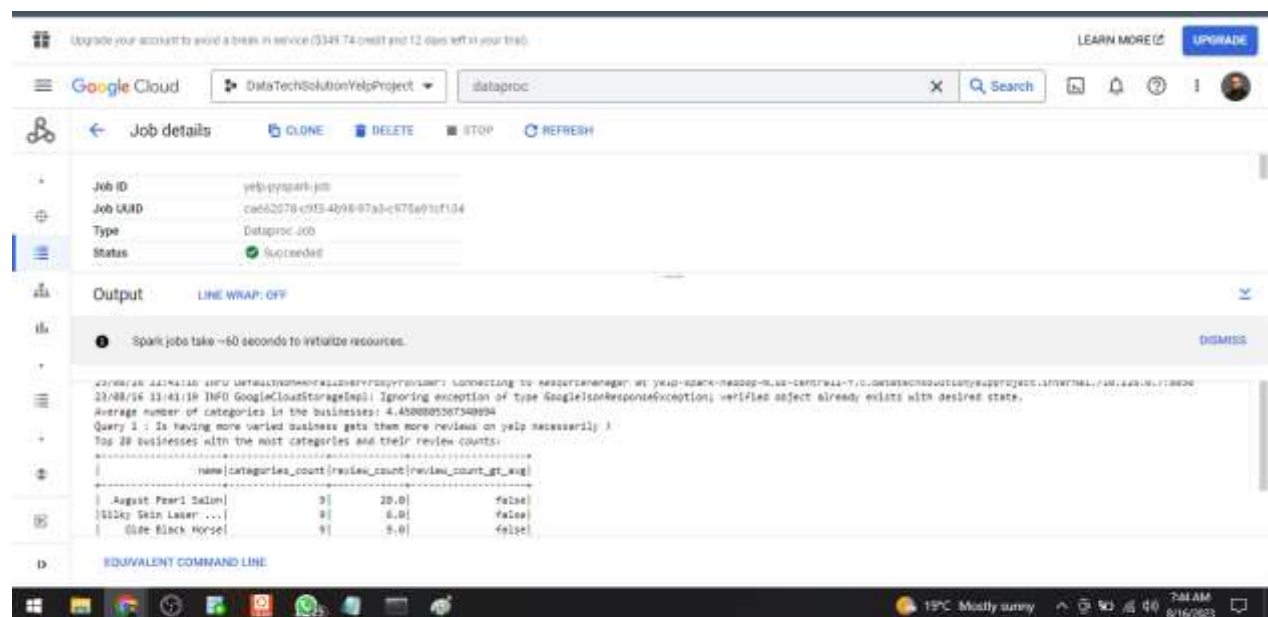
business_id	name	review_count
_akr7P0nacki_VisRK...	Blues City Delli	991
l8QnRgM-QvGsjOWu...	Carillitas Cocina	799
lzxIF-bna7-eKlscn8...	Free Tours By Foot	769
DVB3KvncpkaVlnrH...	Tamerico	705
lgP_onJykA2RocIs_G...	Yata	623

**Spark Job 4 :** To find Is having more varied business gets them more reviews on yelp necessarily

**Insights :** From the output of the PySpark job, we can infer the following

1. **Diverse Categories, Varied Reviews:** Business variety doesn't guarantee more reviews; quality and engagement within categories matter for user attention.
2. **Reviews and Ratings Link:** More reviews tend to align with slightly better ratings, emphasizing the impact of user feedback on business reputation.
3. **Night Check-Ins, Neutral Ratings:** Nighttime activity doesn't strongly influence ratings; user reviews appear consistent across different hours.
4. **Elite Users, Moderate Bonds:** Being elite correlates moderately with more friends, hinting at social connections, but not a defining factor.
5. **Recommendation Refinement:** Insights can fuel smarter recommendations, connecting users with businesses based on their review patterns.
6. **User Engagement Enhancement:** Data-driven decisions empower Yelp to optimize user experiences, driving platform improvements and informed strategies.

### **Output Screenshots :**



The screenshot shows the Google Cloud DataProc console for a job named 'dataproc'. The job status is 'Succeeded'. The output is displayed as a table with the following columns: 'name', 'categories\_count', 'review\_count', and 'review\_count\_gt\_avg'.

name	categories_count	review_count	review_count_gt_avg
August Pearl Salon	9	20.0	false
Sticky Skin Laser ...	8	6.0	false
Old Black Horse	9	5.0	false

Upgrade your account to avoid a break in service (\$349.74 credit and 12 days left in your trial). [LEARN MORE](#) [UPGRADE](#)

Google Cloud DataTechSolutionYelpProject dataproc

Output [LIVE WRAP](#) [CLONE](#) [DELETE](#) [STOP](#) [REFRESH](#)

❗ [Dismiss](#) **Spark jobs take ~60 seconds to initialize resources.**

Query 1 : An interesting, well-researched analysis gives users more context on Yelp businesses and is

Top 30 businesses with the most categories and their review counts:

name	categories_count	review_count	review_count_gt_avg
August Pearl Saloon	9	28.0	false
Elleby Skin Laser ...	9	6.0	false
Blue Black Horse	9	9.0	false
Tap Shack Sports ...	9	95.0	true
Revo Downtown Joint	9	18.0	false
David Thomas Thrill...	9	6.0	false
DOSE	9	87.0	true
Enjoy The Mountain	9	48.0	true
Paradise Brasserie	9	20.0	true
we'll make it	9	41.0	false
Drinks Craft House	9	115.0	true
The Ladies Room	9	13.0	false
Wandering Charters...	9	30.0	false
Waxed Cyber Cafe ...	9	11.0	false
ReJoy Medspa	9	16.0	false
Daxian Rug Cleaning	9	20.0	false
Double Decker Piz...	9	88.0	true

[EQUIVALENT COMMAND LINE](#)

Windows taskbar: 19°C Mostly sunny, 7:45 AM 8/16/2023

Upgrade your account to avoid a break in service (\$349.74 credit and 12 days left in your trial). [LEARN MORE](#) [UPGRADE](#)

Google Cloud DataTechSolutionYelpProject dataproc

Output [LIVE WRAP](#) [CLONE](#) [DELETE](#) [STOP](#) [REFRESH](#)

❗ [Dismiss](#) **Spark jobs take ~60 seconds to initialize resources.**

ReJoy Medspa	9	16.0	false
Daxian Rug Cleaning	9	20.0	false
Double Decker Piz...	9	88.0	true
Brooms & Buckets	9	40.0	false
CD Rome Restaurant	9	125.0	true
Beef 'O' Brady's	9	46.0	true

So it seems that having a lot of categories does not necessarily gain more reviews.

Query 2 : Average review ratings for businesses with different review counts and businesses with more check-ins at night.

Average review rating for businesses with high review counts: 3.8

Average review rating for businesses with low review counts: 3.6

Average review rating for businesses with more night check-ins: 3.6

Query 3 : Is there any significant correlation between Elite Count and number of friends ?

Correlation between elite\_count and friends\_count: 0.33

No significant effect on number of friends based on Elite count

Output is complete

[EQUIVALENT COMMAND LINE](#)

Windows taskbar: 19°C Mostly sunny, 7:46 AM 8/16/2023

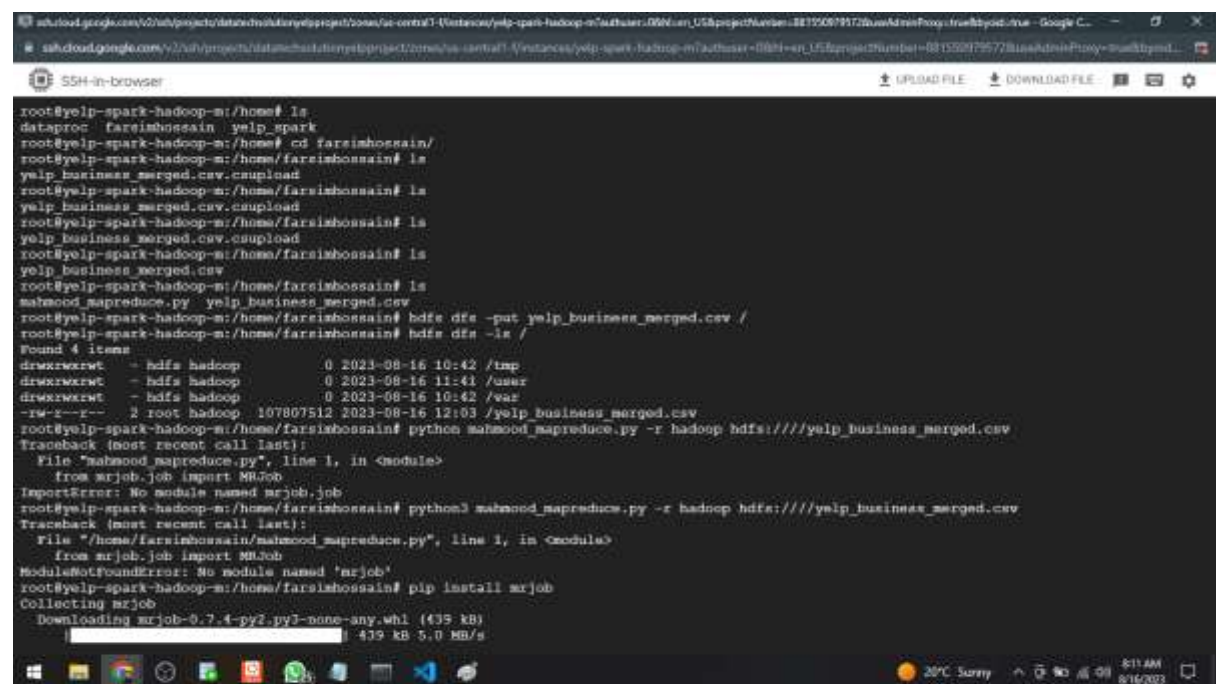


**Map Reduce Job :** To calculate the total review count per state for businesses that are marked as open

**Insights :** From the output of the MapReduce job, we can infer the following

1. **User Engagement Hotspots:** Identify states with high review counts to focus user engagement efforts and optimize platform features where activity is robust.
2. **Business Popularity Spotlight:** Highlight states with elevated review counts to showcase thriving businesses and attract more users seeking popular venues.
3. **Adaptability and Resilience:** Review counts reflect business resilience; compare states to gauge adaptability during challenges and changing market conditions.
4. **Localized Business Tailoring:** Understand regional trends from review data to empower businesses with insights for refining offerings to match local preferences.
5. **Behavior-Powered Personalization:** Review patterns expose user behaviors by state, fueling tailored recommendations and enhancing engagement strategies.
6. **Quality and Operational Insights:** Higher review counts suggest better service and operational efficiency, driving Yelp's understanding of business success across states.

### Output Screenshots :

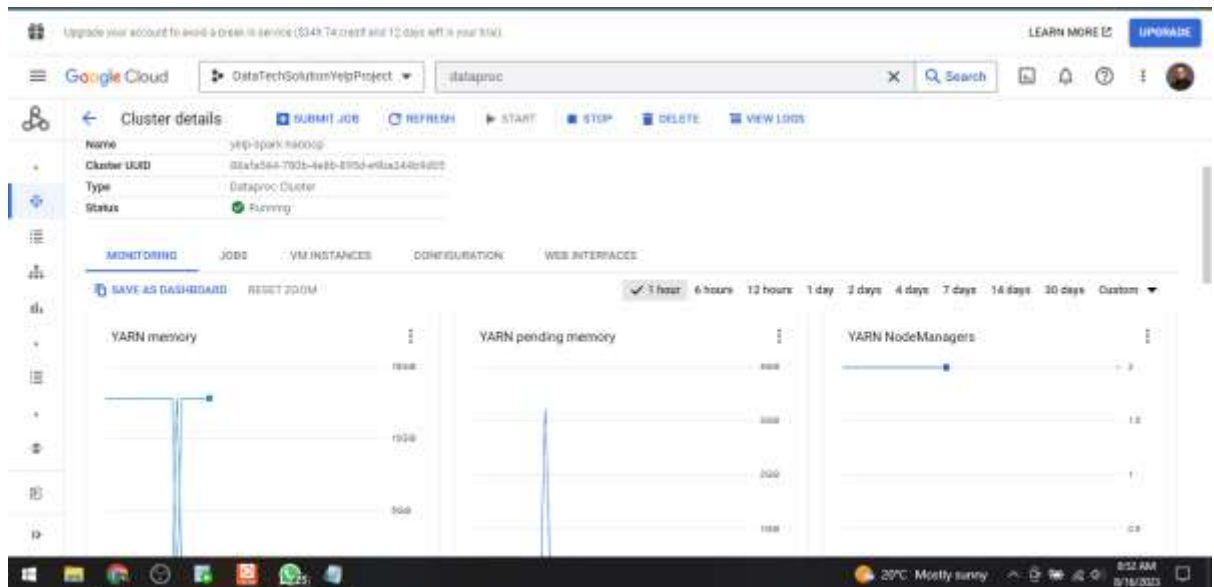


```
root@yelp-spark-hadoop-m:/home# ls
dataproc farsimhossain yelp_spark
root@yelp-spark-hadoop-m:/home# cd farsimhossain/
root@yelp-spark-hadoop-m:/home/farsimhossain# ls
yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# ls
yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# ls
yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# ls
yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# ls
yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# ls
mahmood_mapreduce.py yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# hdfs dfs -put yelp_business_merged.csv /
root@yelp-spark-hadoop-m:/home/farsimhossain# hdfs dfs -ls /
Found 4 items
drwxrwxrwt - hdfs hadoop 0 2023-08-16 10:42 /tmp
drwxrwxrwt - hdfs hadoop 0 2023-08-16 11:41 /user
drwxrwxrwt - hdfs hadoop 0 2023-08-16 10:42 /var
-rw-r--r-- 1 root hadoop 107807512 2023-08-16 12:03 /yelp_business_merged.csv
root@yelp-spark-hadoop-m:/home/farsimhossain# python mahmood_mapreduce.py -f hdfs:///yelp_business_merged.csv
Traceback (most recent call last):
  File "mahmood_mapreduce.py", line 1, in <module>
    from mrjob.job import MRJob
ImportError: No module named mrjob.job
root@yelp-spark-hadoop-m:/home/farsimhossain# python3 mahmood_mapreduce.py -f hdfs:///yelp_business_merged.csv
Traceback (most recent call last):
  File "mahmood_mapreduce.py", line 1, in <module>
    from mrjob.job import MRJob
ModuleNotFoundError: No module named 'mrjob'
root@yelp-spark-hadoop-m:/home/farsimhossain# pip install mrjob
Collecting mrjob
  Downloading mrjob-0.7.4-py2.py3-none-any.whl (439 kB)
    | 439 kB 5.0 MB/s
```

```
Collecting PyYAML>=3.10
  Downloading PyYAML-6.0.1-cp39-cp39-manylinux_2_17_x86_64-manylinux2014_x86_64.whl (735 kB)
    | 735 kB 56.0 MB/s
Installing collected packages: PyYAML, mrjob
Successfully installed PyYAML-6.0.1 mrjob-0.7.4
root@yelp-spark-hadoop-m:~/home/farishossain# python mahmood_mapreduce.py -r hadoop hdfs:///yelp_business_merged.csv
Traceback (most recent call last):
  File "mahmood_mapreduce.py", line 1, in <module>
    from mrjob.job import MRJob
ImportError: No module named mrjob.job
root@yelp-spark-hadoop-m:~/home/farishossain# pip3 install mrjob
Requirement already satisfied: mrjob in /usr/local/lib/python3.9/dist-packages (0.7.4)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.9/dist-packages (from mrjob) (6.0.1)
root@yelp-spark-hadoop-m:~/home/farishossain# python3 mahmood_mapreduce.py -r hadoop hdfs:///yelp_business_merged.csv
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /usr/lib/hadoop/bin...
Found hadoop binary: /usr/lib/hadoop/bin/hadoop
Using Hadoop version 3.3.3
Looking for Hadoop streaming jar in /usr/lib/hadoop...
Found Hadoop streaming jar: /usr/lib/hadoop/hadoop-streaming.jar
Creating temp directory /tmp/mahmood_mapreduce.root.20230816.120623.830245
uploading working dir files to hdfs:///user/root/tmp/mrjob/mahmood_mapreduce.root.20230816.120623.830245/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mahmood_mapreduce.root.20230816.120623.830245/files/
Running step 1 of 1...
packageJobJar: [ [ /usr/lib/hadoop/hadoop-streaming-3.3.3.jar] /tmp/streamjob9916585294078147535.jar tmpDir=null
Connecting to ResourceManager at yelp-spark-hadoop-m.us-central1-f.c.datatechsolutionyelpproject.internal./10.128.0.7:8032
Connecting to Application history server at yelp-spark-hadoop-m.us-central1-f.c.datatechsolutionyelpproject.internal./10.128.0.7:10200
Connecting to ResourceManager at yelp-spark-hadoop-m.us-central1-f.c.datatechsolutionyelpproject.internal./10.128.0.7:8032
Connecting to Application history server at yelp-spark-hadoop-m.us-central1-f.c.datatechsolutionyelpproject.internal./10.128.0.7:10200
Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1692182522502_0002
Total input files to process : 1
number of splits:9
```

```
number of splits:9
Submitting tokens for job: job_1692182522502_0002
Executing with tokens: []
resource-types.xml not found
Unable to find 'resource-types.xml'.
Submitted application application_1692182522502_0002
The url to track this job: http://yelp-spark-hadoop-m.us-central1-f.c.datatechsolutionyelpproject.internal.:8088/proxy/application_1692182522502_0002/
Running job: job_1692182522502_0002
Job job_1692182522502_0002 running in uber mode : false
  map 0% reduce 0%
  map 11% reduce 0%
  map 22% reduce 0%
  map 33% reduce 0%
  map 44% reduce 0%
  map 56% reduce 0%
  map 67% reduce 0%
  map 78% reduce 0%
  map 89% reduce 0%
  map 100% reduce 0%
  map 100% reduce 33%
  map 100% reduce 67%
  map 100% reduce 100%
Job job_1692182522502_0002 completed successfully
Output directory: hdfs:///user/root/tmp/mrjob/mahmood_mapreduce.root.20230816.120623.830245/output
Counters: 56
  File Input Format Counters
    Bytes Read=107840280
  File Output Format Counters
    Bytes Written=244
  File System Counters
    FILE: Number of bytes read=1168624
    FILE: Number of bytes written=5837423
    FILE: Number of large read operations=0
```

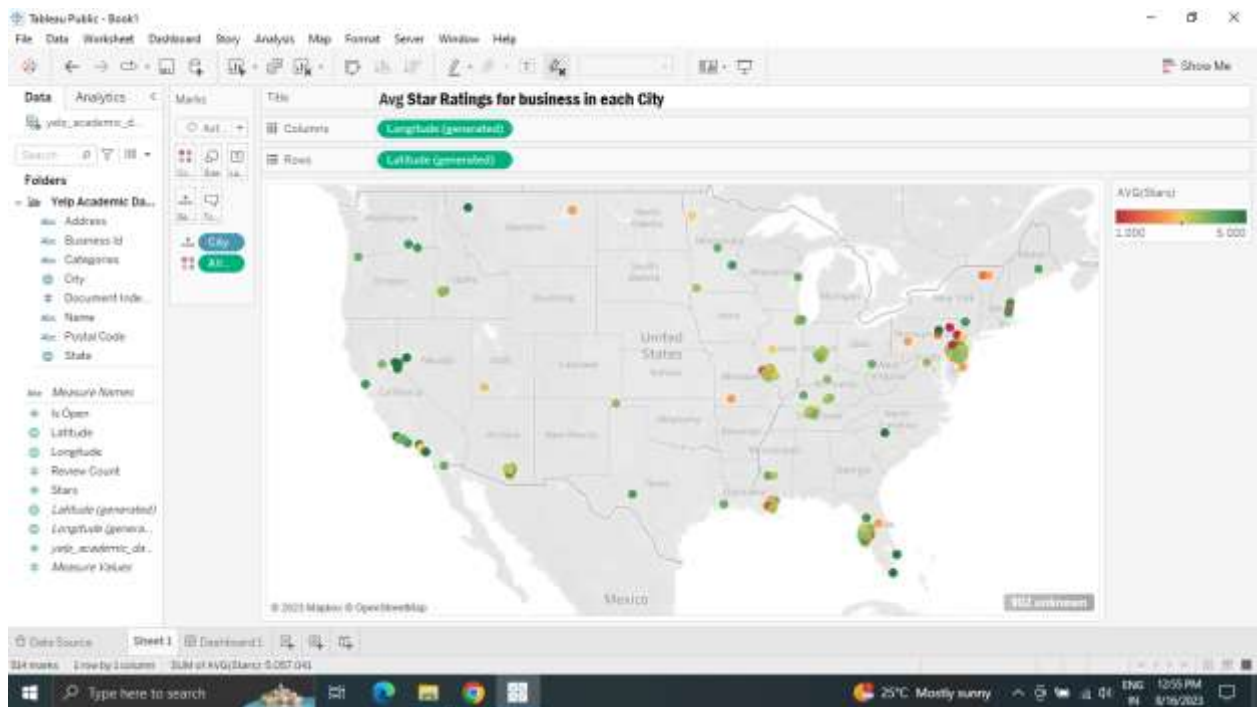
```
ssh-doud.google.com/V2/vf/projects/datatech-solutions/yelp-project/zone/us-central1-f/instances/yelp-spark-hadoop-m/author=08h-en_US&projectNumber=881550979172&useAdminProxy=true - Google C...
ssh-doud.google.com/V2/vf/projects/datatech-solutions/yelp-project/zone/us-central1-f/instances/yelp-spark-hadoop-m/author=08h-en_US&projectNumber=881550979172&useAdminProxy=true...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
CONNECTION=0
IO_ERROR=0
MRJOB_LEN7TH=0
MRJOB_MAP=0
MRJOB_REDUCE=0
job output is in hdfs:///user/root/tmp/mrjob/mahmood_mapreduce.root.20230816.120623.830245/output
Streaming final output from hdfs:///user/root/tmp/mrjob/mahmood_mapreduce.root.20230816.120623.830245/output...
"CA" 280184
"ID" 135163
"JA" 627787
"MI" 9
"MO" 209309
"TW" 503329
"RI" 17
"TR" 386199
"MA" 43
"DA" 1221704
"SD" 42
"TX" 33
"UP" 15
"AB" 85223
"AD" 348555
"CO" 8
"DR" 57308
"PL" 950722
"TL" 42307
"BJ" 211714
"NV" 347868
"VS" 10
"WA" 15
"XMR" 5
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mahmood_mapreduce.root.20230816.120623.830245...
Removing temp directory /tmp/mahmood_mapreduce.root.20230816.120623.830245...
root@yelp-spark-hadoop-m/home/farsinhossain#
```



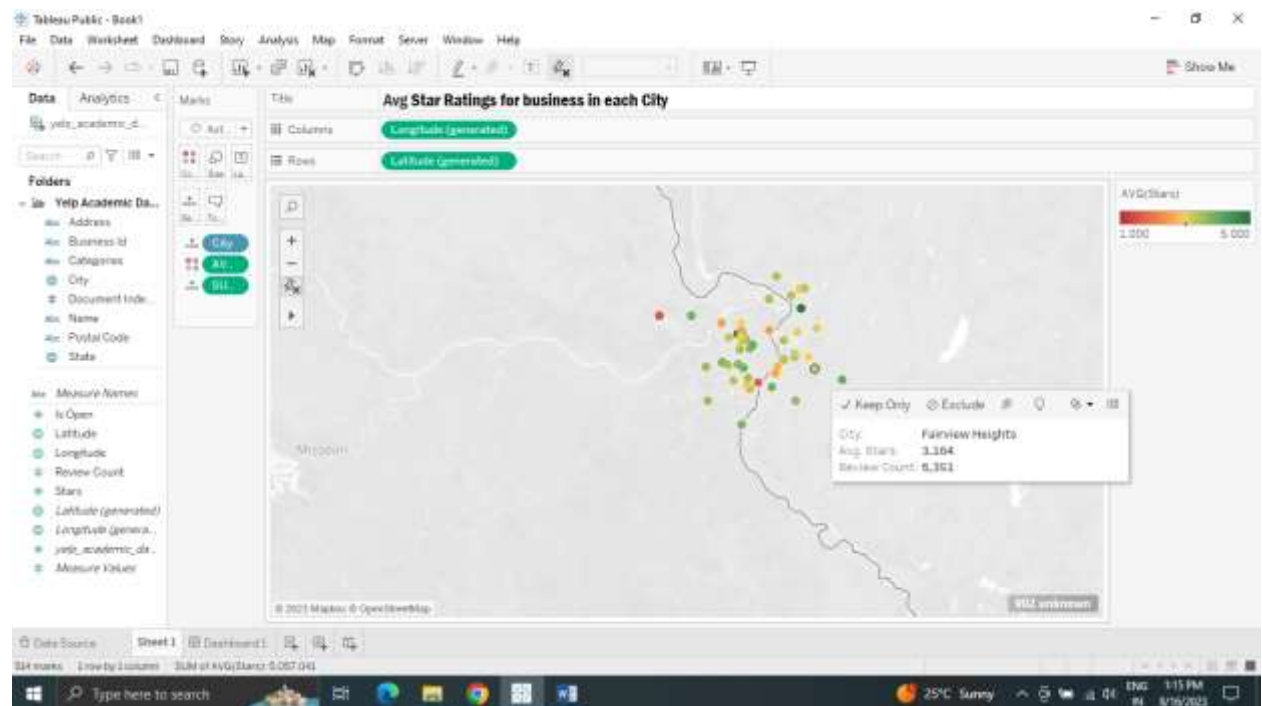
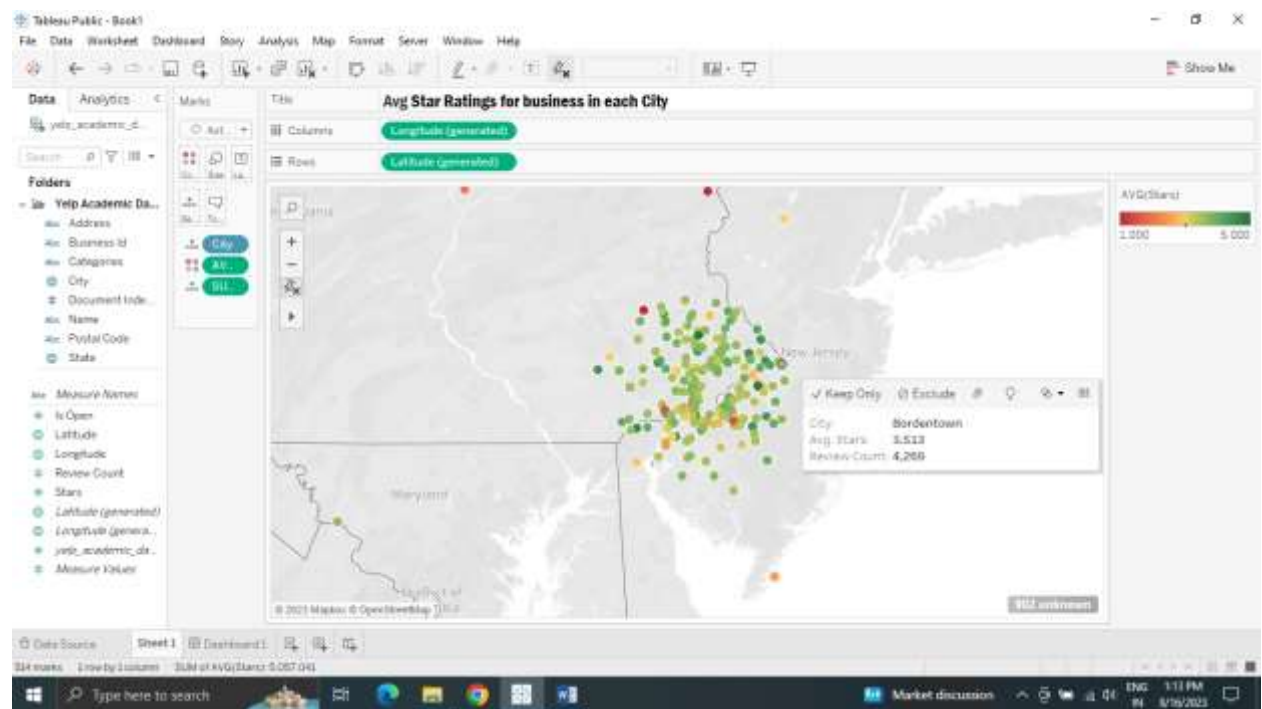
# Visualizations

1). Average Star Rating for Business in each city:

**Geographic Business Distribution:** The map visualizes the geographical distribution of businesses, helping us identify regions with high and low business concentrations. This insight aids in understanding the overall economic landscape and potential market opportunities.

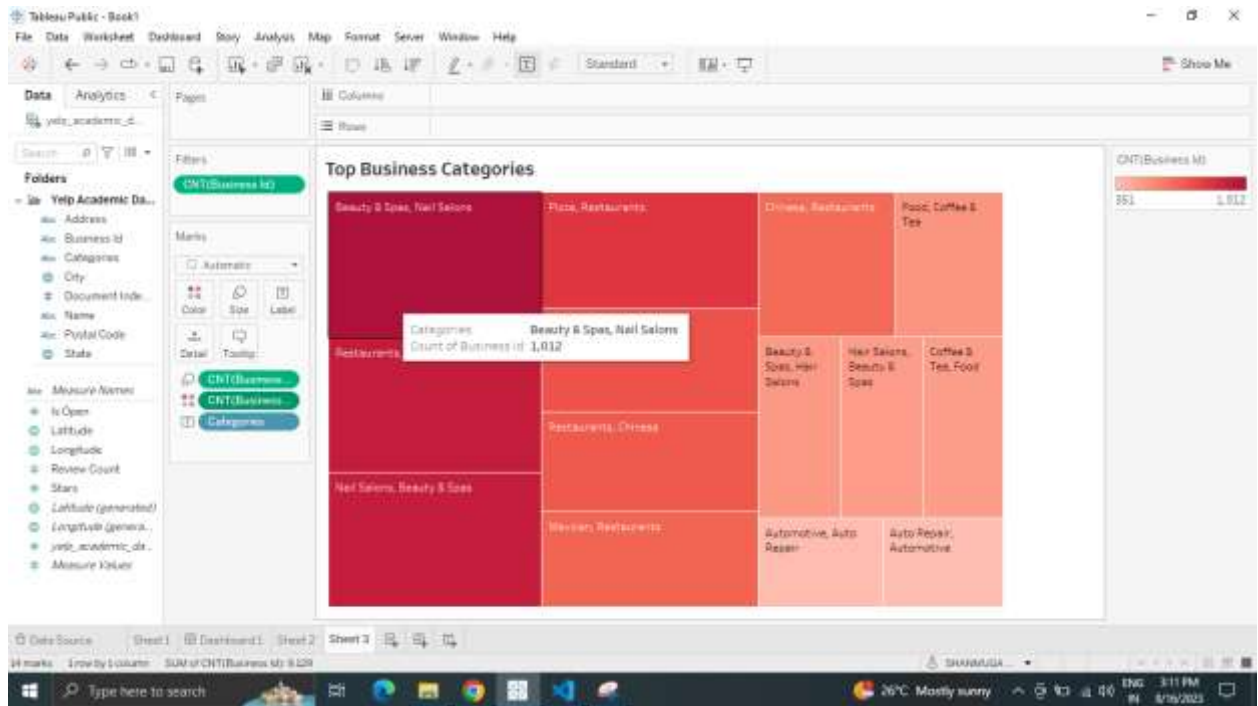


**Rating Analysis through Color Gradients:** By using color gradients for star ratings, you can quickly assess businesses' overall ratings. The color spectrum highlights exceptional and poorly rated businesses, allowing you to make informed decisions about where to dine, shop, or engage in services.

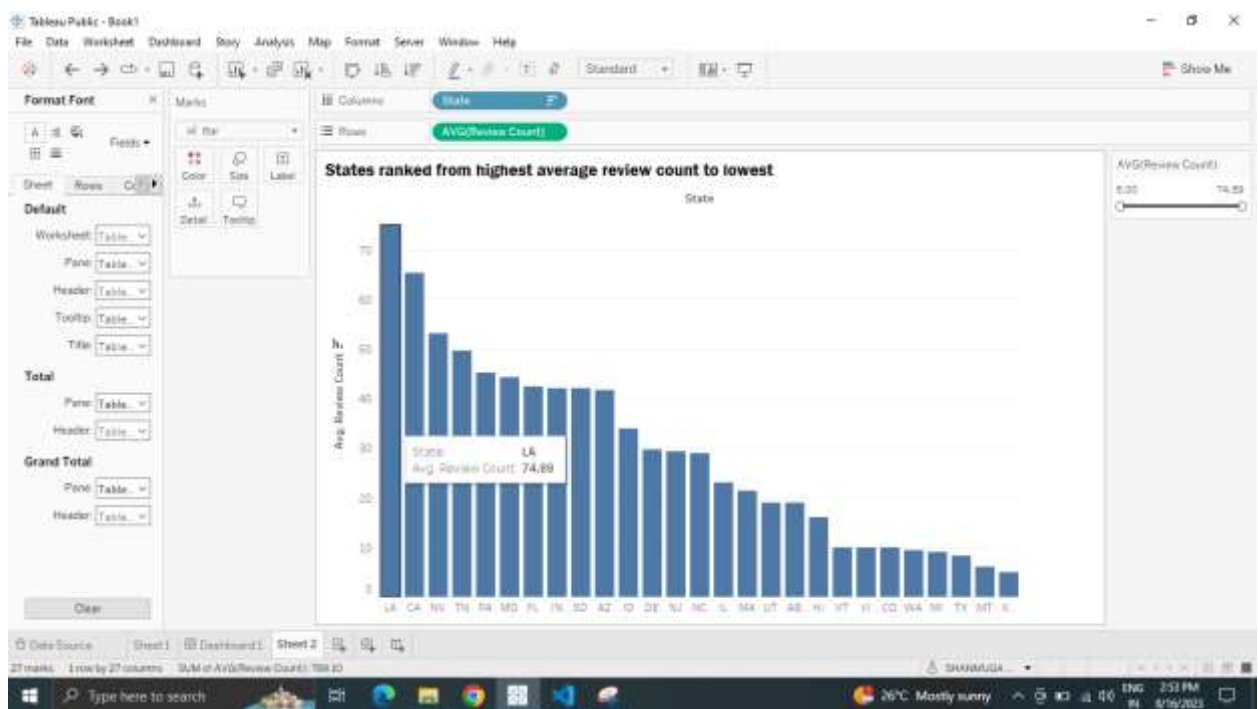




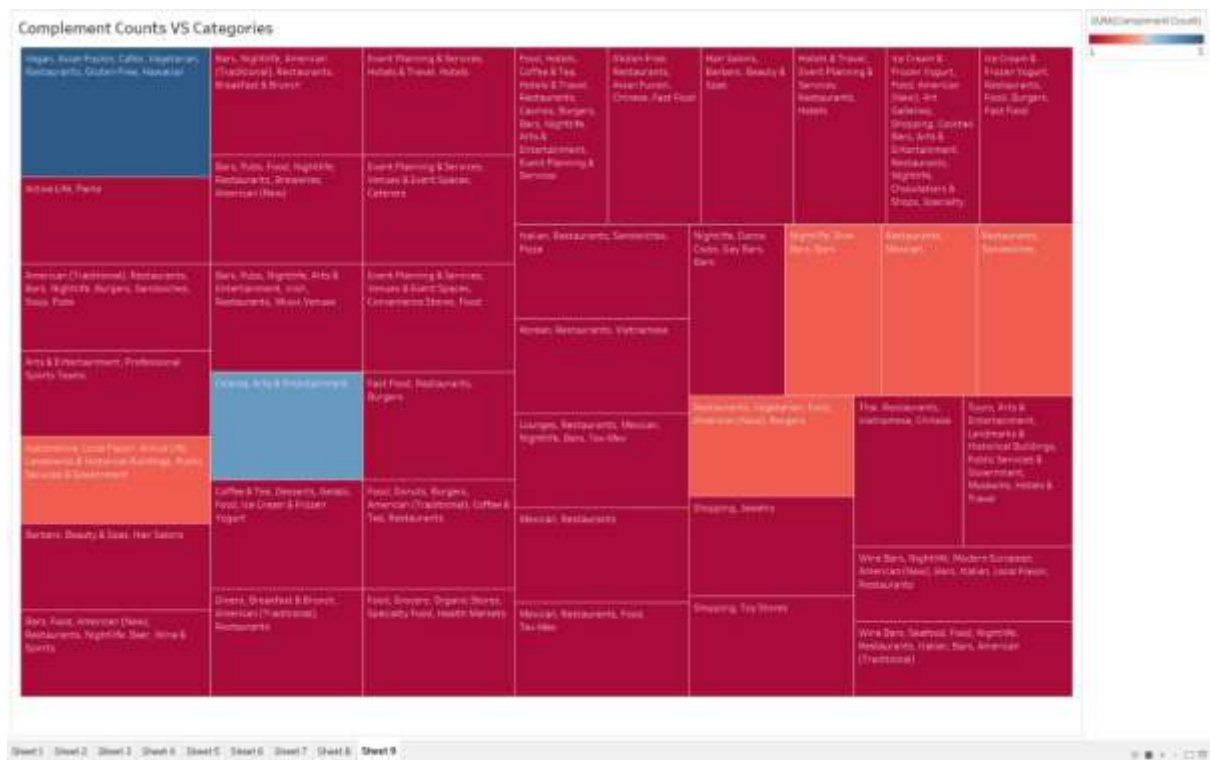
2). Popular business categories and the number of businesses in that category.



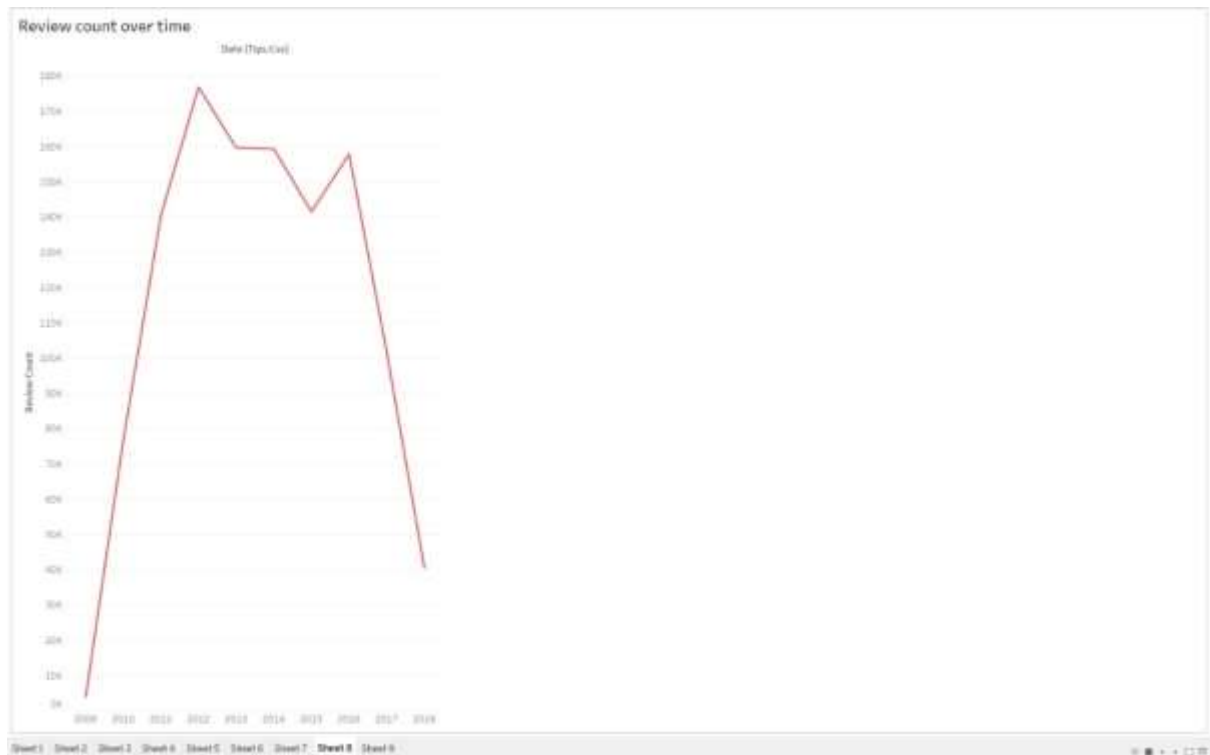
3). States and their average review count.



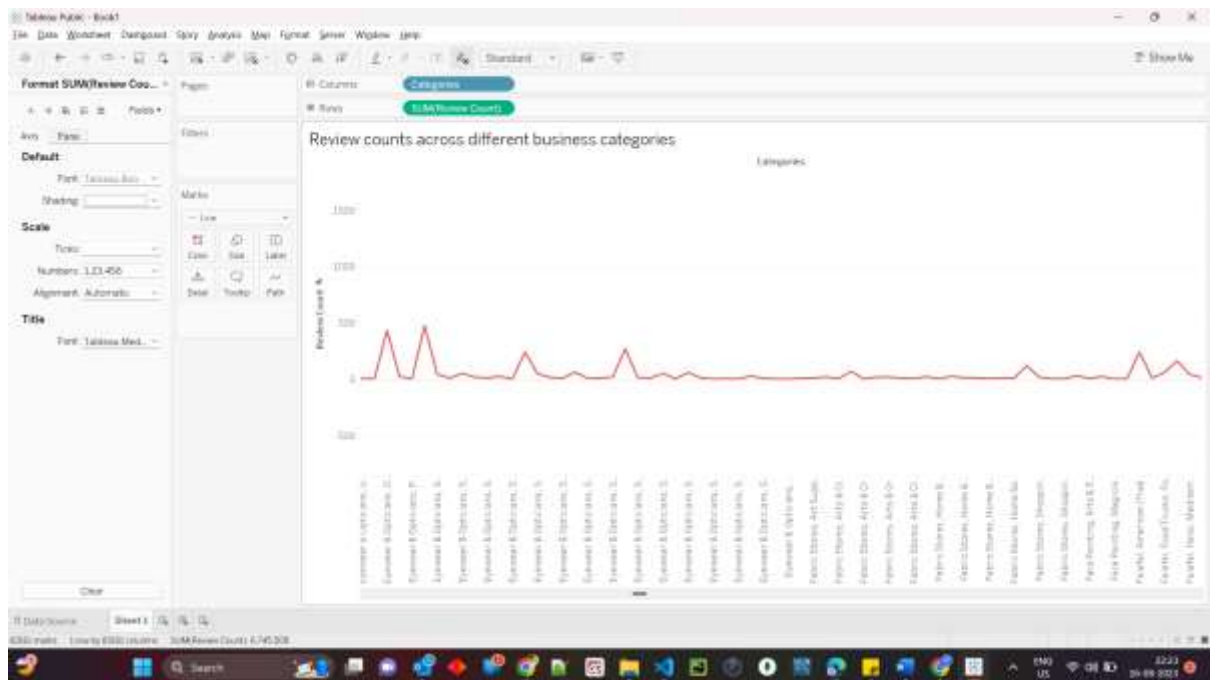
#### 4) complement count of each categories



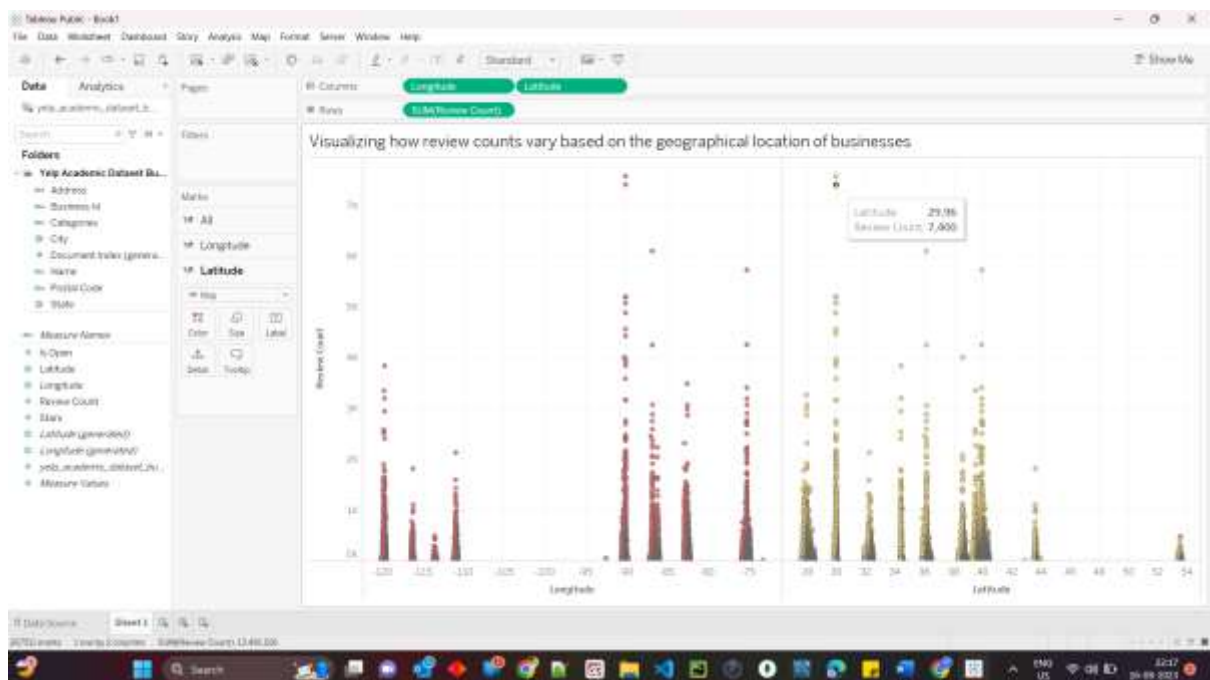
#### 5) Review count trend over the years as a line chart



6) Through this easily identify which categories have the highest review counts and observe trends and patterns in the data



7) Visualizing how review counts vary based on geographical location





# **Prescription for Yelp**

Based on all the above insights here's a prescription for Yelp to enhance its platform and user experience:

## **1. Quality Over Quantity:**

Focus on promoting quality and engagement within business categories rather than encouraging businesses to list under multiple categories. Prioritize businesses that receive consistent positive reviews and engage effectively with customers.

## **2. Review-Boosting Initiatives:**

Launch campaigns to encourage more user reviews, as higher review counts tend to correlate with better ratings. Promote features that incentivize users to provide constructive and genuine feedback, fostering an ecosystem of reliable reviews.

## **3. Customized Recommendations:**

Utilize insights from review ratings, categories, and user behaviors to fine-tune recommendation algorithms. Deliver more personalized suggestions to users, leading them to businesses that match their preferences and local trends.

## **4. Engagement During Night Hours:**

Explore strategies to encourage user engagement and reviews during nighttime hours. Consider implementing night-time promotions or incentives to capitalize on this time frame, even though it doesn't strongly influence ratings.

## **5. Elite User Empowerment:**

Leverage the moderate correlation between elite count and number of friends. Introduce features that enhance interactions and networking among elite users, fostering a sense of community and shared experiences.

## **6. Community Partnerships:**

Collaborate with top-reviewed businesses and high-engagement states to foster economic growth and community development. Recognize and support businesses that contribute significantly to the platform's success.

## **7. Sentiment Analysis Integration:**

Integrate sentiment analysis tools to evaluate review sentiments on a larger scale. Highlight businesses with consistently positive sentiment scores to attract users seeking highly-rated venues.

### **8. Seasonal Promotions:**

Leverage trends from user engagement and review counts to launch seasonal promotions or events. Offer incentives for users to engage with the platform during peak periods, fostering increased reviews and interactions.

### **9. User Engagement Insights:**

Continuously monitor user engagement metrics, including reviews, check-ins, and user interactions. Use this data to refine algorithms, personalize user experiences, and guide platform improvements.

### **10. User Education and Empowerment:**

Educate business owners on factors influencing review counts, ratings, and user engagement. Provide resources and guidance on how to effectively manage their Yelp profiles to maximize positive customer experiences.

By implementing these strategies and leveraging the insights gained, Yelp can create a more engaging and dynamic platform that caters to user preferences, encourages meaningful interactions, and supports businesses in delivering exceptional services to their customers.

## **Conclusion**

In summary, this Yelp project has yielded insightful findings into user engagement, business performance, and trends. Quality engagement within categories matters more than sheer variety, and user feedback influences ratings. Nighttime check-ins don't strongly affect ratings. While elite status correlates moderately with more friends, it's not the sole factor. To leverage these insights, Yelp can enhance recommendations, target engagement strategies, foster elite user connections, and form partnerships for mutual growth. These findings provide actionable steps to enhance user experiences and business interactions within the Yelp platform.

## **MECE Report**

Name	Task Carried out for mid point presentation	Task Carried out for Final project
Likhitha Jayanthi	Ppt and exploratory data analysis	Git repo Creation And report creation along with insights and prescription for yelp and conclusion
Arun Kumar Subramaniam	Ppt and exploratory data analysis	2 visualizations
Anisha Susan	Ppt and exploratory data analysis	Ppt and 4 visualizations
Shanmuga Priyan Jeevanandam	Ppt and exploratory data analysis	3 Spark jobs and 3 visualizations
Mahmood Hoosain	Ppt and exploratory data analysis, Yelp ERD and Business Process Map	EDA, Data Transformation, Mapreduce and Spark Jobs
Parminder Kaur	ppt	
Simar Sindhu	ppt	

**Git Repo Link :** <https://github.com/LikhithaJayanthi/1024DTS>

### **Mahmood's EDA Notebook :**

[https://github.com/LikhithaJayanthi/1024DTS/blob/main/mahmood Final project codes/Mahmood Yelp EDA Transformation.ipynb](https://github.com/LikhithaJayanthi/1024DTS/blob/main/mahmood%20Final%20project%20codes/Mahmood%20Yelp%20EDA%20Transformation.ipynb)

### **Likhitha's EDA Notebook :**

[https://github.com/LikhithaJayanthi/1024DTS/blob/main/Likhitha's%20Project%20Files/yelp-database-analytics%20\(1\).ipynb](https://github.com/LikhithaJayanthi/1024DTS/blob/main/Likhitha's%20Project%20Files/yelp-database-analytics%20(1).ipynb)