

YELP ANALYTICS

GROUP D





Agenda

- Objectives
- Data Source & Model
- Project Challenges
- Strategy & Technology
- Spark Jobs & Map Reduce
- Tableau Visualizations
- Conclusion

Problem Definition

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

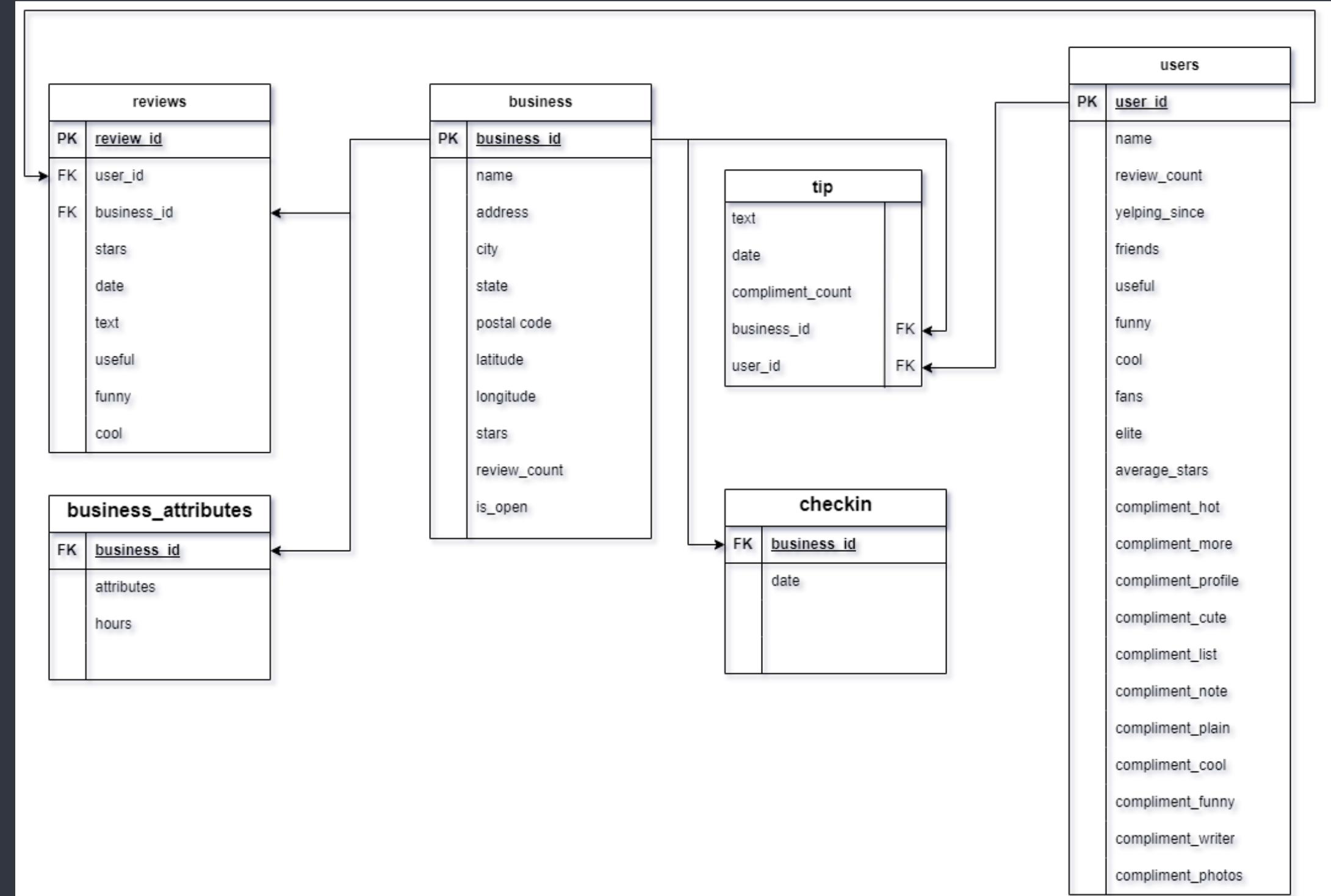
Objectives

- The objective of this project is to help Yelp by analyzing the Yelp dataset and gaining insights into various aspects of businesses, users, and reviews.
- The analysis aims to extract valuable information such as top-rated businesses, user behaviours, popular business categories, geographic trends, and more.
- The project aims to use data exploration, visualization, and statistical analysis techniques to uncover patterns and trends within the Yelp dataset.



Data Source & Model

- 50% target sales increase in the following the data source for this project is the Yelp dataset, which is a collection of data related to businesses, users, and reviews from the Yelp platform.
- The dataset includes business information such as their names, categories, locations, and star ratings. It also contains user profiles with details like user IDs, names, review counts, and average star ratings. Additionally, the dataset includes individual reviews that users have left for businesses, along with associated star ratings.
- Link: <https://www.yelp.com/dataset>
- Size : 9.29 GByear





Project Challenges

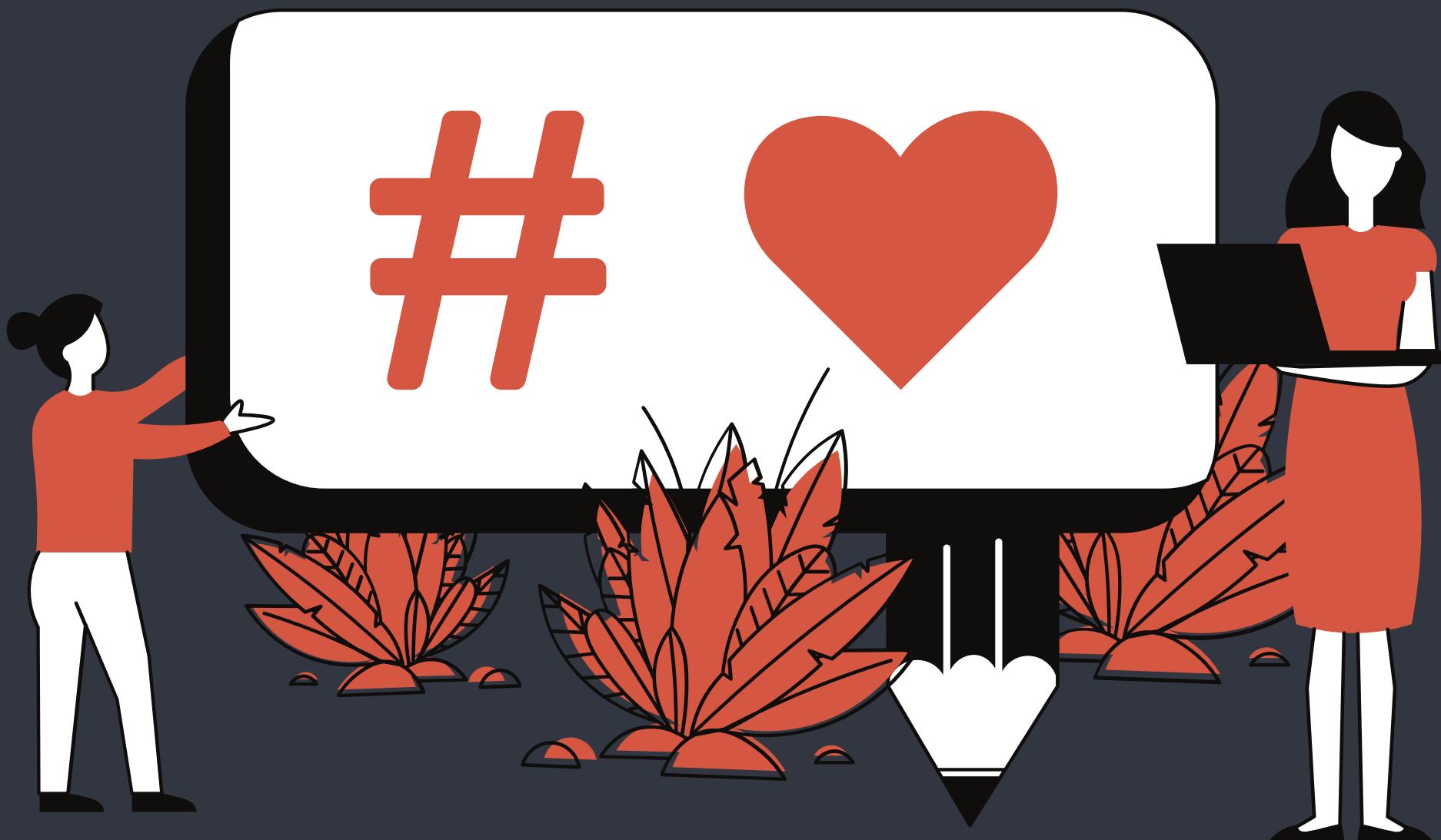
- **LARGE DATASET**
- **DATA CLEANING**
- **JSON FILES**
- **PROCESSING AND STORAGE**
- **SPARK JOBS**
- **MAP REDUCE**



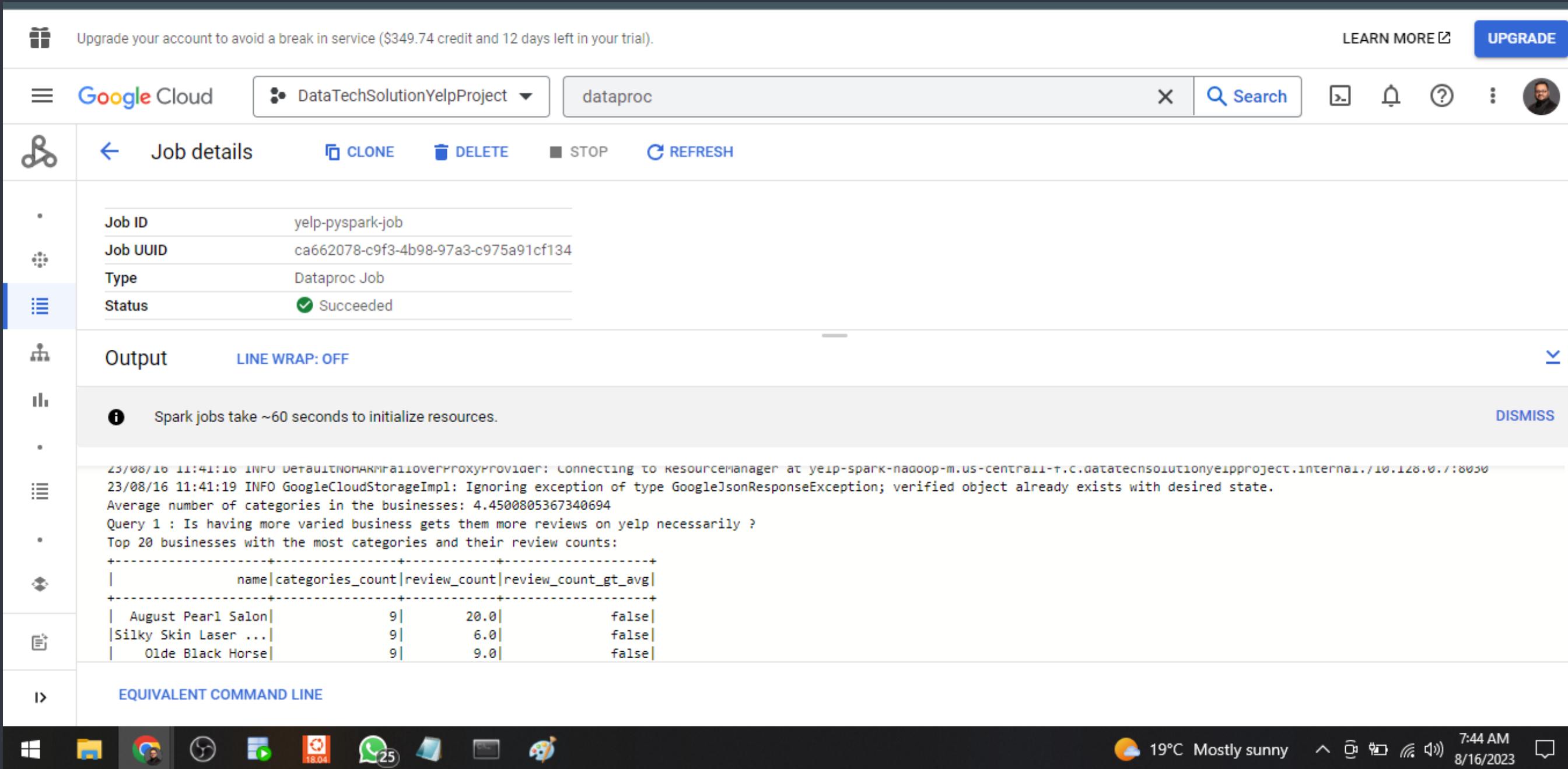
Strategy & Technology

- Cloud Notebooks
 - Basic Exploratory Data Analysis
 - Pin-pointing problem statement
-
- Bigquery
 - Dataproc
 - Tableau
 - PySpark & MapReduce

Spark Jobs & Map Reduce



Top 20 categories with the most number of business and their review counts



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 Search

Job details CLONE DELETE STOP REFRESH

Job ID: yelp-pyspark-job
Job UUID: ca662078-c9f3-4b98-97a3-c975a91cf134
Type: Dataproc Job
Status: Succeeded

Output LINE WRAP: OFF

Spark jobs take ~60 seconds to initialize resources. DISMISS

```
25/08/16 11:41:16 INFO DefaultHadoopFileOverProxyProvider: Connecting to ResourceManager at yelp-spark-naoop-m.us-central1-t.c.datatechsolutionyelpproject.internal./10.128.0.7:18080
23/08/16 11:41:19 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
Average number of categories in the businesses: 4.4500805367340694
Query 1 : Is having more varied business gets them more reviews on yelp necessarily ?
Top 20 businesses with the most categories and their review counts:
+-----+-----+-----+
| name|categories_count|review_count|review_count_gt_avg|
+-----+-----+-----+
| August Pearl Salon| 9| 20.0| false|
| Silky Skin Laser ...| 9| 6.0| false|
| Olde Black Horse| 9| 9.0| false|
```

EQUIVALENT COMMAND LINE

Windows Taskbar icons: File Explorer, Google Chrome, Task View, File History, Task Scheduler, WhatsApp (25 notifications), Paint.

System tray: 18.04, 19°C Mostly sunny, 7:44 AM, 8/16/2023.

Spark jobs take ~60 seconds to initialize resources.

DISMISS

Query 1 : Is having more varied business gets them more reviews or vice versa? If so, how many?

Top 20 businesses with the most categories and their review counts:

	name	categories_count	review_count	review_count_gt_avg
1	August Pearl Salon	9	20.0	false
2	Silky Skin Laser ...	9	6.0	false
3	Olde Black Horse	9	9.0	false
4	Top Shelf Sports ...	9	95.0	true
5	Reno Downtown Joint	9	10.0	false
6	David Thomas Trai...	9	6.0	false
7	DOSC	9	87.0	true
8	Enjoy The Mountain	9	48.0	true
9	Panera Bread	9	58.0	true
10	emPIEnada	9	41.0	false
11	Brixx Craft House	9	115.0	true
12	The Ladies Room	9	13.0	false
13	Windsong Charters...	9	30.0	false
14	Naked Cyber Cafe ...	9	12.0	false
15	ReJuv Medspa	9	16.0	false
16	Zakian Rug Cleaning	9	26.0	false
17	Double Decker Piz...	9	88.0	true

EQUIVALENT COMMAND LINE



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

datapro



Search



Output

LINE WRAP:
CLONE

DELETE

STOP

REFRESH



Spark jobs take ~60 seconds to initialize resources.

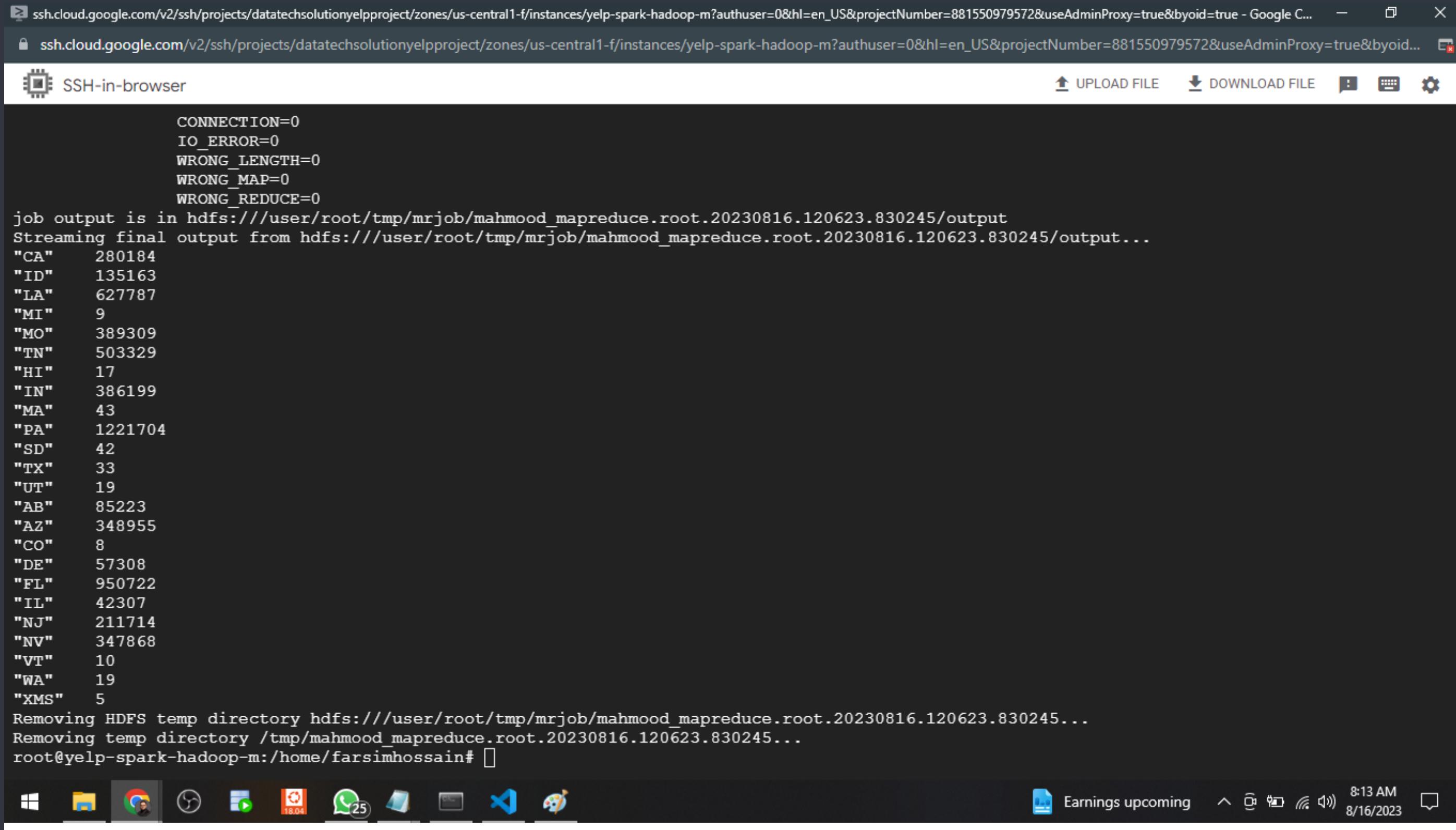
DISMISS

Naked Cyber Cafe ...	9	12.0	false
ReJuv Medspa	9	16.0	false
Zakian Rug Cleaning	9	26.0	false
Double Decker Piz...	9	88.0	true
Brooms N Buckets	9	40.0	false
CD Roma 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

Map Reduce



The screenshot shows an SSH-in-browser session on a Windows desktop. The terminal window displays the results of a MapReduce job. The output includes statistics for various states (CA, ID, LA, MI, MO, TN, HI, IN, MA, PA, SD, TX, UT, AB, AZ, CO, DE, FL, IL, NJ, NV, VT, WA, XMS) and the final streaming output to HDFS.

```
SSH-in-browser
SSH Connection: ssh.cloud.google.com/v2/ssh/projects/datatechsolutionyelpproject/zones/us-central1-f/instances/yelp-spark-hadoop-m?authuser=0&hl=en_US&projectNumber=881550979572&useAdminProxy=true&byoid=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/datatechsolutionyelpproject/zones/us-central1-f/instances/yelp-spark-hadoop-m?authuser=0&hl=en_US&projectNumber=881550979572&useAdminProxy=true&byoid=true

CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_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
"LA"      627787
"MI"       9
"MO"     389309
"TN"     503329
"HI"       17
"IN"     386199
"MA"       43
"PA"     1221704
"SD"        42
"TX"       33
"UT"       19
"AB"     85223
"AZ"     348955
"CO"        8
"DE"     57308
"FL"     950722
"IL"     42307
"NJ"     211714
"NV"     347868
"VT"       10
"WA"       19
"XMS"       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/farsimhossain#
```

Windows taskbar icons include File Explorer, Google Chrome, Task View, File History, Task Scheduler, WhatsApp, Microsoft Edge, and Visual Studio Code. System tray icons show battery level (16:04), signal strength, and volume.

The average review count for businesses in each category and the top 10 categories with the highest average review counts.

The screenshot shows the Google Cloud DataProc job details page. The job is a Dataproc job that has succeeded. The output section displays a table of average review counts for various business categories. The table is as follows:

category	avg_review_count
Israeli	1574.0
Conveyor Belt Sushi	1128.0
Serbo Croatian	479.0
Iberian	412.8
Brasseries	344.4255319148936
Shanghainese	340.11764705882354
Public Markets	335.6521739130435
Hong Kong Style Cafe	267.8333333333333
Cajun/Creole	266.807150595883
Belgian	263.79166666666667

The average review count for businesses in each state, along with the highest average review count state-wise and the top 5 states with the highest average review counts.

The screenshot shows the Google Cloud DataProc Job details page. The job ID is job-b59c67f8, the status is Succeeded, and the type is Dataproc job. The output log displays the results of a data processing job, specifically the average review count for businesses in each state. The output is as follows:

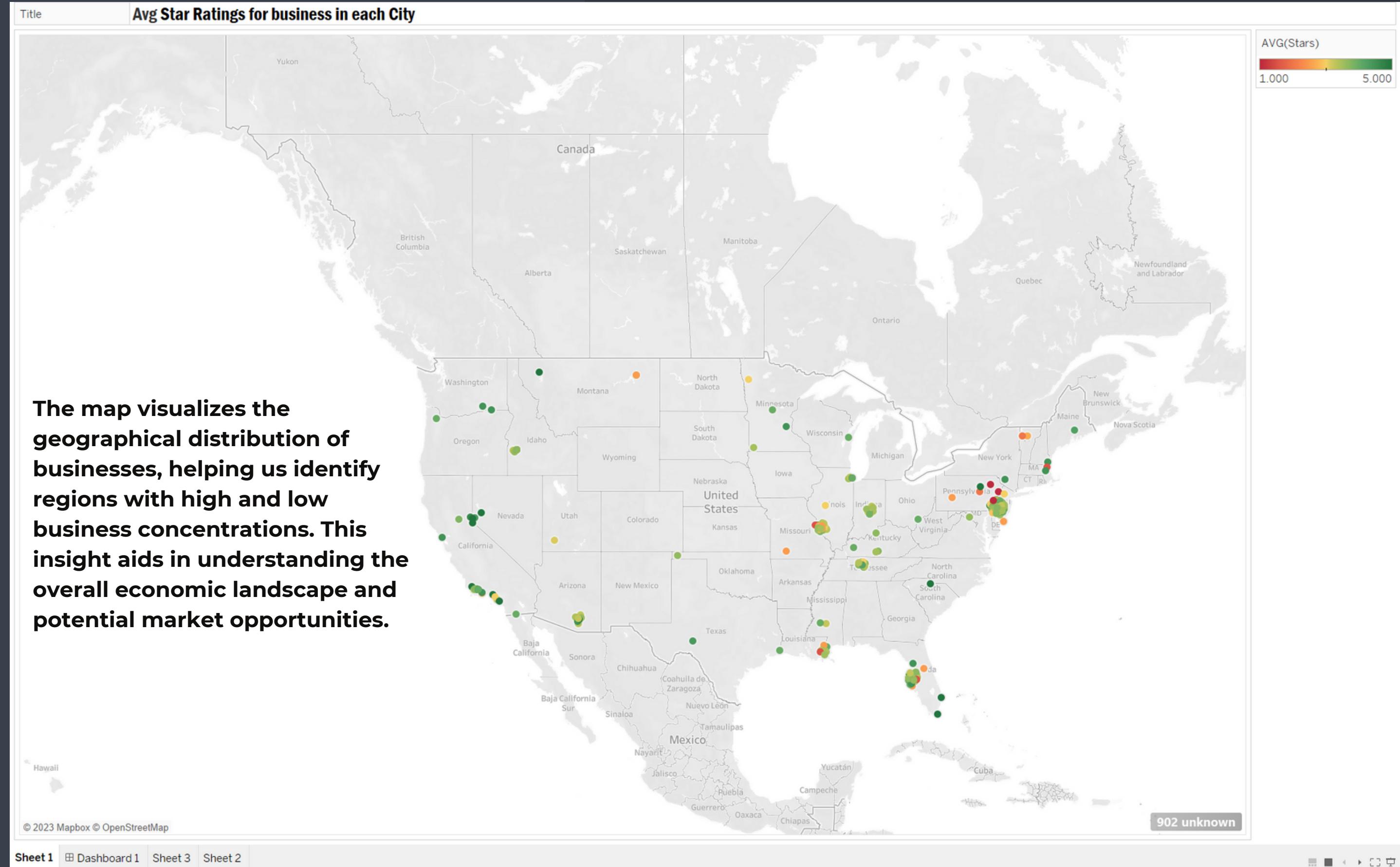
```
23/08/16 05:43:40 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object a.
23/08/16 05:43:40 WARN GfsStorageStatistics: Detected potential high latency for operation op_mkdirs. latencyMs=174; pre
23/08/16 05:43:51 WARN package: Truncated the string representation of a plan since it was too large. This behavior can l
State with the highest average review count: LA
+-----+
|state| avg_review_count|
+-----+
| LA | 74.88673921805723 |
| CA | 65.27714779934654 |
| NV | 53.13674659753727 |
| TN | 49.61803251493033 |
| PA | 45.26543082934281 |
+-----+
```

The top 5 Business with most number of 5 Star rating with the total review.

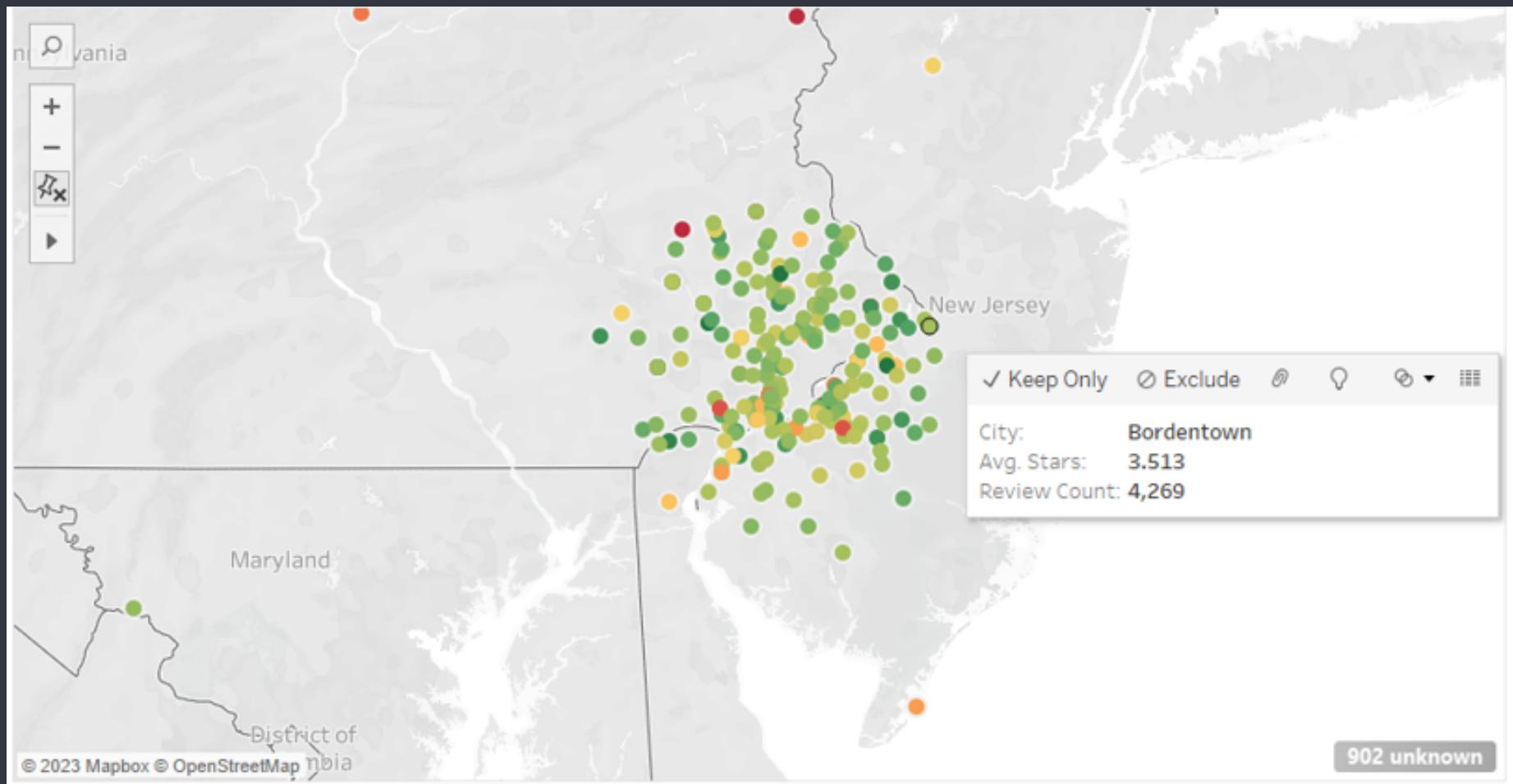
The screenshot shows the Google Cloud DataProc Job details page. The job ID is job-ca417f30, and it has succeeded. The output section displays a table of business reviews:

business_id	name	review_count
_aKr7POnacN_VizRK...	Blues City Deli	991
8QgnRpM-QxGsjDNuu...	Carlillos Cocina	799
zxIF-bnaJ-eKIsznB...	Free Tours By Foot	769
DVBJRvnCpkqaYl6nH...	Tumerico	705
gP_oWJykA2RocIs_G...	Yats	623

Geographic Business Distribution

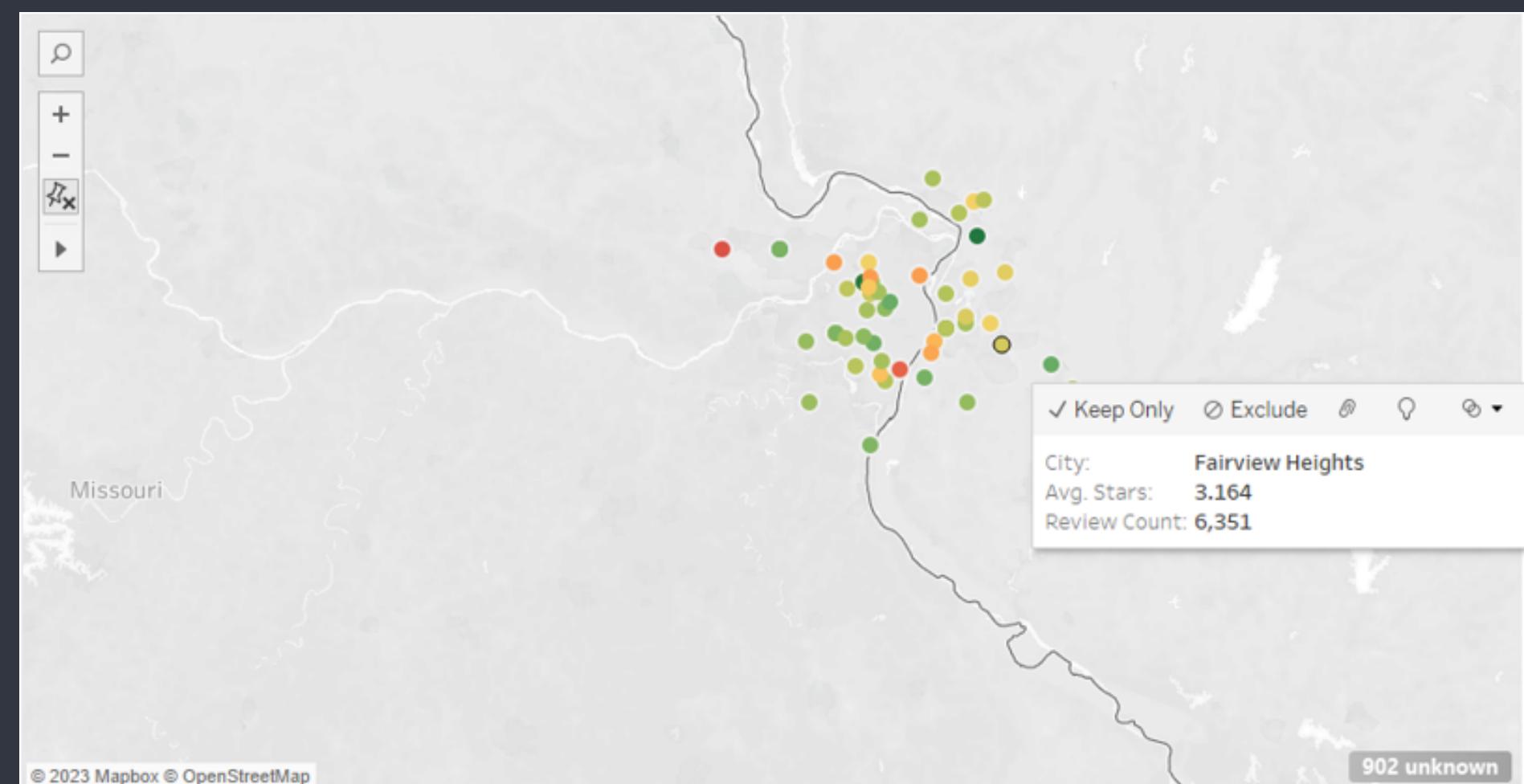


Rating Analysis through Color Gradient

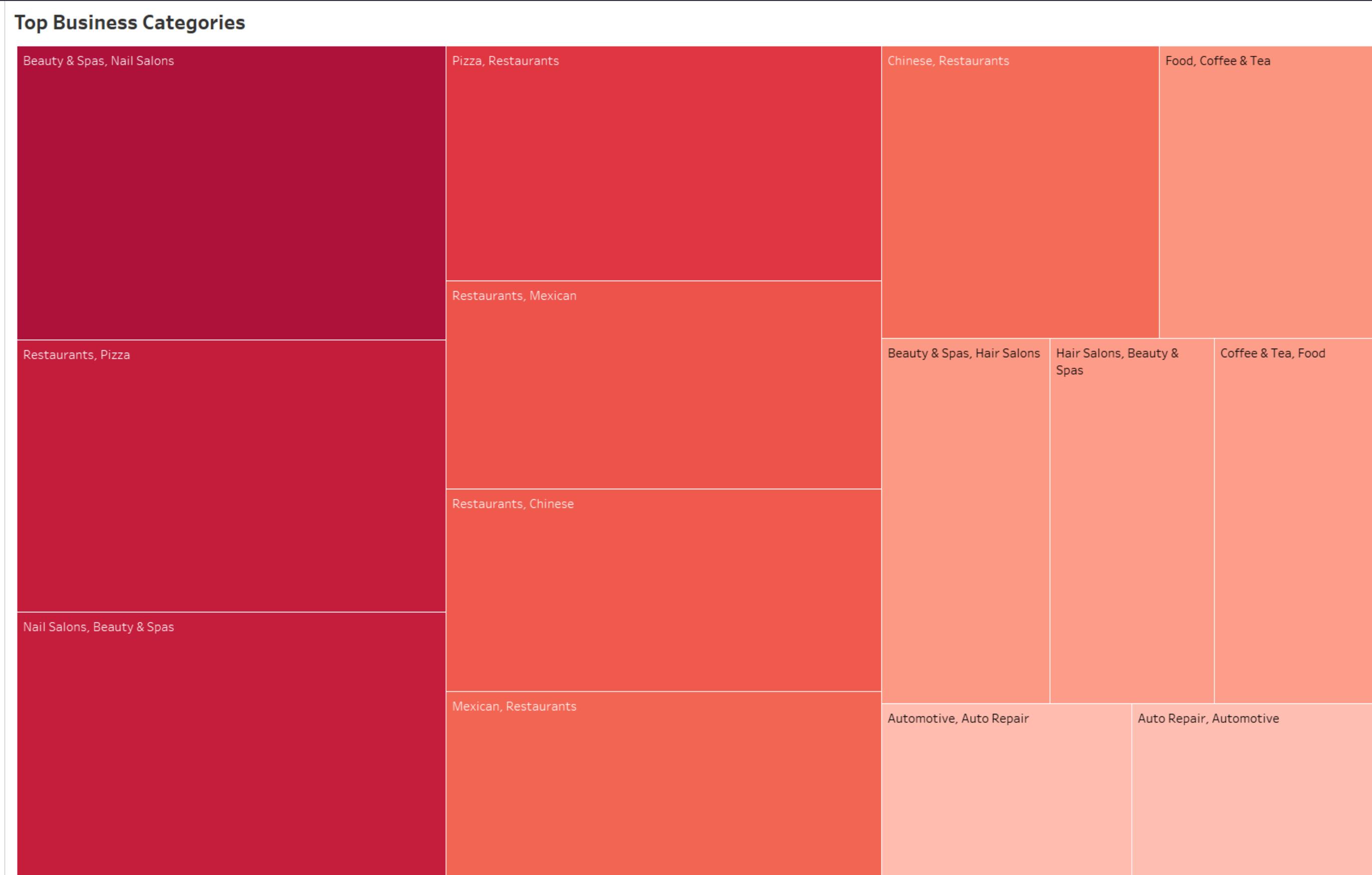
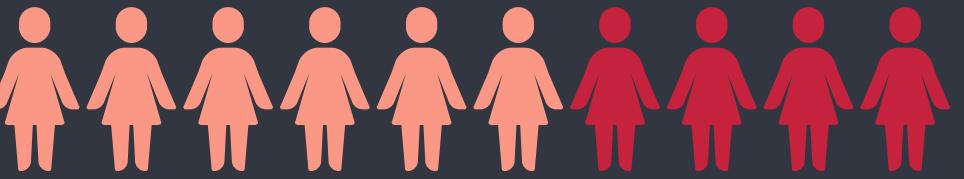


- The color spectrum highlights exceptional and poorly rated businesses, allowing you to make informed decisions about where to dine, shop, or engage in services.

- By using color gradients for star ratings, you can quickly assess businesses' overall ratings.



Top Business Categories



Top Business Categories

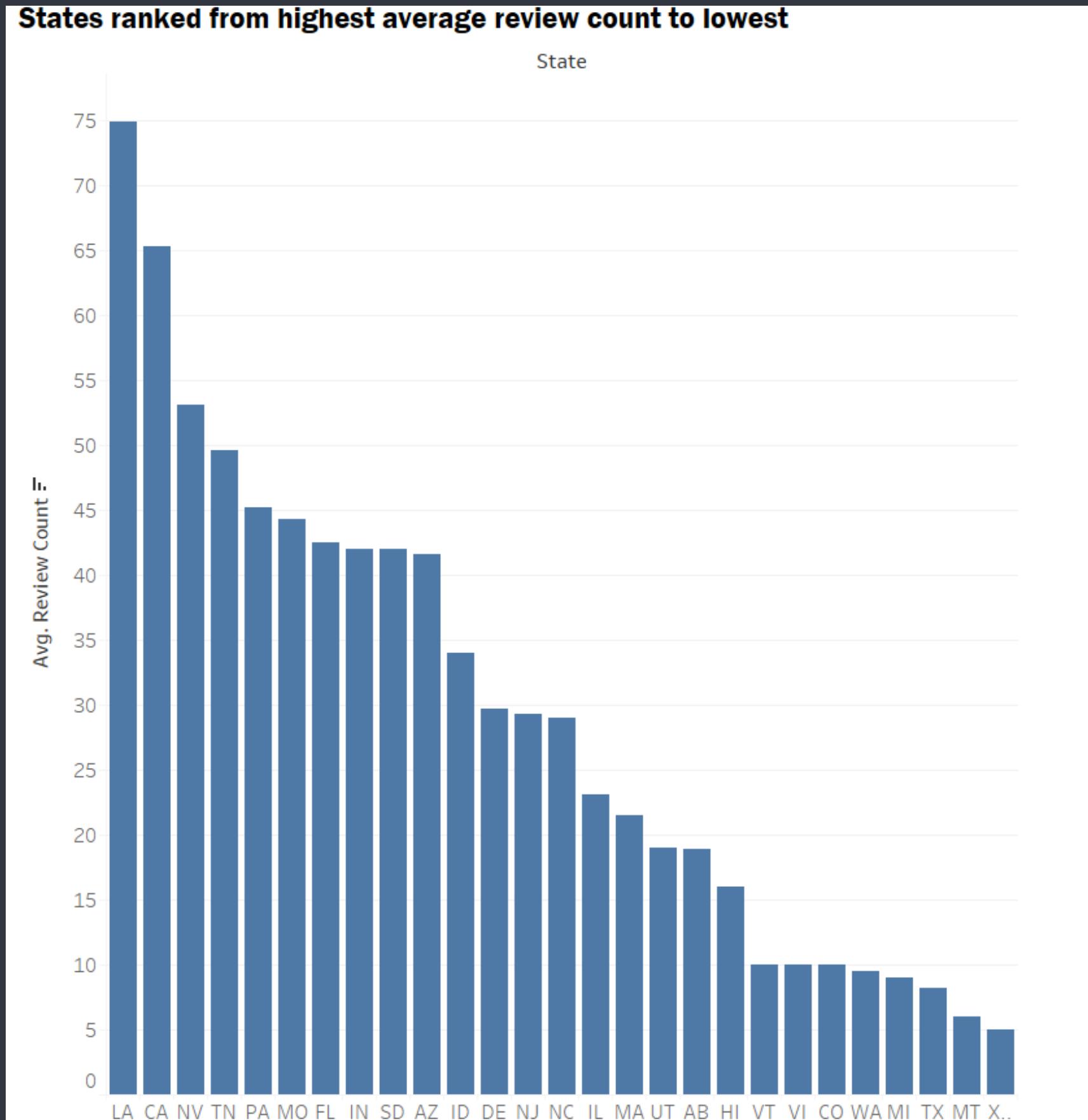
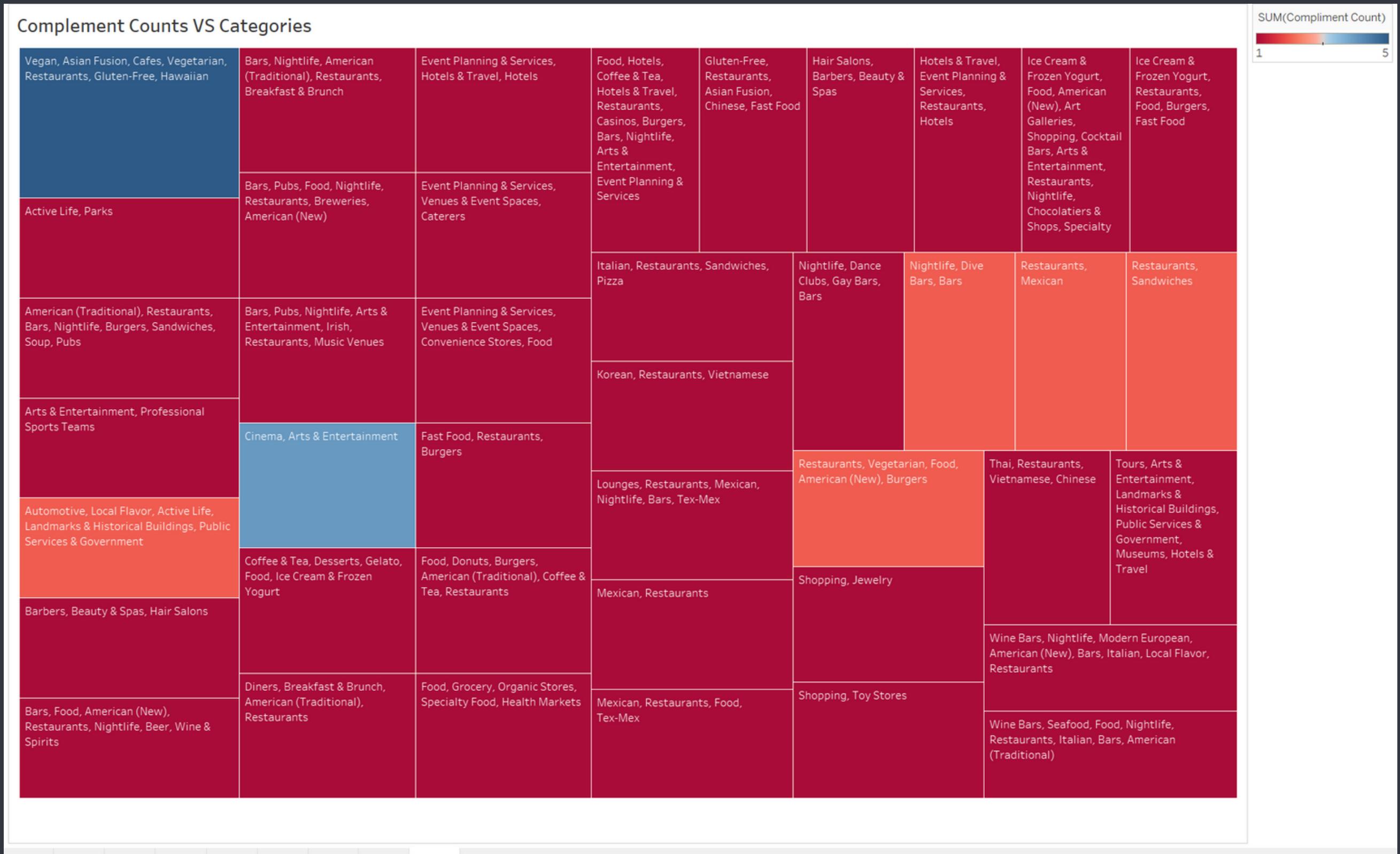
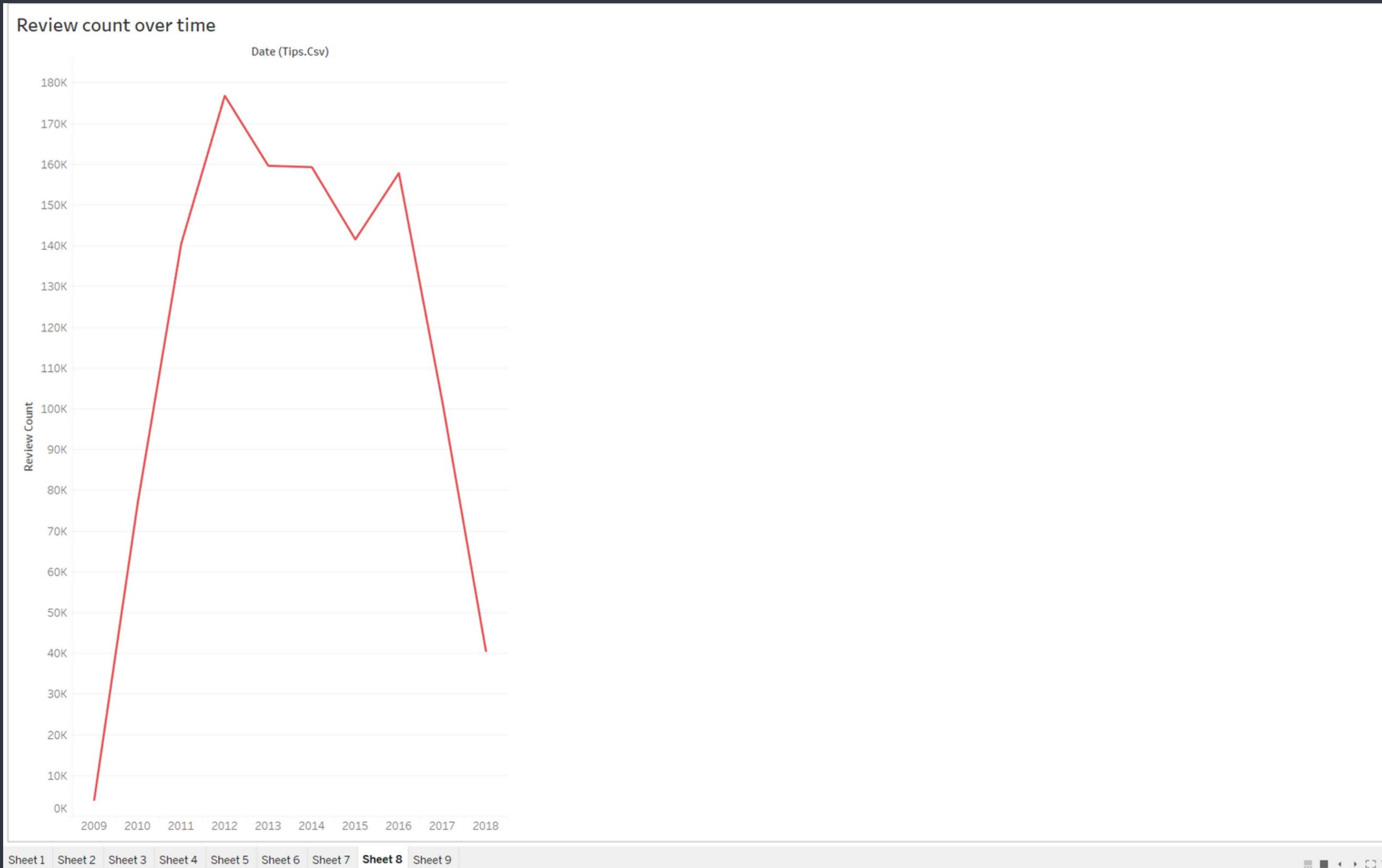


Chart shows different states and their average review count providing valuable insights into how businesses in various states are performing based on customer reviews. This visualization can help identify states with higher or lower average review counts, potentially indicating the overall satisfaction levels of customers.

Complement Count VS Categories



Review count trends over the years





Conclusion

- The outlined prescription for enhancing Yelp's platform and user experience is a well-rounded approach that combines several key strategies. Prioritizing quality over quantity within business categories, fostering genuine user reviews, and refining personalized recommendations are essential steps.
 - Encouraging engagement during nighttime hours and empowering elite users contribute to a sense of community and enrich interactions.
 - Collaborating with top-reviewed businesses and leveraging sentiment analysis adds credibility and guides user choices. Seasonal promotions, continuous user engagement monitoring, and educating business owners further enhance the platform. By applying these strategies and capitalizing on the insights shared, Yelp can evolve into a more vibrant and user-centric platform, catering to preferences, nurturing meaningful connections, and enabling businesses to provide outstanding services.

Thank You

