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.

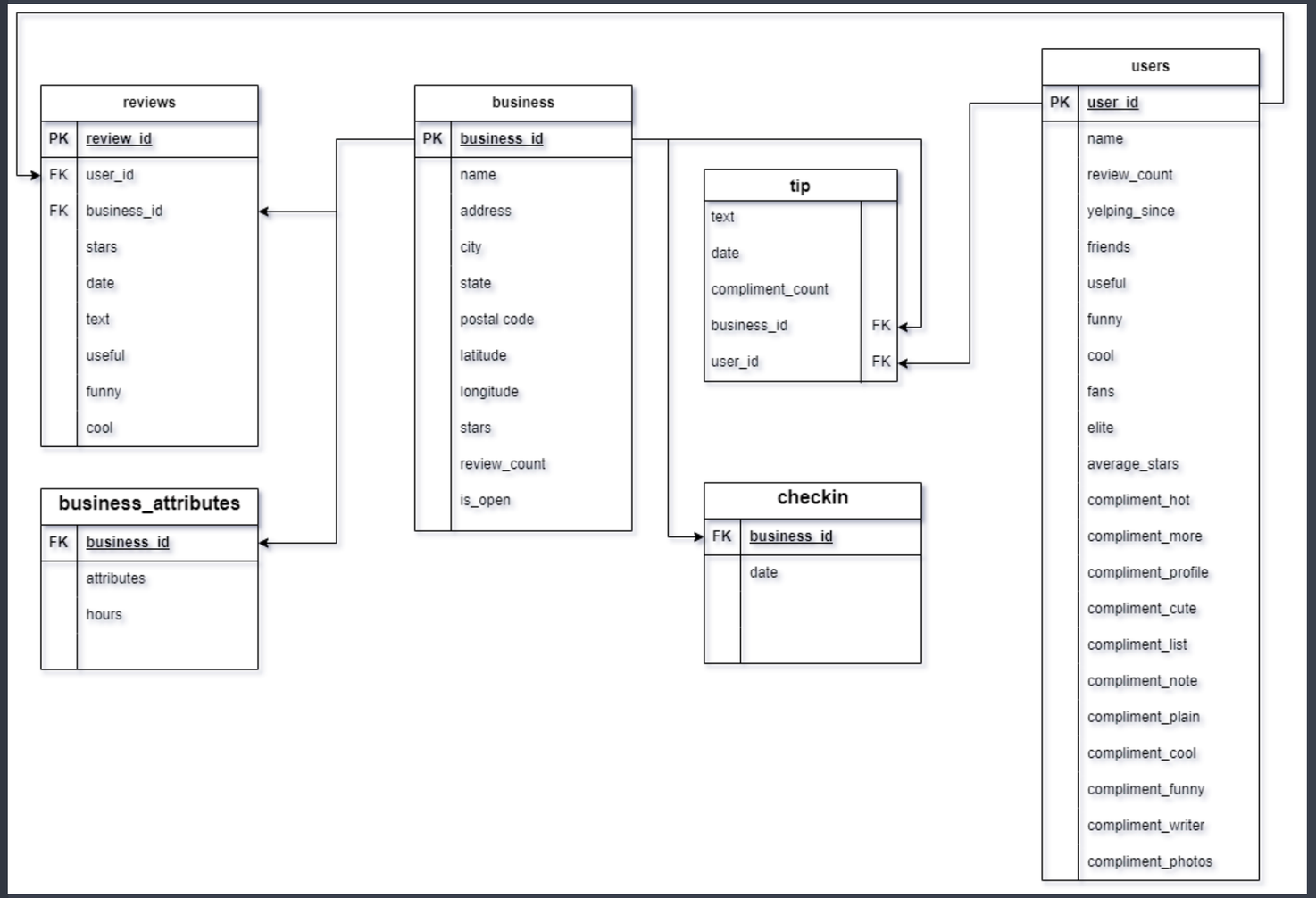
Objectives

- 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

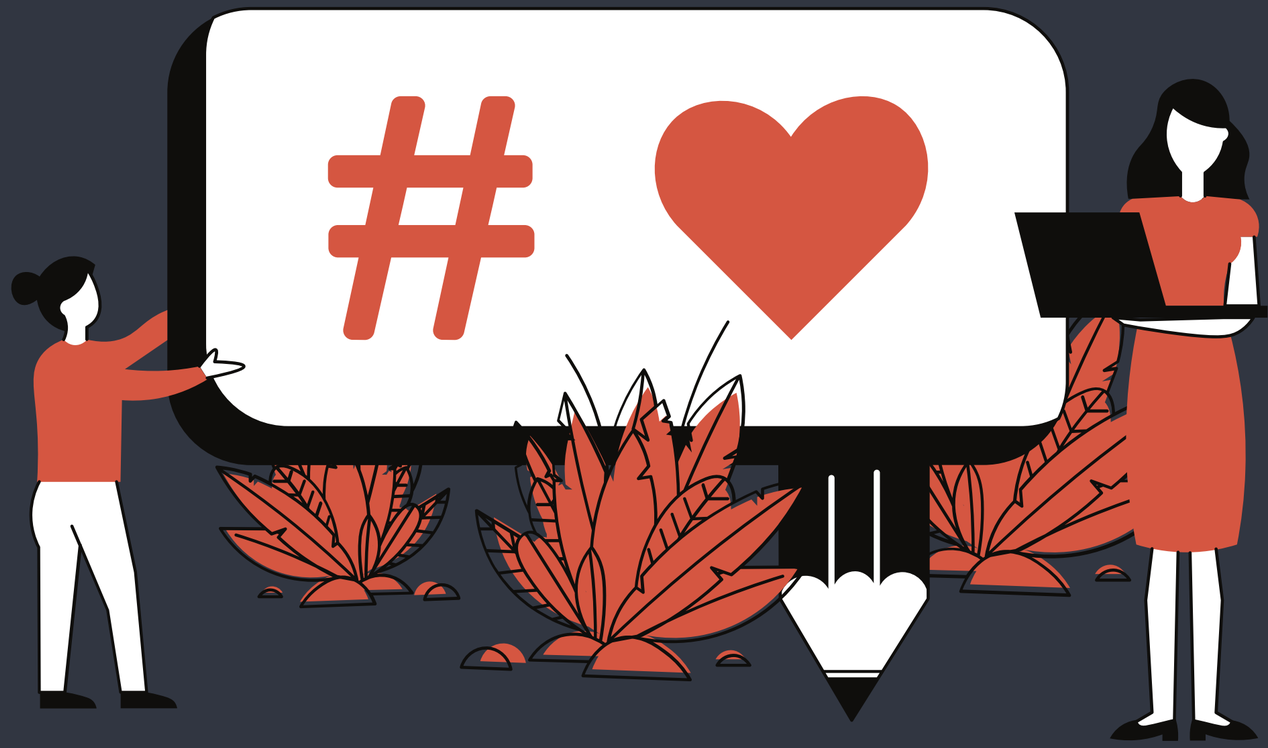
- LARGE DATASET
- DATA CLEANING
- JSON FILES
- PROCESSING AND STORAGE
- SPARK JOBS
- MAP REDUCE

Strategy & Technology

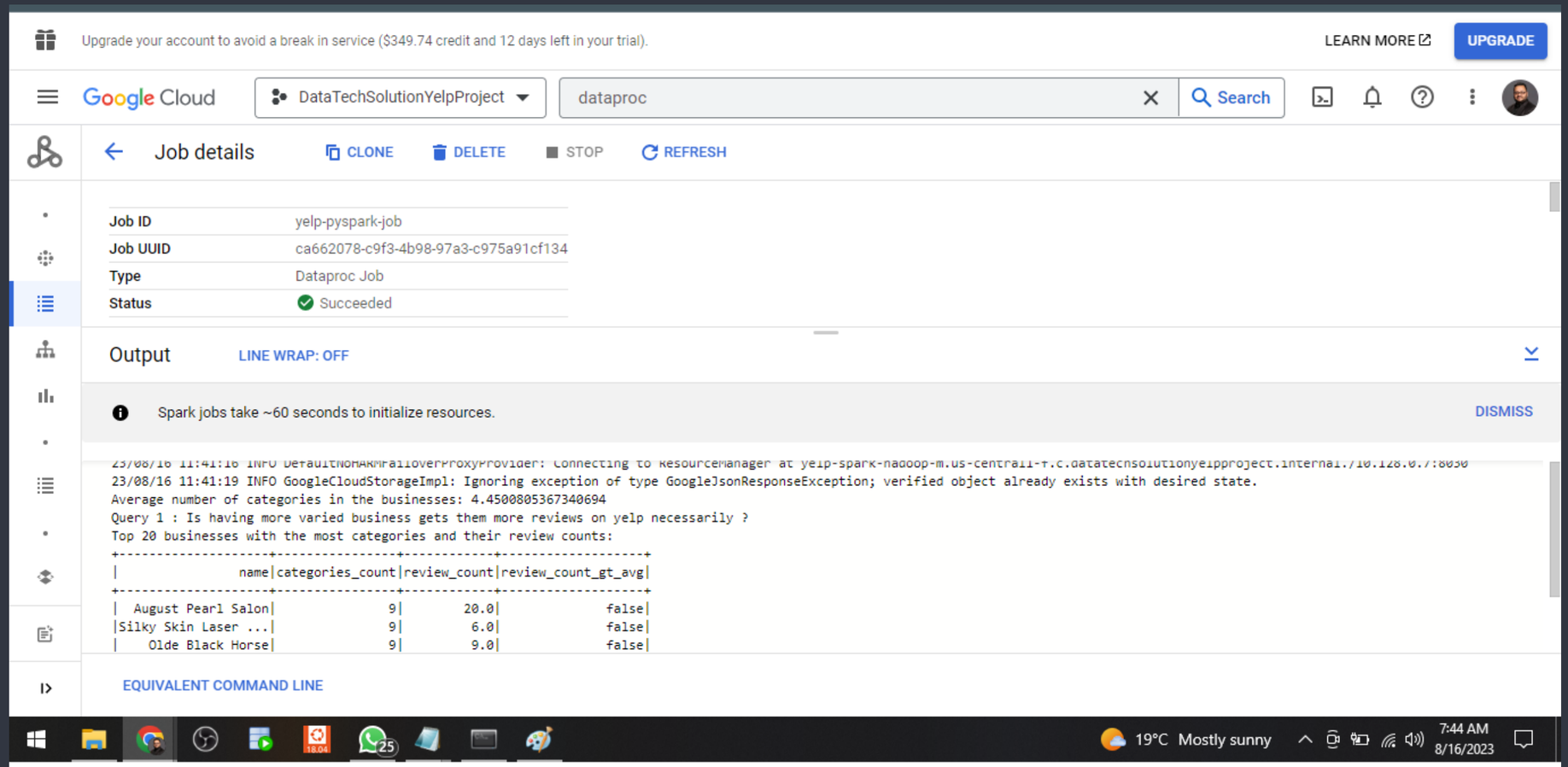
- Cloud Notebooks
- Basic Exploratory Data Analysis
- Pin-pointing problem statement

- Bigquery
- Dataproc
- Tableau
- PySpark & MapReduce

Spark Jobs & Map Reduce



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



Spark jobs take ~60 seconds to initialize resources. DISMISS

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
Top Shelf Sports ...	9	95.0	true
Reno Downtown Joint	9	10.0	false
David Thomas Trai...	9	6.0	false
DOSC	9	87.0	true
Enjoy The Mountain	9	48.0	true
Panera Bread	9	58.0	true
emPIEnada	9	41.0	false
Brixx Craft House	9	115.0	true
The Ladies Room	9	13.0	false
Windsong Charters...	9	30.0	false
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

EQUIVALENT COMMAND LINE

Output LINE WRAP: CLONE DELETE STOP REFRESH

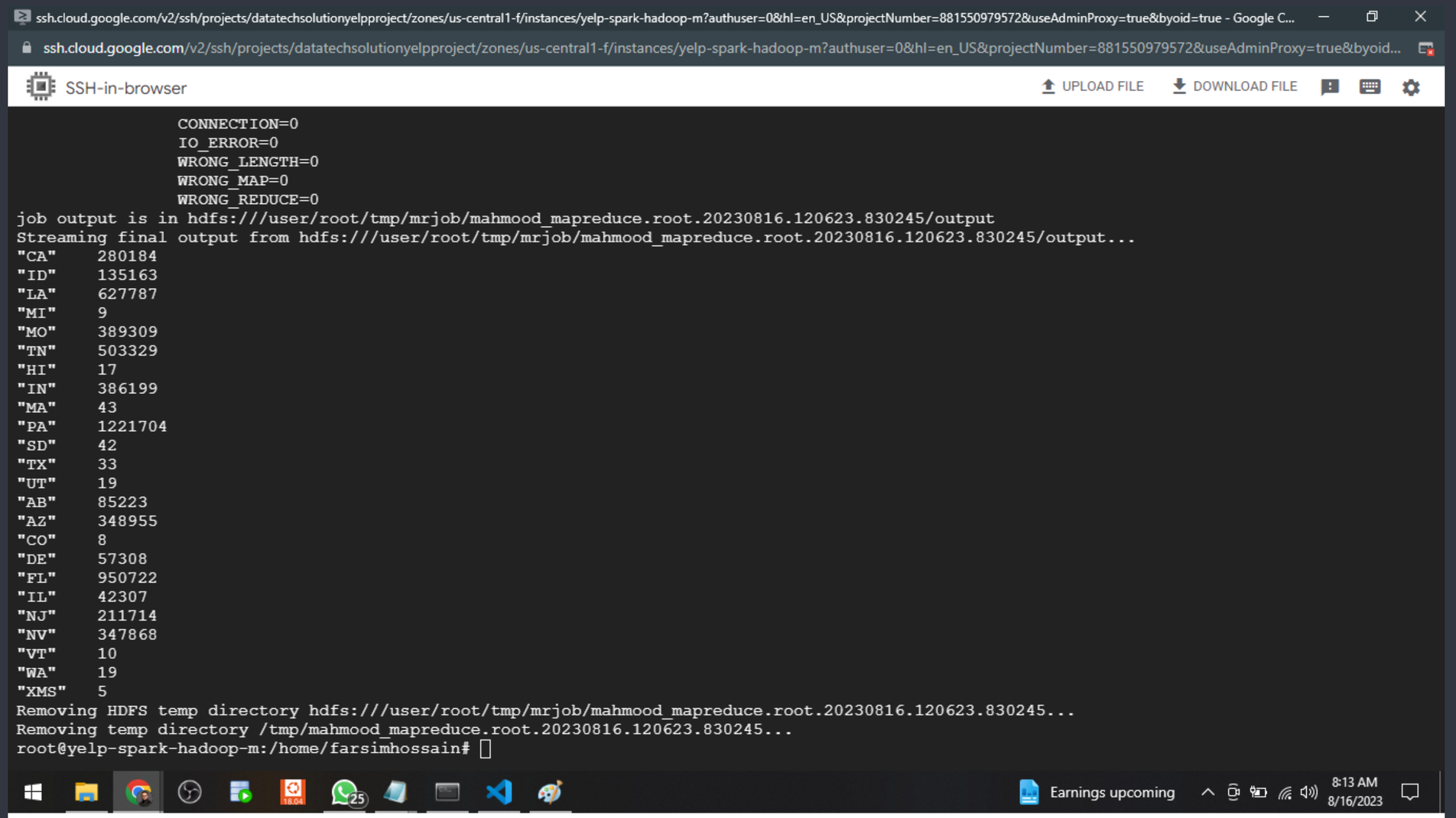
Spark jobs take ~60 seconds to initialize resources. DISMISS

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

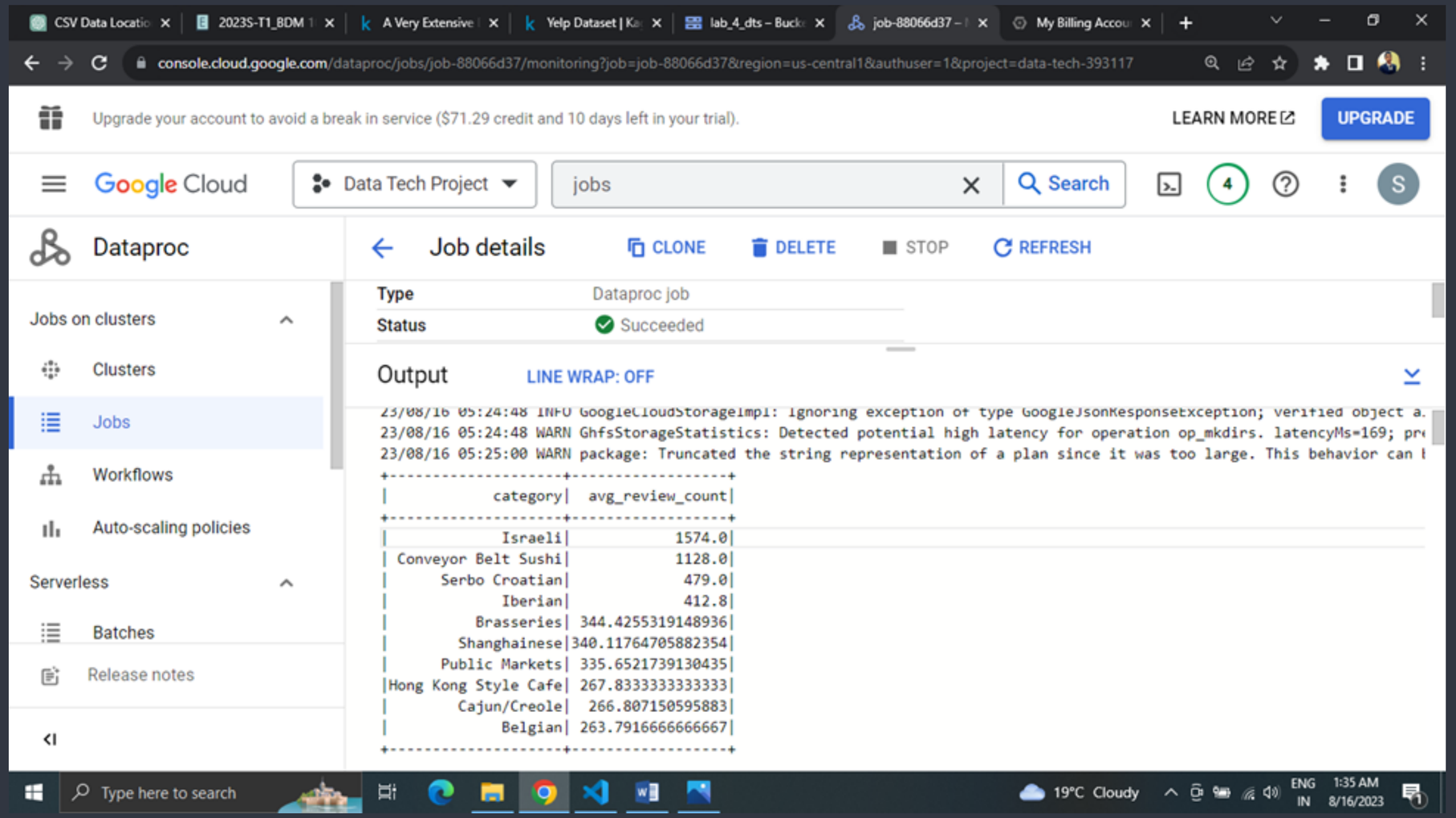
So it seems that having a lot of categories does not necessarily gain more reviews
Query 2 : Average review ratings for businesses with different review counts and businesses with more check-ins at night.
Average review rating for businesses with high review counts: 3.8
Average review rating for businesses with low review counts: 3.6
Average review rating for businesses with more night check-ins: 3.6
Query 3: Is there any signifant correlation betwen 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

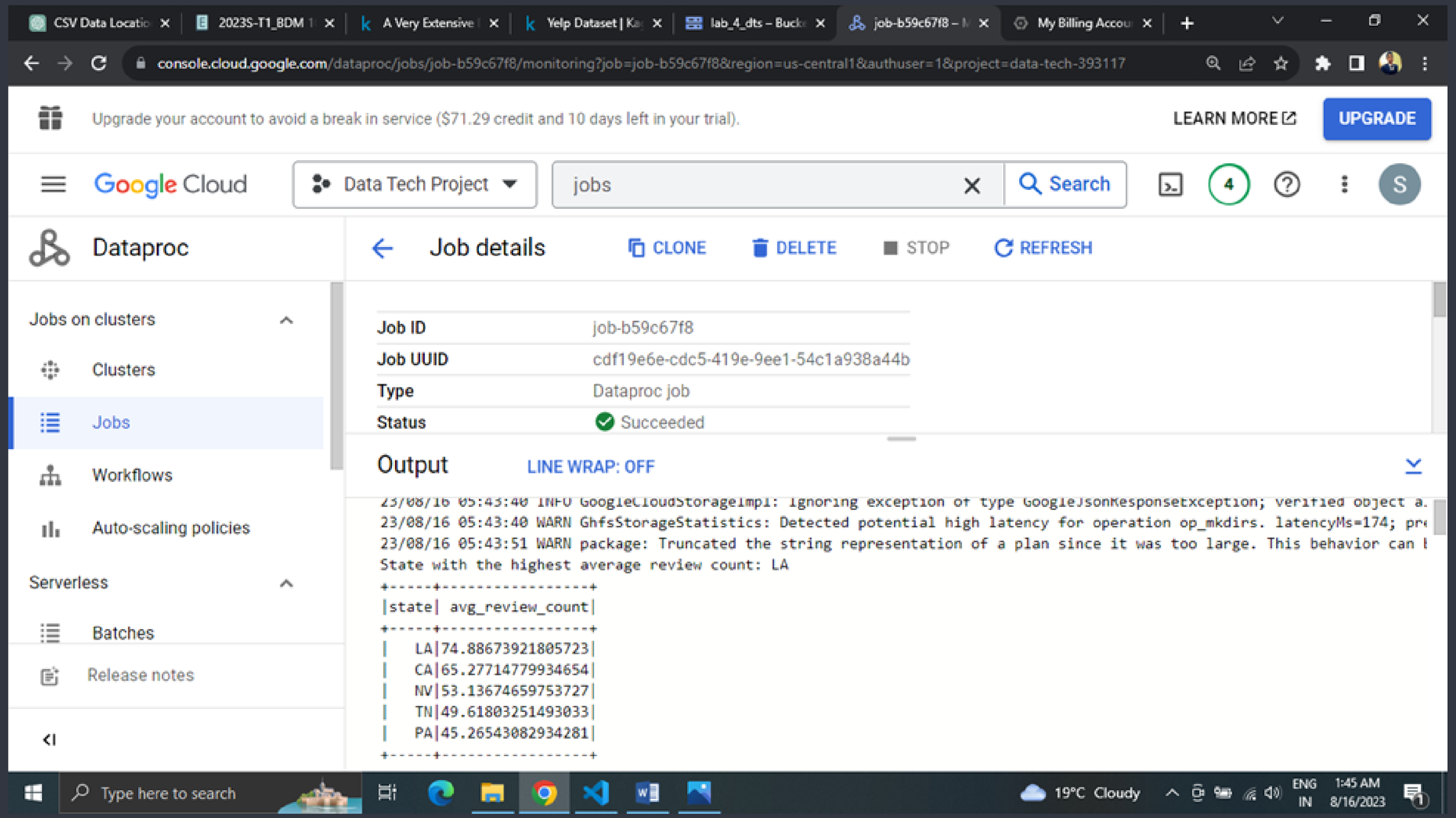
Map Reduce


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 C...
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...
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
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#
Earnings upcoming
8:13 AM
8/16/2023

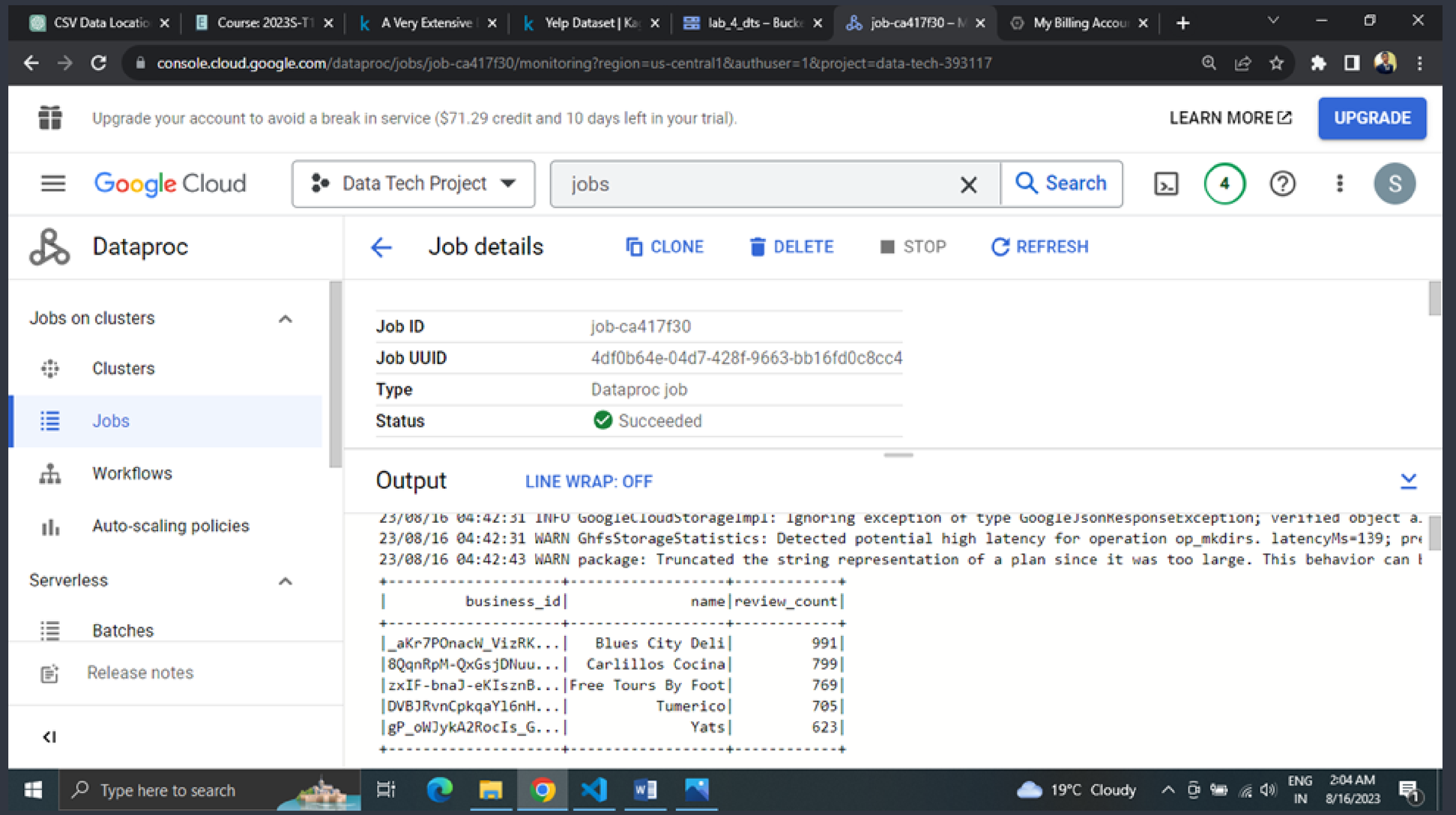
The average review count for businesses in each category and the top 10 categories with the highest average review counts.



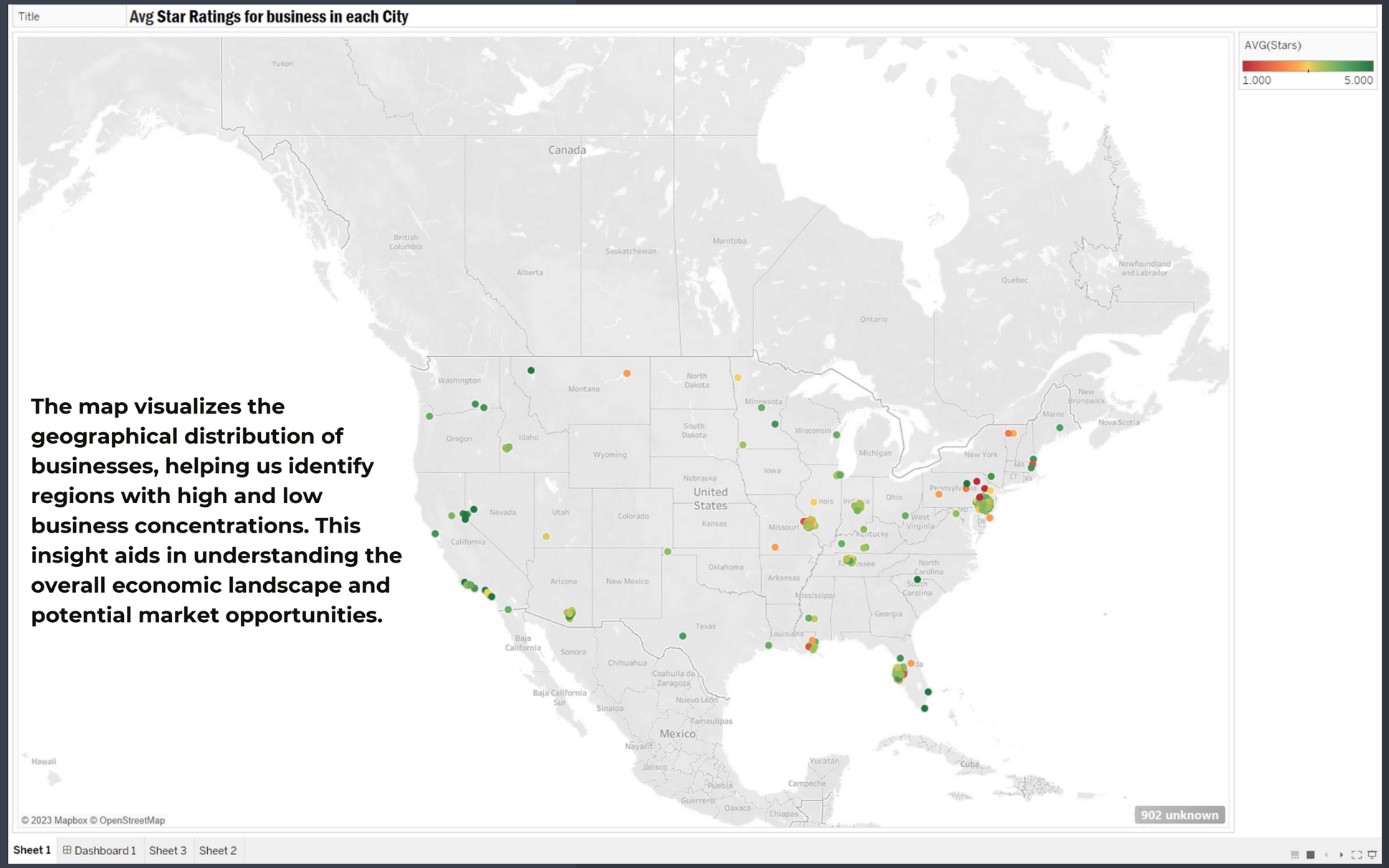
The average review count for businesses in each state, along with the highest average review count state-wise and the top 5 states with the highest average review counts.



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

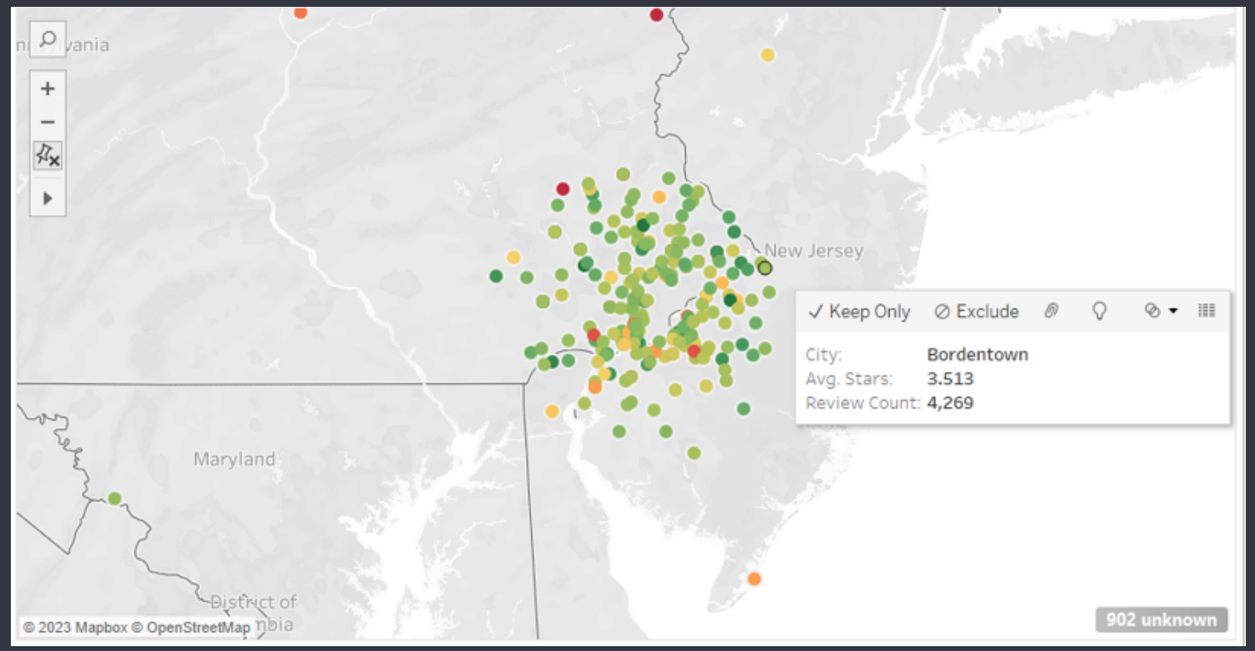


Geographic Business Distribution



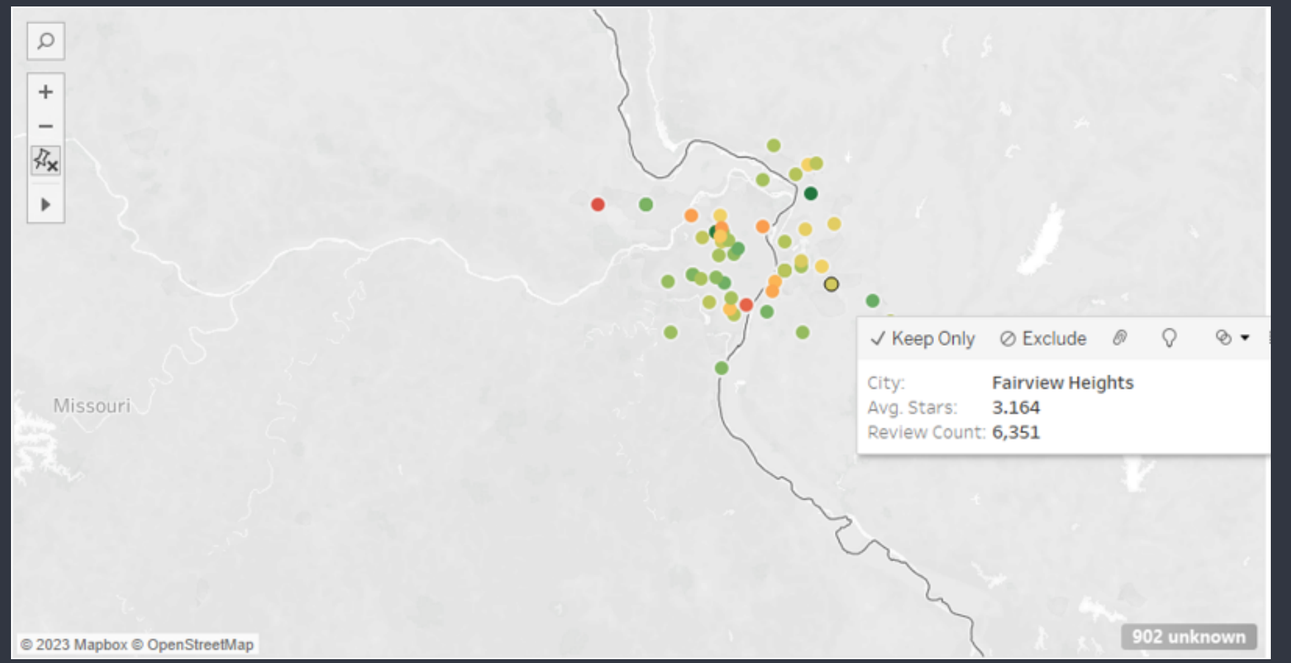
The map visualizes the geographical distribution of businesses, helping us identify regions with high and low business concentrations. This insight aids in understanding the overall economic landscape and potential market opportunities.

Rating Analysis through Color Gradient

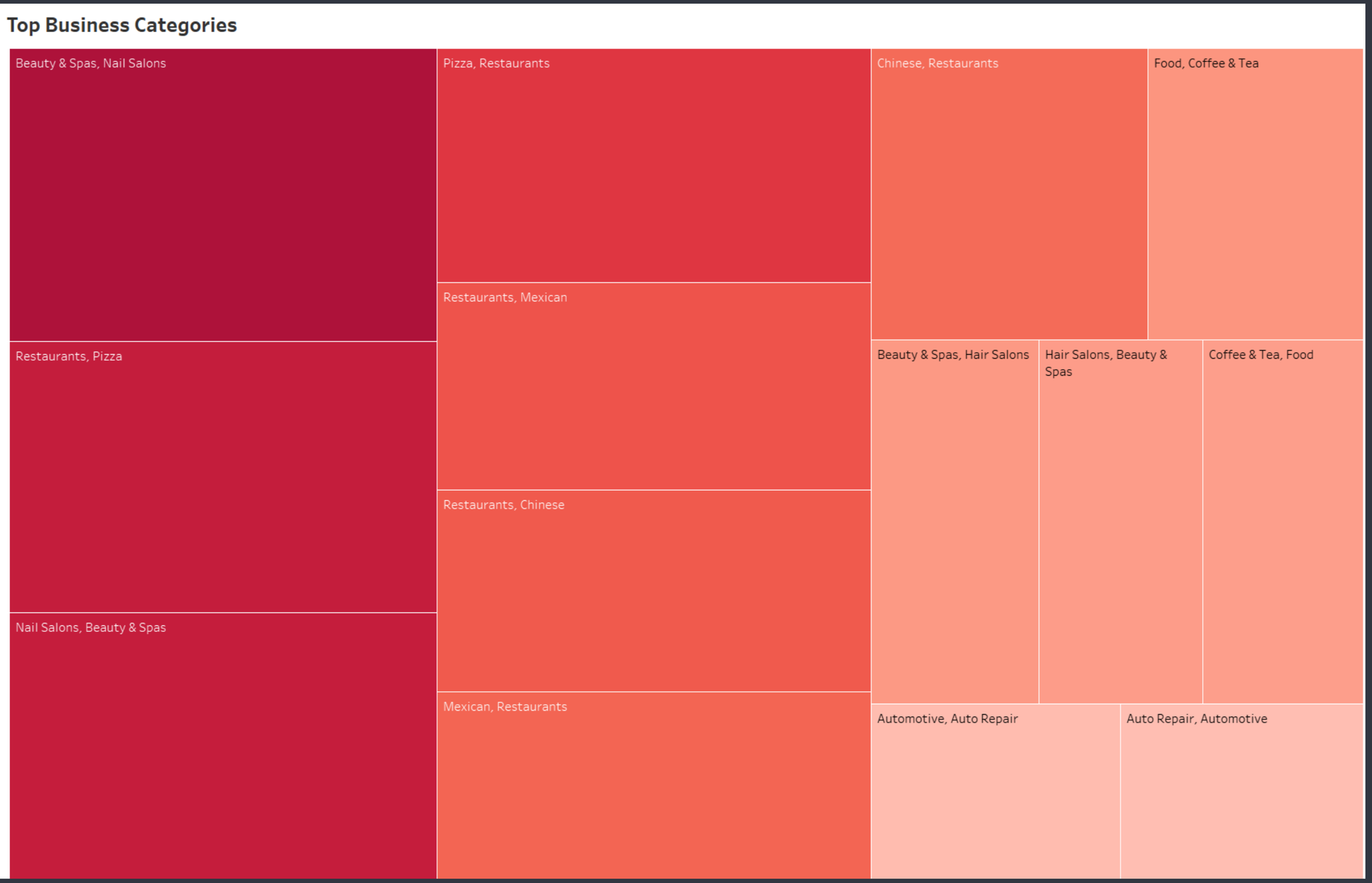


- By using color gradients for star ratings, you can quickly assess businesses' overall ratings.

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



Top Business Categories



Top Business Categories

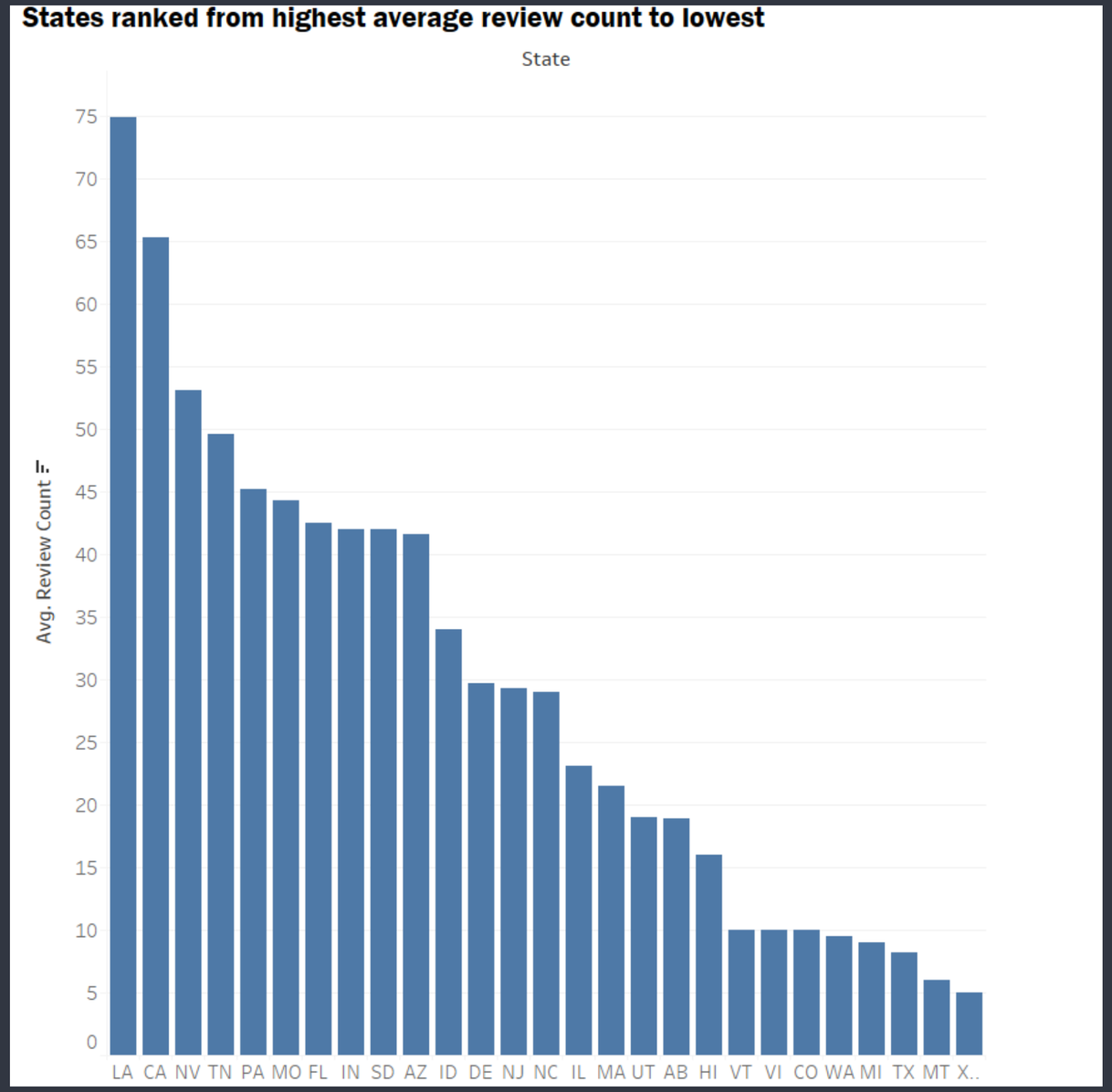
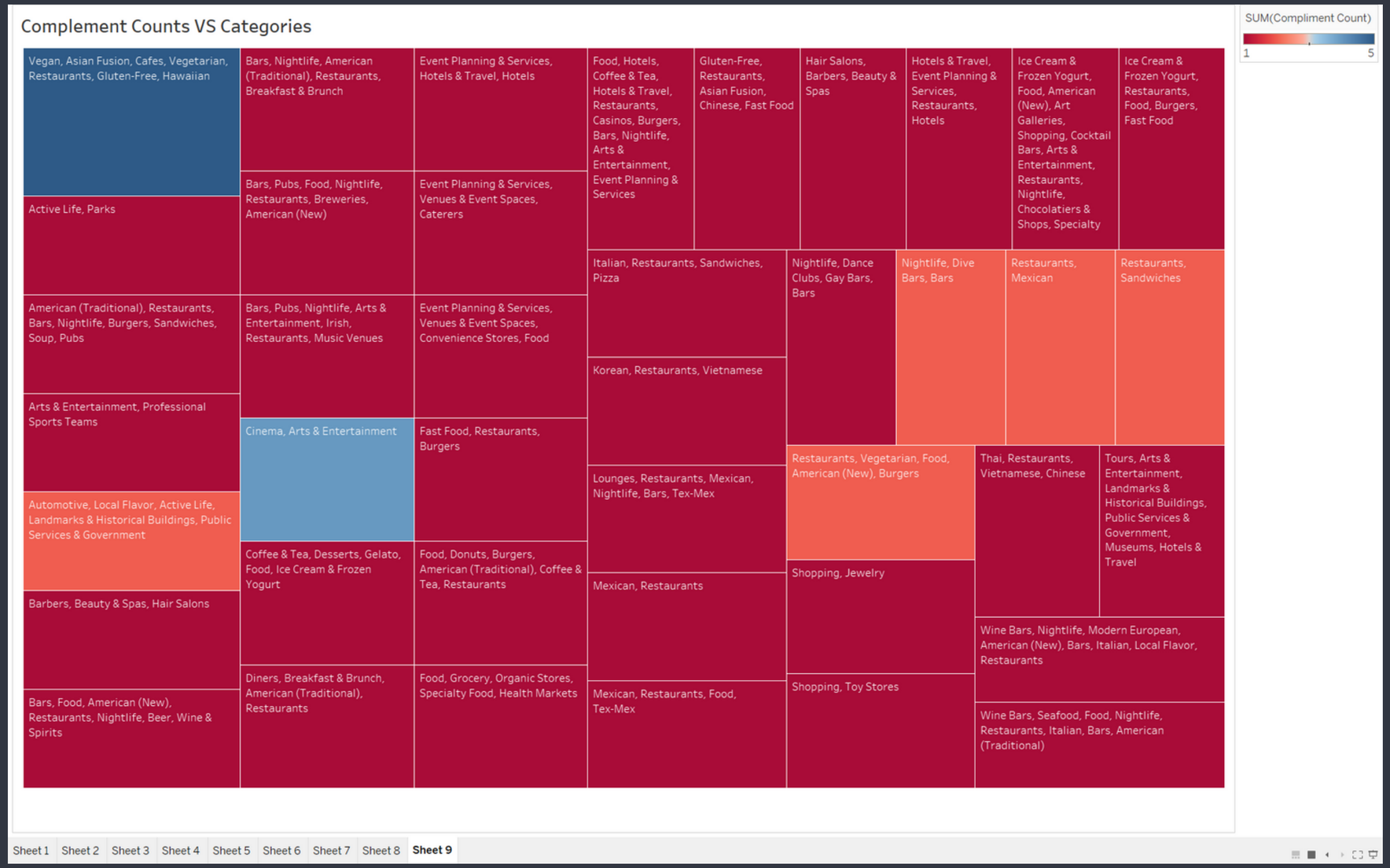
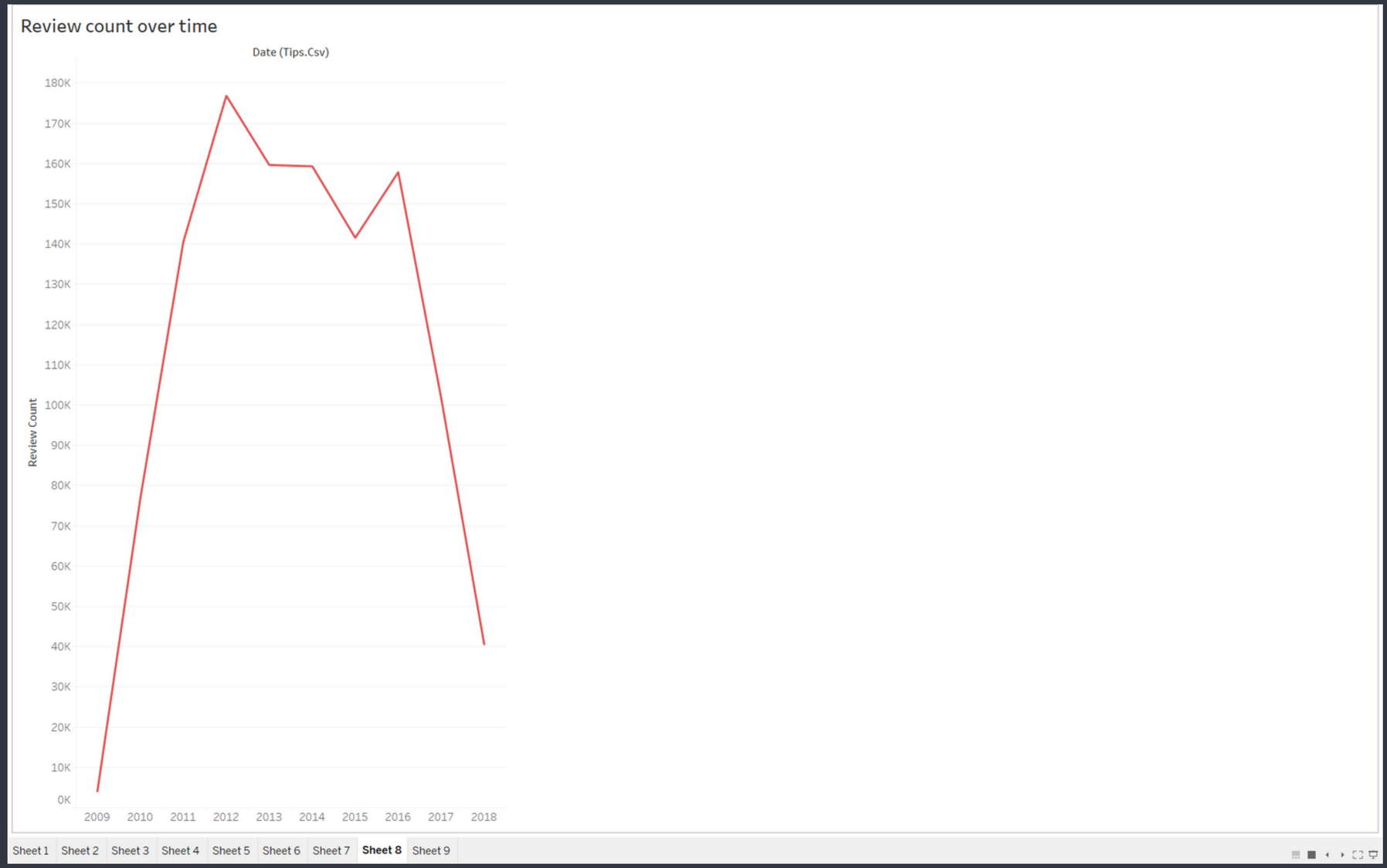


Chart shows different states and their average review count providing valuable insights into how businesses in various states are performing based on customer reviews. This visualization can help identify states with higher or lower average review counts, potentially indicating the overall satisfaction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