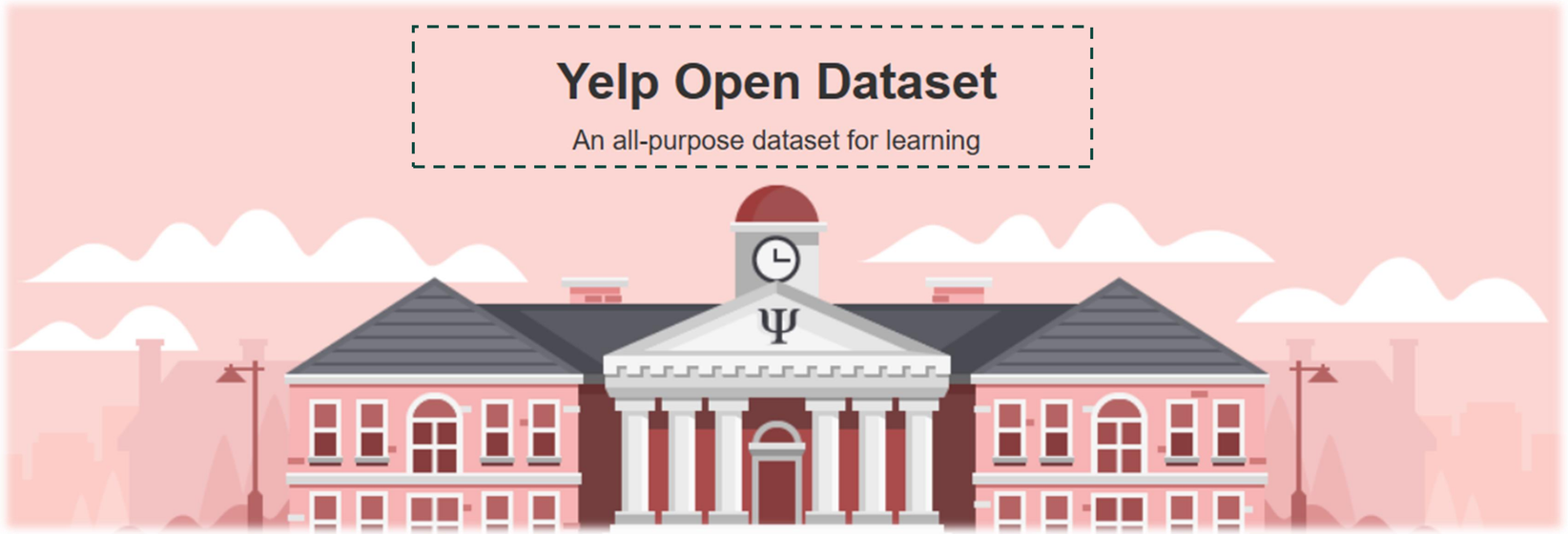
Project Mid-BDM 1024 - Data Technology Solutions



GROUP B
Adriana Marcela Penaranda Baron
Carlos Rey Pinto
Eduardo Roberto Williams Cascante
Haldo Jose Somoza Solis
Ishika Sukhija
Luis Alejandro Gutierrez Hayek
Marzieh Mohammadi Kokaneh
Nilesh Khurana

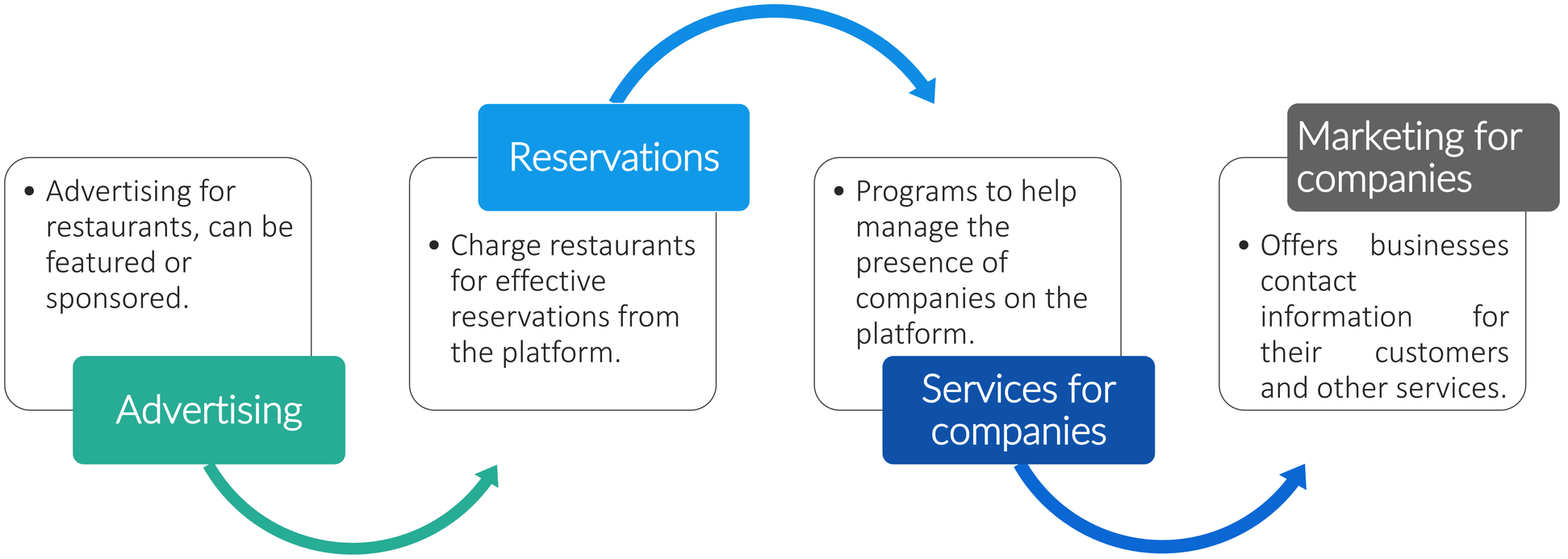
YELP Overview

- Yelp is a community-driven platform that connects people with great local businesses.
- Yelp facilitates effortless discovery, connection, and transactions between consumers and businesses in a wide array of categories, including requesting service quotes and booking restaurant tables.

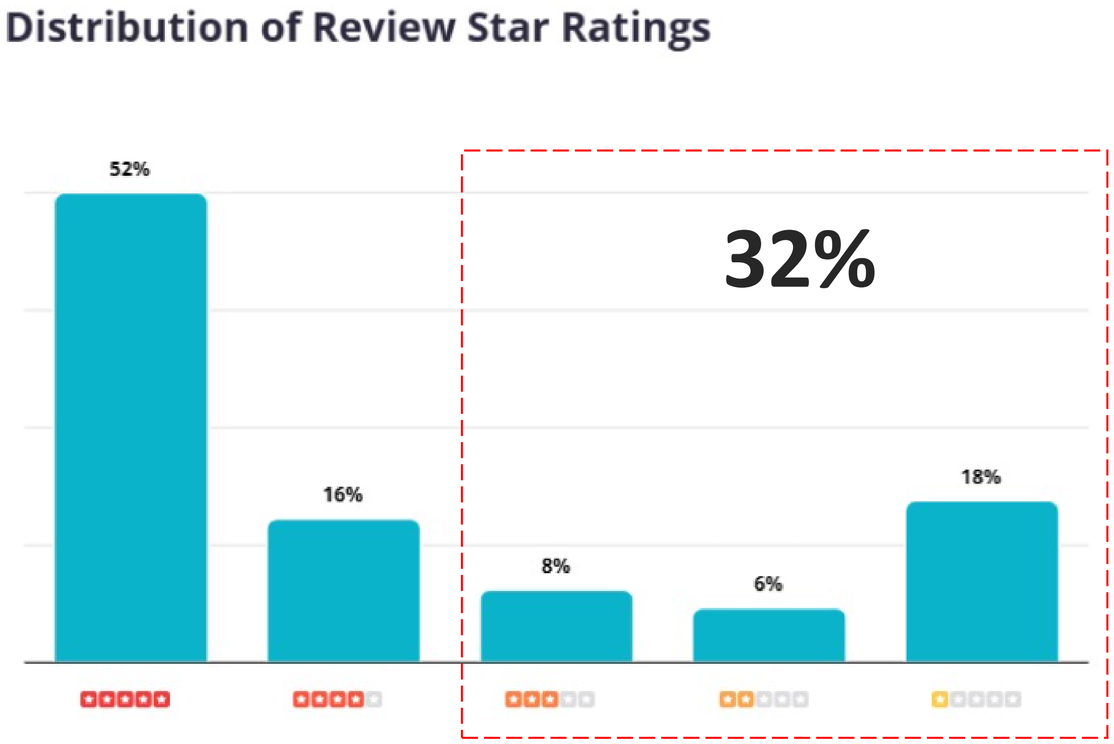
$312M
Q1 2023 Net Revenue

$1M
Q1 2023 Net Loss

YELP main sources of income



Opportunity the project is addressing

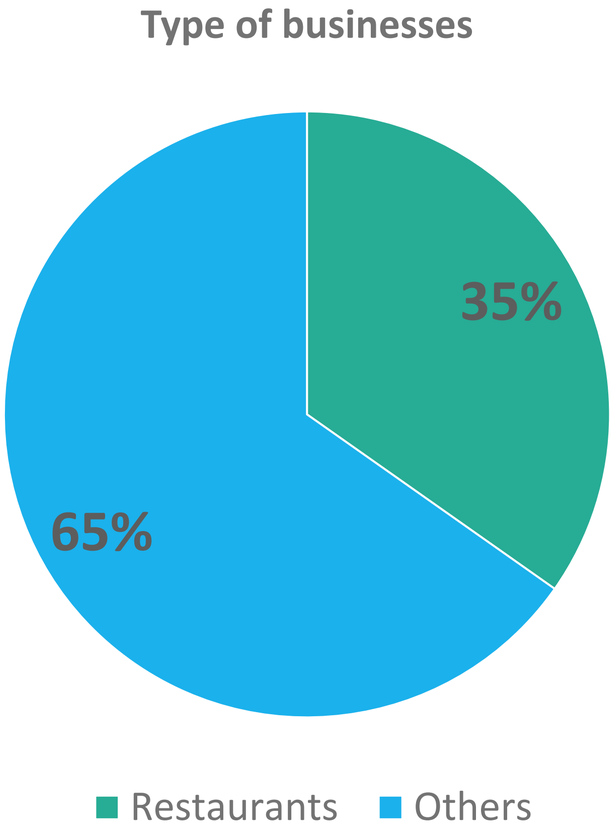


(Yeld 2023). Distribution of Review star ratings.
https://www.yelp-press.com/company/fast-facts/default.aspx

32% of businesses on the Yelp platform have a review start rating under 3, which affects their online reputation.

YELP should be concerned with improving the business rating in order to demonstrate the success of its platform and attract additional customers looking for reputation management solutions. Increasing customer ratings benefits Yelp and businesses, resulting in better revenues.

Pilot



Restaurants represent 35% of the dataset, which is a significant percentage of the categories Yelp manages on its Platform

This data analytics project will offer Yelp the opportunity to focus on helping struggling businesses improve their online reputation by analysing customer feedback, identifying recurring issues, and recommending actionable strategies for enhancement.

Why Yelp should consider us as the best solution?

01

Cloud implementation

Our analytical products could offer the opportunity to be executed in the cloud, which gives Yelp analytics in a short time!

02

Integrates with Yelp's platform

We have the capacity to extract the data from their platform and add it to different analytical tools for faster and better results.

03

Uses the latest technology tools

We use great technology for this Pilot such as Python, PostgreSQL, Cloud from Azure Services, and for the next step we will use Apache Hadoop technologies.

04

Competitive pricing

Our solution implemented in the CLOUD offers "Pay-as-you-go," which means you pay only for what you use, so they won't need to invest in infrastructure, which makes us competitive!

Main Steps

START

1 Dataset evaluation matrix

2 Design the PostgreSQL database (yelp Dataset Dictionary and Model Diagram)

3 Design a python program to transform the information from. json to PostgreSQL database

4 Create the PostgreSQL database and tables with relationships in the Cloud (Azure)

5 Import the information (.json documents) to the Azure database (about 8.9M of rows migrated in 6 tables)

6 Issues when importing the dataset, but We finally overcame them!

7 Dataset cleaning process

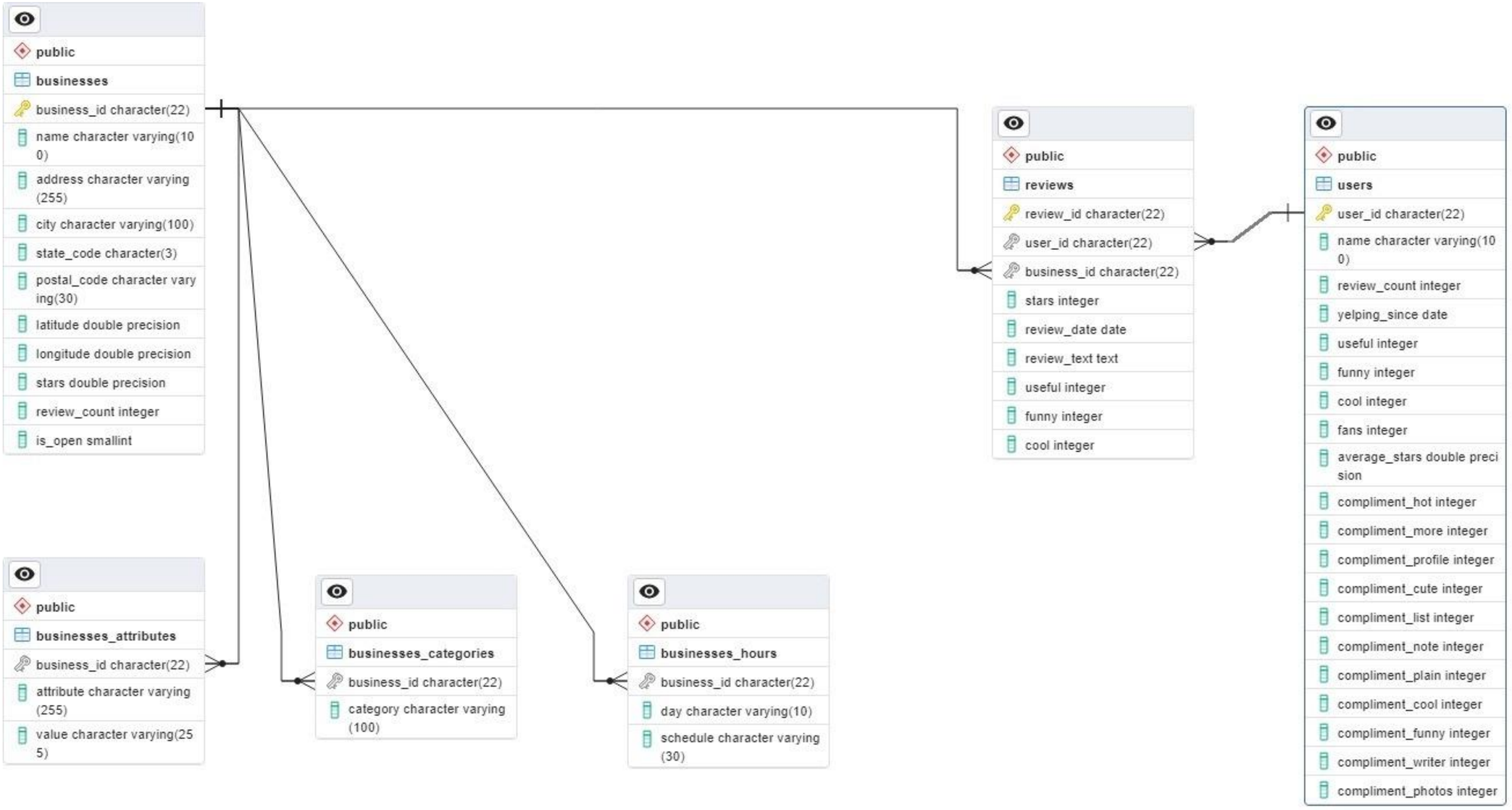
8 Dataset analysis

9 Insigts for yelp!!

END

Data Model Diagram

Entities (11) →Relationships →Decision Tree Analysis → **ER Diagram** (6 entities)



Queries and Insights

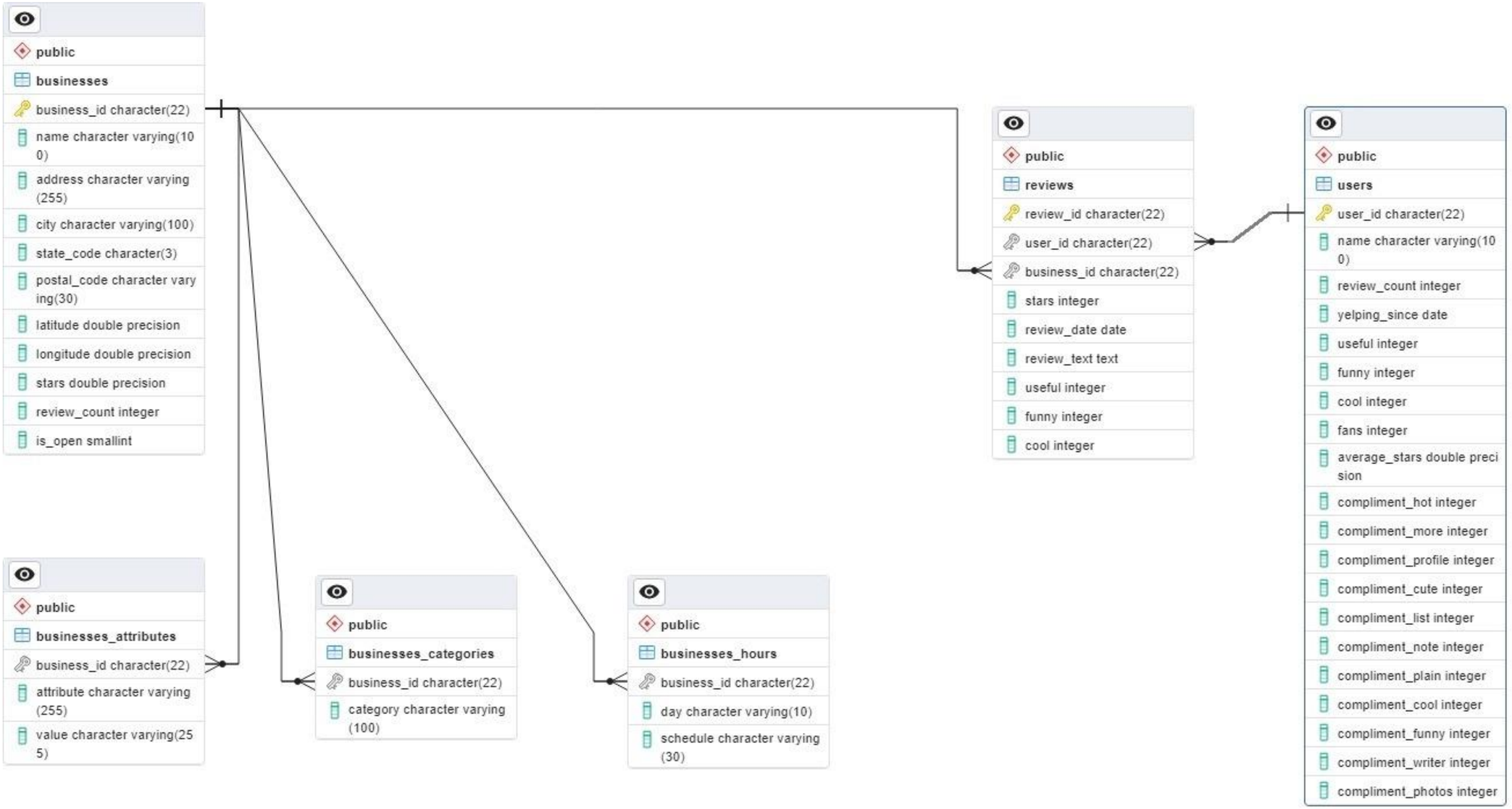
N°	Query	Output	Conclusions
1	Find the restaurants with the highest stars	1553 restaurants with a rating of 5 stars	Identify these restaurants' best practises and encourage businesses with low ratings to replicate them and enhance their reputation.
2	Identify businesses with the worst ratings (<=2) and analyze reviews to identify recurring issues	56,079 restaurant reviews were found with a star rating equal to or less than a score of 2 for 765 restaurants.	This output facilities our product to focus on this 765 restaurants to identify patterns and specific areas where businesses can improve. This may include aspects such as the quality of the products or services, customer service, cleanliness, waiting time, among others.
3	Identify the number of reviews related to the issues "customer service", "clean" or "waiting time"	19,667 of 56,079 restaurant reviews that represents 35%	This result support us in identifying the major issues that users are concerned about, resulting in a negative review and rating, and Yelp needs to work on them.

Queries and Insights

N°	Query	Output	Conclusions
4	Identify the restaurants where comment issues "customer service“ and has more than 10 reviews	58 restaurants	These restaurants are part of the priority of this analysis to identify what is making users give them a bad rating, and the recommendation is that Yelp prioritise executing the improvement strategies on them. The solution may include setting quality guidelines for Yelp-listed restaurants, providing additional training for customer service staff, and encouraging improved cleaning and customer service standards at partner establishments.
5	Identify the restaurants where comment issues “cleanliness“ and has more than 10 reviews	45 restaurants	
6	Identify the restaurants where comment issues “waits“ and has more than 10 reviews	296 restaurants	
7	Find the most active users and reward them with special promotions to build their loyalty	The query shows the first 100 most active users	After implementing the improvements in its services, YELP will be able to recommend the most active users visit these restaurants and rate them in order to improve their score and start a loyalty programme where these same people attract more people.

Data Model Diagram

Entities (11) →Relationships →Decision Tree Analysis → **ER Diagram** (6 entities)



Spark jobs

N°	Spark job	Output
1	Identify what categories are over and under the general average rate	1.qty_categories.csv 2.avg_categories.csv
2	Standard variances among other the categories	3.standard_deviation.csv
3	Relation between cities and categories in terms of high and low ratings or extremes	5.cities_categories.csv

eduwil Term1 folder

Code | Blame | 124 lines (104 loc) · 5.15 KB

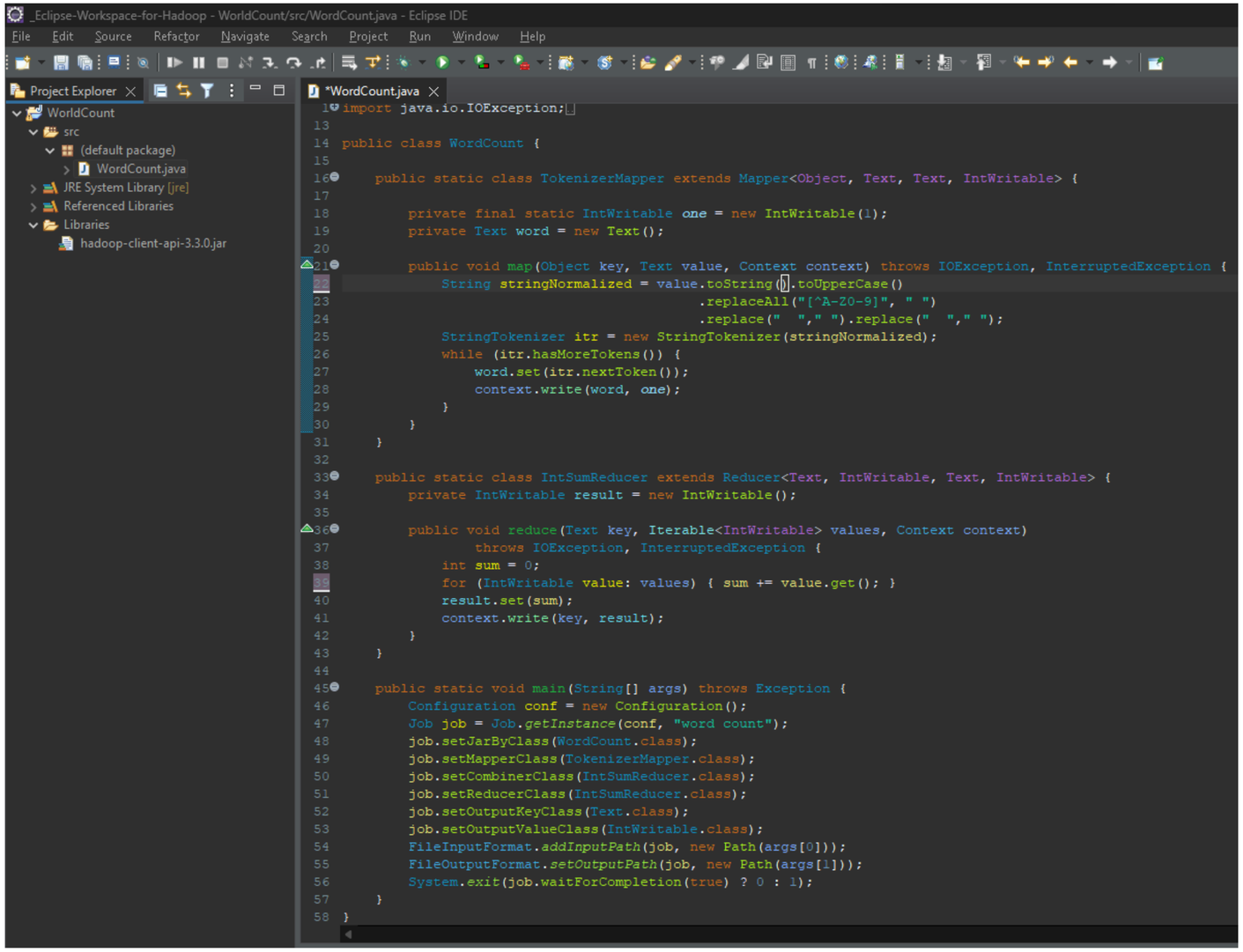
```python
# import libraries
import sys
import pyspark as ps
from pyspark.sql           import SparkSession
from pyspark.sql.functions import *

print ("Importing pyspark libraries...OK")
print ()

# retrieve command line arguments and store them as variables
datadir    = sys.argv[1] # gs://dataproc-staging-us-central1-321442252608-e66zqwhf/yelp/data/
outputfile = sys.argv[2] # gs://dataproc-staging-us-central1-321442252608-e66zqwhf/yelp/results
print ("Retrieving command line arguments and store them as variables...OK")
print ()

# Defining spark/sql context
sqlContext = SparkSession.builder.getOrCreate()
print ("Defining spark/sql context...OK")
print ()

# loading csv files
df_reviews    = sqlContext.read.format('com.databricks.spark.csv').options(header = 'true', inferschema = 'true').load(datadir + 'reviews.csv')
df_categ_rest = sqlContext.read.format('com.databricks.spark.csv').options(header = 'true', inferschema = 'true').load(datadir + 'rest_categories.csv')
df_categ      = sqlContext.read.format('com.databricks.spark.csv').options(header = 'true', inferschema = 'true').load(datadir + 'categories.csv')
print ("Loading csv files...OK")
print ()

# Getting restaurant categories
lst_rest_cat = df_categ_rest.select("category").rdd.flatMap(lambda x: x).collect()
print ("Getting restaurant categories...OK")
print ()

# Filtering by restaurant categories
df_categ = df_categ.where(df_categ.cat_category.isin(lst_rest_cat))
print ("Filtering by restaurant categories...OK")
print ()
```

Map reduce jobs

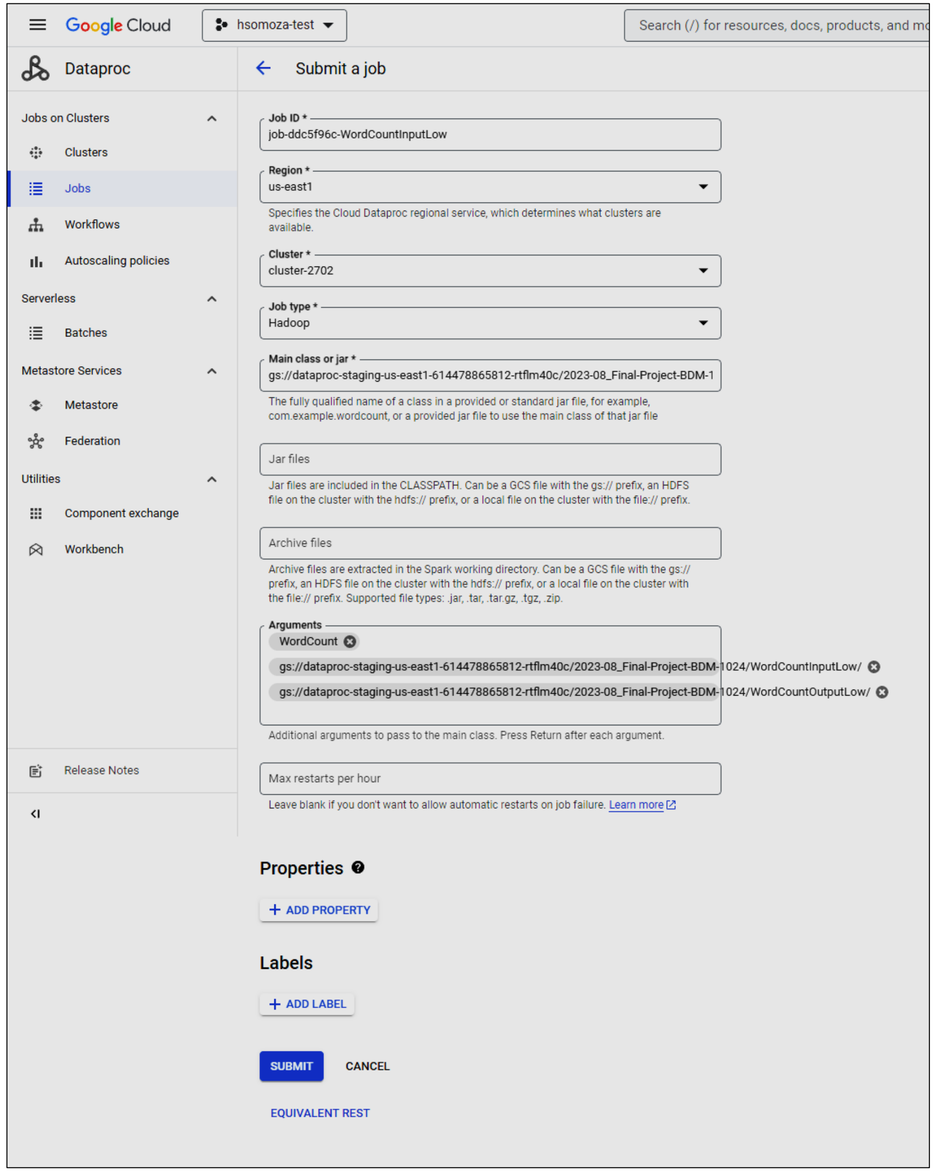
Input: Spark Jobs **results** to identify what categories are over and under the general average rate.

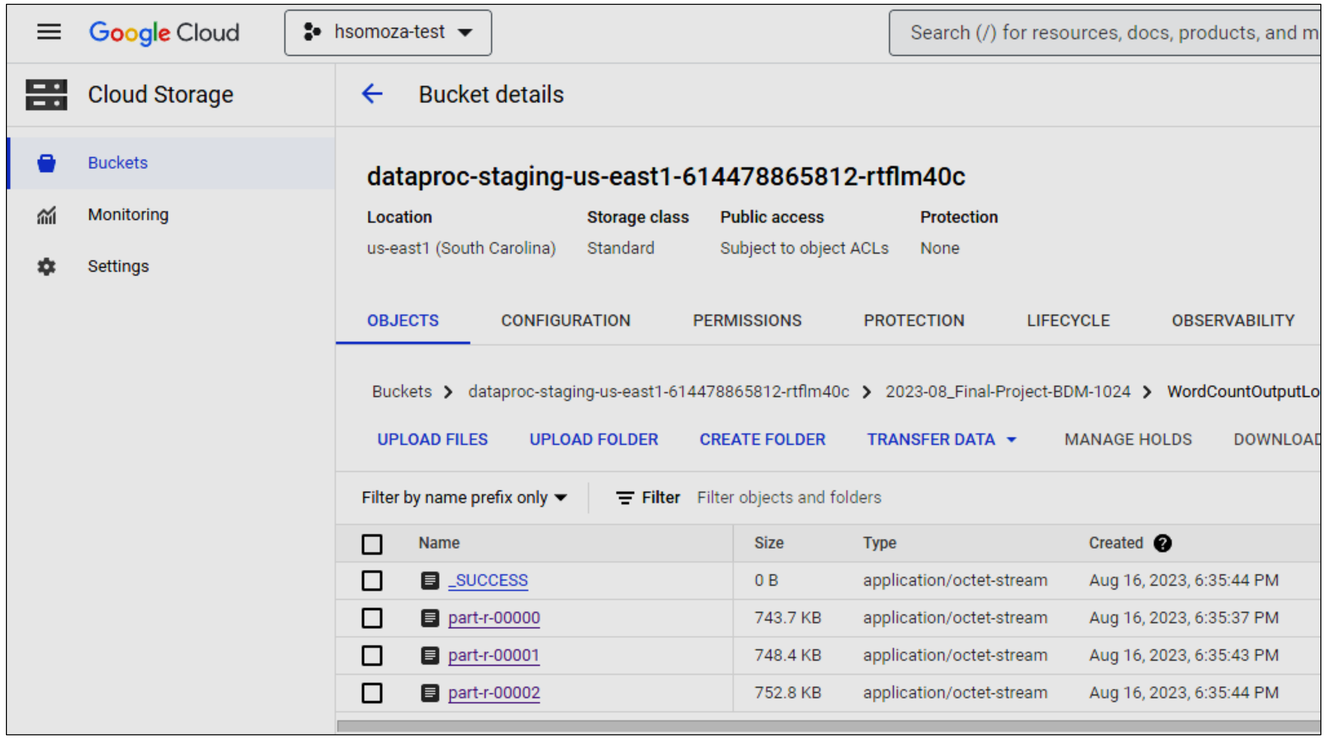
01 **Identify the keywords from reviews of the highest-rated restaurants**

02 **Identify the keywords from reviews of the lowest-rated restaurants**

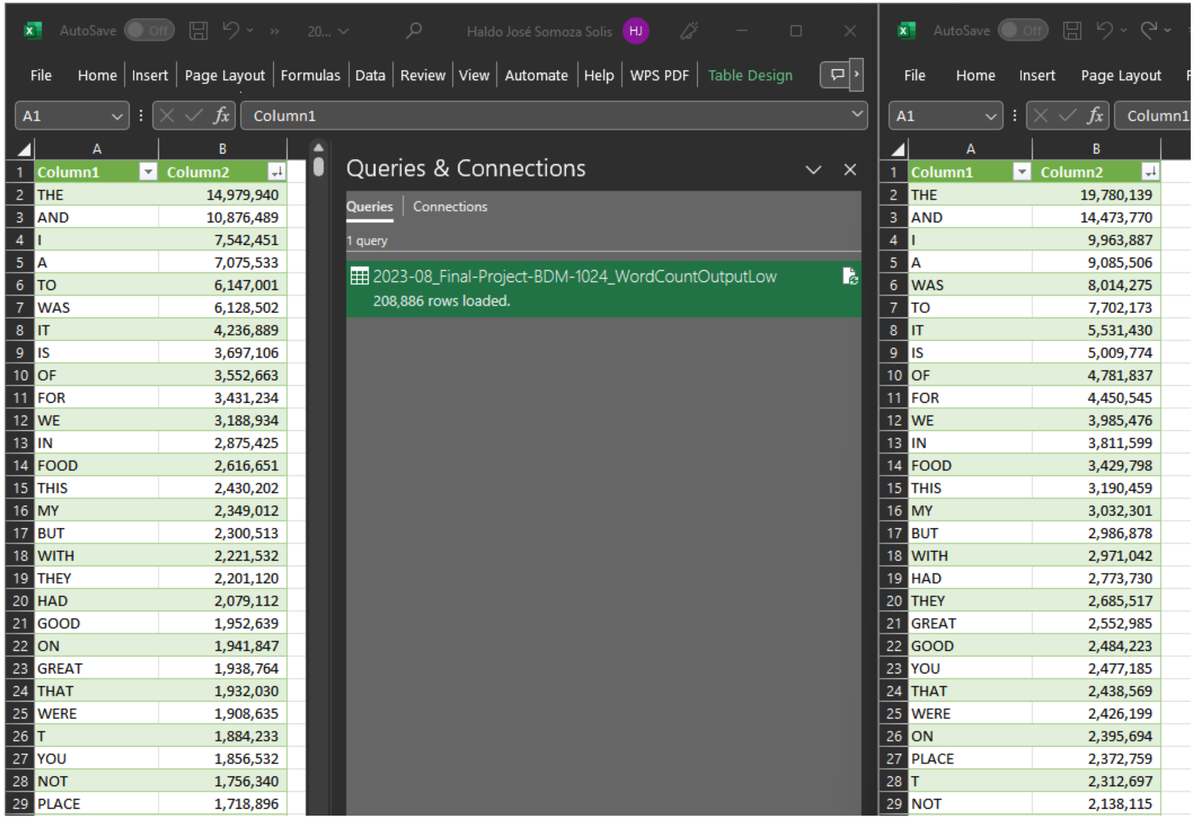


Map reduce jobs





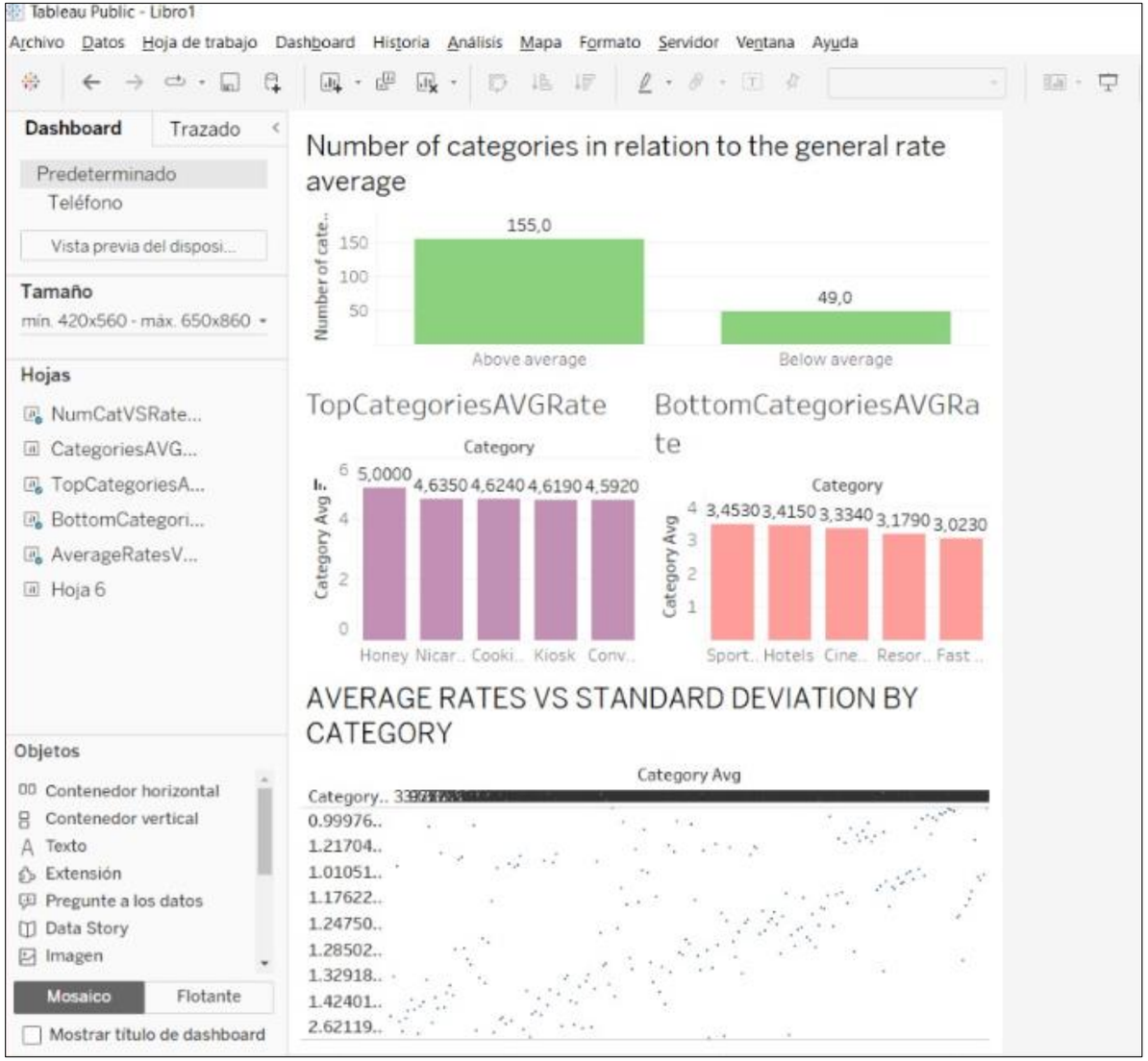
Map reduce jobs



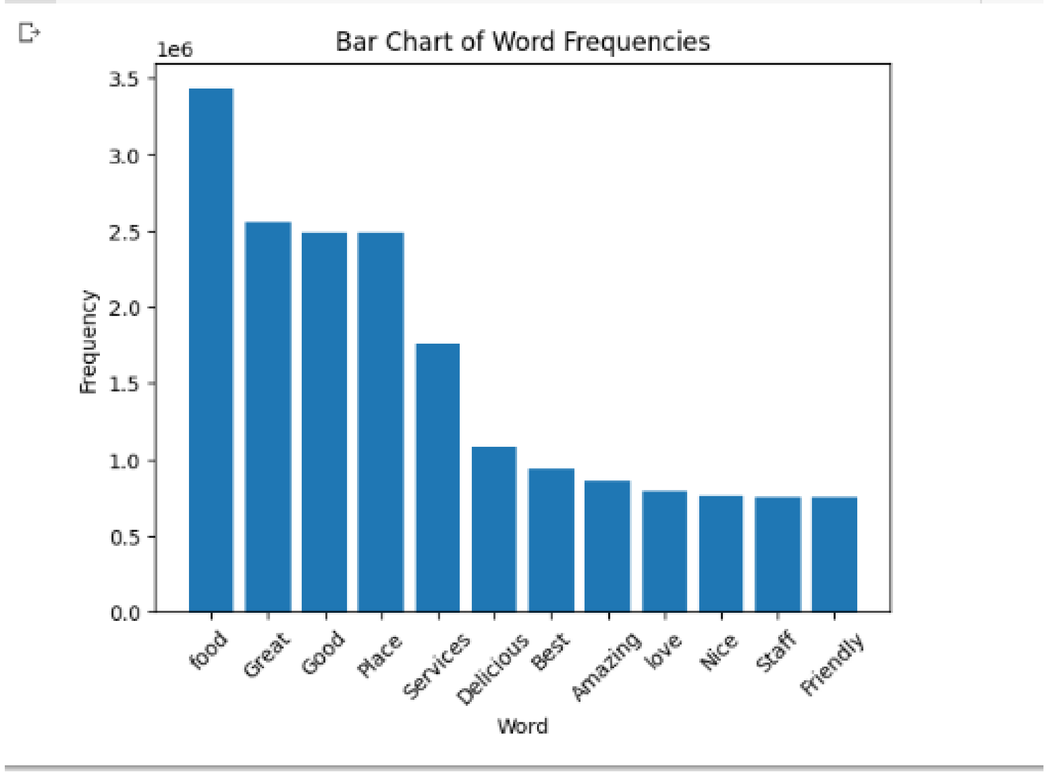
	A	B	C	D
1	Column1	Column2		
14	FOOD	3429798		
20	THEY	2685517		
21	GREAT	2552985		
22	GOOD	2484223		
27	PLACE	2372759		
35	SERVICE	1758807		
50	DELICIOUS	1087390		
56	BEST	943263		
61	AMAZING	857580		
69	LOVE	793568		
70	GOT	779726		
71	VE	768092		
72	NICE	767495		
73	BEEN	762117		
74	DEFINITELY	751551		
75	STAFF	751292		
76	FRIENDLY	751161		
77	CHICKEN	741094		
78	UP	736032		

Visualizations

Input: Spark Jobs results



Input: Map Reduce Jobs results



To conclude **our solution focus** on customer rating enhancement that creates a mutually beneficial environment, **improving Yelp's reputation and generating greater revenue opportunities for businesses and for Yelp as well!**