Yelp Business Reviews Project

Dataset Overview:

Dataset link - https://business.yelp.com/data/resources/open-dataset/

Direct_Download - This download contains 1 compressed TAR file (4.35 GB). Uncompressed, it contains 1 PDF and 5 JSON files (8.65 GB). Documentation included.

Data Preparation & Infrastructure:

1.After downloading and extracting the dataset, split the reviews file into 10 parts (optional), the code for it is posted in the Github Repository.
2.Uploaded the file/files to AWS S3 bucket within the yelp reviews folder.
3.Logged onto snowflake and created a database for the project.
4.First tested out the sentiment analyzer, then created tables by extracting data from the AWS S3 bucket.
5.To extract data directly from aws s3, the user needs to get aws credentials (aws key id and aws secret key).
6.Created table of Business in Yelp dataset on worksheet 2 and table of Yelp reviews on worksheet 3
Once the tables were created, ran some queries on worksheet 4 on snowflake.

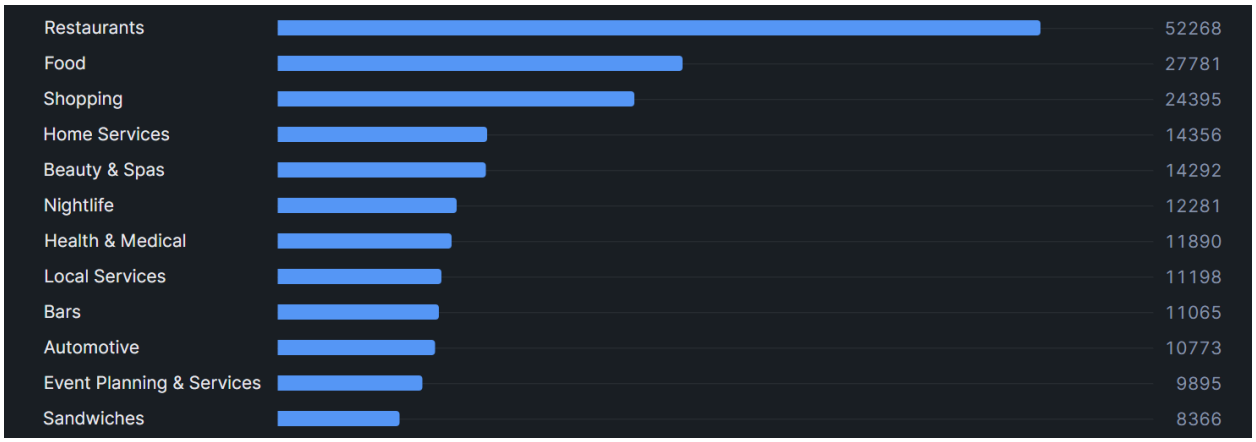
Links for the code -
Code for splitting dataset (optional step)
Sentiment analyzer
Creating yelp business table
Creating yelp reviews table
Queries

Following are some of the queries and insights gathered. Snowflake returns visuals for the queries without needing any extra coding. Just one tap that feature was quite useful for insights compared to offline sql workbenches where the results are often in tabular form and require extra steps to visualize them.

Some results have been restricted to top results only.

1. Finding the number of businesses in each category

Displaying the top results

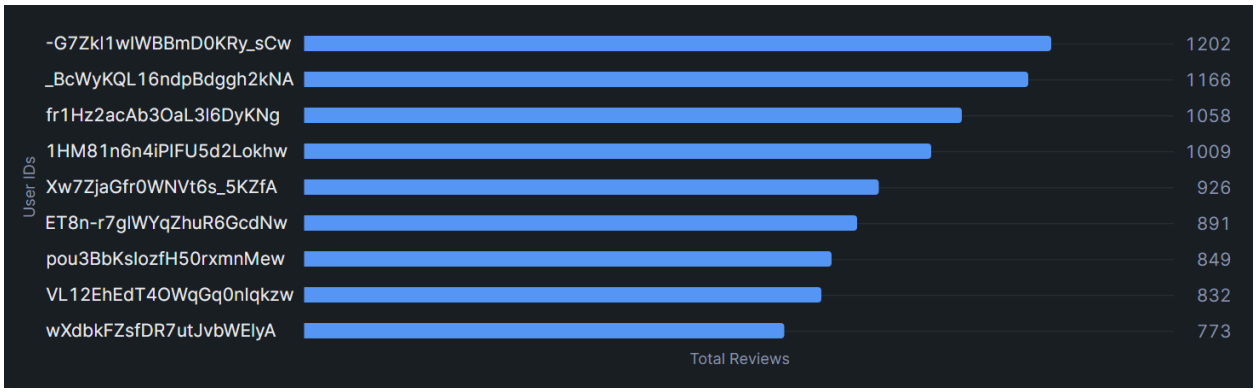


2. Find the top 10 users who have most reviews

These User Ids can be used to find out actual user names from the users table in the dataset, But for our project we have stuck to just user ids .

3. Find the top 10 users who have reviewed the most businesses in the "restaurant" category.

This helps us understand people who have been frequently visiting / rating the restaurants, which can help us provide them with a special badge, etc to promote such users and even as an appreciation for their contribution.



4.Find the most popular categories of businesses (based on the number of reviews)

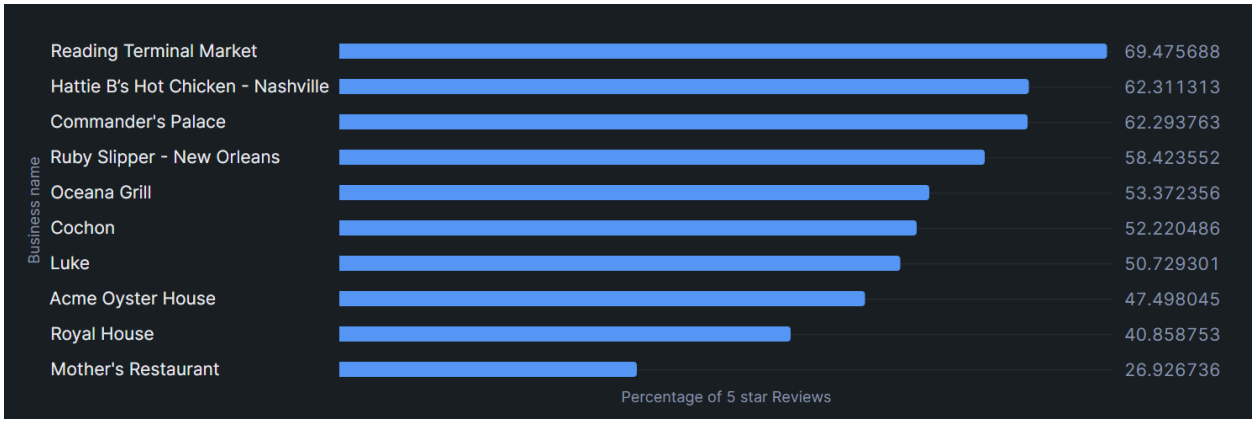
Displaying the top categories here helps us identify what sorts of business get reviewed most frequently, As we can see most of them include some sort of cuisine or happen to be restaurants, reflecting that users drop reviews for restaurants and food related businesses more frequently than any other category.

6. Find the month with highest number of reviews

Months of July, August, January and June are the top 4 months during which users write a review. June, july and August being the summer months in the US shows us that people tend to review places when they are on a vacation.

7. Find the percentage of 5 star reviews for each business.

Here we have identified the businesses with the highest percentage of 5 star reviews. This helps us identify top performers and publicly well perceived businesses.



8. Find the top most reviewed cities

This insight helps in understanding cities with the highest amount of reviews being posted indicating a high usage of the app and helps in identifying cities where app could be marketed / reviewers could be incentivised.

9. Find the average rating of businesses which have at least 100 reviews

This query returns a list of top rated businesses with at least 100 reviews. The list is in the descending order of AVG_RATING column.

	BUSINESS_ID	NAME	TOTAL_REVIEWS	AVG_RATING
1	_ab50qdWOk0DdB6XOrBitw	Acme Oyster House	7674	4.124967
2	ac1AeYqs8Z4_e2X5M3if2A	Oceana Grill	7517	4.146202
3	GXFMD0Z4jEVZBCsbPf4CTQ	Hattie B's Hot Chicken - Nashville	6161	4.446356
4	ytynqOUb3hjKeJfRj5Tshw	Reading Terminal Market	5779	4.605468
5	oBNrLz4EDhiscSlbOI8uAw	Ruby Slipper - New Orleans	5265	4.291358
6	iSRTaT9WngzB8JJ2YKJUig	Mother's Restaurant	5255	3.438820
7	VQcCL9PiNL_wkGf-uF3fjg	Royal House	5147	3.786672
8	_C7QiQQc47AOEv4PE3Kong	Commander's Palace	4970	4.292153
9	GBTPC53ZrG1ZBY3DT8Mbcw	Luke	4662	4.177392

10. List the top 10 users who have written most reviews and the businesses they have reviewed

This list helps in understanding the top reviewers on the platform and the businesses they have reviewed.

	USER_ID	BUSINESS_ID
1	04qAdlAl5WXGHLa7fvN_wQ	dRTzm4V5hcbkYgLlV6kC4w
2	G8y49dNr1sGXvglqEM9DXA	Sm5vf0m1EA4n7sMakmWTsg
3	O_nvi6zqeEg1k8-2jUev1A	0WrdiQ4qJBu2Gc7wRclfdw
4	P7QOegOMb0QW1TjzhkUR3w	4w_TG6uA5blxnC5tpkOb3g
5	Pbk3hZ3Ap3TvaBEVTmn2xQ	0N2D4HtF2gLJvWxjlfiZDg
6	TEPZ3CQ9YxDX4J4d-oNiQA	im1dUk7C1YoqT-KPKBF4Ng
7	cCbzNIHAWmpxpYZRstmeqw	UUpwoAAZcJyUWOqp81_3mA
8	g4TyvMhL20-DhXpxlpjQpg	rbddEDXRlVu7MyTyEEYp9g
9	m_QRydHQ7eFHaQB2TznUzg	Swjkmg5qAHT6hwXQRWQoHQ
10	x7Cr2uvyc3GOllH_yifx4g	haPszlAbTLcUbpwHplOPsg

12.Find top 10 businesses with highest positive sentiment reviews

Here the sentiment analyzer is used to identify the overall sentient of reviewers towards the business by letting the sentiment analyzer read through the words mentioned in the reviews posted.