Exploring Yelp Toronto Businesses and Their Reviews in Relation to Nearby Attractions

Naura Izzah Taufiq (1009713669)

A. Introduction

Tourism plays a major role in shaping the economy of Toronto, Canada's largest city and a popular tourist destination. Understanding how proximity to attractions influences business success can offer valuable insights for business owners, urban planners, and tourism boards. This study investigates whether businesses near Toronto's tourist attractions perform better in ratings and review volume compared to those farther away. It also examines whether certain attractions are linked to higher ratings and more positive reviews. Additionally, the project explores how review sentiment differs between businesses located near and far from popular attractions, along with analyzing Toronto businesses' overall review sentiment.

This analysis uses three datasets. The first is the Places of Interest and Toronto Attractions dataset from Toronto Open Data, sourced from the Toronto Tourist Information Centre. The other two are Yelp datasets from Kaggle, originally from Yelp Open Dataset. These include Yelp Business data (business details like name, location, rating, review counts, categories) and Yelp Review data (filtered to only include Toronto business reviews).

By merging these datasets and applying data analysis techniques, this study aims to explore the relationship between business performance and proximity to attractions, while also analyzing general sentiment trends in Toronto business reviews.

B. Methods

B.1 Loading and Merging Datasets

The datasets for this project were obtained from the Toronto Open Data API and Kaggle (Yelp Open Data subsets). The Toronto Open Data API was accessed via the CKAN interface to retrieve the "Places of Interest and Toronto Attractions" dataset, which provides information

such as attraction name, category, and geographical coordinates as well as some other details like contacts, websites and a short bio. The Yelp Business and Review datasets were downloaded then read. The Business dataset provides information about business names, locations and coordinates, ratings, categories, and review counts. Finally, the Review dataset offers textual reviews, ratings, and dates as well as the user and business information.

Before proceeding with creating new variables and merging the datasets, it is crucial to address duplicate entries within each dataset to prevent inflation (double duplicates) of the merged dataset. For the attractions dataset, duplicates were identified by comparing entries with the same name. For the business dataset, duplicates were detected by examining rows with matching business names, addresses, neighborhoods, and postal codes since a chain will have the same names but diffent locations. Finally, for the reviews dataset, duplicates were identified by comparing user IDs, review texts, business IDs, and ratings as one user may write the same texts for other businesses.

The most significant issue was found in the business dataset, where 356 duplicate rows were identified. Upon observation, it seemed like this might be caused by update in business information (perhaps the opening of a new renovation/closing of the business, or updating the review count). The most significant issue was 356 duplicate rows in the business dataset, likely due to updates in business information. We retained the most recent data based on whether the business is open or has the highest review count. This reduced the total entries from 174,567 to 174,387 (note that this is still not exclusive to Toronto only).

WWe filtered Yelp Business data to only include Toronto businesses and merged it with the attractions dataset by finding the nearest attraction for each business. Missing values were first addressed by removing entries lacking coordinates information since it was pertinent for this step. Then, to identify the nearest tourist attraction for each business, we constructed a KDTree spatial index using the attraction coordinates. Each business was linked to its nearest attraction, and the distance between them was calculated. This result in a dataset that contains Toronto business details and the nearest attraction information for that business (id, name, category and its distance from the business).

We also merged the Toronto business data with the reviews data such that now we have more details on the business that the user reviewed upon. That concluded this loading and merging dataset step (primarily using pandas), and we are ready to do exploratory data analysis.

B.2 Data Wrangling and Exploratory Data Analysis

We primarily used Pandas for data wrangling. We first observed the two datasets (Toronto business data with attractions information, Toronto reviews data with business details) by its dimensions and columns. The business data has 17,199 observations with 17 variables: business ID, name of the business, its neighborhood, its full address, city (Toronto), state (ON), postal code, latitude, longitude, average ratings, review count (total number of reviews), is_open (indicating whether the business is open or not), categories (which categories do the business

offer; may be restaurant/cafe and its cuisine_type, the services, etc.) alongside the nearest attraction details (IDs, name, category, and the distance to that business). Most variables are character/string object, except some numerical variables like coordinates (continuous), review counts (discrete), attraction IDs (discrete), is_open (binary number) and attraction distance in kilometer (continuous).

The reviews Data Shape has 430,985 reviews and 11 variables: review IDs, user ID that wrote that review, business ID and business name of the intended review, rating, date of the review, the review text, reactions (number of upvote for useful/funny/cool) and the categories of the business. Again, most variables are character/string object, except some numerical variables like rating (discrete from 1-5) and reactions. Note that we also decided to rename some variables for readability.

Upon observing the summary of the datas, we found some missing values for some columns like business neighborhood and postal code that we decided to ignore since it did not affect much of our later analysis. However, we realized there were some variables that needed further check.

1. Geographical filtering was applied to validate the coordinates, ensuring that businesses were located within Toronto's boundaries (latitude: 43.5-43.9, longitude: -79.6-79.2). We decided to remove anything outside since they were no longer in Toronto city.
2. There were two unique state so corrections were made for state code errors ('AB' to 'ON') to maintain consistency. It turned out to be a mistake since the address for that business was in Toronto.
3. There were duplicates in the review texts. We decided to perform extensive cleaning. First, we identified several texts occurring multiple times. To further investigate, we checked for exact duplicates where the same review text, business ID, and review ID were repeated. No exact duplicates were found, but some reviews were associated with multiple user IDs. Thus, we sorted reviews by date and retained only the oldest instance of each duplicated review text, assuming the earliest review was the original. We also removed cases where the same review text was associated with multiple users. Note that we did not remove the duplicated text for different business IDs as the rationale would be that it is possible for one user to use the same review for another business.

This process reduced the dataset from 430,985 to 430,906 entries, ensuring that the dataset accurately reflects genuine customer feedback for reliable analysis.

Finally, we conducted summary statistics of some variables and created exploratory graphs using matplotlib and seaborn libraries to see insights surrounding the initial questions. We defined businesses within 1 kilometer of a tourist attraction as "near tourist attractions" and those beyond this distance as "regular neighborhoods" and explored if there is a difference in business metrics. We also see if that has anything to do with the attraction type.

To gain insights from the review text data, we performed sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool from the nltk library. For each review, we produced a dictionary containing four sentiment metrics: positive, negative, neutral, and compound scores. The compound score, which ranges from -1 (most negative) to +1 (most positive), was extracted for further analysis.

Based on the compound score, reviews were labeled as positive (score > 0.05), neutral (score between -0.05 and 0.05), or negative (score < -0.05). This allowed us to analyze the relationship between sentiment distribution and user rating. The sentiment data was also merged with the business dataset to link reviews with their nearest attraction categories. The average sentiment score was calculated for each attraction category, also for those near attractions/regular neighbourhoods.

3. Preliminary Results

3.1 Businesses Near Tourist Attractions vs. Regular Neighborhoods

The analysis revealed interesting patterns in the data. Both businesses in regular neighborhoods and those near tourist attractions have similar average ratings (3.51 and 3.48, respectively). The boxplot (Plot 1) illustrates this. However, businesses near attractions receive significantly more reviews on average (29.93 vs. 17.37). Additionally, nearly twice as many businesses are located within 1 km of attractions (10,572) compared to those farther away (6,487) as written in Table 1. Plot 2 illustrates that businesses near tourist attractions not only have higher average review counts but also include more outliers with extremely high review counts (Table 1 also indicates higher review average). Some businesses near tourist attractions have received over 1,400 reviews, while the maximum for businesses in regular neighborhoods is around 450. This suggests that proximity to tourist attractions drives increased visibility and customer engagement, even if it doesn't necessarily lead to higher ratings.

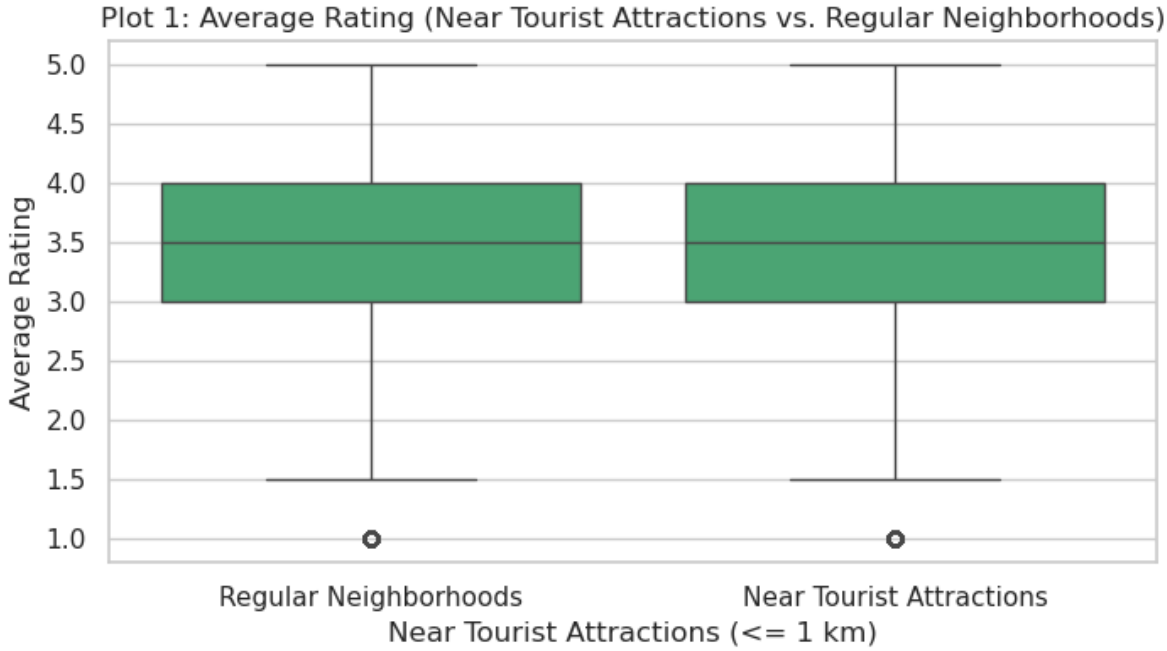
Table 1: Businesses Near Tourist Attractions vs. Regular Neighborhoods

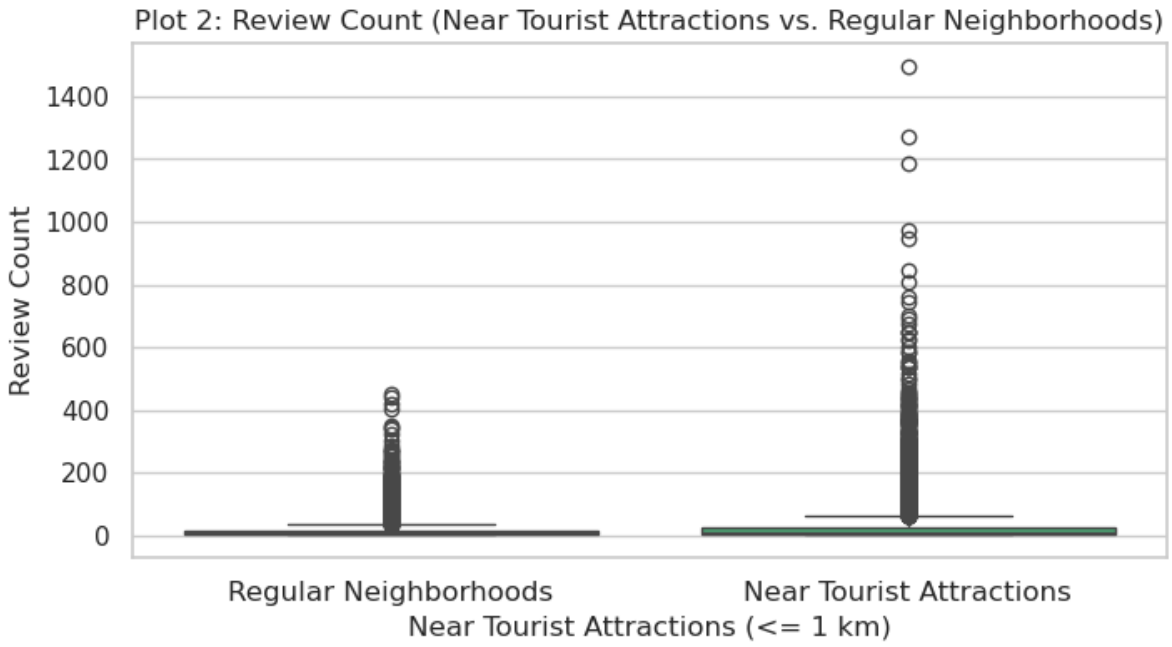
	Metric	Near Tourist Attractions (<= 1 km)	Regular Neighborhoods (> 1 km)
0	Average Rating	3.48	3.51
1	Average Review Count	29.93	17.37
2	Number of Businesses	10572.00	6487.00

	Metric	Near Tourist Attractions (<= 1 km)
0	Average Rating	3.48
1	Average Review Count	29.93

2	Number of Businesses	10572.00

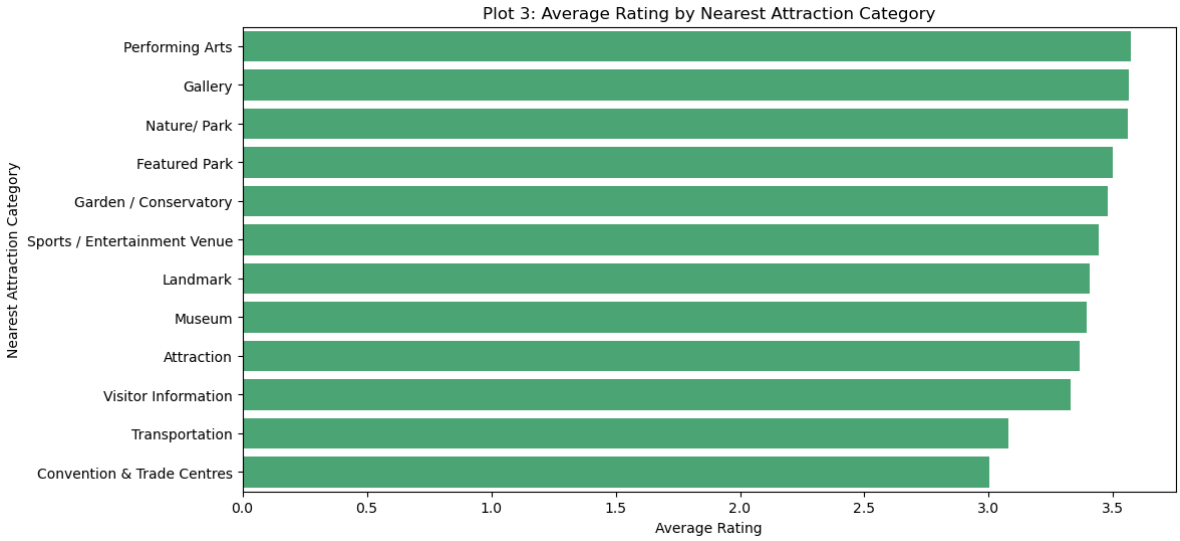
	Regular Neighborhoods (> 1 km)
0	3.51
1	17.37
2	6487.00





3.2 Business Performance by Attraction Category

Furthermore, businesses near cultural and nature-oriented attractions, such as performing arts, galleries, and parks, generally receive higher ratings and more positive sentiments compared to businesses near transportation hubs and convention centers (Plot 3). Businesses near performing arts have the highest average ratings (approximately 3.7), followed by businesses near galleries and nature/parks (both above 3.5). In contrast, businesses near transportation hubs and convention/trade centers have relatively lower average ratings (around 3.0-3.2). These findings suggest that certain types of attractions may be more conducive to positive customer experiences, possibly due to the relaxing or aesthetically pleasing environments they provide.



3.3 Sentiment Analysis of Reviews

The sentiment analysis of review texts revealed a strongly positive skew in the data. We can see that the majority of reviews express positive sentiment, and the minority is neutral (Plot 4). On Table 1, the exact count confirms this, where we have 373,503 reviews that express positive sentiment, making up 86.6% of the data we have (illustrated by Plot 5). This distribution shows the general tendency of customers to leave positive reviews on Yelp, whereas 12.1% express negative sentiment, and only 1.3% are neutral reviews.

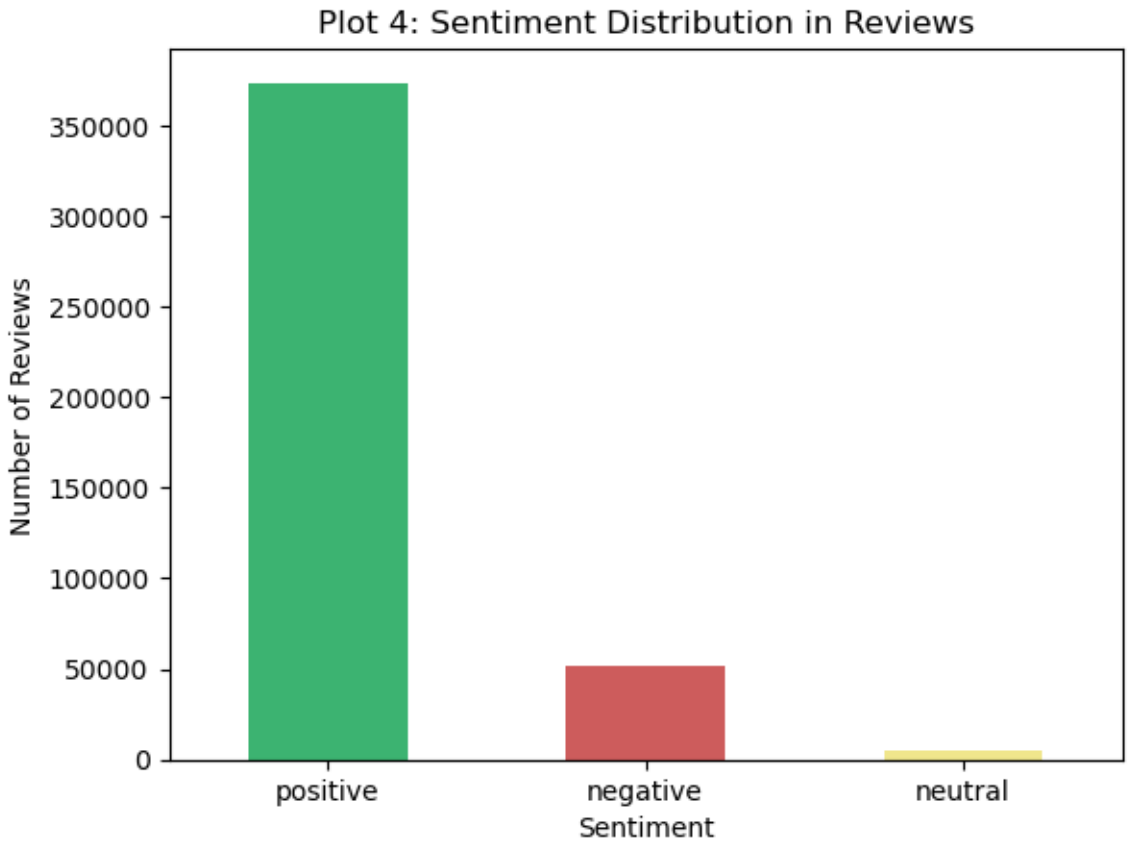
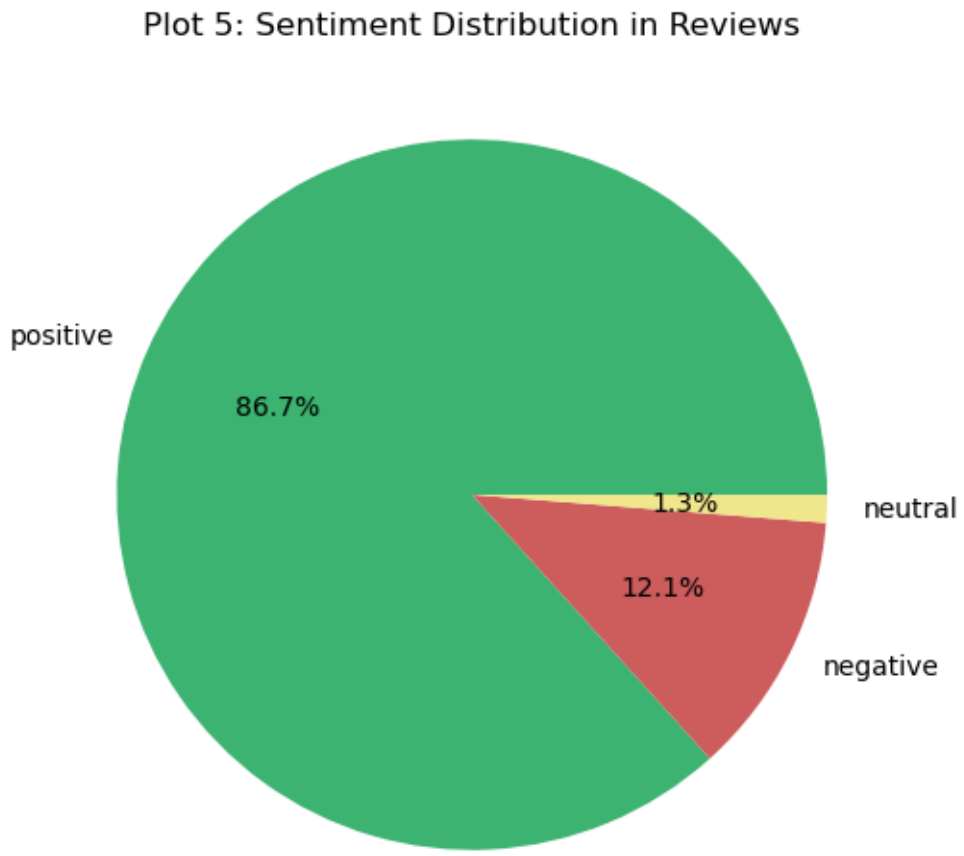
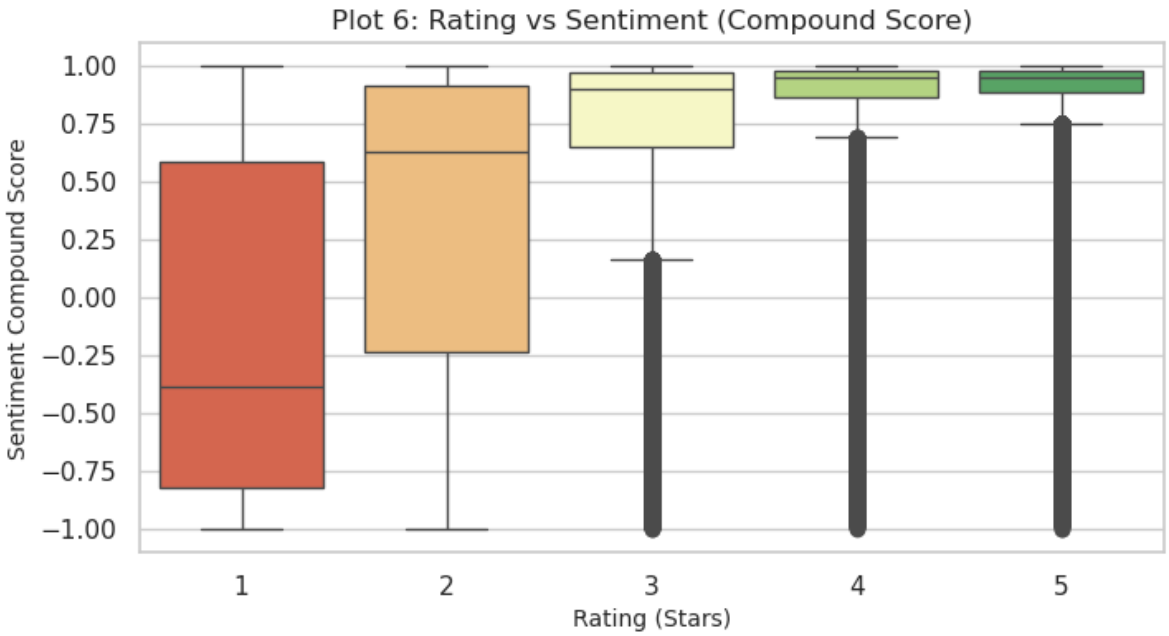


Table 1: Sentiment Distribution

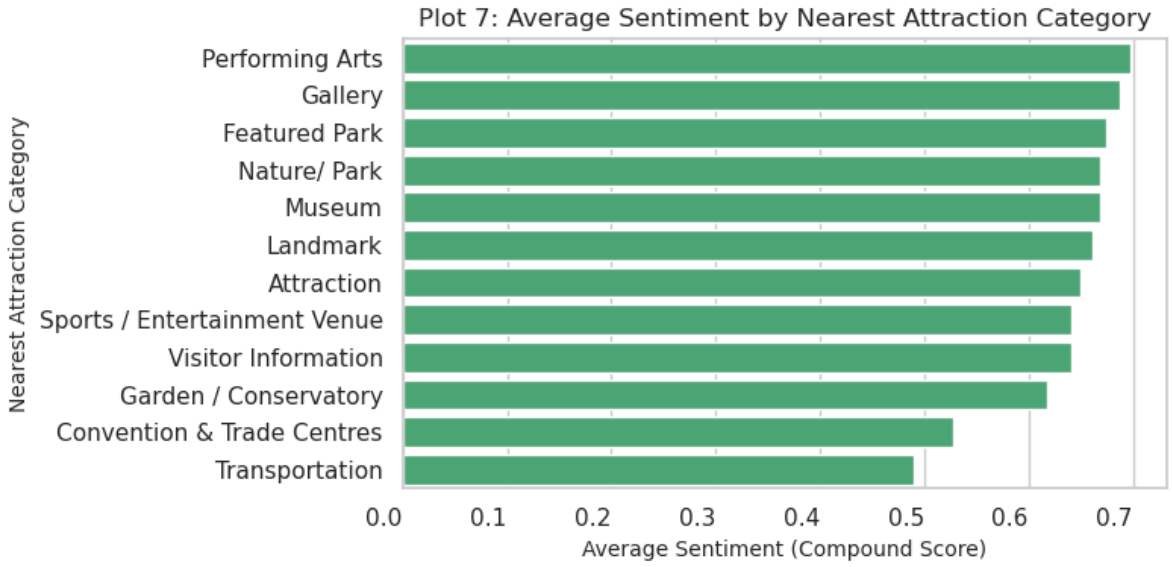
sentiment_label	Count	Percentage
positive	373503	86.678533
negative	51956	12.057386
neutral	5447	1.264081



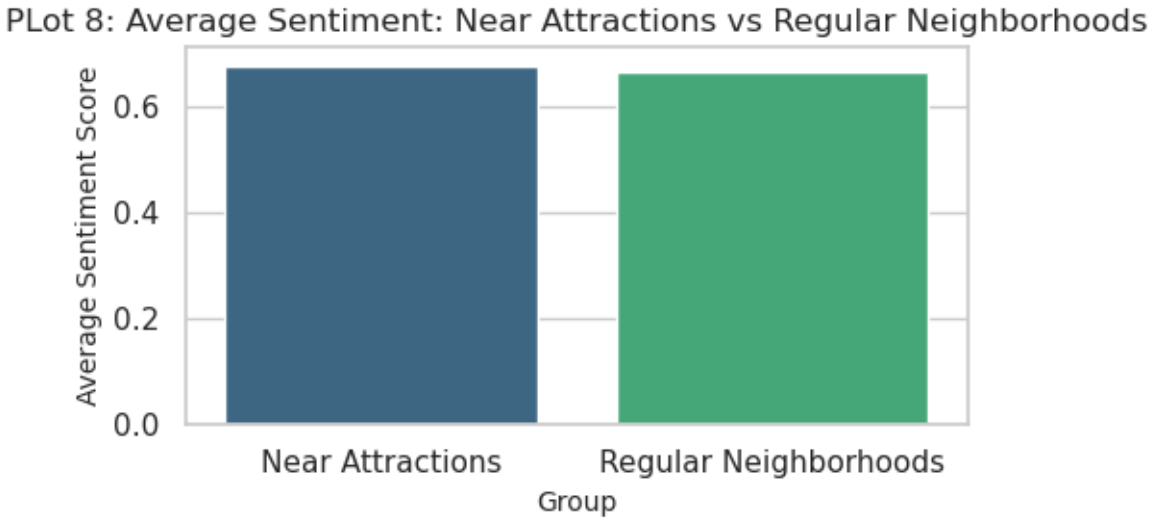
We also found out that as star ratings increase from 1 to 5, sentiment scores generally become more positive and less variable (as expected, shown on Plot 6). One-star ratings mostly show negative sentiment (median around -0.4), while five-star ratings consistently show high positive sentiment (median above 0.9). Interestingly, even 3-star ratings have a generally positive sentiment median despite being considered "average" ratings, though they have notable negative outliers, so this matched with our finding above that the reviews tend to skew positively.



Businesses with the nearest attractions being cultural venues like performing Arts, galleries, and parks generate the highest sentiment scores for their reviews (with businesses near Performing Arts on average reached almost 0.7 in score), while businesses near transportation facilities and convention/trade centers receive the lowest sentiment scores (around 0.5-0.55). This matched our previous finding of which businesses category have the higher ratings, and our analysis above that generally higher rating produces higher sentiment score (more positive review). However, it is worth mentioing that most attraction categories maintain positive sentiment scores overall, suggesting visitors generally have favorable experiences across business with any attractions nearby.



Knowing that businesses near attractions generally have positive reviews, we wanted to compare between those which are really closed by the attractions (under or equal to 1 km of distance near the attractions) and those located in regular neighbourhoods. Surprisingly, there is not much of a difference; those in regular neighbourhoods still have on average positive reviews, similar to the really-near attractions ones as shown in Plot 8.



4. Conclusion and Next Step

This analysis has provided several key insights into how tourist attractions influence business performance in Toronto. Businesses near tourist attractionsgenerally receive more reviews, but their ratings are not necessarily higher than those of businesses located further away. Additionally, cultural and nature-oriented attractions appear to be associated with more positive sentiments and higher ratings compared to other types of attractions. Most of the reviews are positive, and they become more positive as the user rating on that review increase.

For the next step in the final project, we plan to implement tokenization, word cloud generation, and Latent Dirichlet Allocation (LDA) topic modeling, to identify common topics in reviews for businesses near different types of attractions. Predictive modeling techniques will also be applied to develop models that predict business success (high ratings, positive sentiment, high amount of reviews) based on location features, including proximity to tourist attractions. This could include regression models to predict ratings or classification models to predict sentiment categories.