CIS 509 - Final Presentation

Yelp Dataset Project - Identifying Prime Restaurant Locations and Cuisines

Presented By Team 107:

Deeksha Lingaraju
Reeve George
Manish Reddy
Shriya
Yashwanth Guruswamy

Business Context/Problem

- Choosing the right restaurant location is critical for success in the competitive food industry, as poor site selection is a major reason for restaurant failures.
- Our Goal : Leverage Yelp data to identify prime locations/cuisines for new restaurant openings based on demand, competition, and customer sentiment.
- Evaluate the number of competitors, their reputation, and market saturation to help highlight opportunities for new establishments.
- By analyzing restaurant distribution and customer density, we can determine which cuisines have high demand but low availability, creating market opportunities.
- Provide data-driven insights to entrepreneurs and investors to minimize risk and maximize restaurant success.

Restaurant Industry Overview

- The restaurant industry is a multi-billion dollar market, with steady growth driven by consumer demand for dining out and food delivery services.
- High competition among restaurants, with success heavily dependent on location, customer experience, and online reputation.
- Increasing reliance on online reviews, social media, and food delivery apps influences restaurant choices and business success.
- High failure rates for new restaurants, but we believe data-driven site selection can improve success by identifying high-demand, low-competition areas.
- Businesses increasingly use customer reviews, sentiment analysis, and competitor analysis to make strategic location and menu decisions.

Research Questions & Purpose

- What factors contribute to the success of a new restaurant location?

  → We want to understand the key drivers of restaurant success (e.g., customer sentiment, competitor density, and foot traffic) that allow entrepreneurs to make informed site selection decisions, reducing the risk of business failure.

- Which restaurant categories or cuisines are underserved in high-demand areas?

  → We want to identify gaps in the market that help entrepreneurs and investors target cuisines with strong potential while avoiding oversaturated markets.

- Where are the best locations for opening new restaurants based on customer demand, sentiment, and competition?

  → By combining geospatial clustering with sentiment and demand analysis, we can pinpoint prime locations where new restaurants have the highest likelihood of success.

Dataset Overview and Summary

Dataset Selection:

- From the Yelp data, we chose the business, reviews, and users dataset to perform analysis on our project.
- For the business dataset, various food filters like Restaurant, Pizzeria, Sushi, Cafe, Fast Food and 10 others were selected. We also filtered it based on the current open status of the location, timeframe of the business, and reviews.
- Based on the chosen businesses, we extracted the relevant reviews from the reviews dataset.
- We also filtered the unique users who wrote the reviews in the selected timeframe (past 1-2 years).

Sample Size Justification:

- We utilized a 10000-row sample from each dataset to ensure a statistically significant analysis while maintaining computational efficiency and scalability to perform predictive modelling.
- The dataset represents multiple regions and restaurant types, making it generalizable for market trends.
- By focusing on reviews from the past 1-2 years, the dataset remains up-to-date and relevant for business decision-making.

Dataset Overview and Summary

- Suitability of Data for Business Questions

The selected dataset effectively supports various business questions regarding restaurant industry analysis, including:

- Customer Demand Analysis: The dataset includes 10,000 reviews, allowing for sentiment analysis to determine customer satisfaction and demand for different restaurant types.
- Competitor Analysis: The dataset contains 10,000 businesses, enabling insights into competition in different regions.
- Reputation & Sentiment Analysis: The text column in review.csv was processed using NLP techniques to extract customer sentiment and trends.
- Underserved Cuisine & Market Gaps: The category field in business.csv helped analyze the density of different cuisines across regions, identifying underserved markets.

Data Preprocessing

Data Cleaning:

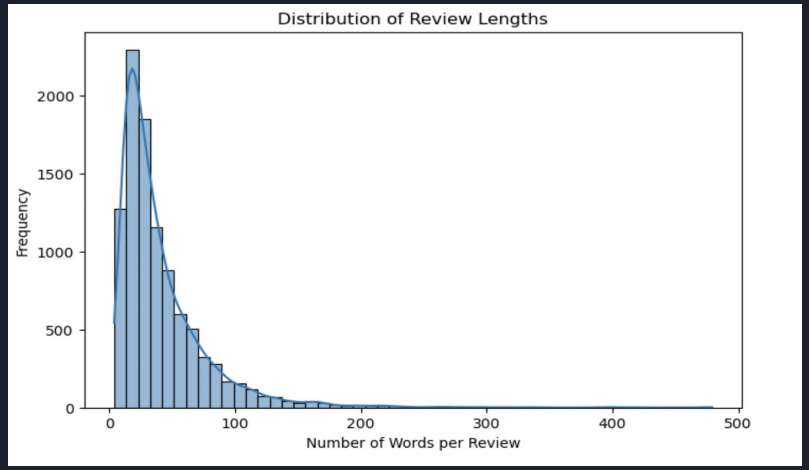
- Handled missing values.
- Removed duplicate records to avoid redundancy in analysis.

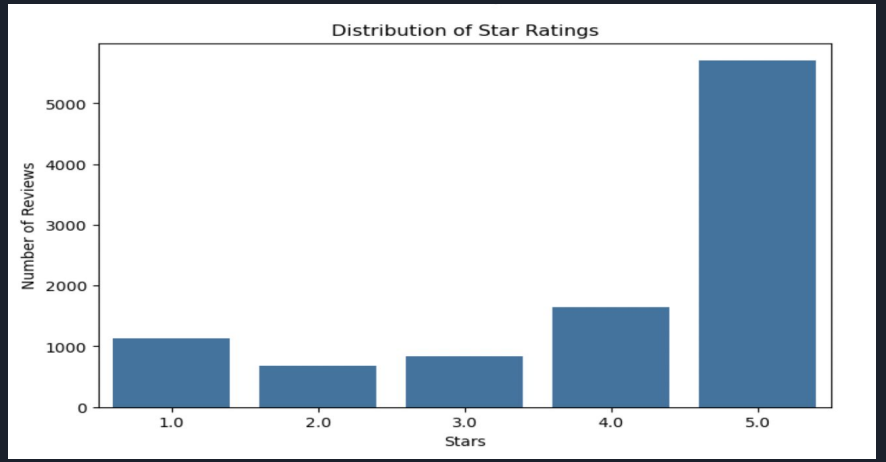
Text Cleaning and Tokenization:

- Converted text to lowercase to maintain uniformity.
- Removed stopwords using NLTK and Spacy.
- Removed non-alphabetic characters.
- Tokenized text and applied lemmatization.

Summary Statistics:

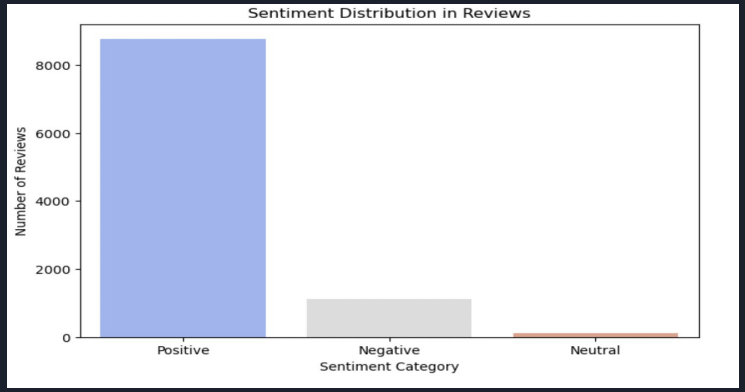
Metric	Value
Number of Reviews	10000
Total Number of Tokens	413152
Number of Unique Words	17607
Average Review Length	41.3152
Unique Customers	9290
Unique Businesses	1000
Average Stars per Review	4.0130
Average Votes per Review	1.1493

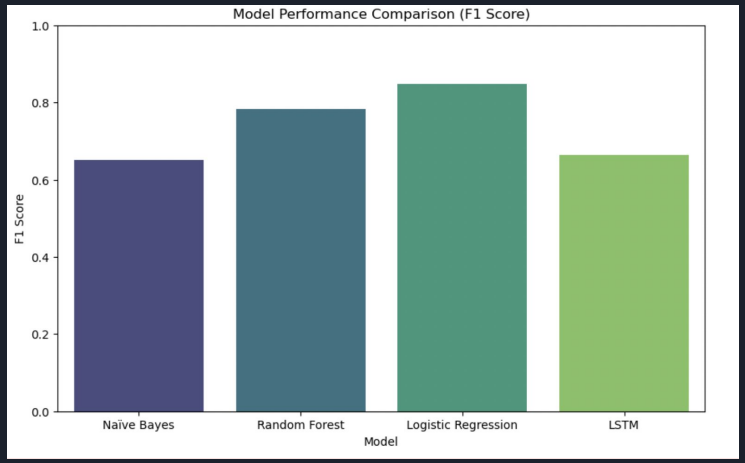




Analytical Methods and Techniques

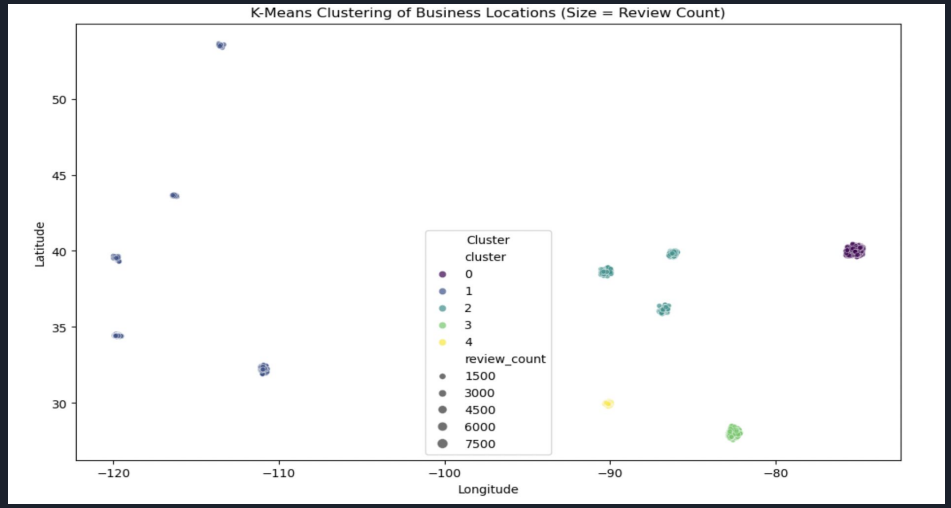
- We used VADER to categorize reviews as positive, negative, or neutral.
- We implemented four sentiment analysis models along with TF-IDF Vectorization, Naive Bayes, Random Forest, Logistic Regression, and LSTM to understand the sentiment of the reviews.
- Logistic Regression outperformed all other models in both accuracy and F1 Score.
- LSTM did not perform as expected, indicating that deep learning may not always be necessary for text classification.
- Naïve Bayes had the lowest F1 Score, struggling with neutral and negative sentiment detection.
- Random Forest performed well but had lower precision than Logistic Regression.

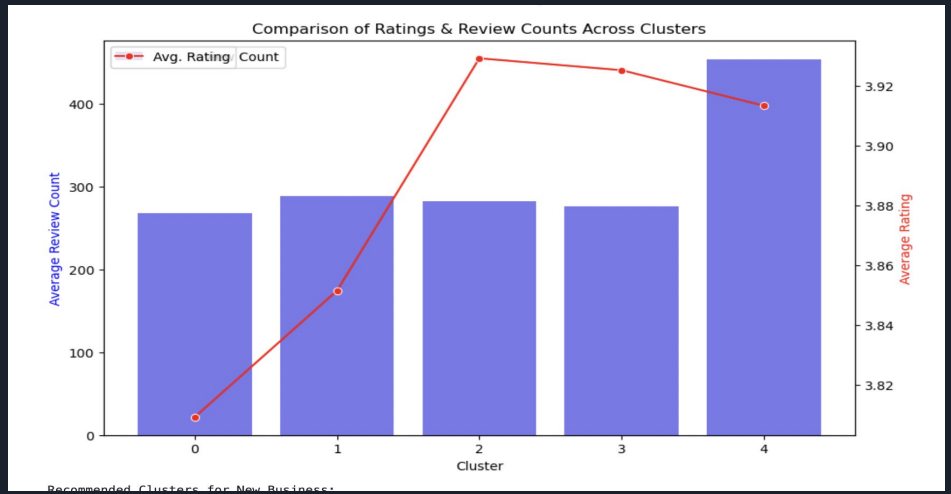




Geospatial Analysis

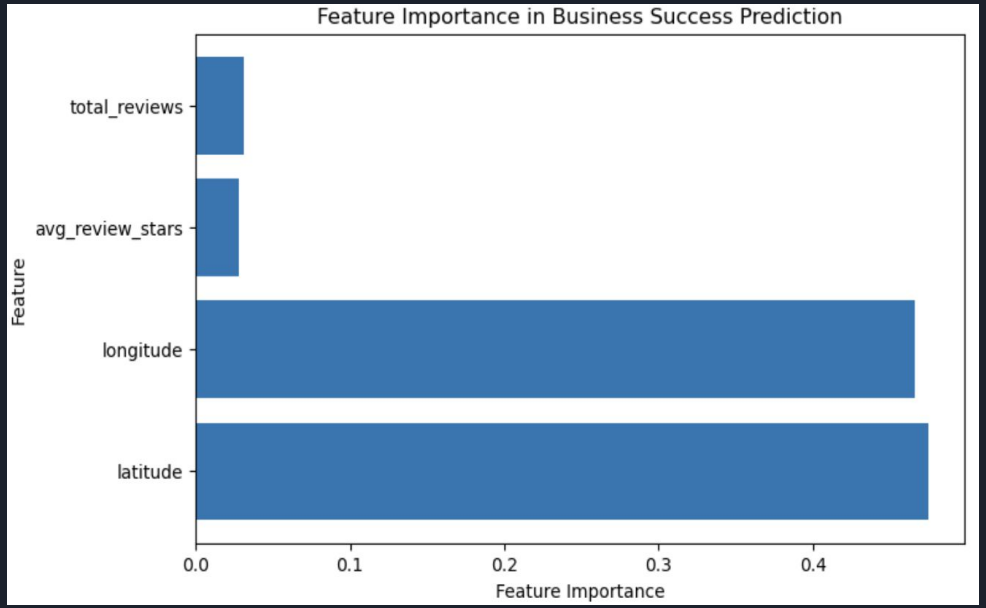
- Applied K-Means clustering (5 clusters) on business latitude, longitude, review count, and ratings to group locations.
- Sorted clusters based on business success (higher review counts and star ratings) to identify high-potential areas.
- Used OpenStreetMap API to retrieve real-world locations (cities/counties) for the best business clusters.
- Recommended top 3 clusters (i.e Cluster 2, 3 and 4) to open new business and some high-potential locations under them.
- **Martin County, Indiana**
- **Hillsborough County, Florida**
- **New Orleans, Louisiana**





Key Factors Driving Business Success: Feature Importance Analysis

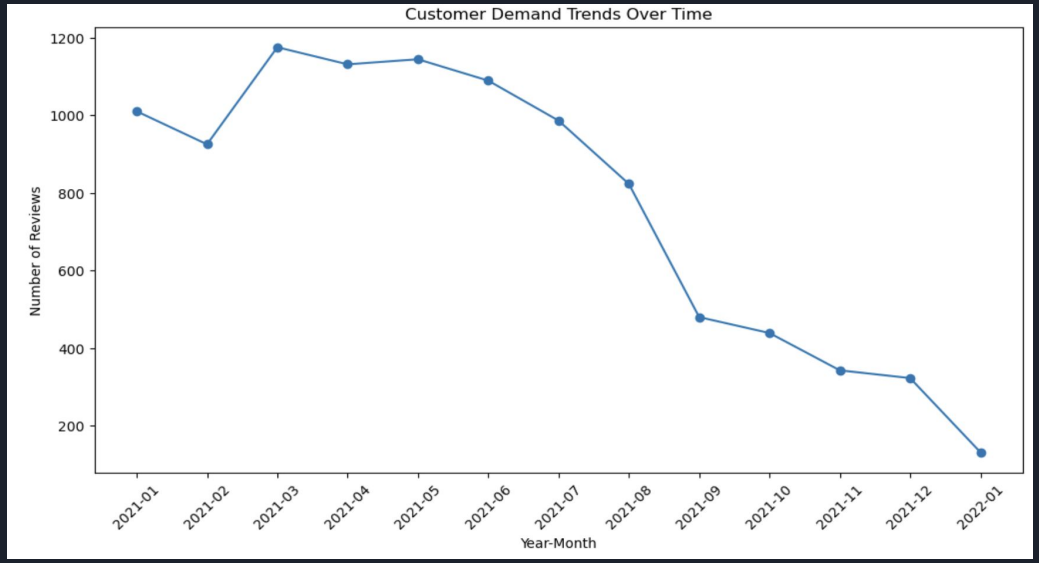
- Business Success Prediction – Random Forest model predicts success using location, reviews, and ratings.
- Feature Importance – Latitude & Longitude are the strongest predictors, followed by review count and ratings.
- High-Potential Locations – KNN recommends best business locations based on competitor success.
- **Business Name Suggestion** – Finds popular names in similar categories and locations.
- **Insights** – Helps entrepreneurs pick ideal locations & names using data-driven decisions



Analysis of Business Popularity and Demand Trends

- We group Businesses by Category and compute three key metrics for each category:
  i) avg_stars: Average rating (customer satisfaction).
  ii) total_reviews: Total reviews (demand indicator).
  iii) business_count: Number of businesses (competition indicator).

- Calculating competitor density=business count/total_reviews
- Identified underserved cuisines based on high ratings (avg_stars > 4.0) and low competition (competitor_density < 0.01).
- Analyzed demand trends by aggregating monthly review counts.



Underserved Cuisines:

	categories	avg_stars
3	Acai Bowls, Restaurants, American (New), Food,...	4.5
5	Active Life, Beer Gardens, Grocery, Middle Eas...	4.5
7	Active Life, Event Planning & Services, Hotels...	4.5
10	Active Life, Nightlife, Sports Bars, Dog Parks...	4.5
13	Active Life, Restaurants, Botanical Gardens, T...	4.5

Takeaways:
✅ Low-competition, high-rating categories present business potential
✅ Monitoring demand trends helps strategic decision-making

High-Growth Restaurant Location Analysis

- We compute Sentiment Analysis by Location: Groups reviews by business_id to calculate:

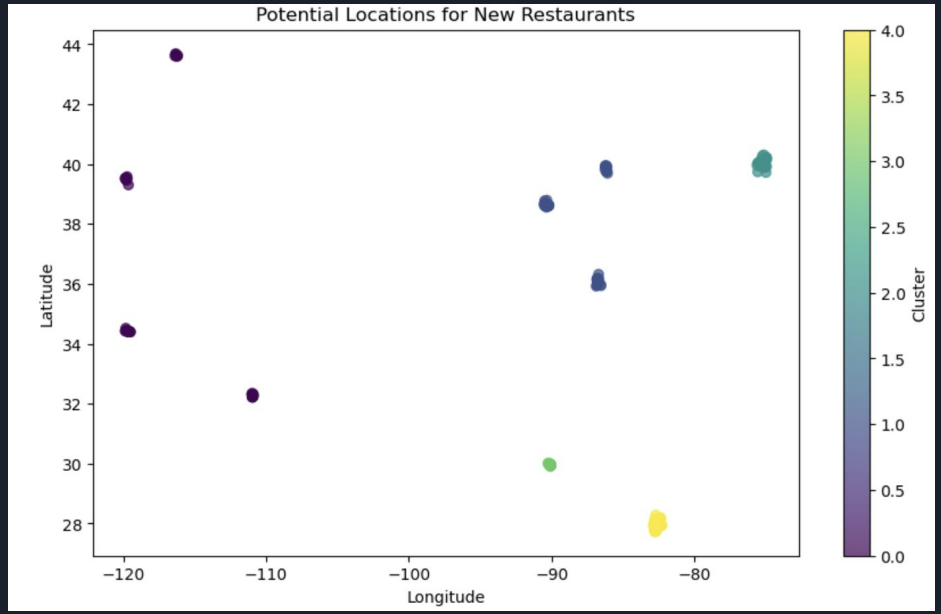
  (i) avg_sentiment → The average customer rating for each restaurant (mean).
  (ii) total_reviews → The number of reviews each restaurant has received.
- Merge Sentiment Data with Business Data: Adds average sentiment ratings and review counts to business data.
- Filtered high-growth businesses with sentiment > 4.0 and above-median reviews (high customer interest & high potential for expansion)
- Applied K-Means clustering (n=5) on latitude & longitude to group locations.

Output Obtained:

- Clusters of high-potential restaurant locations displayed on a scatter plot.
- Top restaurants with high ratings & reviews identified in a table.
- Geospatial insights on optimal business expansion areas.

High Growth Restaurant Locations:

	business_id	name
8	GBTPC53ZrG1ZBY3DT8Mbcw	Luke
60	UCMSWPqzXjd7QHq7v8PJjQ	Prep & Pastry
62	vN6v8m4DO45Z4pp8yxxF_w	Surrey's Café & Juice Bar
65	pSmOH4a3HNNpYM82J5ycLA	The Pancake Pantry
66	8uF-bhJFgT4Tn6DTb27viA	District Donuts Sliders Brew



Competitor Analysis

- Identified similar businesses in proximity to analyze competition using KNN.
- Mapped oversaturated vs. underserved business types based on total businesses and review counts.
- Evaluated competitor performance using review counts and star ratings.
- Determined which business types have strong customer demand but few competitors.

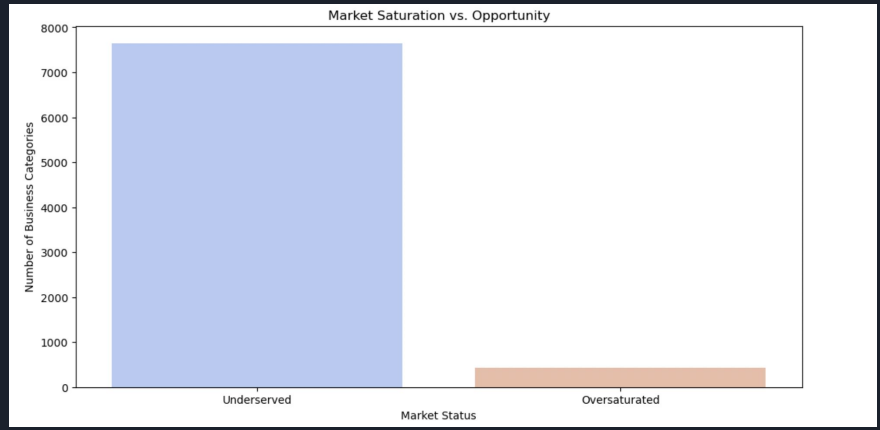
Key Insights & Findings

✅ Oversaturated Business Types (High Competition)

- Example: "American (New), Restaurants" has 31 businesses with high review counts.
- Implication: High competition → Difficult for new businesses to stand out.

✅ Underserved Business Types (High Potential)

- Example: "Acai Bowls, Juice Bars & Smoothies" has only 1 business but a high rating (4.5 stars).
- Implication: Strong customer demand but low competition, making it an ideal opportunity for new entrants.



Managerial Insights & Recommendations

Optimal Business Locations Identified:

- High-growth restaurant areas include Martin County, Indiana, Hillsborough County, Florida, and New Orleans, Louisiana (Top 3).
- These areas show high review volume and above-average predicted star ratings, making them strong candidates for new restaurant investments.

Underserved Cuisine Opportunities:

- The analysis highlights cuisines with high potential but low competition, such as Live/Raw Food, Chicken Shops, Chicken Wings, Seafood, and Cajun/Creole.
- Other underserved categories include Acai Bowls, Asian Fusion, Juice Bars & Smoothies, and Specialty Food.
- Opening restaurants in these categories in high-growth areas can help meet unmet demand.

Competitive Landscape & Market Saturation:

- Some restaurant categories, such as American (New), Bars/Nightlife, and Traditional American Restaurants, are oversaturated.
- Investing in these categories without differentiation may lead to high competition and lower success rates.

Concluding Remarks

Data-Driven Expansion Strategy:

- Investing in locations with high demand and low competition can maximize profitability.
- Cities with clusters of successful restaurants indicate potential for new entrants with strategic positioning.

Leverage Customer Sentiment & Market Trends:

- Sentiment analysis helps uncover customer dissatisfaction points, allowing for improvements in service, pricing, or menu offerings.
- Businesses can differentiate by focusing on underrepresented cuisines that align with positive sentiment trends.

Key Factors for Restaurant Success:

- Locations with high foot traffic, strong review volume, and positive customer sentiment tend to have greater success and correlate with better ratings and engagement.
- Restaurants near established high-rated businesses can leverage spillover customer traffic but should ensure differentiation.

Future Considerations & Additional Data Needs:

- Additional factors like demographics and economic indicators could further refine site selection.
- Seasonal demand trends suggest timing of openings could impact initial success—launching in high-review months might maximize visibility.

Thank You For Your Time!

Any Questions?