CSC 575- Intelligent Information Retrieval

Final Project

Section- 801

Student Name(s): Nachiketh Reddy, Aniket Surve

TASK 1: Business Recommendation by Table Lookup

Introduction:

The dataset included details on numerous companies located in various cities. Using this dataset, queries were to be run in order to obtain business recommendations based on predetermined standards. To show off the recommendation system's capabilities, the dataset was put into a table and a number of queries were run.

Data Exploration:

The dataset comprises business details such as name, city, categories, stars, review count, and attributes. It contains businesses from multiple cities, including Philadelphia, Nashville, Tampa, Indianapolis, and others.

Methodology:

We used Pandas DataFrame operations for querying in order to filter and sort the dataset according to particular parameters like city, categories, and attributes. The questions were created to highlight the recommendation system's adaptability by addressing various degrees of specificity and complexity. The recommendation of adjacent establishments was made possible by the use of geolocation data to compute the distances between landmarks and businesses.

Queries and Recommendations:

Best restaurants in Nashville: We identified the top restaurants in Nashville based on their star ratings, review counts, and alphabetical order of names. The query resulted in recommendations such as Big Al's Deli, Tasty And Delicious, and others.

Best Chinese restaurants in Tampa: This query focused on finding the highest-rated Chinese restaurants in Tampa. Recommendations included Fuzion Spice, Soul Of Korea, and others.

Pubs in Philadelphia that are Wheelchair Accessible: We identified pubs in Philadelphia that are wheelchair accessible. Recommendations included Lucky's Last Chance, SouthHouse, and others.

Best restaurants in Nashville:

	stars	name	review_count
95158	5.0	Big Al's Deli	390
31772	5.0	Tasty And Delicious	260
51628	5.0	D'Andrews Bakery & Cafe	212
73120	5.0	Edessa Restaurant Kurdish Turkish Cuisine	198
37626	5.0	The CafÃ□Â© at Thistle Farms	189

Best Chinese restaurants in Tampa:

	stars	name	review_count
56096	5.0	Fuzion Spice	80
61094	5.0	Pure Zen Mobile Massage	8
53908	5.0	Luxe Beauty Studio	7
85221	5.0	Anna Health	6
34261	4.5	Soul Of Korea	238

Pubs in Philadelphia that are Wheelchair Accessible:

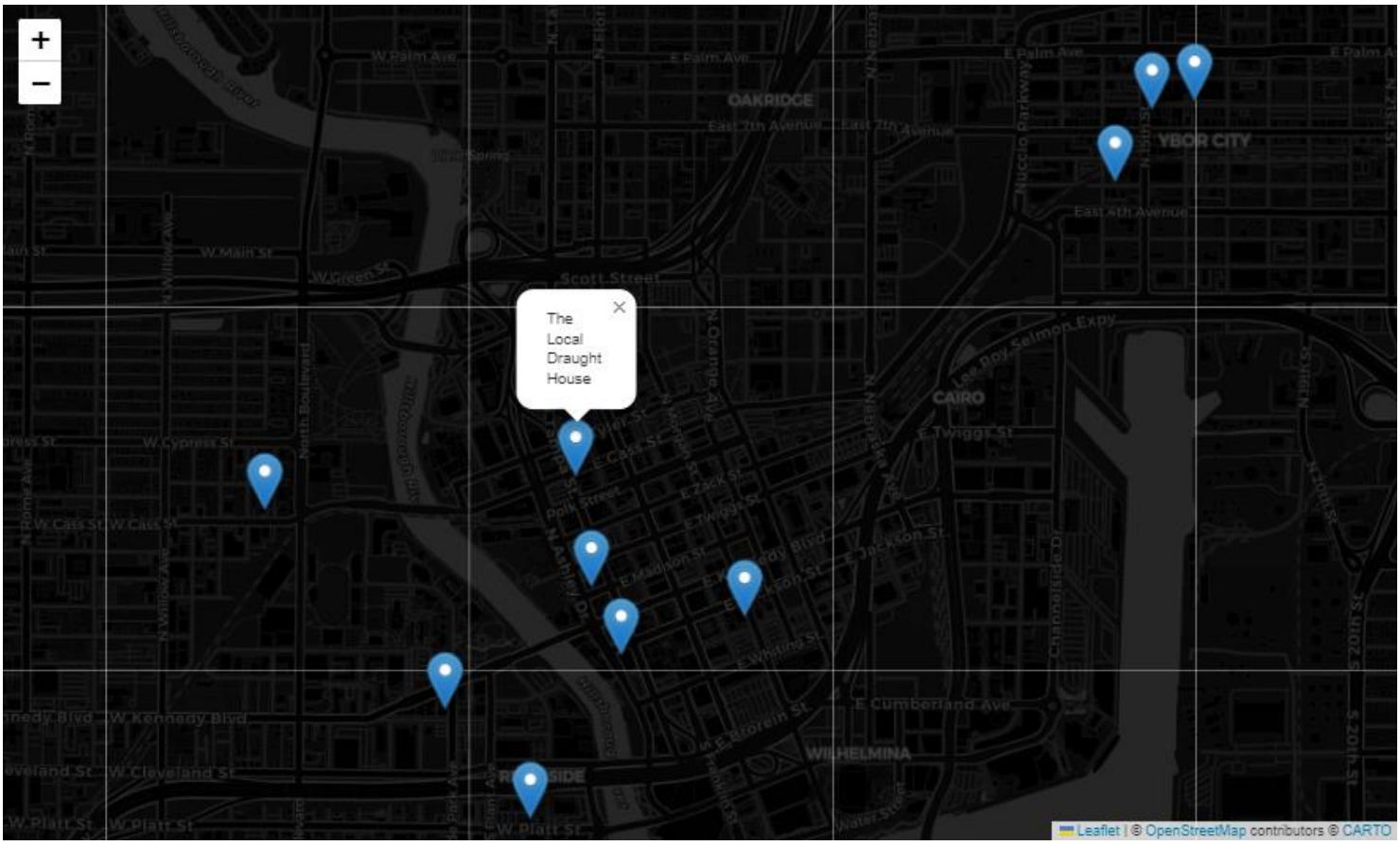
	stars	name	review_count
94105	4.5	Lucky's Last Chance	579
17602	4.5	SouthHouse	158
90322	4.5	Silence Dogood's Tavern	111
78757	4.5	Original 13 Ciderworks	65
56738	4.0	White Dog Cafe	1301

Business hours for "DeSandro on Main" in Philadelphia for Friday: We extracted the business hours for "DeSandro on Main" in Philadelphia specifically for Fridays. The result was 17:00 to 00:30.

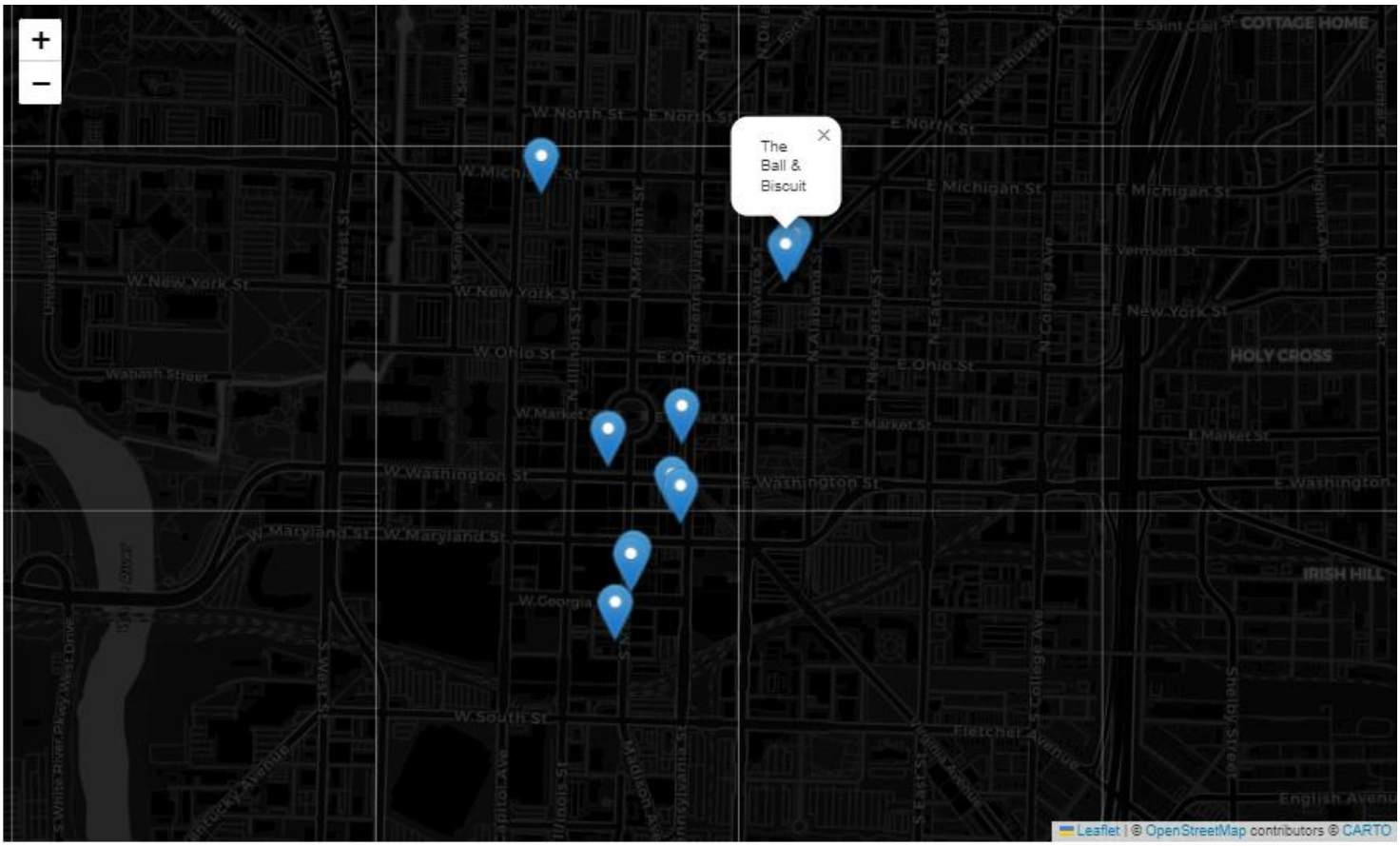
Business hours for 'DeSandro on Main' in Philadelphia for Friday:
17:0-0:30

Pubs near a landmark in Tampa and Indianapolis: Leveraging geolocation information, we identified pubs located within a 5-mile radius of a specified landmark in Tampa and Indianapolis. We then plotted the top 5 closest pubs on interactive maps using latitude and longitude coordinates.

Nearby Pubs in Tampa:



Nearby Pubs in Indianapolis:



TASK 2: Recommender System using Embeddings

Data Preprocessing:

Tokenization: The review texts are tokenized to extract individual words, resulting in a vocabulary of unique tokens.

Text Cleaning: Punctuation and stopwords are removed from the tokens to ensure that only meaningful words contribute to the embeddings.

Combining Reviews: All reviews written by a user or about a business are aggregated into a single string or list of strings, representing the collective sentiment or characteristics associated with that user or business.

Embedding Creation:

GloVe Embeddings: We utilize pre-trained GloVe embeddings with a vector length of 50. These embeddings provide dense representations of words based on their co-occurrence statistics in large text corpora.

Mean Embedding: For each user and business, we compute the mean vector of GloVe embeddings for the tokens present in their aggregated review texts. This process captures the semantic representation of the user's preferences or the business's attributes.

```
Number of words in embeddings: 400000
Number of users: 65400
Number of businesses: 9406
```

Modeling and Prediction:

The predicted star rating for the user-business pair is calculated by taking the dot product of the user and business embeddings.

The squared error between the predicted and actual star ratings is computed and accumulated to calculate the total squared error. The mean squared error (MSE) is calculated by dividing the total squared error by the number of user-business pairs in the test dataset.

The square root of the MSE is then computed to obtain the root mean squared error (RMSE), which is a measure of the average deviation of predicted ratings from actual ratings across all user-business pairs. The calculated RMSE is printed to evaluate the performance of the recommendation system. In this case, the RMSE is approximately 2.44.

Linear Regression: Initially, a linear regression model is trained using the dot product of user and business embeddings to predict star ratings. This simple model provides a baseline for comparison.

XGBoost and Gradient Boosting: More complex models such as XGBoost and Gradient Boosting are employed to capture nonlinear relationships between embeddings and star ratings. These models leverage ensemble learning techniques to improve predictive accuracy.

Evaluation:

Root Mean Squared Error (RMSE):

Linear Regression: RMSE on the testing dataset is approximately 1.20, indicating moderate predictive accuracy.

XGBoost: After hyperparameter tuning, the tuned XGBoost model achieves an RMSE of around 1.20, comparable to linear regression but potentially offering better generalization.

Gradient Boosting: The Gradient Boosting model yields an RMSE of approximately 1.19, slightly outperforming both linear regression and XGBoost.

Training vs. Testing RMSE:

Linear Regression: The training RMSE is similar to the testing RMSE, suggesting adequate model generalization.

XGBoost and Gradient Boosting: These models also exhibit minimal overfitting, with training and testing RMSE values in close proximity.

Model Selection and Tuning:

Hyperparameter Tuning: Grid search is employed to identify the optimal hyperparameters for the XGBoost model, resulting in a learning rate of 0.1, maximum depth of 5, and 300 estimators.

Model Comparison: The performance of different models is compared based on their RMSE values, with Gradient Boosting demonstrating slightly superior predictive accuracy compared to linear regression and XGBoost.

Insights and Further Enhancements:

Interpretation of Results: Analyzing the predictions and RMSE values provides insights into the strengths and limitations of the recommender system. For example, the relatively modest RMSE suggests that the system can provide reasonably accurate star rating predictions.

Further Enhancements: To enhance the recommender system, future iterations could explore advanced embedding techniques (e.g., contextual embeddings), incorporate additional features (e.g., user demographics, business attributes), and experiment with deep learning architectures for more nuanced predictions.

TASK 3: Item-based Collaborative Recommendation using Embeddings

Methodology:

We need to identify businesses common to the dataset and the embeddings utilizing the business id as key: It iterates through the business IDs stored in the business_embeddings dictionary to identify businesses categorized under the specified business type and location. For each business that matches the specified categories and location, the code extracts essential information such as business ID, name, and city. The extracted information for each relevant business is stored in a list for further querying.

Query 1: Five businesses similar to 'Spring Chinese Restaurant' in Philadelphia of type Restaurants

Utilizes cosine similarity to compare embeddings of restaurants in Philadelphia with 'Spring Chinese Restaurant' and ranks them based on similarity scores. The code first filters restaurants in Philadelphia from the original businesses table. Then, it computes cosine similarity between the embeddings of these restaurants and the target restaurant. Finally, it ranks the similar restaurants and presents the top five.

```
Target business in Philadelphia of type Bars: Queen Sheba

Five businesses similar to 'Queen Sheba' in Philadelphia of type Bars using Jaccard simil
arity:
1. Name: Queen Sheba, Stars: 4.0, Categories: Nightlife, Restaurants, Sports Bars, Bars,
Ethiopian, Similarity: 1.0000
2. Name: Smiley's Cafe, Stars: 4.5, Categories: Juice Bars & Smoothies, Food, Mediterrane
an, Greek, Restaurants, Sandwiches, Similarity: 0.6124
3. Name: Prohibition Taproom, Stars: 4.0, Categories: American (Traditional), Bars, Night
life, Restaurants, Beer Bar, Gastropubs, Breakfast & Brunch, Tapas/Small Plates, Similari
ty: 0.5704
4. Name: Pub & Kitchen, Stars: 3.5, Categories: Restaurants, Pubs, Food, Bars, American (
New), Nightlife, Similarity: 0.5684
5. Name: Mac's Tavern, Stars: 3.5, Categories: Nightlife, Restaurants, American (New), Ba
rs, Pubs, Similarity: 0.5617
6. Name: The Black Sheep Pub & Restaurant, Stars: 3.5, Categories: Bars, Pubs, Irish Pub,
Nightlife, Sports Bars, Restaurants, American (New), Similarity: 0.5602
```

Query 2: Five businesses similar to 'Queen Sheba' in Philadelphia of type Bars

Similar to Query 1, but focuses on bars in Philadelphia. Utilizes cosine similarity to compare embeddings and rank similar bars. The code follows a similar process as Query 1, but with a focus on bars. It extracts bar businesses from the original dataset, computes cosine similarity, and ranks the results.

```
Target business in Philadelphia of type Bars: Queen Sheba

Five businesses similar to 'Queen Sheba' in Philadelphia of type Bars using Jaccard s
imilarity:
1. Name: Queen Sheba, Stars: 4.0, Categories: Nightlife, Restaurants, Sports Bars, Ba
rs, Ethiopian, Similarity: 1.0000
2. Name: Smiley's Cafe, Stars: 4.5, Categories: Juice Bars & Smoothies, Food, Mediter
ranean, Greek, Restaurants, Sandwiches, Similarity: 0.6124
3. Name: Prohibition Taproom, Stars: 4.0, Categories: American (Traditional), Bars, N
ightlife, Restaurants, Beer Bar, Gastropubs, Breakfast & Brunch, Tapas/Small Plates,
Similarity: 0.5704
```

```
4. Name: Pub & Kitchen, Stars: 3.5, Categories: Restaurants, Pubs, Food, Bars, Americ
an (New), Nightlife, Similarity: 0.5684
5. Name: Mac's Tavern, Stars: 3.5, Categories: Nightlife, Restaurants, American (New)
, Bars, Pubs, Similarity: 0.5617
6. Name: The Black Sheep Pub & Restaurant, Stars: 3.5, Categories: Bars, Pubs, Irish
Pub, Nightlife, Sports Bars, Restaurants, American (New), Similarity: 0.5602
```

Query 3: Five businesses similar to 'Soho Wellness & Med Spa' in Tampa of types ['Health & Medical', 'Beauty & Spas']

Targets health, medical, beauty, and spa businesses in Tampa. Utilizes cosine similarity to compare embeddings and identify similar businesses. Similar to previous queries, this implementation filters relevant businesses, computes cosine similarity, and ranks the results. It employs modular functions and data manipulation tools for efficiency.

```
Target business in Tampa of types ['Health & Medical', 'Beauty & Spas']: Soho Wellnes
s & Med Spa

Five businesses similar to 'Soho Wellness & Med Spa' in Tampa of types ['Health & Med
ical', 'Beauty & Spas'] :
1. Name: Soho Wellness & Med Spa, Stars: 4.5, Categories: Health & Medical, Beauty &
Spas, Similarity: 1.0000
2. Name: Massage Studio, Stars: 4.5, Categories: Health & Medical, Beauty & Spas, Sim
ilarity: 0.9931
3. Name: Massage Studio, Stars: 4.5, Categories: Health & Medical, Beauty & Spas, Sim
ilarity: 0.9875
4. Name: Love Nail & Spa, Stars: 4.0, Categories: Health & Medical, Beauty & Spas, Si
milarity: 0.9875
5. Name: Lecada Medical Artistry, Stars: 4.5, Categories: Health & Medical, Beauty &
Spas, Similarity: 0.9871
6. Name: Manhattan Nails Salon, Stars: 4.0, Categories: Health & Medical, Beauty & Sp
as, Similarity: 0.9860
```

Query 4: Five businesses similar to 'Safelite AutoGlass' in Tampa of types ['Automotive', 'Windshield Installation & Repair']

Focuses on automotive and windshield repair businesses in Tampa. Computes cosine similarity between embeddings to identify similar businesses. Similar to previous queries, this implementation preprocesses data, computes cosine similarity, and presents the results. It emphasizes data processing and similarity computation for accurate recommendations.

```
Target business in Tampa of types ['Automotive', 'Windshield Installation & Repair']:
Safelite AutoGlass

Five businesses similar to 'Safelite AutoGlass' in Tampa of types ['Automotive', 'Win
dshield Installation & Repair'] :
1. Name: Safelite AutoGlass, Stars: 3.0, Categories: Automotive, Windshield Installat
ion & Repair, Similarity: 1.0000
2. Name: Courtesy Hyundai Tampa, Stars: 3.0, Categories: Automotive, Windshield Insta
llation & Repair, Similarity: 0.9924
3. Name: Audio Itch, Stars: 3.5, Categories: Automotive, Windshield Installation & Re
pair, Similarity: 0.9892
4. Name: Sears Auto Center, Stars: 2.0, Categories: Automotive, Windshield Installati
on & Repair, Similarity: 0.9884
```

```
5. Name: Wheel Tec, Stars: 4.0, Categories: Automotive, Windshield Installation & Rep
air, Similarity: 0.9876
6. Name: Ed Morse Cadillac Tampa, Stars: 3.5, Categories: Automotive, Windshield Inst
allation & Repair, Similarity: 0.9873
```

Query 5: Five businesses similar to 'Sociale Italian Tapas & Pizza Bar' in Tampa of types ['Pizza', 'Bars']

Targets pizza and bar businesses in Tampa. Utilizes cosine similarity to compare embeddings and rank similar businesses. Following a similar pattern as previous queries, this implementation focuses on pizza and bar businesses. It preprocesses data, computes cosine similarity, and presents the top similar businesses.

```
Target business in Tampa of types ['Pizza', 'Bars']: Sociale Italian Tapas & Pizza Ba
r

Five businesses similar to 'Sociale Italian Tapas & Pizza Bar' in Tampa of types ['Pi
zza', 'Bars'] :
1. Name: Sociale Italian Tapas & Pizza Bar, Stars: 4.0, Categories: Pizza, Bars, Simi
larity: 1.0000
2. Name: The C House, Stars: 4.0, Categories: Pizza, Bars, Similarity: 0.9959
3. Name: Timpano Hyde Park, Stars: 3.5, Categories: Pizza, Bars, Similarity: 0.9952
4. Name: American Social, Stars: 3.5, Categories: Pizza, Bars, Similarity: 0.9951
5. Name: Precinct Pizza, Stars: 3.5, Categories: Pizza, Bars, Similarity: 0.9950
6. Name: Sake House, Stars: 3.0, Categories: Pizza, Bars, Similarity: 0.9950
```

CONCLUSION:

To sum up, our investigation and use of intelligent information retrieval methods in the context of recommendation systems and business recommendations have produced insightful findings. We showed in Task 1 how well Pandas DataFrame operations work for querying and filtering datasets to produce customized business recommendations. We demonstrated the system's versatility and usefulness in a variety of settings by utilizing a range of inquiries, such as locating wheelchair-accessible pubs or the best restaurants in particular towns. By identifying surrounding businesses based on predetermined landmarks, the recommendation system's capabilities were further improved by utilizing geolocation data.

Going on to Task 2, the creation of a recommender system with embeddings demonstrated how important model selection and data pretreatment are to getting precise predictions. Semantic representations of user preferences and business attributes were efficiently captured by tokenizing, cleaning, and using pre-trained GloVe embeddings in review texts. We found the best method for star rating prediction by comparing and evaluating several models, such as Gradient Boosting, XGBoost, and linear regression. In addition, the review process yielded significant insights that guided future upgrades, like investigating more sophisticated embedding techniques and incorporating more features.

In Task 3, we turned our attention to embeddings for item-based collaborative recommendation. We used cosine similarity to find firms that were comparable based on their embeddings. Through the execution of diverse queries aimed at distinct business categories and regions, we showcased the system's capability to provide pertinent and individualised suggestions catered to customer inclinations. We made sure the recommendation system was successful and scalable in a variety of scenarios by implementing modular functions and processing data efficiently. All in all, our thorough investigation and application of intelligent information retrieval methods highlight the possibility of improving user experiences and streamlining decision-making across a range of fields.

REFLECTION:

We've explored the usefulness of constructing intelligent information systems through exercises like business recommendation and developing recommender systems with embeddings. These practical projects have given us invaluable knowledge on how to use data to efficiently meet information needs in the real world. Our ability to customize recommendations to improve user experiences has been developed through the analysis of user preferences and business attributes. Our understanding of intelligent information systems has expanded as a result of these tasks, which have also given us useful tools for solving challenging problems. Equipped with the knowledge and abilities acquired from these experiences, we feel more assured in our capacity to make a significant contribution to initiatives both inside and outside of our coursework.

TEAM	CONTRIBUTIONS
ANIKET SURVE	Data Exploration, Methodology, Queries, and Recommendations: This section involved exploring the dataset, devising querying methodologies, formulating specific queries, and interpreting results to provide tailored business recommendations. It encompassed tasks such as examining various attributes, filtering and sorting data, and identifying top businesses based on predetermined criteria. Additionally, it included tasks like extracting business hours and identifying nearby establishments using geolocation data. Collaborated with Nachiketh on some aspects.
NACHIKETH REDDY	Embedding Creation, Modeling, Evaluation, and Conclusion: This section encompassed tasks related to preprocessing data, creating embeddings, building predictive models, evaluating model performance, and summarizing findings. It involved steps such as tokenization, cleaning text data, aggregating embeddings, training models, tuning hyperparameters, and analyzing results. We worked together to ensure thorough data processing, model building, and interpretation of results for the recommender system. Collaborated with Aniket on other aspects as well.