Analyzing Business Integrity and Consumer Trends

Member: Phunsok Norboo, Quan Nguyen, Yiyou Chen, Ruoxian Zhang

Introduction

Dataset

DBSCAN

K-Means

LDA

Isolation Forest

Conclusion

Dataset (Google Reviews)

Dataset Source:

UC San Diego McAuley Lab

Datasets Used:

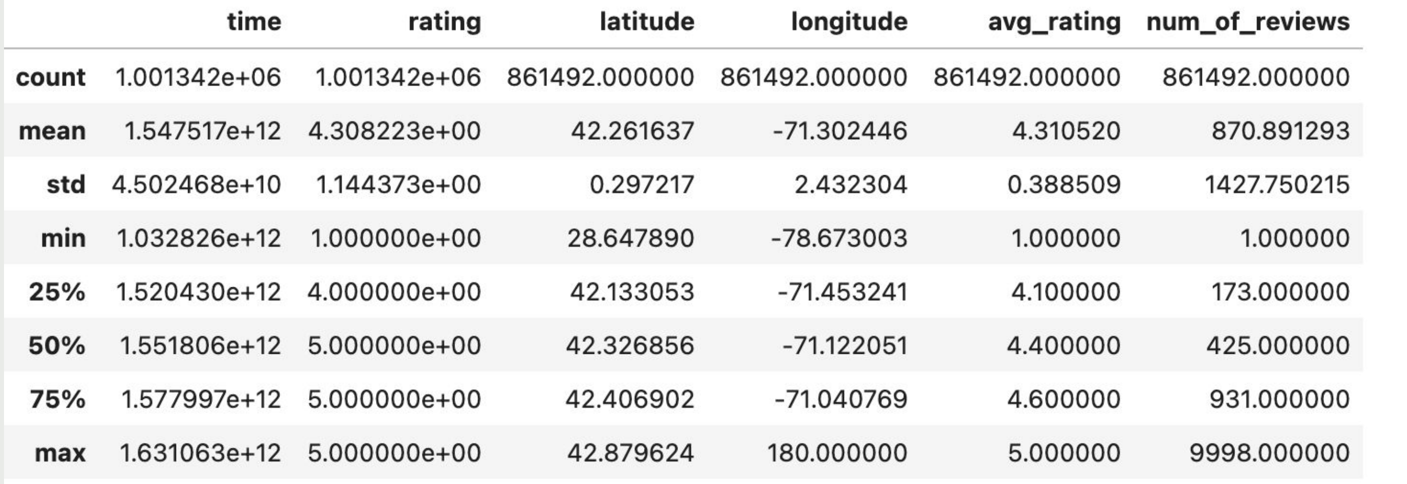
Review Dataset: 10.4+ million records (Ratings, review text, business responses)

Meta Dataset: 92,520 records (Category, location, average ratings, review counts)

Data Selection:

Focused on businesses in Massachusetts (MA) to refine the analysis scope

EDA



	time	rating	latitude	longitude	avg_rating	num_of_reviews
count	1.001342e+06	1.001342e+06	861492.000000	861492.000000	861492.000000	861492.000000
mean	1.547517e+12	4.308223e+00	42.261637	-71.302446	4.310520	870.891293
std	4.502468e+10	1.144373e+00	0.297217	2.432304	0.388509	1427.750215
min	1.032826e+12	1.000000e+00	28.647890	-78.673003	1.000000	1.000000
25%	1.520430e+12	4.000000e+00	42.133053	-71.453241	4.100000	173.000000
50%	1.551806e+12	5.000000e+00	42.326856	-71.122051	4.400000	425.000000
75%	1.577997e+12	5.000000e+00	42.406902	-71.040769	4.600000	931.000000
max	1.631063e+12	5.000000e+00	42.879624	180.000000	5.000000	9998.000000

Predominantly high ratings (4-5 stars) with a slight positive skew

Outliers in review counts and abrupt rating shifts suggest potential manipulation

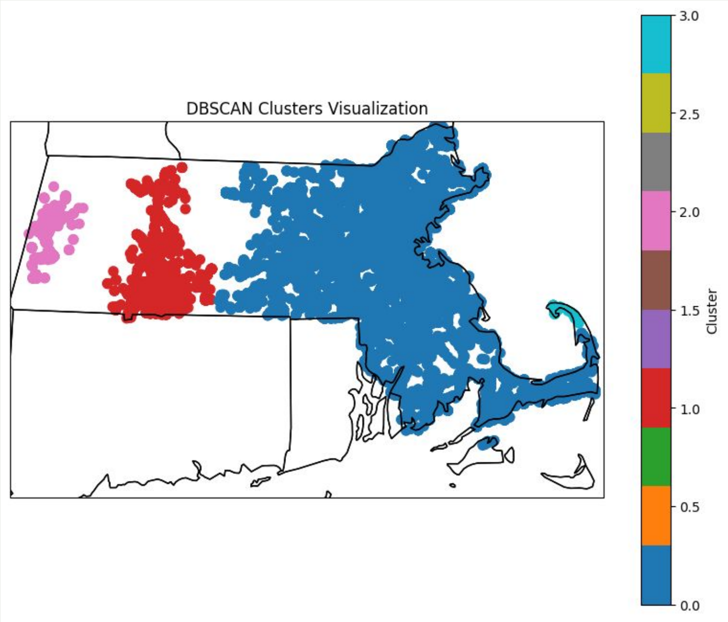
DBSCAN – Cluster businesses based on the longitude and latitude.

Group points that are within a 10 km distance and only form clusters if at least 50 points are closely grouped together – best result with the highest silhouette score = 0.42 ✅

```
dbscan = DBSCAN(eps= 10000, min_samples= 300,
metric='euclidean').fit(coords_m)
```

All businesses in MA:

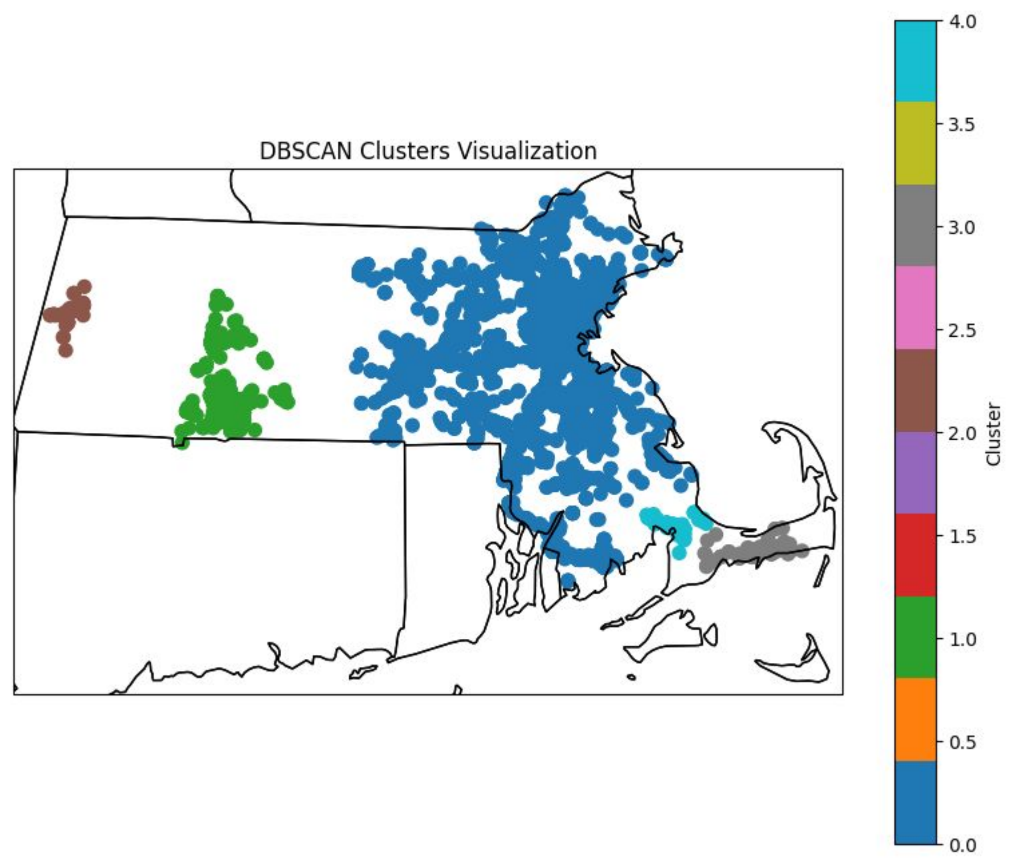
- Great Boston Area
- Worcester
- Springfield
- Cape Cod



DBSCAN

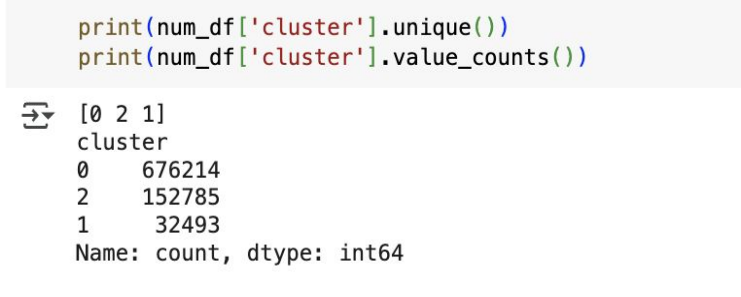
Suspicious businesses in MA:

- Great Boston Area
- Worcester
- Springfield
- Cape Cod
- Sagamore Beach



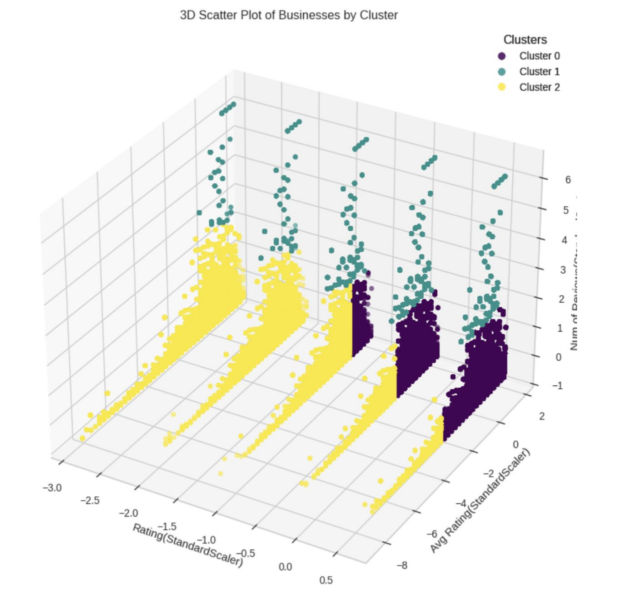
Insights of K-Means

- **Cluster 0 (Normal Businesses):** likely represents the majority of typical businesses.
- **Cluster 2 (Moderate Variation):** may indicate a slightly different pattern in customer engagement.
- **Cluster 1 (Highly Suspicious Group):** these businesses might be fraudulent or have manipulated reviews.

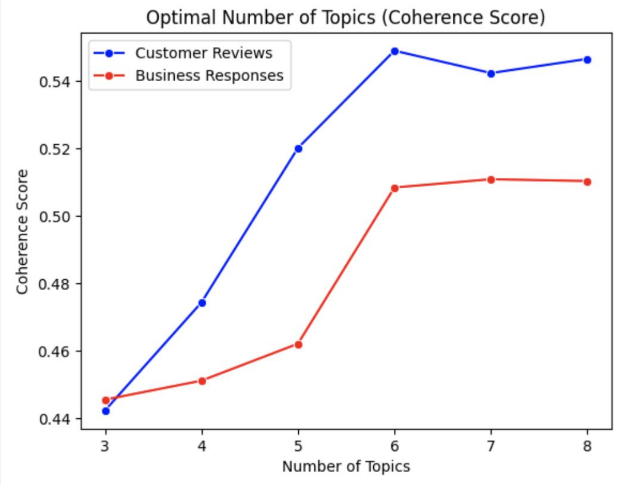


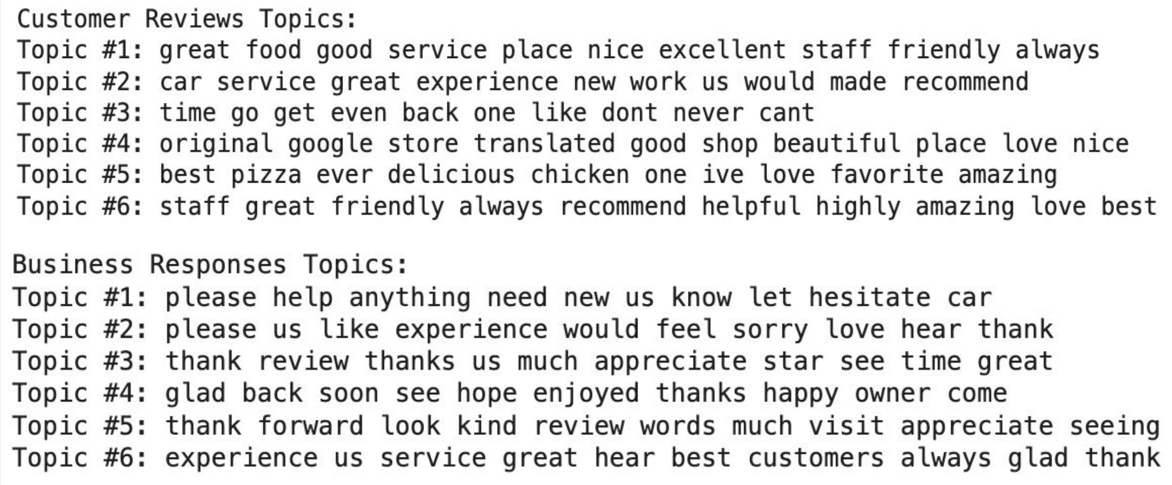
```
print(num_df['cluster'].unique())
print(num_df['cluster'].value_counts())
```

```
[0 2 1]
cluster
0    676214
2    152785
1     32493
Name: count, dtype: int64
```



LDA (Topic Modelling)



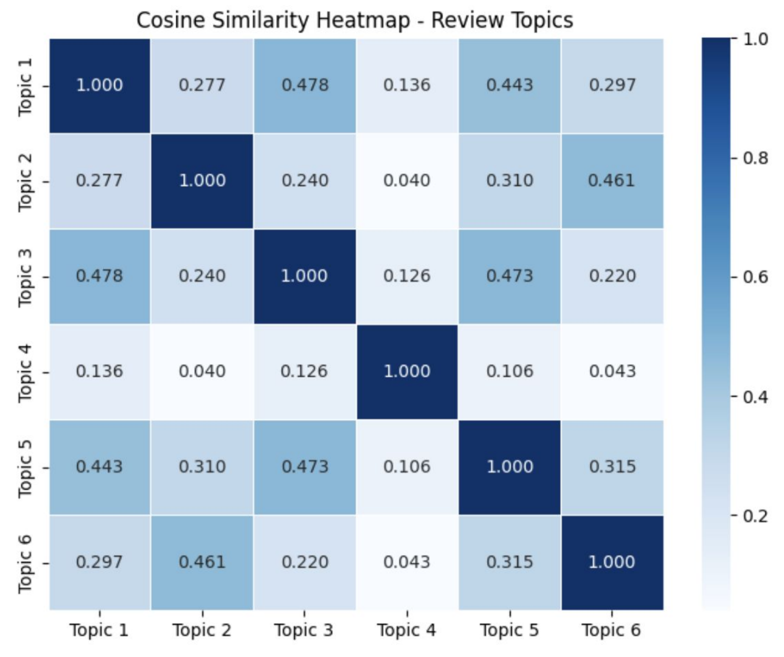


```
Customer Reviews Topics:
Topic #1: great food good service place nice excellent staff friendly always
Topic #2: car service great experience new work us would made recommend
Topic #3: time go get even back one like dont never cant
Topic #4: original google store translated good shop beautiful place love nice
Topic #5: best pizza ever delicious chicken one ive love favorite amazing
Topic #6: staff great friendly always recommend helpful highly amazing love best

Business Responses Topics:
Topic #1: please help anything need new us know let hesitate car
Topic #2: please us like experience would feel sorry love hear thank
Topic #3: thank review thanks us much appreciate star see time great
Topic #4: glad back soon see hope enjoyed thanks happy owner come
Topic #5: thank forward look kind review words much visit appreciate seeing
Topic #6: experience us service great hear best customers always glad thank
```

Topics are way too redundant!!!

LDA (Topic Modelling)



Cosine Similarity Heatmap - Review Topics

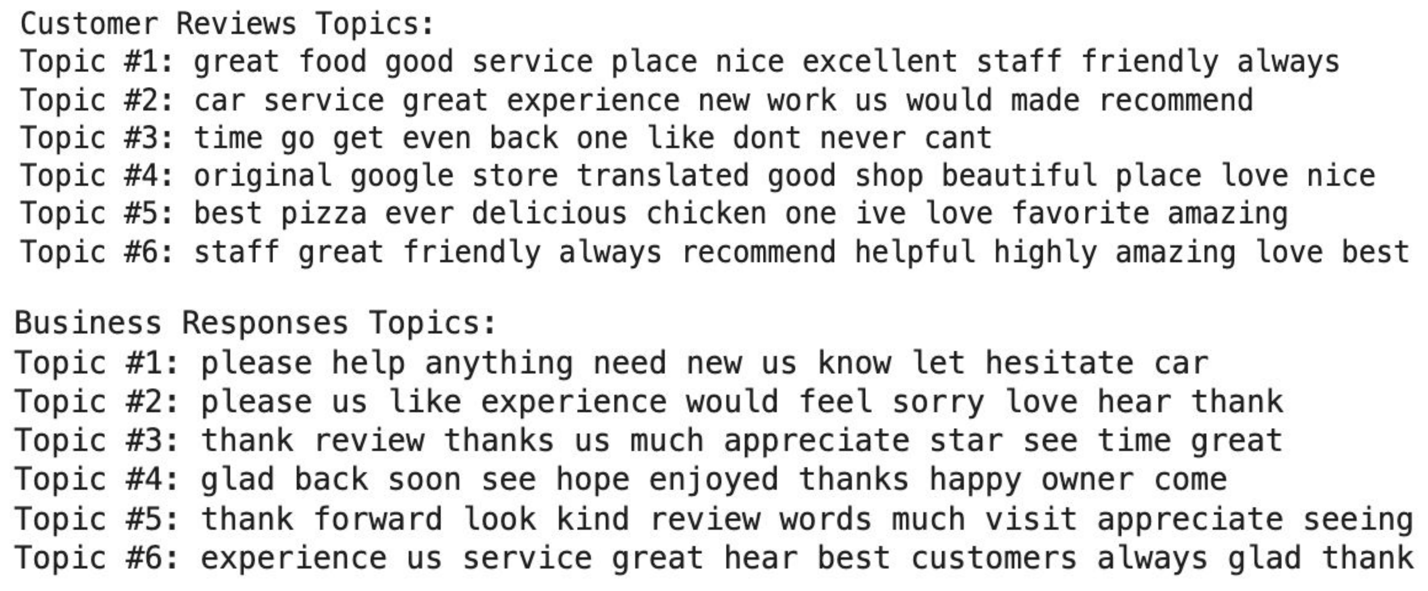
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Topic 1	1.000	0.277	0.478	0.136	0.443	0.297
Topic 2	0.277	1.000	0.240	0.040	0.310	0.461
Topic 3	0.478	0.240	1.000	0.126	0.473	0.220
Topic 4	0.136	0.040	0.126	1.000	0.106	0.043
Topic 5	0.443	0.310	0.473	0.106	1.000	0.315
Topic 6	0.297	0.461	0.220	0.043	0.315	1.000

Cosine similarity is the answer!

- Set threshold
- Group similar topics into major categories

LDA (Topic Modelling)

FROM THIS

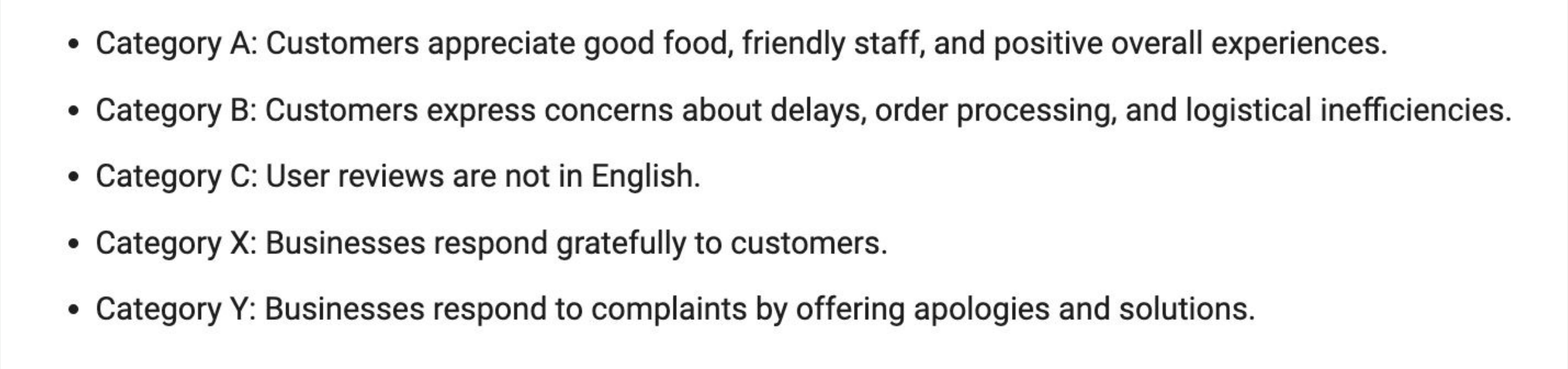


```
Customer Reviews Topics:
Topic #1: great food good service place nice excellent staff friendly always
Topic #2: car service great experience new work us would made recommend
Topic #3: time go get even back one like dont never cant
Topic #4: original google store translated good shop beautiful place love nice
Topic #5: best pizza ever delicious chicken one ive love favorite amazing
Topic #6: staff great friendly always recommend helpful highly amazing love best

Business Responses Topics:
Topic #1: please help anything need new us know let hesitate car
Topic #2: please us like experience would feel sorry love hear thank
Topic #3: thank review thanks us much appreciate star see time great
Topic #4: glad back soon see hope enjoyed thanks happy owner come
Topic #5: thank forward look kind review words much visit appreciate seeing
Topic #6: experience us service great hear best customers always glad thank
```

LDA (Topic Modelling)

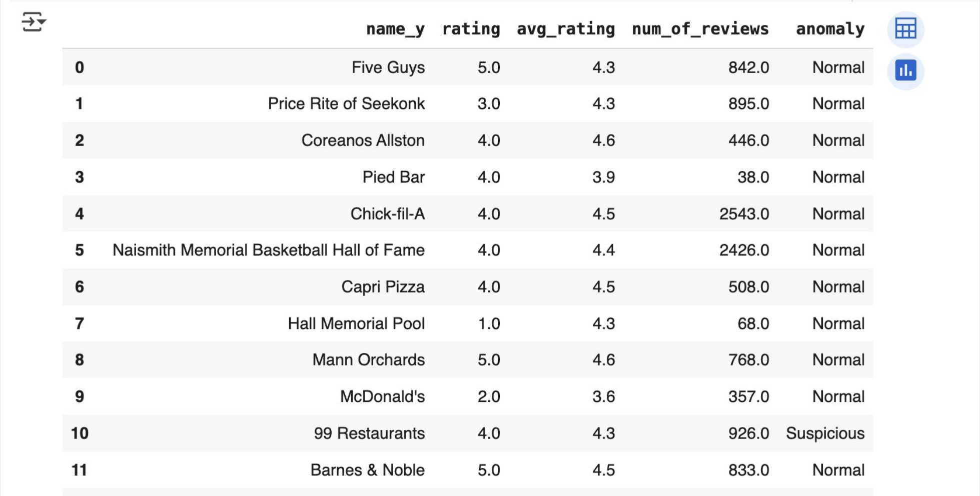
TO THIS



- Category A: Customers appreciate good food, friendly staff, and positive overall experiences.
- Category B: Customers express concerns about delays, order processing, and logistical inefficiencies.
- Category C: User reviews are not in English.
- Category X: Businesses respond gratefully to customers.
- Category Y: Businesses respond to complaints by offering apologies and solutions.

Isolation Forest (Anomaly Detection)

Isolation Forest is a machine learning method used to **find unusual or suspicious data points** in a dataset. Instead of learning what "normal" looks like, it focuses on **how easy it is to separate** a data point from the rest.



	name_y	rating	avg_rating	num_of_reviews	anomaly
0	Five Guys	5.0	4.3	842.0	Normal
1	Price Rite of Seekonk	3.0	4.3	895.0	Normal
2	Coreanos Allston	4.0	4.6	446.0	Normal
3	Pied Bar	4.0	3.9	38.0	Normal
4	Chick-fil-A	4.0	4.5	2543.0	Normal
5	Naismith Memorial Basketball Hall of Fame	4.0	4.4	2426.0	Normal
6	Capri Pizza	4.0	4.5	508.0	Normal
7	Hall Memorial Pool	1.0	4.3	68.0	Normal
8	Mann Orchards	5.0	4.6	768.0	Normal
9	McDonald's	2.0	3.6	357.0	Normal
10	99 Restaurants	4.0	4.3	926.0	Suspicious
11	Barnes & Noble	5.0	4.5	833.0	Normal

Manual tuning of Isolation Forest Results

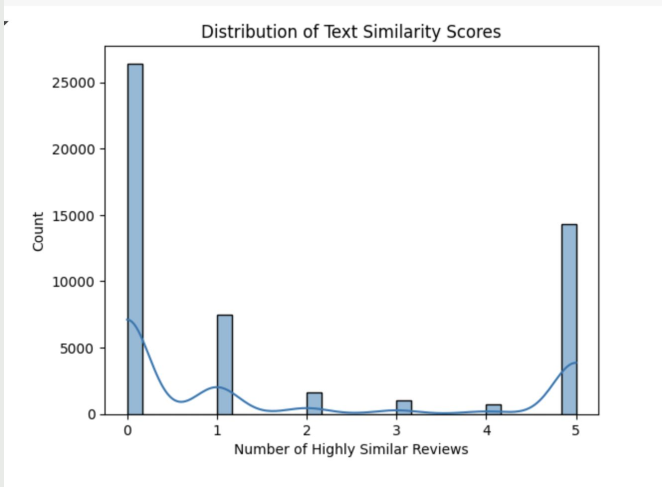
Hyperparameter Tuning

Best Parameters: {'contamination': 0.05, 'max_samples': 1860, 'n_estimators': 64}

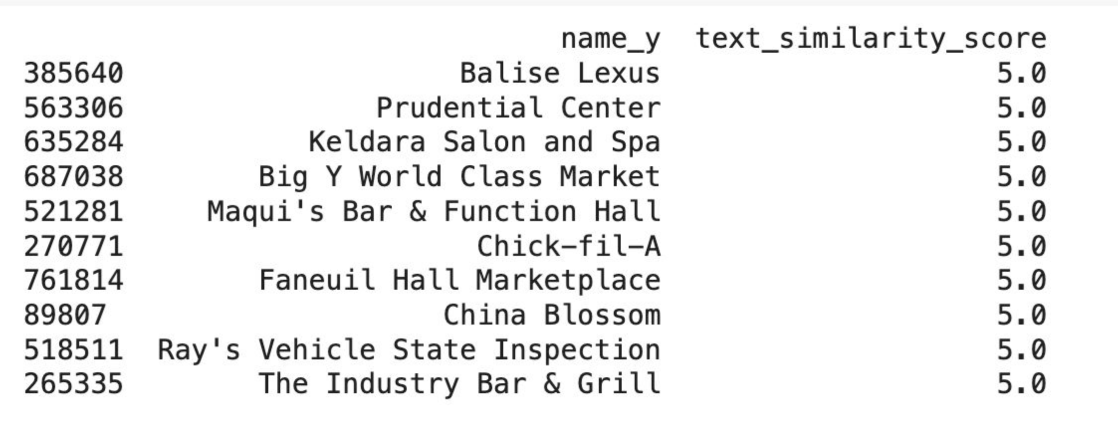
RandomizedSearchCV:

- contamination=0.05 → Higher contamination than our initial 0.03, meaning more businesses are flagged as anomalies.
- max_samples=1860 → Much lower than our manually chosen 5000, suggesting that a smaller sample size works better for this dataset.
- n_estimators=64 → Lower than our initial 100, meaning fewer trees were sufficient for optimal performance in this tuning run, balancing detection accuracy and computation time.

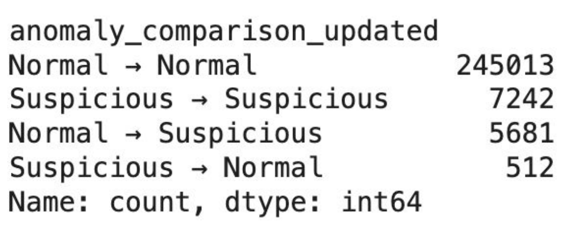
Results of Hyperparameter Tuning



Text Similarity Score

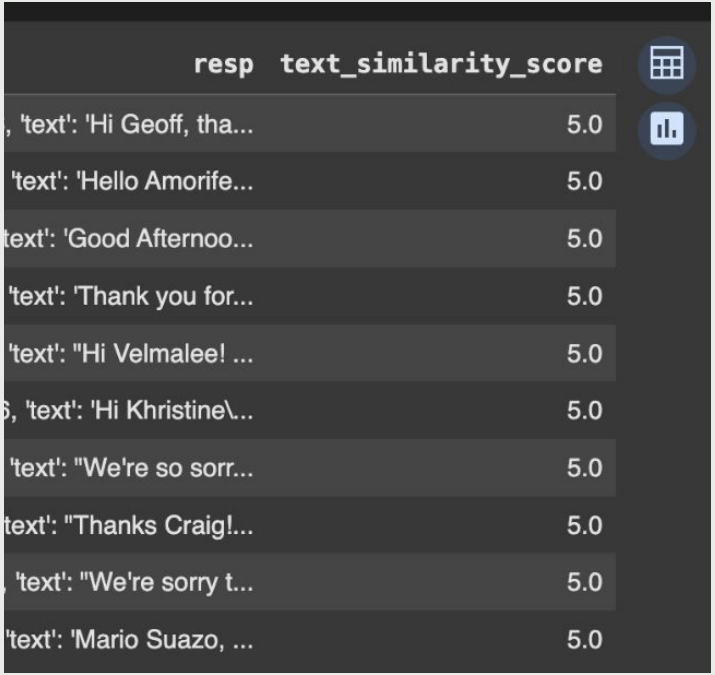


```
                                    name_y  text_similarity_score
385640                         Balise Lexus                    5.0
563306                   Prudential Center                    5.0
635284               Keldara Salon and Spa                    5.0
687038              Big Y World Class Market                    5.0
521281         Maqui's Bar & Function Hall                    5.0
270771                          Chick-fil-A                    5.0
761814              Faneuil Hall Marketplace                    5.0
89807                         China Blossom                    5.0
518511    Ray's Vehicle State Inspection                    5.0
265335              The Industry Bar & Grill                    5.0
```



```
anomaly_comparison_updated
Normal → Normal               245013
Suspicious → Suspicious         7242
Normal → Suspicious             5681
Suspicious → Normal              512
Name: count, dtype: int64
```

Optimized Model Results

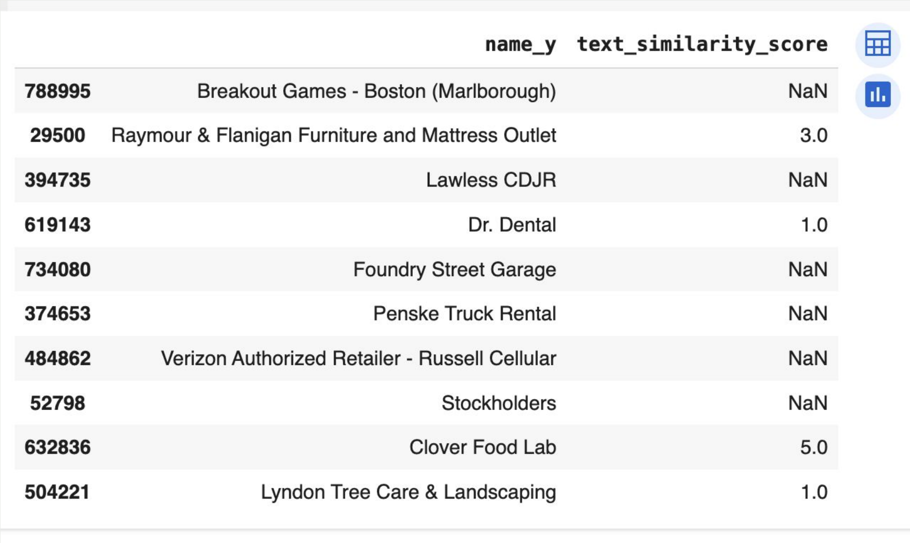


resp	text_similarity_score
, 'text': 'Hi Geoff, tha...	5.0
'text': 'Hello Amorife...	5.0
text': 'Good Afternoo...	5.0
'text': 'Thank you for...	5.0
'text': "Hi Velmalee! ...	5.0
5, 'text': 'Hi Khristine\...	5.0
'text': "We're so sorr...	5.0
text': "Thanks Craig!...	5.0
, 'text': "We're sorry t...	5.0
'text': 'Mario Suazo, ...	5.0

Similar to the topics found in LDA model

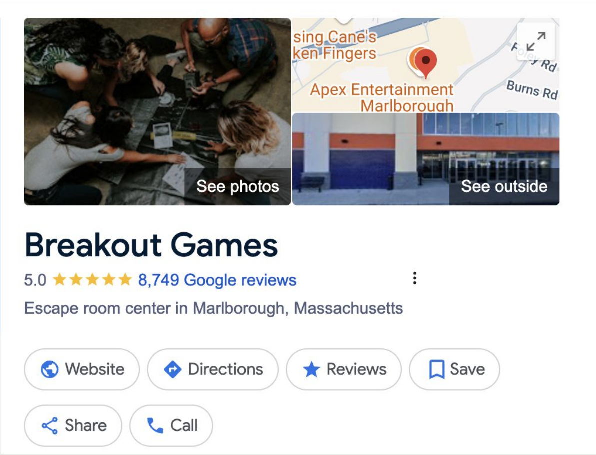
“Thank you, Thanks, Hello, Sorry, etc”

EXAMPLES



	name_y	text_similarity_score
788995	Breakout Games - Boston (Marlborough)	NaN
29500	Raymour & Flanigan Furniture and Mattress Outlet	3.0
394735	Lawless CDJR	NaN
619143	Dr. Dental	1.0
734080	Foundry Street Garage	NaN
374653	Penske Truck Rental	NaN
484862	Verizon Authorized Retailer - Russell Cellular	NaN
52798	Stockholders	NaN
632836	Clover Food Lab	5.0
504221	Lyndon Tree Care & Landscaping	1.0

List of businesses found suspicious through all the models.



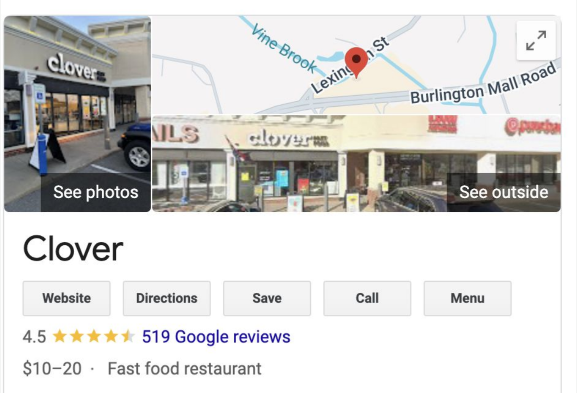
Breakout Games

5.0 ★★★★★ 8,749 Google reviews

Escape room center in Marlborough, Massachusetts

9000 reviews, 5 star rating

hmmm



Clover

4.5 ★★★★★ 519 Google reviews

$10–20 · Fast food restaurant

500 reviews,

Found suspicious through Text similarity score of 5.0

Challenges

Sampling Strategy:

Sampled 10%-20% of the data, improving performance while retaining reliable insights. But we may lose some information from the dataset.

Computationally Expensive:

Our methods cost us long time to run the coding blocks, especially with our large dataset.

Feature Adjustment: (K-Means)

Removed **latitude** and **longitude** features and re-adjusted the model

Limitations

Overfitting:(K-Means, LDA)

Might be overfitting based on the clear clusters visualization(bc of StandardScaler)

Parameter sensitivity: (DBSCAN, Isolation Forest)

The results are highly dependent on the choice of **parameters**.

Challenges with Text Similarity Analysis: (LDA, Isolation Forest)

Real-World Applications

Fraud Detection:

Identifying businesses with potentially fake reviews to enhance platform integrity.

Reputation Management:

Helping legitimate businesses analyze and improve their customer engagement strategies.

Consumer Protection:

Assisting regulatory bodies in maintaining a transparent and fair marketplace.

Thank You!