DATA ANALYST PROJECT

ON

YELP

Ranjit Kumar Singh

USER ENGAGEMENT ANALYSIS FOR RESTAURANT SUCCESS

About YELP

Yelp is a web and Mobile platform that functions as a crowd-sourced local business review site. Users can submit reviews, photos, and tips about businesses, while also browsing information and rating left by others.

AGENDA

- Problem Statement
- Research Objectives
- Hypothesis
- Data Overview
- Analysis and Findings
- Recommendations

Problem Statement

In a competitive market like the restaurant industry, understanding the factors that influence business is crucial for stakeholders. Utilizing the Yelp dataset, this project aims to investigate the relationship between user engagement (reviews, tips, and check-ins) and business success metrics (review ,count, ratings) for restaurants.

Research Objectives

- Quantify the correlation between user engagement(reviews , tips, check-ins)and review count/average star rating: This will help us determine if restaurant with higher user engagement a corresponding increase in review and ratings.

- Analyze the impact of sentiment on review count and average star rating: We will investigate if positive sentiment in review and tips translates to higher star rating and potentially influences the total number of reviews left.

- Time trends in User Engagement We will explore if consistent user engagement over time is a stronger indicator of long-term success compared to sporadic bursts of activity.

Hypothesis Testing

• Higher levels of user engagement (more reviews, tips, and check-ins)correlate with higher review counts and ratings for restaurant.

• Positive sentiment expressed in reviews and tips contributes to higher overall rating counts for restaurants.

• Consistent engagement over time is positively associated with sustained business success for restaurants.

Data Overview

• This dataset is a subset of Yelp and has information about business across 8 metropolitan areas in the USA and Canada.

• The original data is shared by Yelp as JSON files.

• The five JSON files are business, review, user, tip and check-in.

• The JSON files are stored in the database for easy retrieval of data.

Database Creation

[1]:
```
import pandas as pd
import json
from sqlalchemy import create_engine
```

[5]:
```
with open('yelp_academic_dataset_business.json', 'r', encoding='utf-8') as f:
    business_data = [json.loads(line) for line in f]
business_df = pd.DataFrame(business_data)

with open('yelp_academic_dataset_checkin.json','r') as f:
    checkin_data = [json.loads(line) for line in f]
checkin_df = pd.DataFrame(checkin_data)
```

[7]:
```
with open('yelp_academic_dataset_tip.json','r',encoding='utf-8') as f:
    tip_data = [json.loads(line) for line in f]
tip_df = pd.DataFrame(tip_data)
```

[9]:
```
with open('yelp_academic_dataset_user.json','r',encoding='utf-8') as f:
    user_data = [json.loads(line) for line in f]
user_df = pd.DataFrame(user_data)
```

[3]:
```
import pandas as pd
import dask.dataframe as dd

# Load a manageable sample using Pandas
df = pd.read_json('yelp_academic_dataset_review.json', lines=True, nrows=100000)
```

[3]:
```python
import pandas as pd
import dask.dataframe as dd

# Load a manageable sample using Pandas
df = pd.read_json('yelp_academic_dataset_review.json', lines=True, nrows=100000)

# Save as Parquet for efficient future access
df.to_parquet('yelp_reviews.parquet')

# Now load the Parquet file with Dask
review_df = dd.read_parquet('yelp_reviews.parquet')

# Example processing: compute the number of reviews
total_reviews = review_df.shape[0].compute()
print(f'Total number of reviews: {total_reviews}')
```

```
Total number of reviews: 100000
```

[11]:
```python
print(business_df.shape)
print(checkin_df.shape)
print(review_df.shape)
print(tip_df.shape)
print(user_df.shape)
```

```
(150346, 14)
(131930, 2)
(<dask_expr.expr.Scalar: expr=ReadParquetFSSpec(32d3e87).size() // 9, dtype=int32>, 9)
(908915, 5)
(1987897, 22)
```

[13]:
```python
business_df.drop(['attributes','hours'],axis = 1, inplace = True)
```

[3]:
```
import dask.dataframe as dd
from sqlalchemy import create_engine

# Create a SQLAlchemy engine for SQLite
engine = create_engine('sqlite:///Yelp.db')

def load_dataframe(df, table_name, engine):
    # Convert Dask DataFrame to Pandas DataFrame
    df_pd = df.compute()  # This computes the Dask DataFrame into a Pandas DataFrame
    df_pd.to_sql(table_name, con=engine, if_exists='replace', index=False)

# Define a function to read the JSON file in chunks
def load_json_to_dask_dataframe(file_path):
    return dd.read_json(file_path, lines=True, blocksize='64MB')  # Adjust blocksize as needed

# Load JSON files as Dask DataFrames
business_df = load_json_to_dask_dataframe('yelp_academic_dataset_business.json')
checkin_df = load_json_to_dask_dataframe('yelp_academic_dataset_checkin.json')
review_df = load_json_to_dask_dataframe('yelp_academic_dataset_review.json')
tip_df = load_json_to_dask_dataframe('yelp_academic_dataset_tip.json')
user_df = load_json_to_dask_dataframe('yelp_academic_dataset_user.json')

# Load each DataFrame into a separate table
load_dataframe(business_df, 'business', engine)
load_dataframe(checkin_df, 'checkin', engine)
load_dataframe(review_df, 'review', engine)
load_dataframe(tip_df, 'tip', engine)
load_dataframe(user_df, 'user', engine)
```

Analysis and Findings

[27]:
```
# What is the descriptive stats for review count and star rating for business?
# Avg, Min, Max, Median

pd.read_sql_query(f"""SELECT
AVG(review_count) AS Average_review_count,
MIN(review_count) AS Min_review_count,
Max(review_count) AS Max_review_count,
(SELECT review_count FROM business ORDER BY review_count LIMIT 1 OFFSET (SELECT COUNT(*) FROM Business) / 2) AS Median_review_count,

AVG(stars) AS Average_star_rating,
MIN(stars) AS Min_star_rating,
MAX(stars) AS Max_star_rating,
(SELECT stars FROM business ORDER BY stars LIMIT 1 OFFSET (SELECT COUNT(*) FROM business) / 2) AS Median_star_rating

FROM business
WHERE business_id IN {tuple(business_id['business_id'])};
""",conn).transpose()
```

[27]:

	0
Average_review_count	55.975426
Min_review_count	5.000000
Max_review_count	248.000000
Median_review_count	15.000000
Average_star_rating	3.477281
Min_star_rating	1.000000
Max_star_rating	5.000000
Median_star_rating	3.500000

- Out of 150K business, 35k are restaurants business and are open.
- Table showing distribution of business success metrics (review count and average rating)

Highest Rating

[29]:

	name	review_count	avg_rating
0	McDonald's	16490	1.868702
1	Chipotle Mexican Grill	9071	2.381757
2	Taco Bell	8017	2.141813
3	Chick-fil-A	7687	3.377419
4	First Watch	6761	3.875000
5	Panera Bread	6613	2.661905
6	Buffalo Wild Wings	6483	2.344828
7	Domino's Pizza	6091	2.290210
8	Wendy's	5930	2.030159
9	Chili's	5744	2.514706

Highest Review count

[31]:

	name	review_count	avg_rating
0	ā café	48	5.0
1	two birds cafe	77	5.0
2	the brewers cabinet production	13	5.0
3	taqueria la cañada	17	5.0
4	la bamba	44	5.0
5	la 5th av tacos	24	5.0
6	el sabor mexican and chinese food	21	5.0
7	eat.drink.Om...YOGA CAFE	7	5.0
8	d4 Tabletop Gaming Cafe	8	5.0
9	cabbage vegetarian cafe	12	5.0

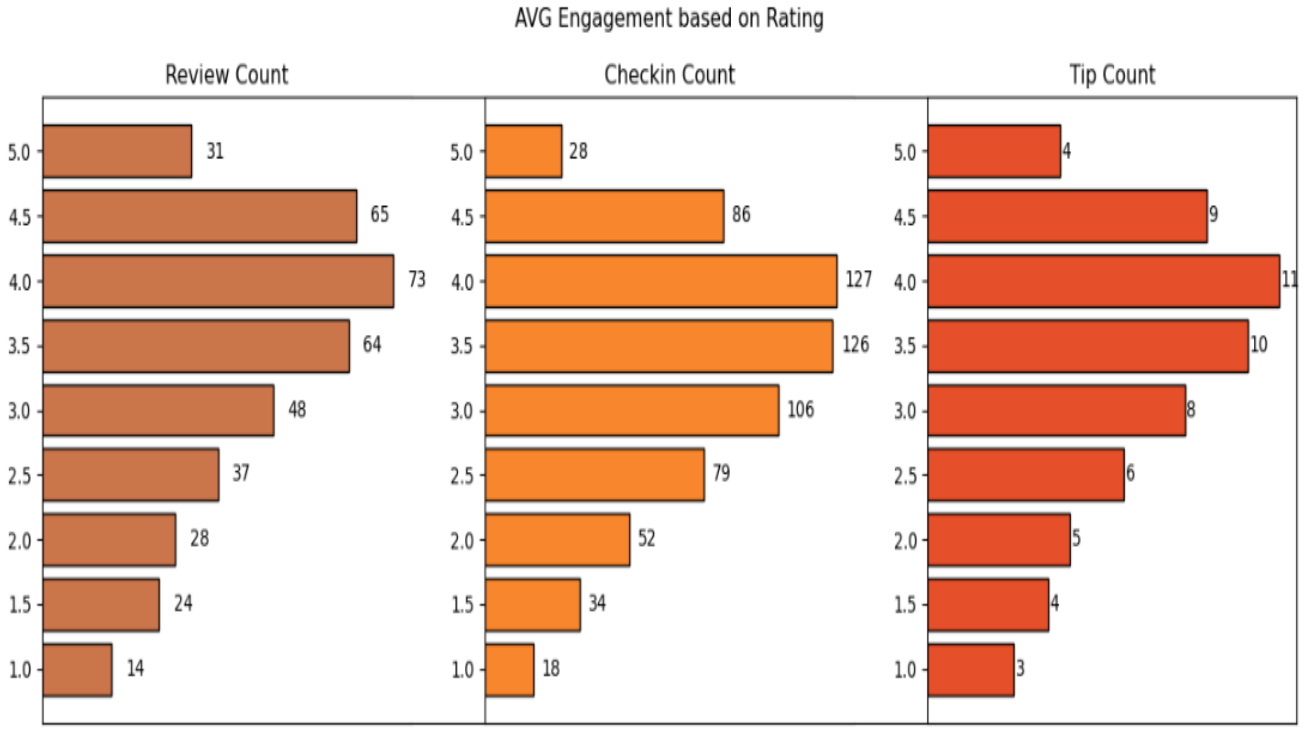
- Highest rating do no guarantee a higher review count, or vice versa.
- Success of Restaurants is not solely determined by rating or review counts.
- Review count reflects user engagement but not necessarily overall customer satisfaction or business performance.

Do restaurants with higher engagement tend to have higher ratings?

• Data shows a general increase in average review, check-in, and tip counts as rating improve from 1 to 4 stars.

• Restaurants rated 4 stars exhibit the highest engagement and shows a downward trend for rating above 4.

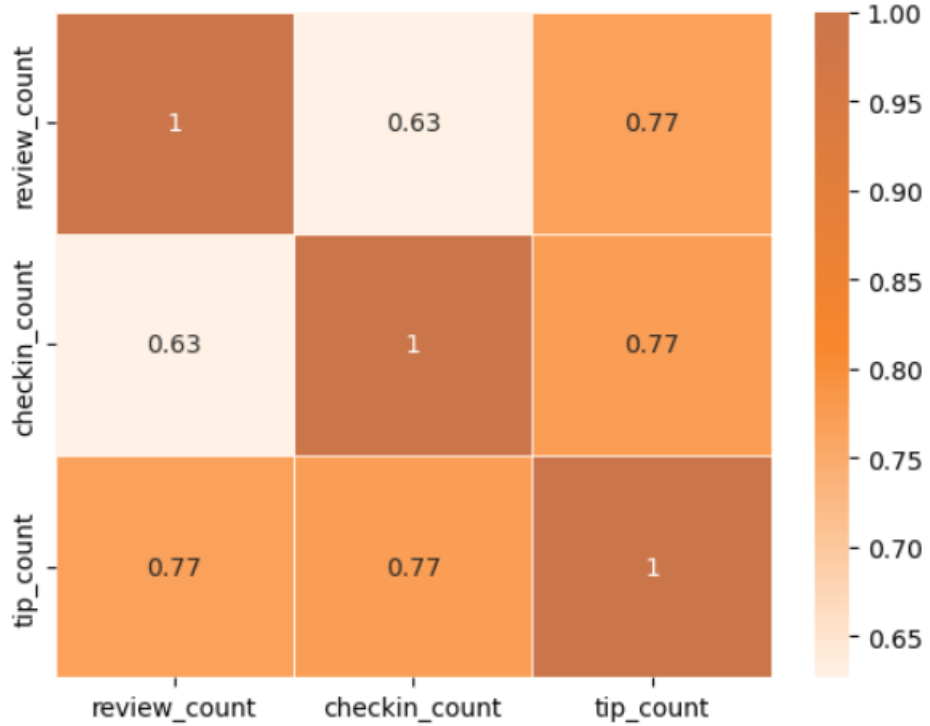
• The drop in engagement at 5.0 stars might suggest either a saturation point where fewer customers feel compelled to add their review, or a selectivity where only a small, satisfied audience frequents there establishments.



Is there a correlation between the number of reviews, tips, and check-ins for a business?

• These correlations suggest that user engagement across different platforms(reviews, tips, and check-ins) is interlinked; higher activity in one area trend to be associated with higher activity in others.

• Business should focus on strategies that boost all types of user engagement, as increase in one type of engagement are likely to drive increases in others, enhancing overall visibility and interaction with customers.



Is there a difference in the user engagement between high-rated and low-rated business?

- Data indicates a clear correlation between higher rating and increased user engagement across reviews, tips and check-ins.

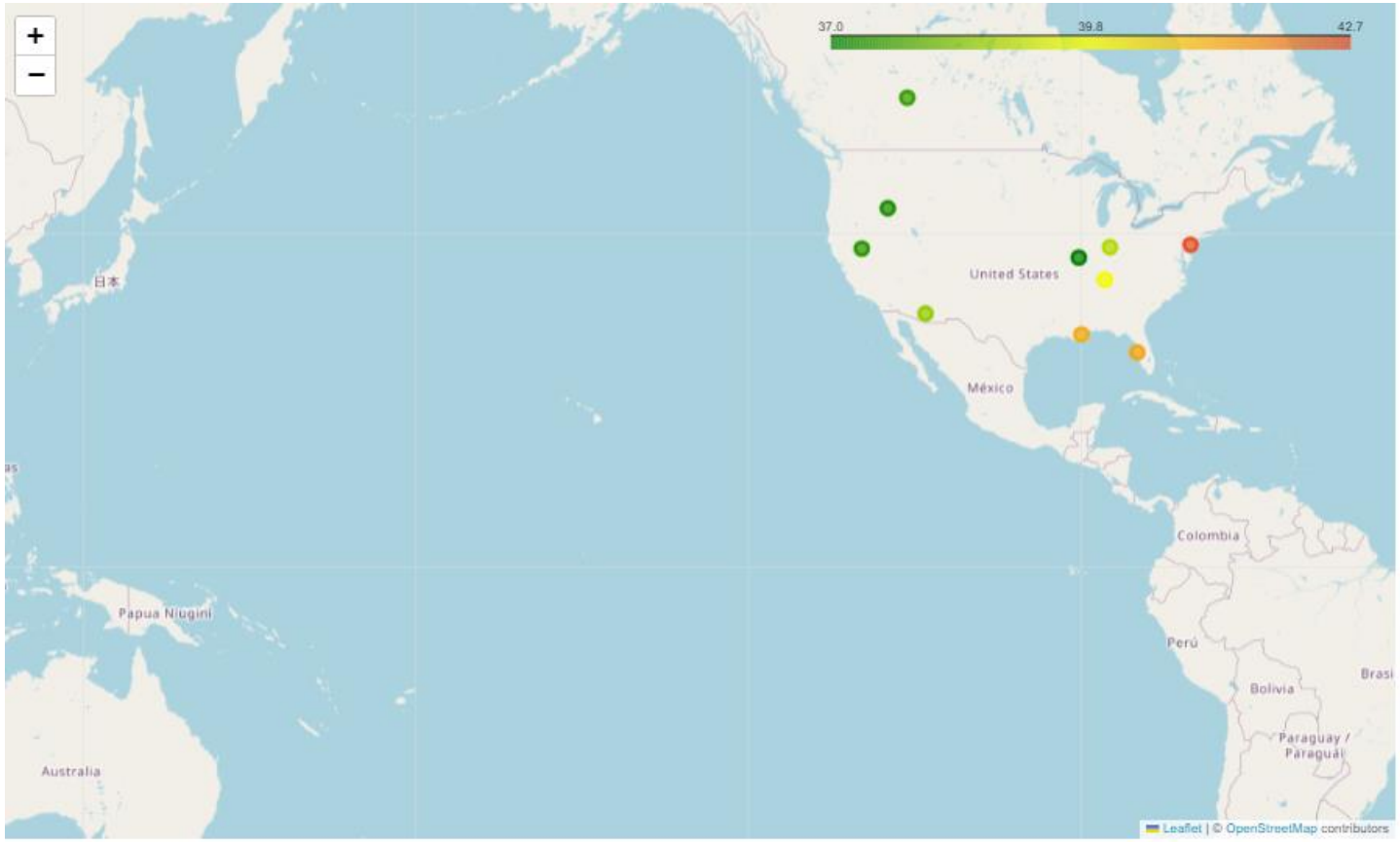
- This pattern underscores the importance of maintaining high service and quality standards, as these appear to drive more review, check-ins, and tips, which are critical metrics of customer engagement and satisfaction.

category	review_count	tip_count	checkin_count
High-Rated	72.291062	10.162766	122.066641
Low-Rated	42.123420	6.541689	88.880828

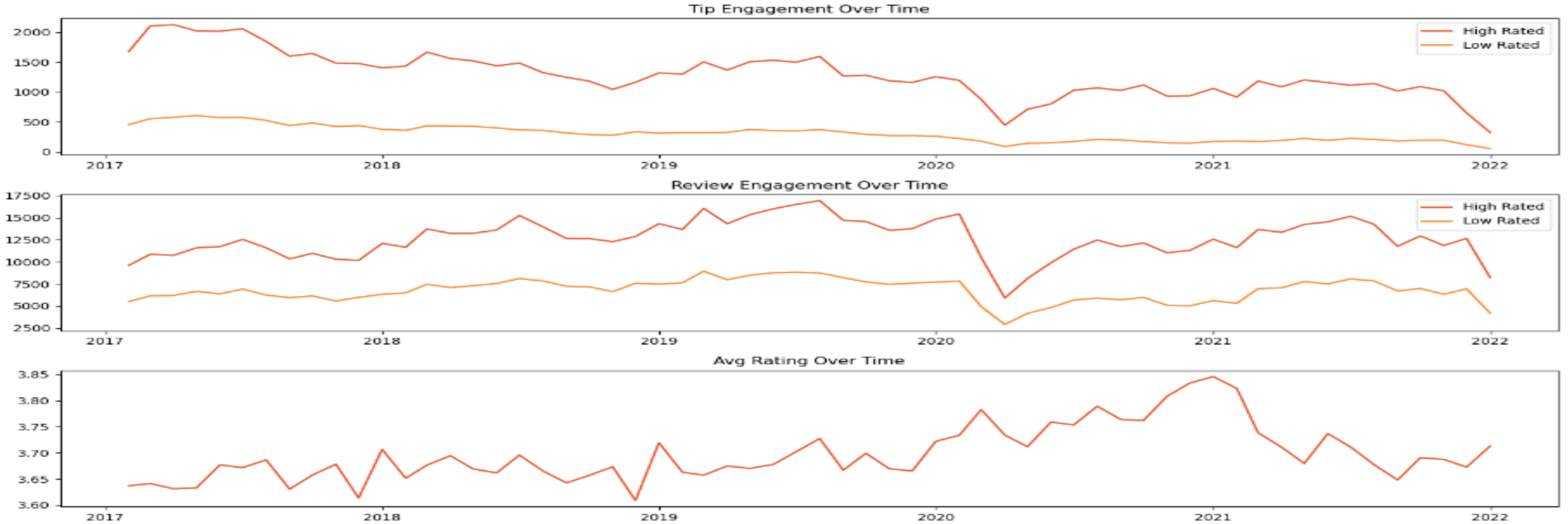
How do the sucess metrics of restaurants vary across different states and cities?

• Philadelphia emerges as the top city with the highest success score, indicating a combination of higher ratings and active user engagement.

•Following Philadelphia, Tampa, Indian polis, and Tucson rank among the top cities with significant success scores, suggesting thriving restaurant sences in these areas.



Are there any patterns in user engagement over time for successful business compared to less successful ones?

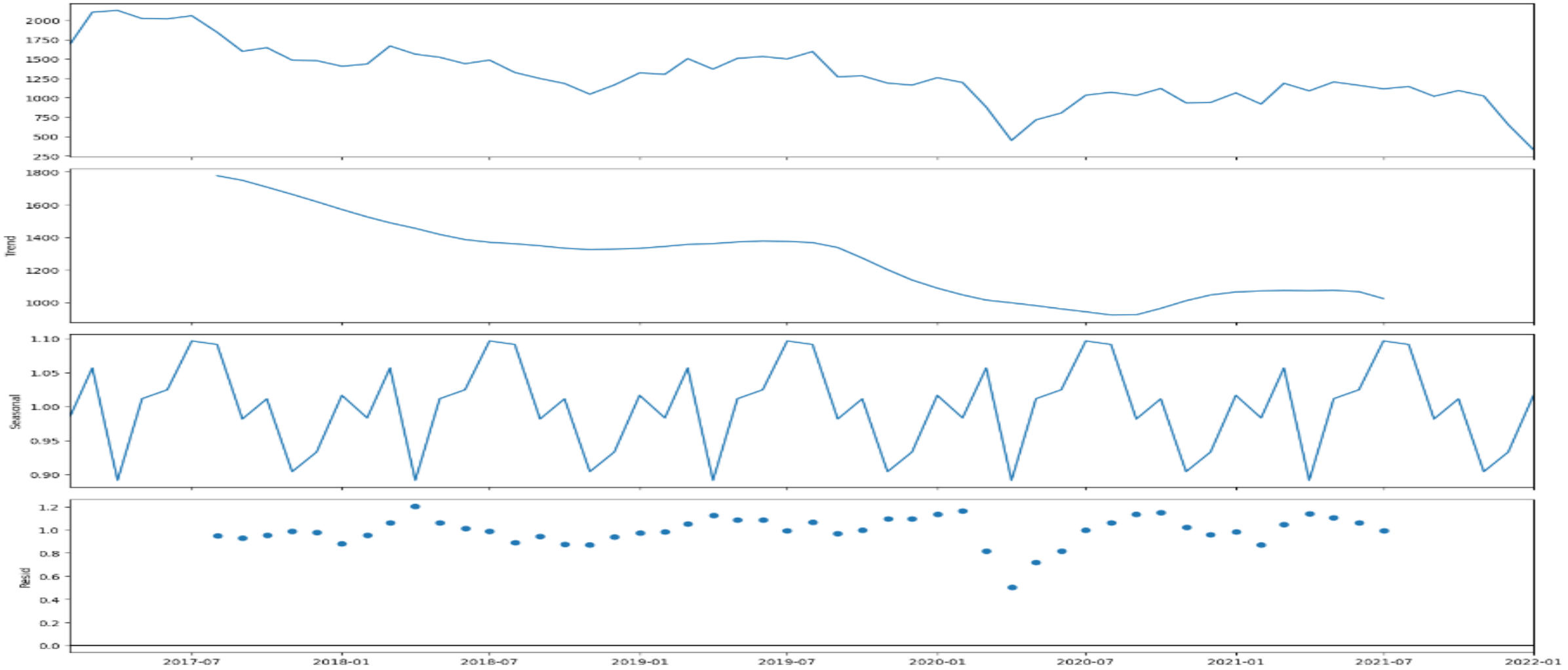


•Successful business, particularly those with higher ratings (above 3.5), exhibit consistent user engagement over time.

•High rated restaurants maintain a steady or growing level of user engagement over time, reflecting ongoing customer interest and satisfaction.

```
tip_high_rated = high_rated_engagement[['month_year','tip_count']].set_index('month_year')
review_high_rated = high_rated_engagement[['month_year','review_count']].set_index('month_year')
rating_df = time_rating[['month_year','avg_rating']].set_index('month_year')
```

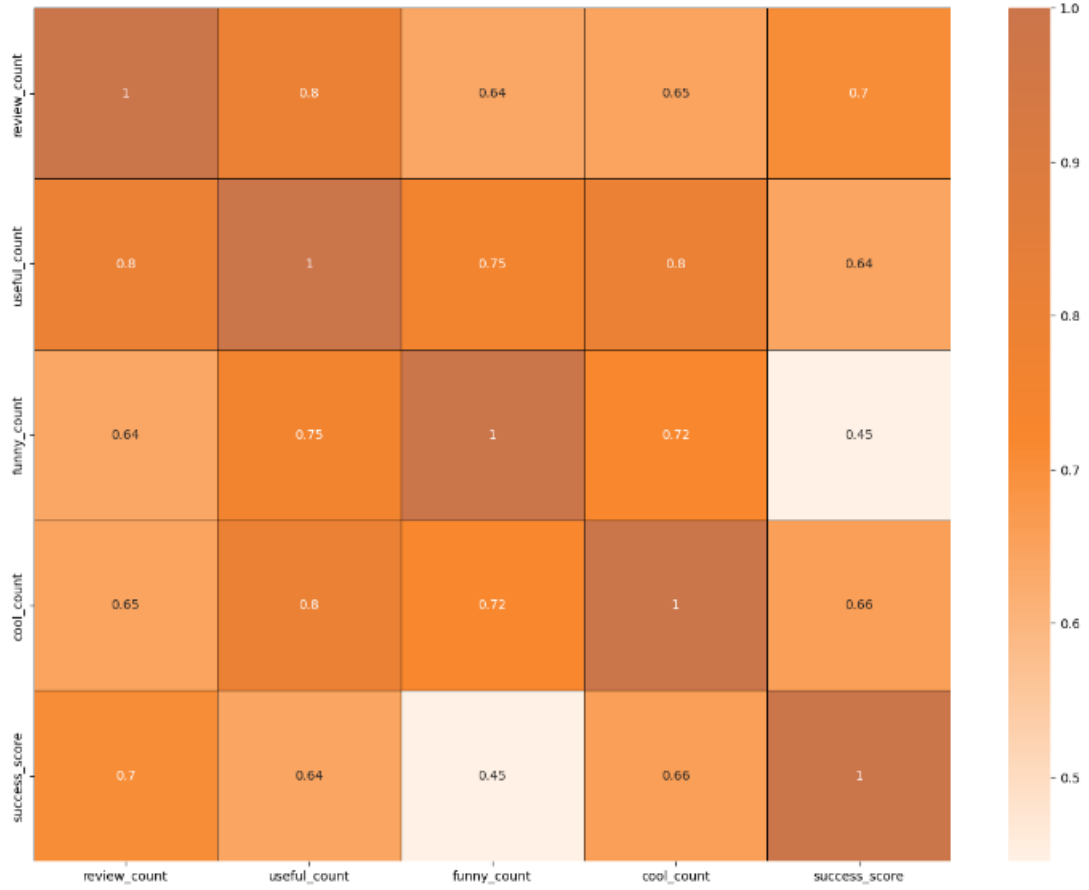
```
from statsmodels.tsa.seasonal import seasonal_decompose
multiplicative_decomposition = seasonal_decompose(tip_high_rated,
                                                  model='multiplicative', period = 12)
plt.rcParams.update({'figure.figsize': (16,12)})
multiplicative_decomposition.plot()
plt.show()
```



How does the sentiment of reviews and tips (useful, funny, cool) correlate with the success metrics of restaurants?

- "Useful", "funny", and "cool" are attributes associated with user reviews. They represent the feedback provided buy users about the usefulness, humor, or coolness of a particular review.

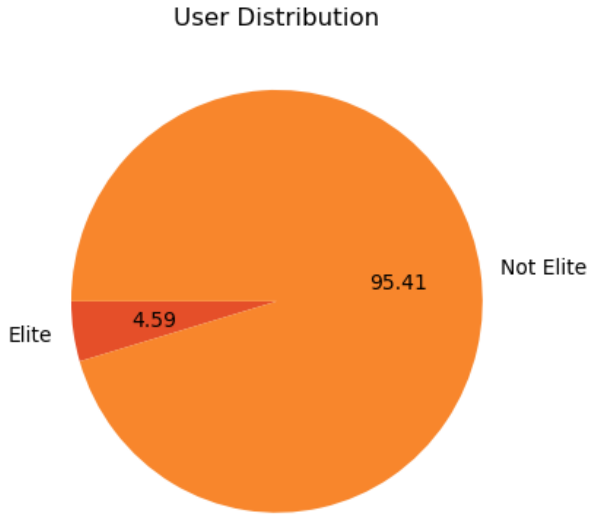
- Higher counts of useful, funny, and cool reviews suggest greater user engagement and satisfaction, which are key factors contributing to a restaurants success.

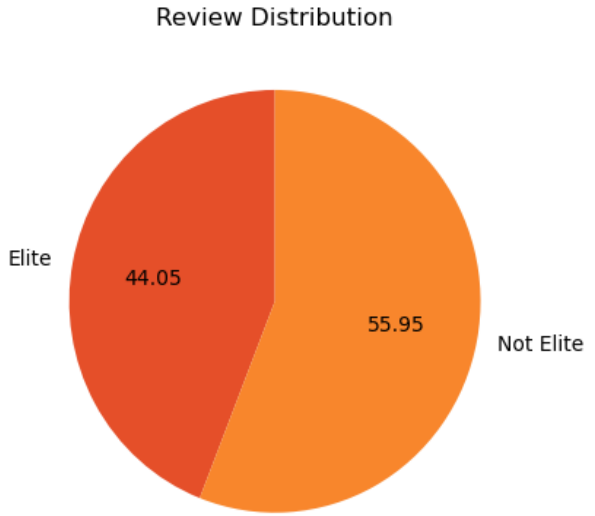


Is there any difference in engagement of elite users and non elite users?

• Elite users are individuals who have been recognized and awarded the "Elite" status by Yelp for their active and high-quality contributions.

• Elite users, despite being significantly fewer in number, contribute a substantial proportion of the total review count compared to non-elite users.





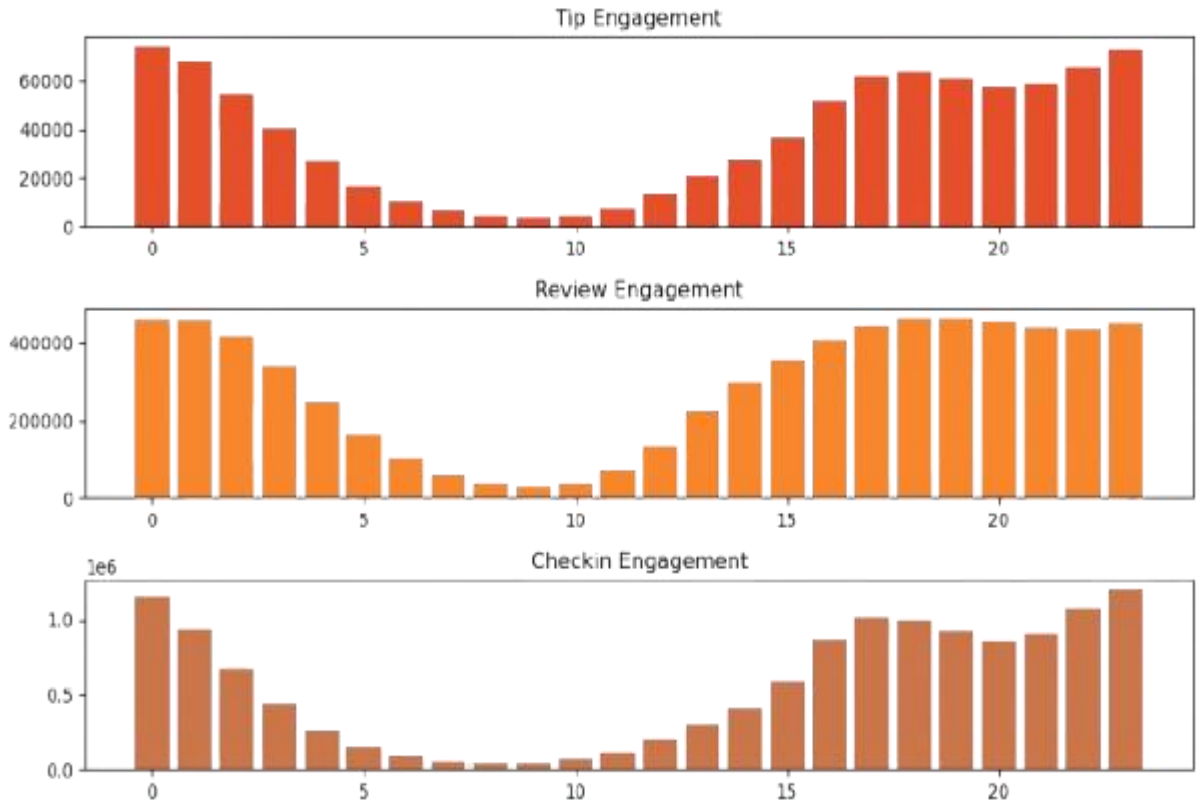
• Establishing a positive relationship with elite users can lead to repeat visits and loyalty, as they are more likely to continue supporting businesses they have good experiences with.

Busiest Hours

• The busiest hours for restaurants, based on user engagement , span from 4 pm to 1 am.

• Knowing the peak hours allows business to optimize their staffing levels and resource allocation during these times to ensure efficient operations and quality service delivery.

• The concentration of user engagement during the evening and night suggests a higher demand for dining out during these times , potentially driven by factors such as work schedules, social gatherings, and leisure activities.



Recommendations

• Utilizing insights from the analysis of various metrics such as user engagement, sentiment of reviews, peak hours, and the impact of elite users, business can make informed decisions to drive success.

• Collaborating with elite users and leveraging their influence can amplify promotional efforts, Increase brand awareness , and drive customer acquisition.

• Businesses can adjust their operating hours or introduce special promotions to capitalize on the increased demand during peak hours.

• Less successful businesses may need to focus on strategies to enhance user engagement over time, such as improving service quality , responding to customer feedback.

• Cities with high success score present opportunities for restaurant chain to expand or invest further.

THANK YOU