WEB SCRAPING PROJECT: OPENRICE

Airin Konno

Project Outline

1. Introduction
2. Aim
3. Data Collection, Preprocessing
4. Exploratory Data Analysis
5. Insights
6. Challenges and Limitations
7. Conclusion

Introduction

- OpenRice is currently the most popular food dining guide in Hong Kong, providing the most comprehensive restaurant information.
- As a consumer, it would be an interesting topic to explore the most popular dishes, prices, and districts for the hottest restaurants.
- For businesses, it would be useful to have data in order to formulate strategies to improve their website rating, increase profits, as well as look at competitor data.

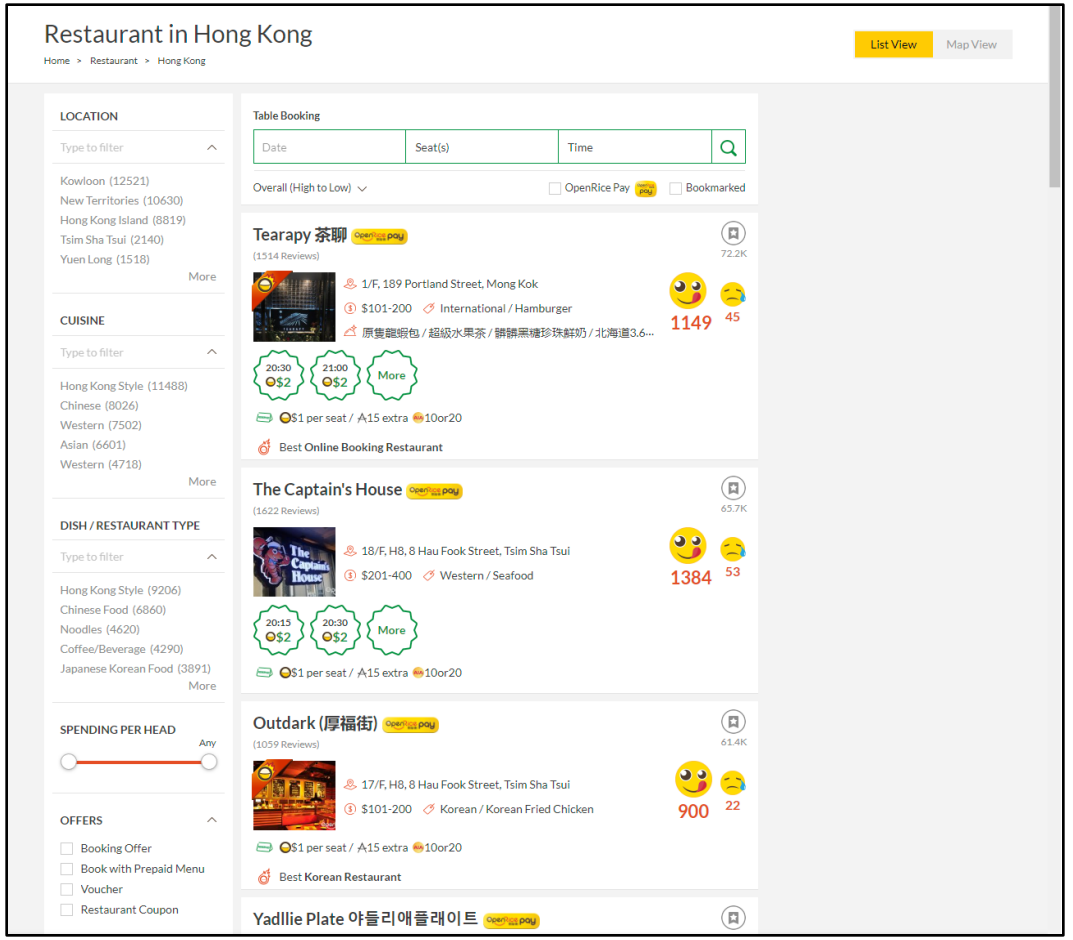
Aim

This project aims to:

1. Explore the dataset using pandas.
2. Create data visualizations using matplotlib and seaborn.
3. Perform data analysis to explain the data visualizations that have been created.

Data Collection

- Data was scraped from openrice.com using BeautifulSoup and Selenium.
- The restaurants were sorted by "Overall (High to Low)", where restaurants with the most engagement (ratings, bookmarks, reviews) were ranked the highest.
- Data scraping was limited to 17 pages (250 restaurants), and could not be scraped further.



Dataset Description

- Structure of the dataset

Variable Name	Description	Sample Data
restaurant_name	Name of restaurant	Tearapy 茶聊; Yadllie Plate 야들리애플래이트
price_range_per_head	Pricing per head	$101-200; $201-400…
good_ratings	Number of good ratings	1149; 1384…
bad_ratings	Number of bad ratings	45; 53…
full_address	Restaurant address	1/F, 189 Portland Street, Mong Kok…
district	Restaurant district	Mong Kok; Tsim Sha Tsui…
cuisine	Type of cuisine	International; Western…
dish	Type of dish in particular	Hamburger; Steak House…
written_reviews	Number of reviews written	1514; 1622…
bookmarks	Number of bookmarks	72200; 65700…

Dataset Info: Initial Data Frame

```
df.head()
```

	restaurant_name	price_range_per_head	good_ratings	bad_ratings	full_address	district	cuisine	dish	written_reviews	bookmarks
0	Tearapy 茶聊	$101-200	1149	45	1/F, 189 Portland Street, Mong Kok	Mong Kok	International	Hamburger	1514	72200
1	The Captain's House	$201-400	1386	54	18/F, H8, 8 Hau Fook Street, Tsim Sha Tsui	Tsim Sha Tsui	Western	Seafood	1625	65700
2	Outdark (厚福街)	$101-200	900	22	17/F, H8, 8 Hau Fook Street, Tsim Sha Tsui	Tsim Sha Tsui	Korean	Korean Fried Chicken	1059	61500
3	Yadllie Plate 야들리 애플래이트	$101-200	1183	66	11/F, CTMA Centre, 1 Sai Yeung Choi Street, Mo...	Mong Kok	Korean	Korean Fried Chicken	1571	117800
4	Outdark (飛達商業中心)	$201-400	1175	28	2/F, Fee Tat Commercial Centre, 613 Nathan Roa...	Mong Kok	Korean	NaN	1370	67800

```
df.describe()
```

	good_ratings	bad_ratings	written_reviews	bookmarks
count	250.000000	250.000000	250.000000	250.000000
mean	651.116000	21.400000	773.088000	30584.400000
std	254.393504	19.119157	307.036189	18936.366615
min	242.000000	0.000000	337.000000	5000.000000
25%	442.000000	9.000000	531.500000	16500.000000
50%	575.000000	16.000000	693.000000	25700.000000
75%	807.500000	27.000000	952.250000	38450.000000
max	1386.000000	149.000000	1671.000000	117800.000000

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   restaurant_name       250 non-null    object
 1   price_range_per_head  250 non-null    object
 2   good_ratings          250 non-null    int64
 3   bad_ratings           250 non-null    int64
 4   full_address          250 non-null    object
 5   district              250 non-null    object
 6   cuisine               250 non-null    object
 7   dish                  228 non-null    object
 8   written_reviews       250 non-null    int64
 9   bookmarks             250 non-null    int64
dtypes: int64(4), object(6)
memory usage: 19.7+ KB
```

Dataset Pre-Processing: Null Values

- To ensure high quality of data, we must clean dirty data.
- Handling null values:
  - There was 22 null values under the column 'dish'.
  - Based on this, missing values in 'dish' will be omitted.
  - We end up with 228 rows of data.

```
df_nan = df
```

```
df_nan.isnull().sum()
```

```
restaurant_name          0
price_range_per_head     0
good_ratings             0
bad_ratings              0
full_address             0
district                 0
cuisine                  0
dish                    22
written_reviews          0
bookmarks                0
dtype: int64
```

Dataset Pre-Processing: Adding New Columns

- For EDA purposes, the overall rating score will be needed.
  - To do this, we must change the data type of 'good_ratings' and 'bad_ratings' to integers.
- New columns:
  - Sum_of_ratings = good_ratings + bad_ratings
  - Overall_rating_percentage_score = good_ratings / sum_of_ratings

```
df.dtypes

restaurant_name          object
price_range_per_head     object
good_ratings             object
bad_ratings              object
full_address             object
district                 object
cuisine                  object
dish                     object
written_reviews           int64
bookmarks                 int64
dtype: object
```

```
df.dtypes

restaurant_name                 object
price_range_per_head            object
good_ratings                     int64
bad_ratings                      int64
full_address                    object
district                        object
cuisine                         object
dish                            object
written_reviews                  int64
bookmarks                        int64
sum_of_ratings                   int64
overall_rating_percentage_score  float64
dtype: object
```

Dataset Pre-Processing: Feature Engineering

- New columns will be created by extracting features present in the dataset for further analysis.
- New columns:
  - Region: Created by grouping districts.
    - Hong Kong Island
    - Kowloon
    - New Territories
  - Average price per head: Created by getting the average of the price per head range.

Dataset Info: Final Data Frame

```
df.head()
```

full_address	district	cuisine	dish	written_reviews	bookmarks	sum_of_ratings	overall_rating_percentage_score	average_price_per_head	region
1/F, 189 Portland Street, Mong Kok	Mong Kok	International	Hamburger	1514	72200	1194	96.231156	150.0	Kowloon
18/F, H8, 8 Hau Fook Street, Tsim Sha Tsui	Tsim Sha Tsui	Western	Seafood	1625	65700	1440	96.250000	300.0	Kowloon
17/F, H8, 8 Hau Fook Street, Tsim Sha Tsui	Tsim Sha Tsui	Korean	Korean Fried Chicken	1059	61500	922	97.613883	150.0	Kowloon
11/F, CTMA Centre, 1 Sai Yeung Choi Street, Mo...	Mong Kok	Korean	Korean Fried Chicken	1571	117800	1249	94.715773	150.0	Kowloon
G/F, Chinachem Cameron Centre, 42 Cameron Road...	Tsim Sha Tsui	Japanese	Sushi/Sashimi	1398	46000	1256	96.815287	300.0	Kowloon

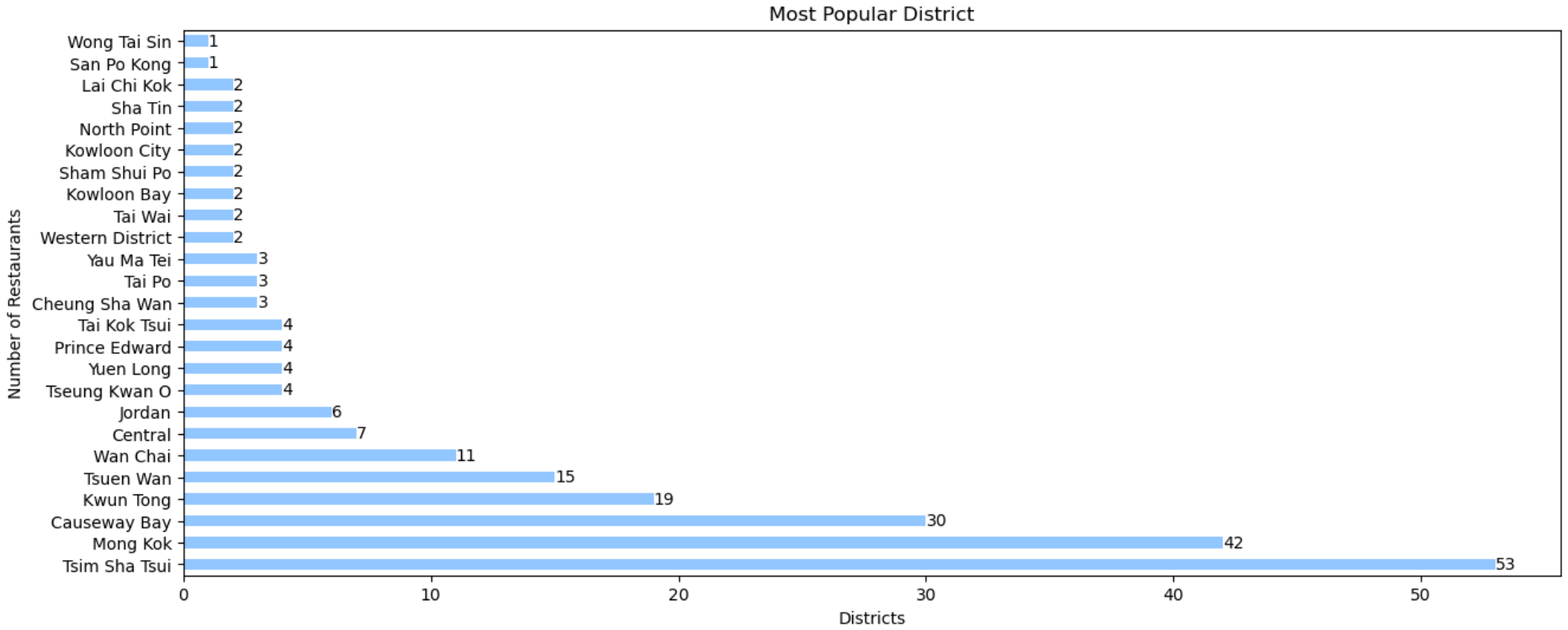
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 226 entries, 0 to 249
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   restaurant_name                 226 non-null    object
 1   price_range_per_head            226 non-null    object
 2   good_ratings                    226 non-null    int64
 3   bad_ratings                     226 non-null    int64
 4   full_address                    226 non-null    object
 5   district                        226 non-null    object
 6   cuisine                         226 non-null    object
 7   dish                            226 non-null    object
 8   written_reviews                 226 non-null    int64
 9   bookmarks                       226 non-null    int64
 10  sum_of_ratings                  226 non-null    int64
 11  overall_rating_percentage_score 226 non-null    float64
 12  average_price_per_head          226 non-null    float64
 13  region                          226 non-null    object
dtypes: float64(2), int64(5), object(7)
memory usage: 26.5+ KB
```

```
df.describe()
```

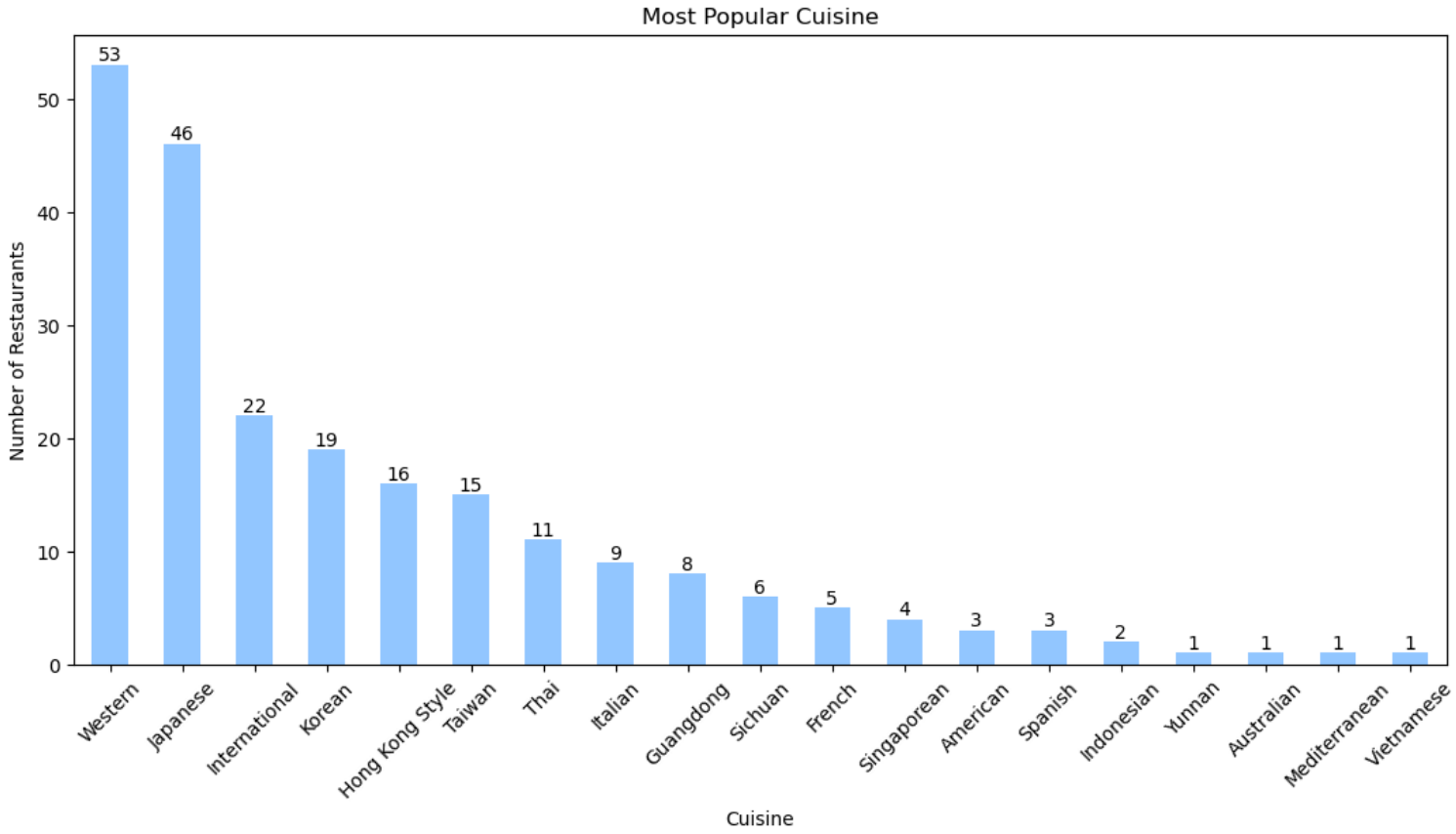
	good_ratings	bad_ratings	written_reviews	bookmarks	sum_of_ratings	overall_rating_percentage_score	average_price_per_head
count	226.000000	226.000000	226.000000	226.000000	226.000000	226.000000	226.000000
mean	651.721239	20.641593	773.884956	31040.707965	672.362832	97.005071	194.137168
std	252.782941	17.815345	305.871847	19226.195728	262.405761	2.198710	105.994900
min	242.000000	0.000000	337.000000	5000.000000	258.000000	81.155015	25.000000
25%	443.000000	9.000000	529.500000	16825.000000	460.250000	96.157051	150.000000
50%	595.500000	16.000000	701.500000	26400.000000	606.500000	97.444262	150.000000
75%	807.500000	26.000000	952.250000	38500.000000	839.500000	98.424393	300.000000
max	1386.000000	149.000000	1671.000000	117800.000000	1452.000000	100.000000	600.000000

Exploratory Data Analysis: Most Popular District



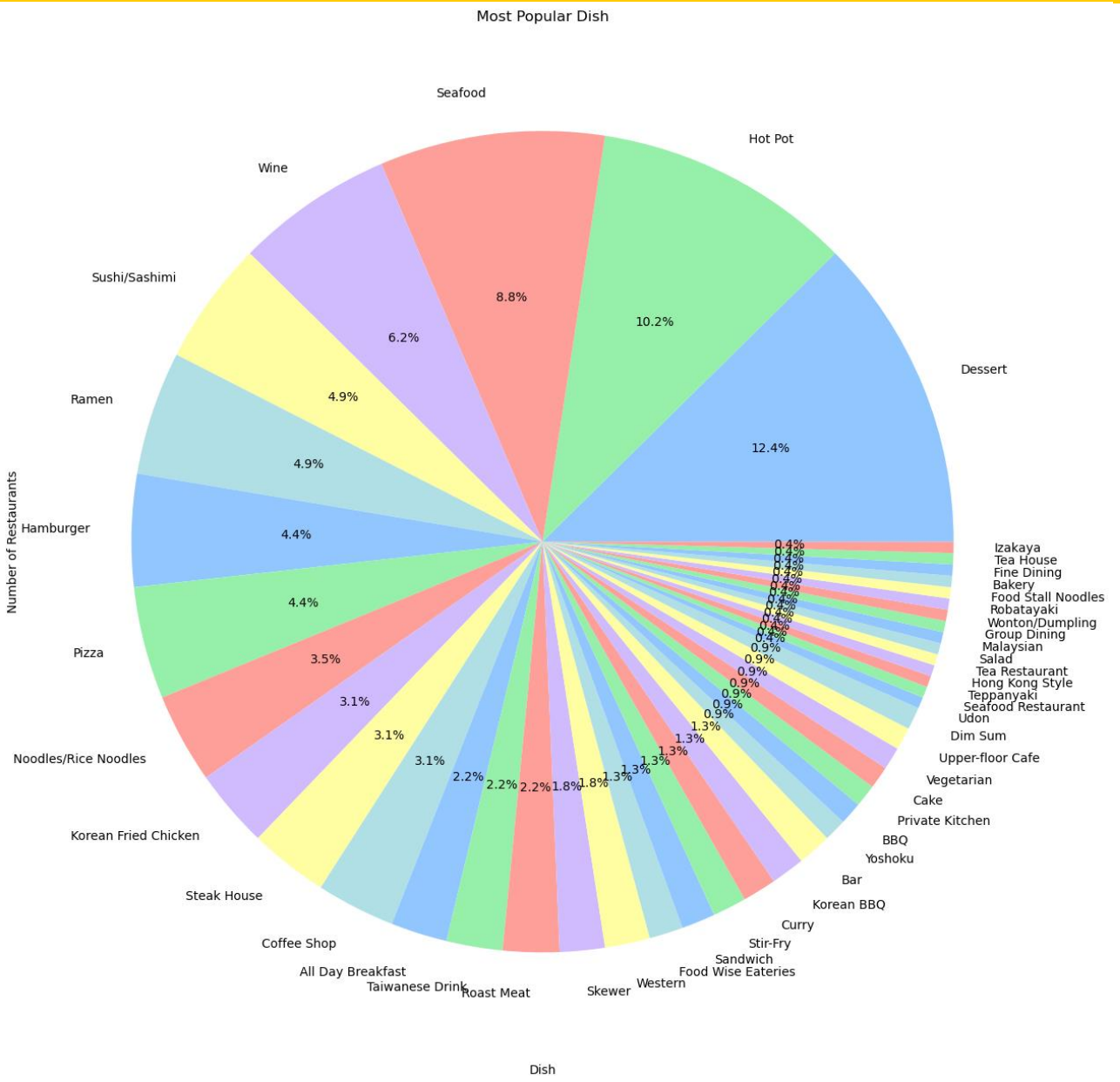
- The top areas for finding good restaurants are in Tsim Sha Tsui, Mong Kok and Causeway Bay

Exploratory Data Analysis: Most Popular Cuisine



- The top restaurants serve Western, Japanese, and International cuisine.

Exploratory Data Analysis: Most Popular Dishes

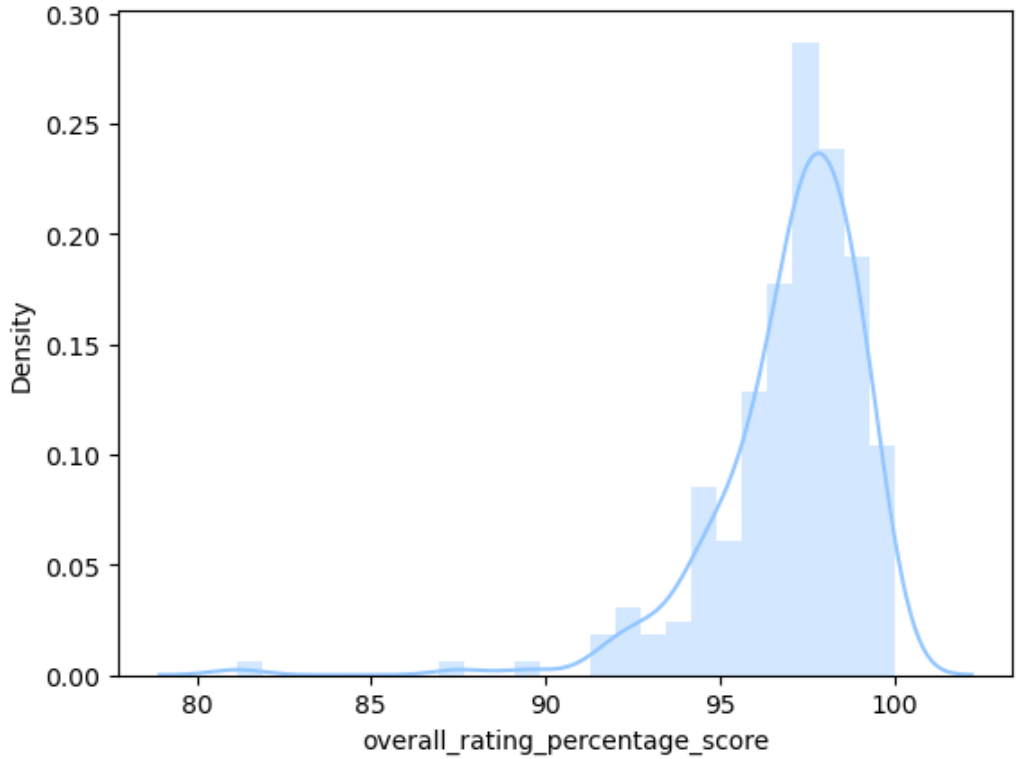


- The top restaurants are mostly populated by dessert, hot pot and seafood restaurants.
- Although Japanese cuisine is ranked second, it is surprising to see Sushi/Sashimi and Ramen ranked 5th and 6th in the most popular dishes.

Exploratory Data Analysis: Overall Rating Percentage Score

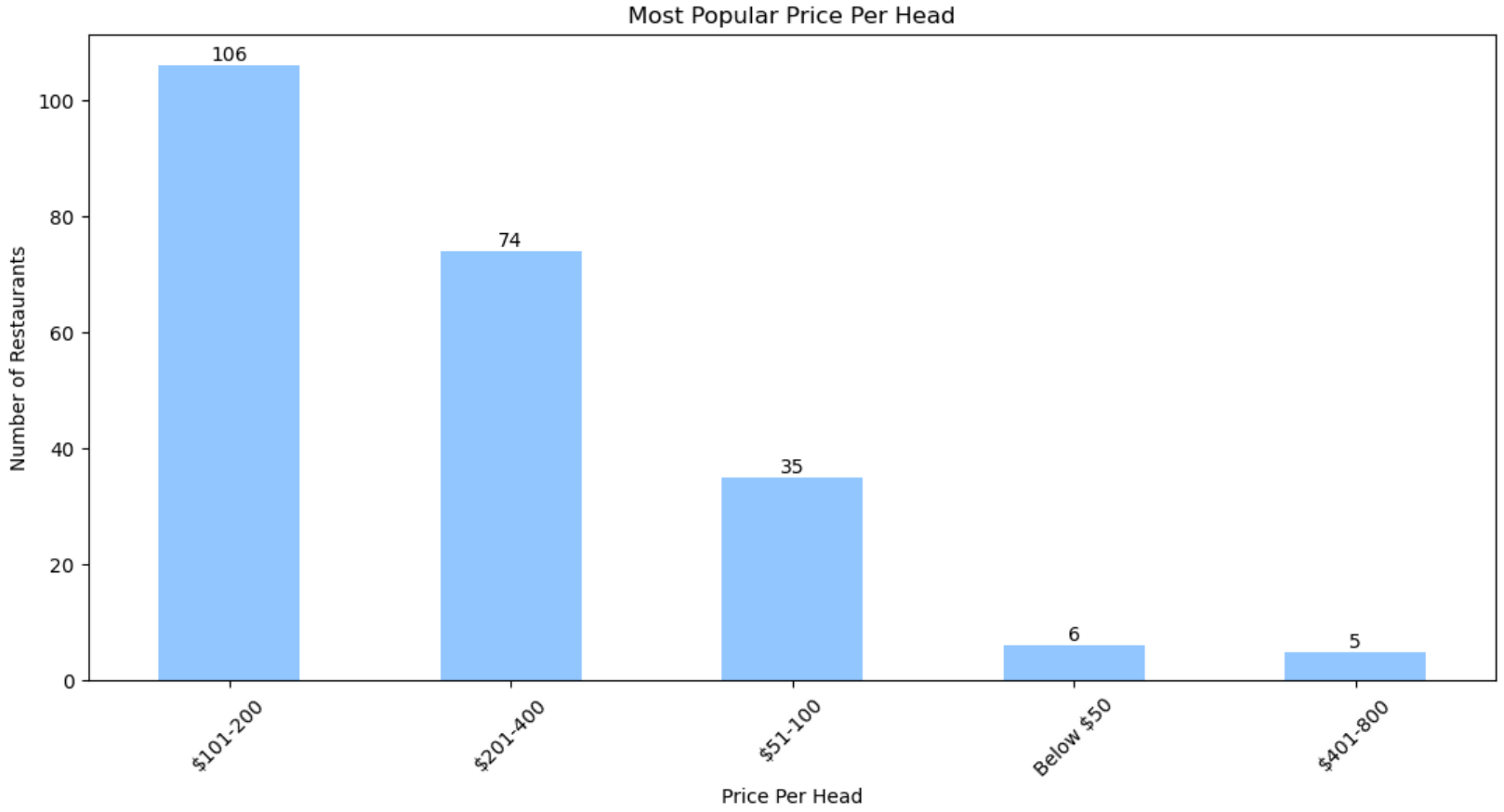
```
df.overall_rating_percentage_score.describe()
```

```
count    226.000000
mean      97.005071
std        2.198710
min       81.155015
25%       96.157051
50%       97.444262
75%       98.424393
max      100.000000
```



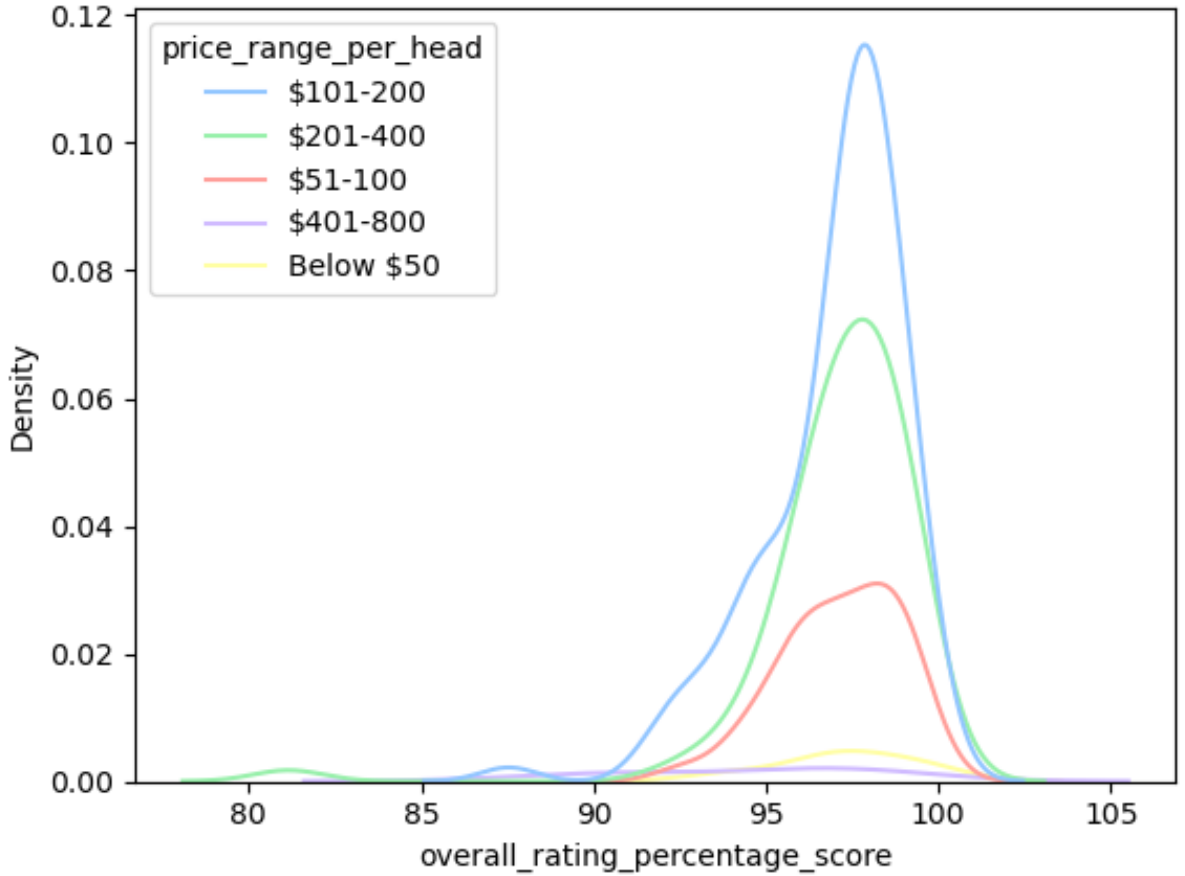
- In the top restaurants, the mean overall rating is 97%, with the min being 81.2% and the max being 100%.

Exploratory Data Analysis: Most Popular Price Per Head



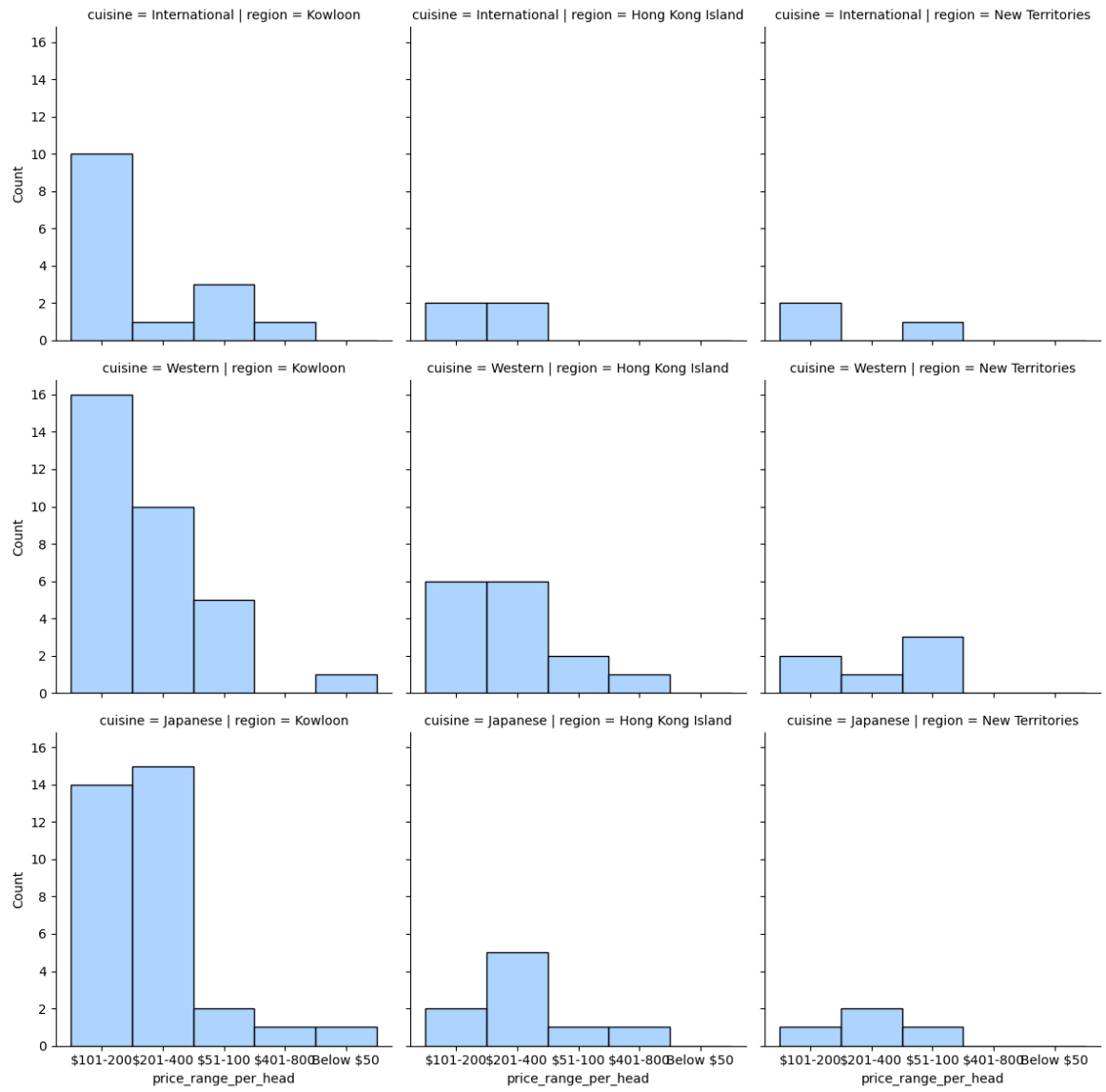
- There are the most restaurants within the $101-200pp price range.
- There are the least restaurants in the opposite of each spectrum listed in the top restaurants.

Exploratory Data Analysis: Price Per Head vs Overall Rating



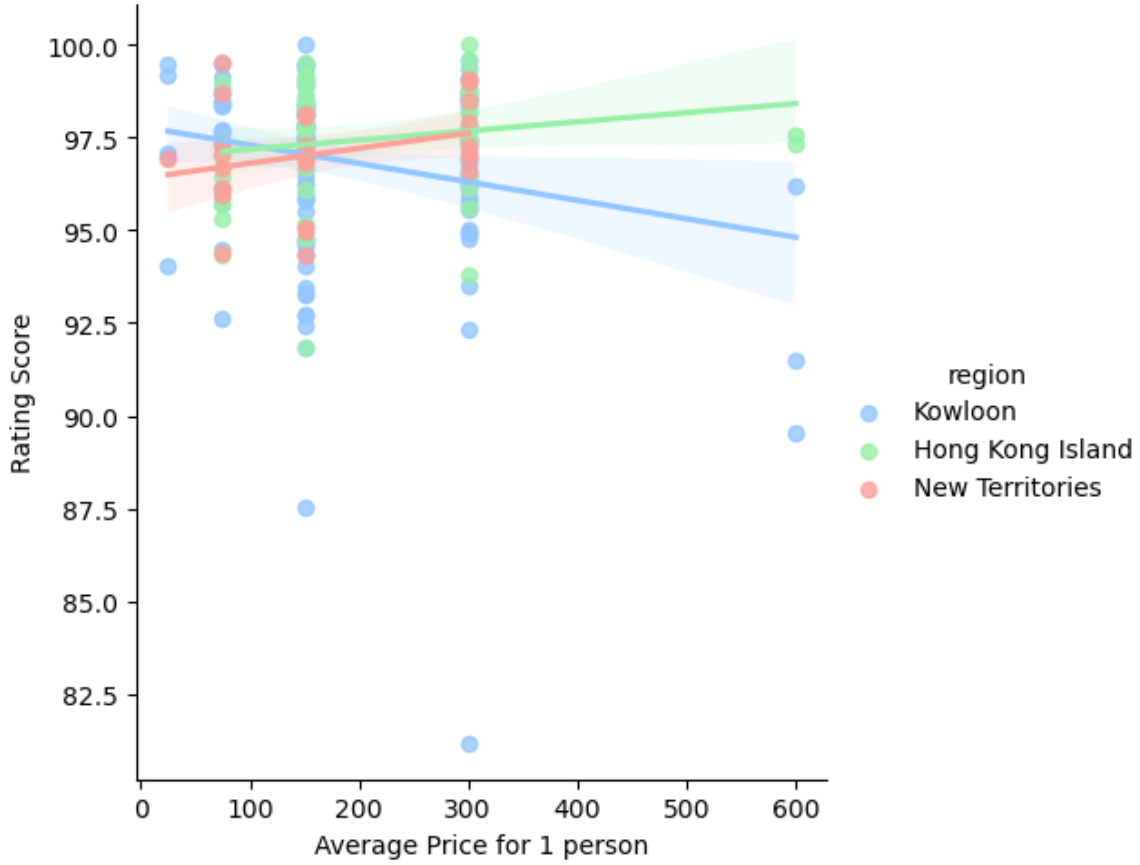
- There are more restaurants in the price range of $101-200 with a high overall rating compared to any other price range.
- All price ranges follow a similar pattern where they peak at 97-98% overall rating.

Exploratory Data Analysis: Where To Go On A Budget



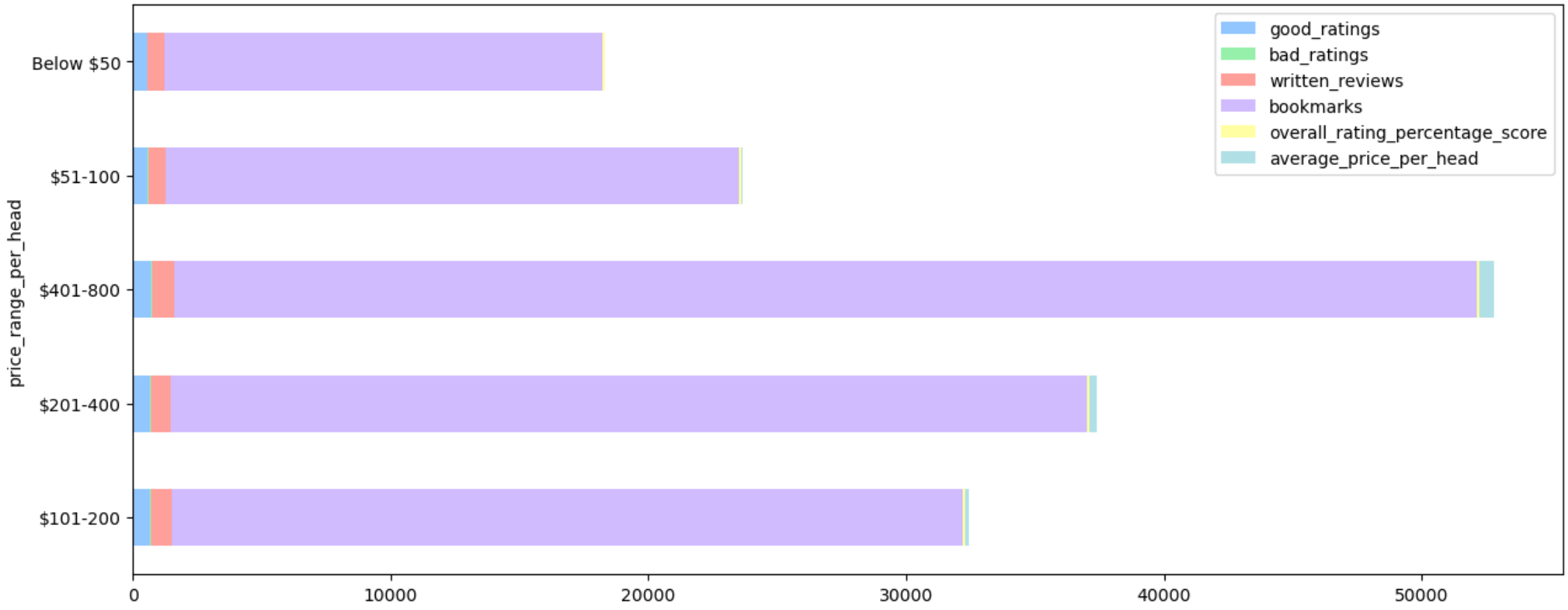
- Price Distribution for Top 3 Cuisine In Different Regions
- A meal at Kowloon and Hong Kong Island would more likely be in the price ranges of $101-200, and $201-400.
- In New Territories, you can find good restaurants in the $51-100 range.
- If you want Japanese, Western, or International food, Kowloon is a hot spot for all of these food.

Exploratory Data Analysis: Rating vs Price vs Region



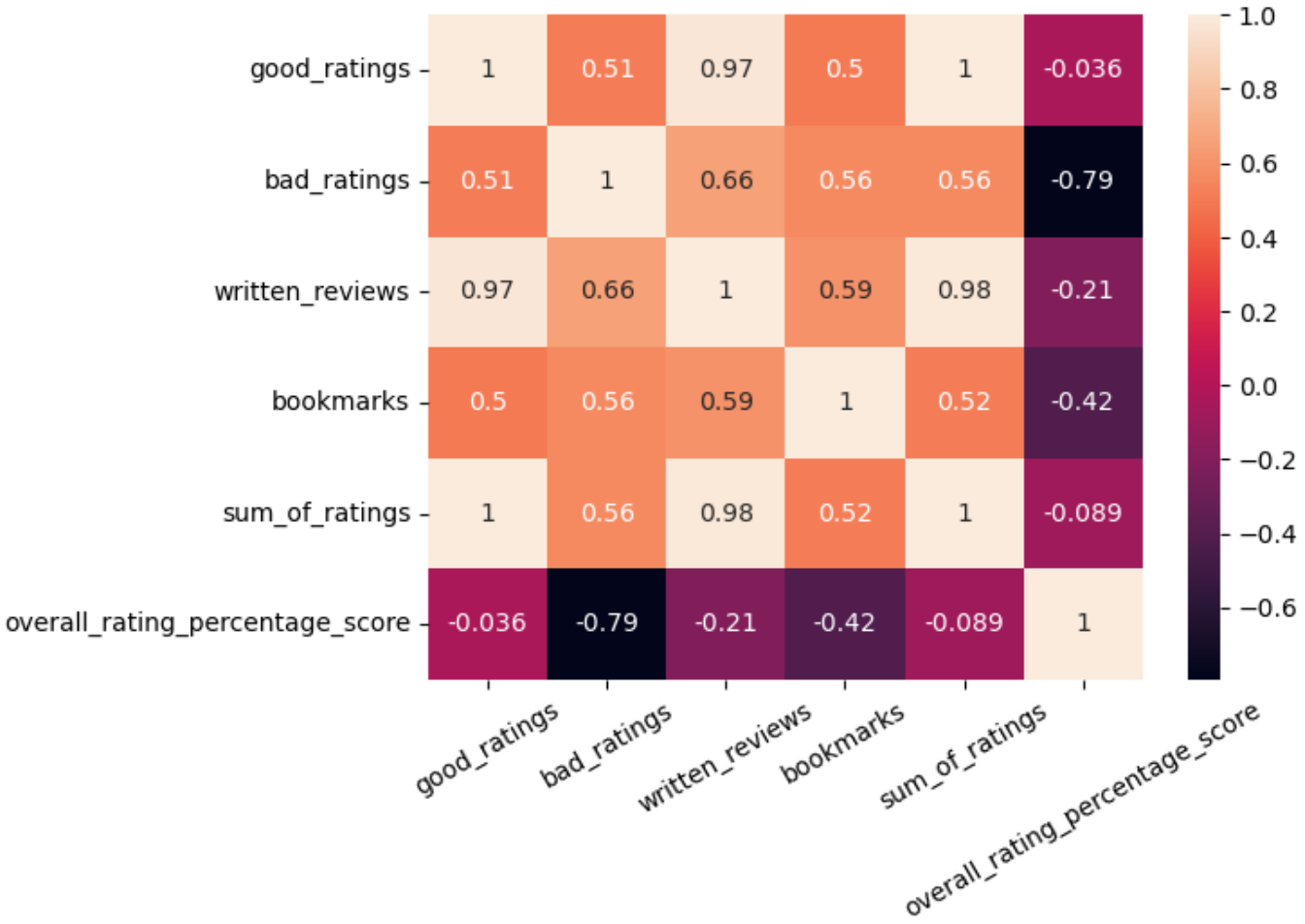
- Kowloon: The more you pay, the lower the restaurant scored.
- HK Island: The more you pay, the higher the rating.
- New Territories: The more you pay, the higher the rating.
  - Data is limited up to an average price of $300.

Exploratory Data Analysis: Price Per Head vs Engagement Stats



Number of ratings is similar across all price ranges, but restaurants in the $401-800 range have the most bookmarks.

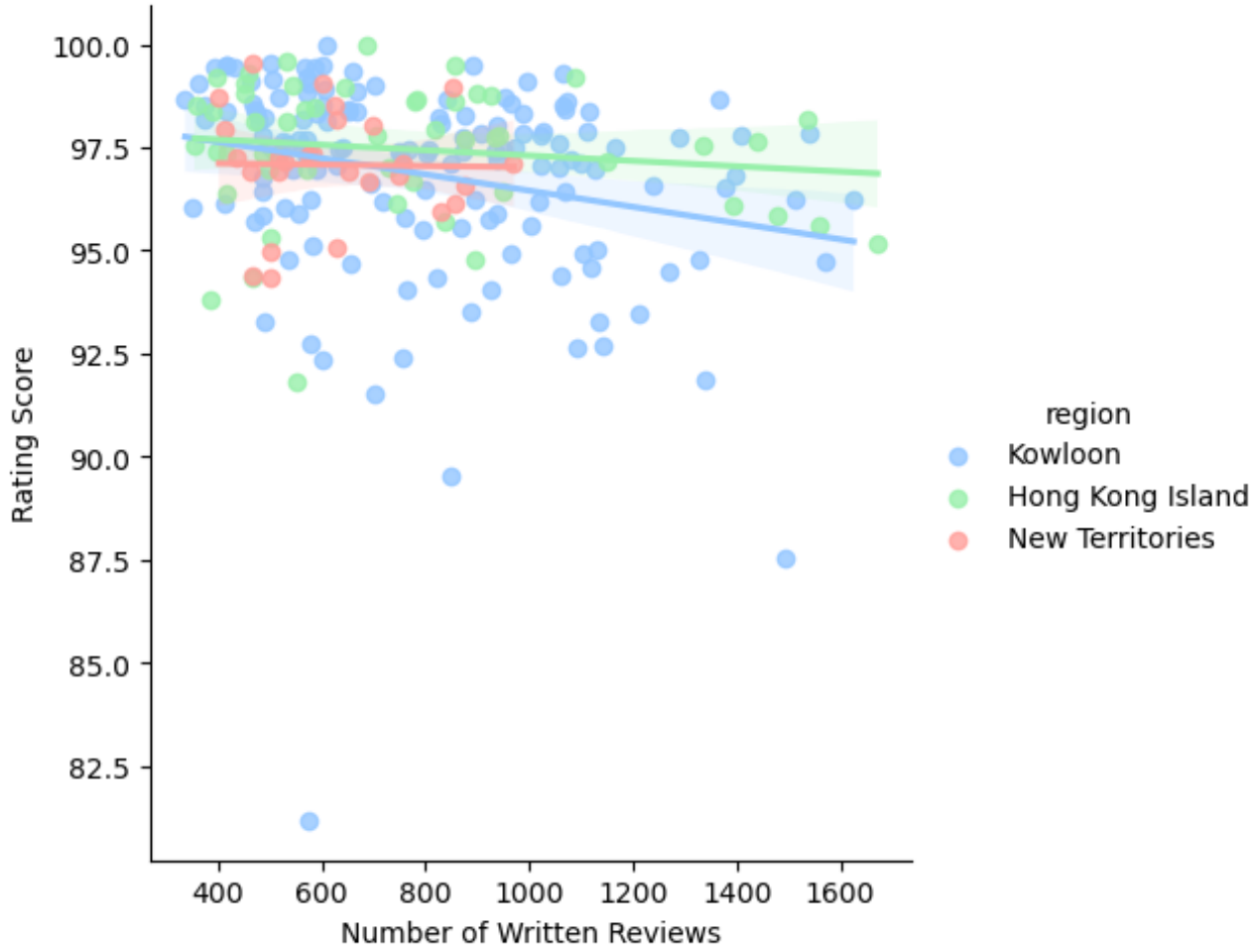
Exploratory Data Analysis: Correlation Between Rating and Engagement



- This heat map shows that there is a negative correlation between overall rating and all other types of engagements.

Exploratory Data Analysis: Number of Reviews vs Rating Score

- It is surprising to see that there is a negative correlation between number of reviews and rating score.
  - Despite having more written reviews, the rating score decreases.



Insight Analysis

- Customers:
  - The most popular cuisine amongst OpenRicers are Western, Japanese, and International cuisine, with the most popular dishes being Dessert, Hotpot and Seafood.
  - Most of the top restaurants are situated in Kowloon (TST and MK).
  - If you want to spend $100-200, you'd find more good options in Kowloon.
  - If you want to spend below $100, you'd find good options in New Territories.
  - You'd want to go to HK Island for more expensive restaurants ,as the more you pay, the higher the rating.
- Businesses:
  - Restaurants in the $401-800 range have the most bookmarks.
    - You could do a lot of promotions and offers on special days (e.g. Birthday, Weekends, Christmas) to entice these customers to come.
  - More written reviews, total number of ratings, and bookmarks may actually negatively impact your rating score.
    - Customers may be more likely to write a complaint than a compliment when writing a review.

Challenges and Limitations

- Data is limited to number of web pages available to scrape.
  - OpenRice only goes up to page 17 (250 restaurants), whereas the website says they have 1.2m restaurants.
- Restaurants may not be categorized correctly on OpenRice(e.g. "Western" appearing on both "Cuisine" and "Dish").
- Null values for some data.
- Slow speed of scraping.

Conclusion

- In this project, I looked at popular food categories, areas, and compared price to regions and ratings.
- For consumers, I hope this analysis will help you make a better decision of where to eat in the future.
- And for businesses, I hope this data helps you find opportunities for growth.