Dining and Dollars

City Income Influences on Yelp Ratings and Restaurant Categories

Group 2:

Ambika Burramukku, Yun-Hsuan (Betty) Chien, Lainah Mangwiza, Duo-Jia Huang, Ariella Breyer

Agenda

1. Business Question and Hypotheses
2. Data Sources and Data Processing
   - API Application
3. Analysis
4. Insights
5. Key Challenges
6. Suggestions/ Further Questions

Business Question:

How does the median income of a city associate with higher Yelp ratings for restaurants, and how does this association differ between categories?

1

Hypotheses

Hypotheses

- **Higher Median Income, Higher Ratings:**
  - Restaurants in high-income regions are more likely to have higher average Yelp ratings.

- **City-Level Variations Across Categories:**
  - The relationship between income and Yelp ratings differs by category and across cities.

2

Data Sources

&

Data Preparation

Data Sources

1

Primary Data Source:

Yelp

(business: business_id, city, stars, category)

- **Transforming from JSON to CSV**

2

Secondary Data Source:

Census Data

(American Community Survey: median income, city, year)

- **Using API Integration**

Data Cleaning & Transformation

- **Data Preparation**
  - Utilized MySQL Workbench, Google Colab, and Excel
  - Columns from Yelp data: business_id, name, city, categories and stars
  - Split the categories columns, made them binary and filtered for restaurants
  - Removed null values, and dropped unnecessary columns

- **Joining Tables:**
  - Merged the Census data with the cleaned Yelp data, used city as the key

Extracting Through API

United States Census Bureau

Census Developer Site

Many public data sets are currently available via API

Choosing Data Set

American Community Survey (ASC)

GenAI x Python

Utilized GenAI to generate Python code that pulls the information we want

Conversion to CSV

Converted to CSV and merged with other data

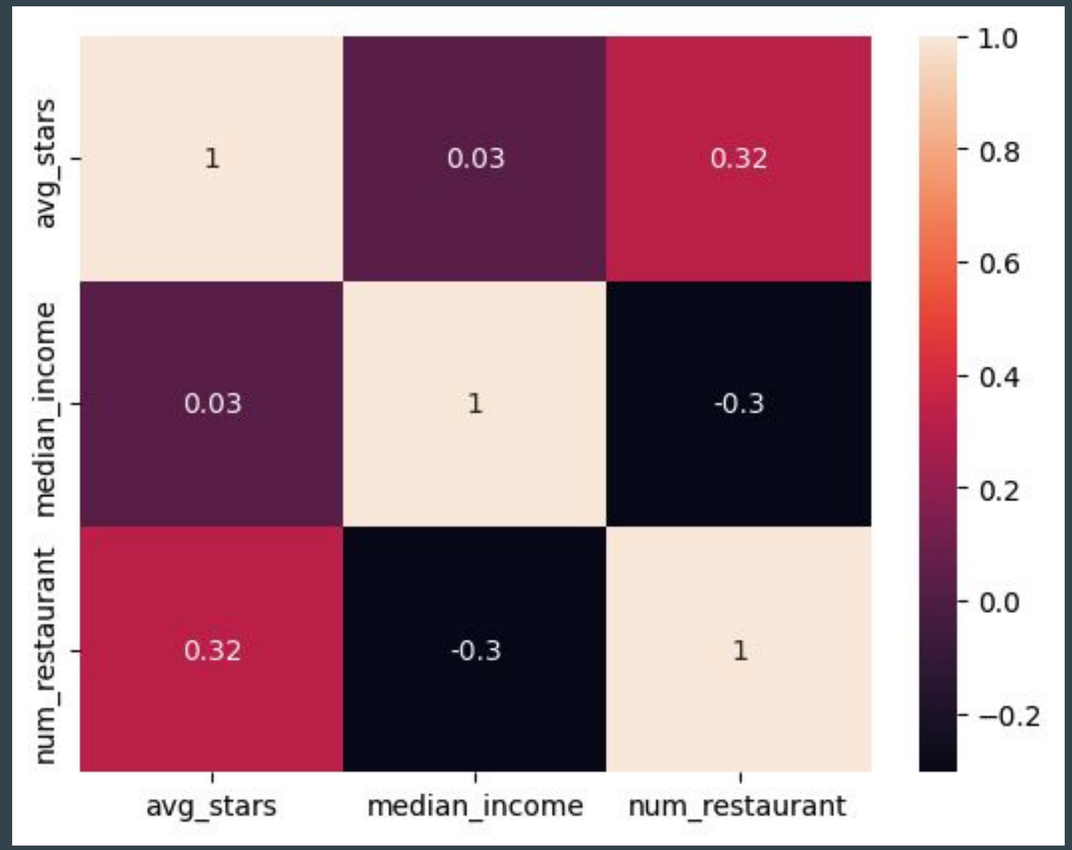
3

Analysis

Analysis

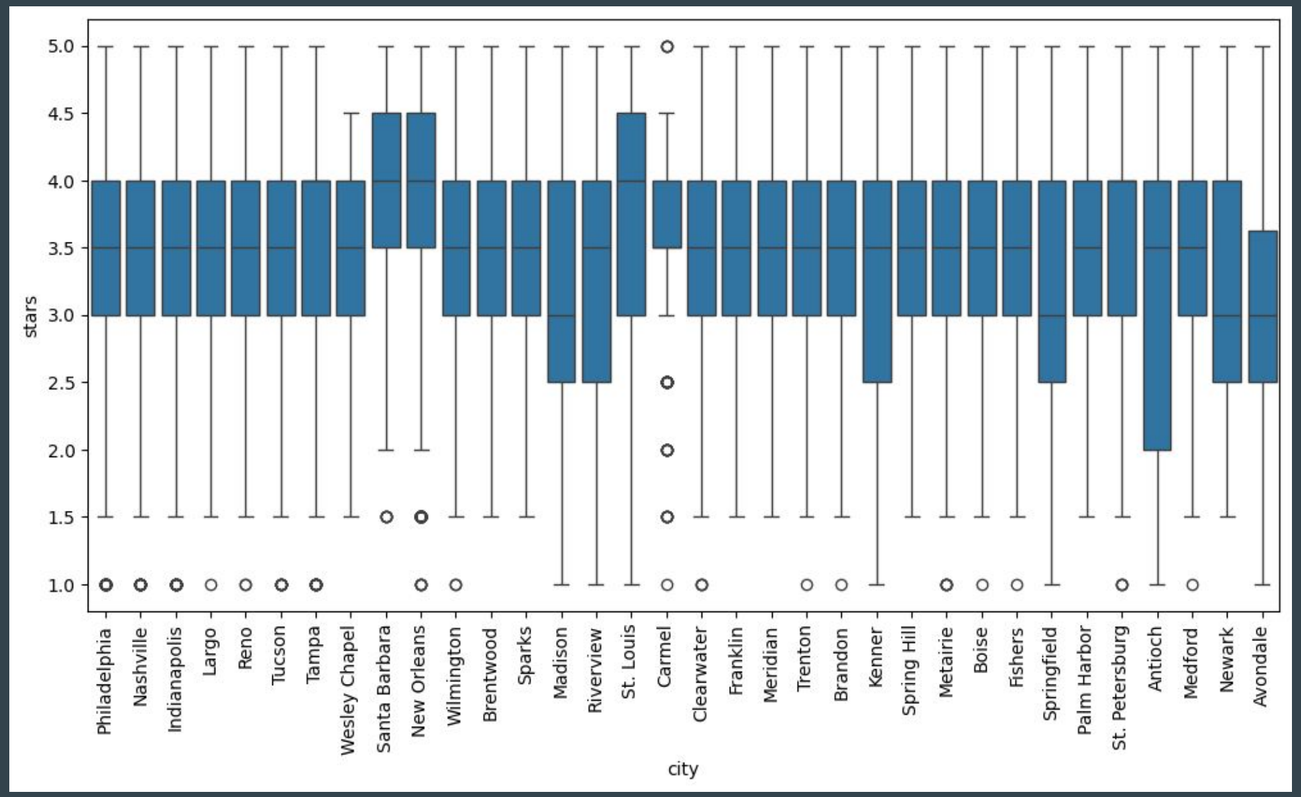
city	avg_stars	median_income	number
Brentwood city	3.421196	132610	184
Carmel city	3.603448	130332	290
Fishers city	3.552239	125159	201
Medford city	3.486111	113253	72
Franklin city	3.426773	108354	437
Santa Barbara city	3.824037	104001	753
Wesley Chapel CDP	3.335484	102188	155
Antioch city	3.196078	100178	102
Riverview CDP	3.356410	98470	195
Meridian city	3.435275	89683	309
Moreno Valley city	3.575531	86909	1271
Palm Harbor CDP	3.660099	85256	203
Sparks city	3.408133	81512	332
Boise City city	3.588836	81425	833
Brandon CDP	3.447040	75294	321
Metairie CDP	3.379110	74528	517
Madison city	3.063158	73647	95

Top Cities by median income



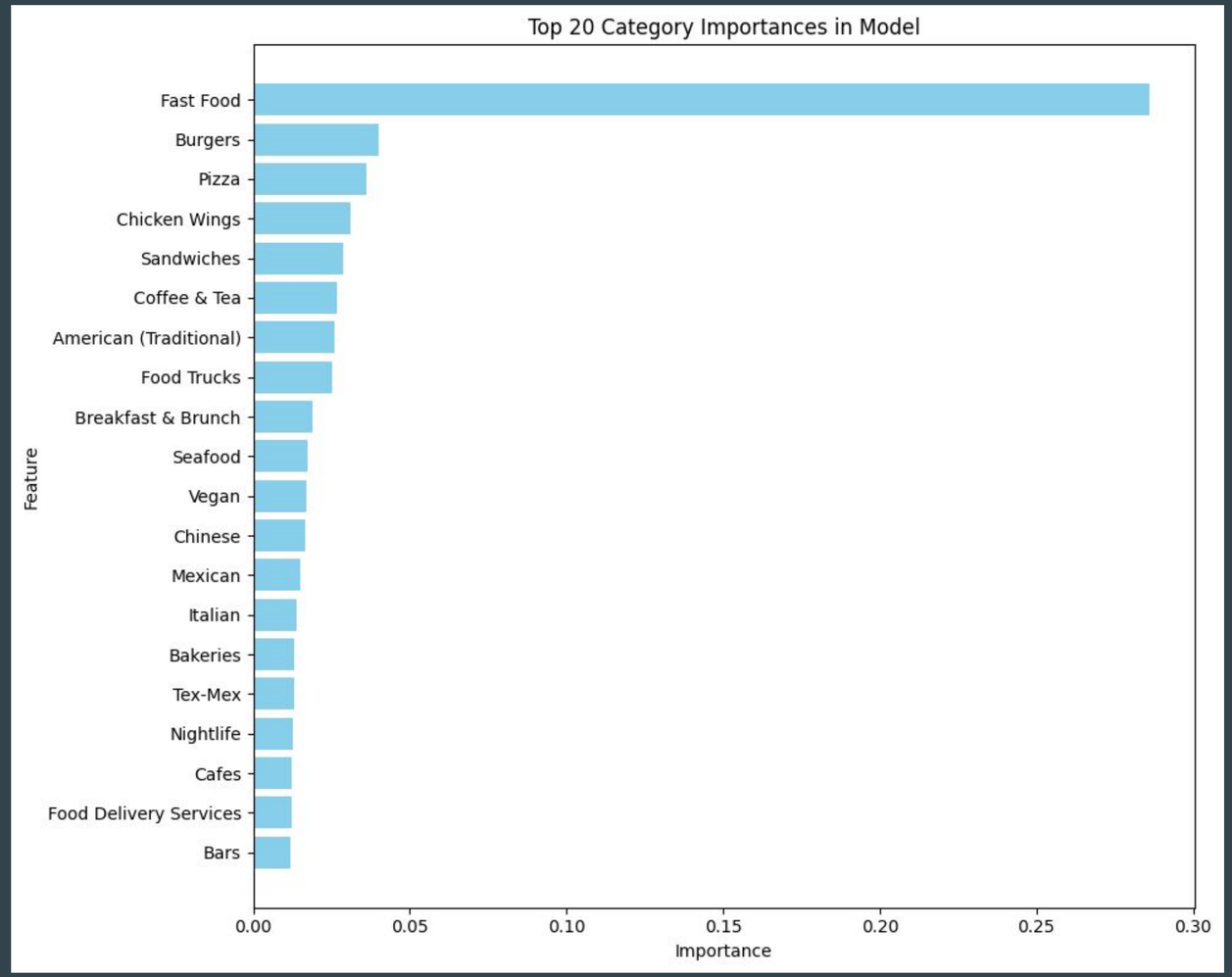
- Used the merged data
- Generated a Heat Map
- Ran a regression (coefficient: 1.44410763e-07)

Categorical Analysis



- Used ANOVA testing to find the difference in means
- p-value is < 0.05, indicates a strong association, we reject Null
- Santa Barbara, New Orleans, and St. Louis had relatively high median ratings while Madison and Springfield had relatively lower medians
- Important to note population differences

Categorical Analysis



- Random Forest Regression to find variance on the restaurant categories and ratings
- R^2 value of 0.259, possibly due to the nature of social/consumer data
- Fast Food had the highest category importance that affects ratings at 0.286 while Breweries and Music Venues had the lowest
- Important to note that this doesn't explain positive or negative correlation

Snippets of Code

```python
# Run ANOVA to check if there is a significant difference in stars by city
# Step 1: Create the model
model = ols('stars ~ C(city)', data=filtered_yelp_data_subset).fit()

# Step 2: Perform ANOVA
anova_results = sm.stats.anova_lm(model, typ=2)

# Step 3: Display results
print(anova_results)
```

```
                sum_sq       df         F        PR(>F)
C(city)     345.099966     32.0  16.28123  1.064442e-88
Residual  16858.915396  25452.0       NaN           NaN
```

ANOVA

```python
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize the random forest regressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)

# Fit the model
rf_model.fit(X_train, y_train)

# Predict and evaluate the model
y_pred = rf_model.predict(X_test)
print(f'Mean Squared Error: {mean_squared_error(y_test, y_pred)}')

# Calculate R² (coefficient of determination)
r_squared = rf_model.score(X_test, y_test)
print(f'R²: {r_squared}')

# Get feature importance (how each feature contributes to the model)
feature_importance = rf_model.feature_importances_
for feature, importance in zip(X.columns, feature_importance):
    print(f'{feature}: {importance}')
```

Random Forest Regression

4

Insights

Insights

Correlation is not super strong between median income and ratings, disproving our hypothesis

Santa Barbara and New Orleans had relatively high median ratings, Madison had lower median ratings. When comparing median income, our hypothesis was disproved but some cities have positive correlation. EX: Santa Barbara (104,001) to Madison (73,647)

Fast Food had the highest importance with ratings amongst categories, when taking into account median_income

5

Key Challenges

Key Challenges

- **Data Source**
  - Find useful secondary data to support our hypotheses.
- **APIs Extraction**
  - Find the right column names to rename
  - Extract data for informative analysis
- **Data Cleaning**
  - Split attribute and category columns using one-hot encoding.
  - Find efficient ways to merge Yelp and Census data tables. (e.g. filter unnecessary data to get smaller dataset for analysis.)
- **Data Analysis**
  - Attempt to find correlations based on our hypotheses.
  - Filter data to find both statistically and practically significant

6

Suggestions/ Further Questions

Suggestions/Further Questions

Recommendations on how to proceed:

We suggest looking further into if there is a correlation between categories specific to cities to see if there is a strong relationship. Perform more feature engineering to categorize cities in income ranges or population and then run analysis for ratings. Utilize the reviews for restaurants to see if there are certain commonalities between ratings and categories.

Questions we recommend looking into:

- What is the relationship between price ranges and Yelp ratings?
- Do income ranges have an effect on categories or stars?
- What are some similarities in reviews between higher and lower rated cities?

Thank you!

Any Questions?