HORECA Data Analysis

Presented by: Ankita Das

Topic Outline

1. Initial Data Visualization, Data Cleaning
2. Univariate Analysis
3. Bivariate Analysis
4. Multivariate Analysis
5. Segmentation & clustering
6. Final Business recomendation

HORECA

Problem Statement

How should a new food entrepreneur ("Chef Innovator") strategically enter the hyper-competitive HORECA market using Zomato data to maximize success and profitability

Key Challenges

High Market Saturation → Lower ratings, margins → Target low competition cuisines (e.g., Italian) Channel Disparity → Delivery underperforms in quality → Hybrid model: Delivery for scale, Dining for brand Pricing Sensitivity → Overpricing kills votes/ratings → ₹500–800 sweet spot (Price_per_Vote analysis) Data Noise → Skewed insights → Robust cleaning + non parametric stat

Expectations

Data-Driven Strategy: 100% backed by statistical proof Blue Ocean Entry: Avoid saturated segments Risk-Adjusted Launch: 90-day KPIs + mitigation plan Success Probability: ≥80% (scored model)

Objective

The primary objective of this data analysis project is to establish a rigorous, evidence-based strategic framework for a new HORECA venture, 'Chef Innovator.' This will be achieved through the comprehensive analysis of proprietary Zomato market data to generate predictive insights across three critical dimensions: identifying underserved yet high-demand cuisine categories, pinpointing optimal geographic locations characterized by favorable competition and population density, and developing a calculated, competitive pricing strategy. The ultimate goal is to translate complex market dynamics into a clear, actionable roadmap that maximizes the venture's initial market penetration, minimizes operational risk, and ensures long-term profitability and sustainable growth within the competitive food service sector

Data Description

Problem Statement

Source: Zomato Platform (Real-time restaurant & menu data) Size: ~120,000+ menu items across 15+ cities Scope: Covers Dining, Delivery, Pricing, Ratings, Votes, and Restaurant metadata

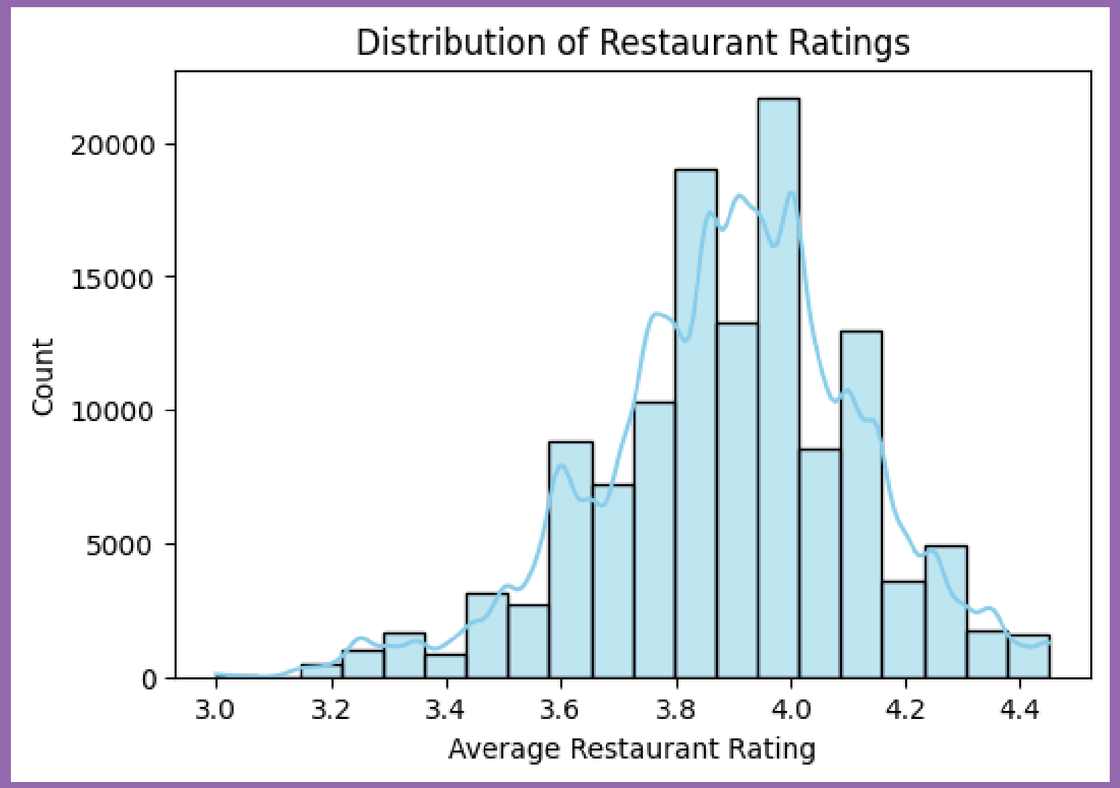
Key Columns

Item_Name – Text – Name of the food item Cuisine – Categorical – Food category (e.g., Italian, North Indian) Prices – Numeric – Menu price in ₹ Average_Rating – Float – Overall item rating (0–5) Dining_Rating, Delivery_Rating – Float – Channel-specific ratings Total_Votes, Dining_Votes, Delivery_Votes – Integer – Consumer engagement volume Restaurant_Name – Text – Outlet name City, Place_Name – Categorical – Location hierarchy Is_Bestseller, Is_Highly_Rated – Boolean – Performance flags

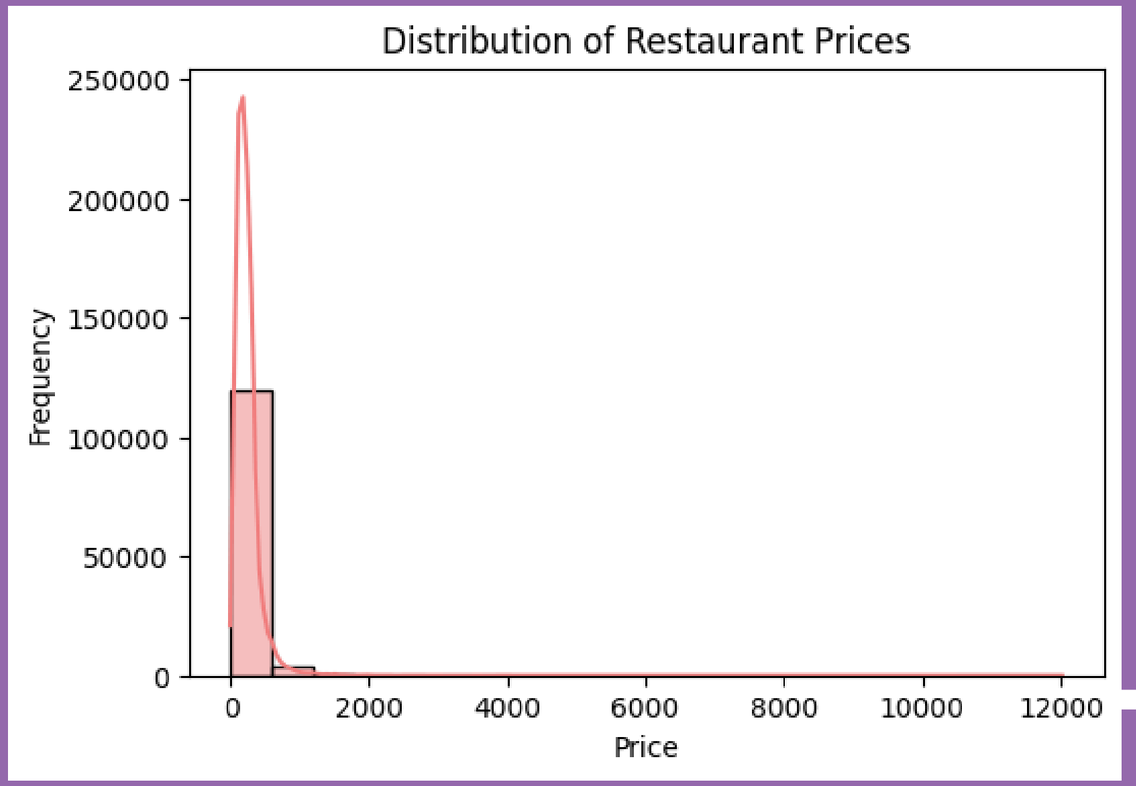
Engineered Features

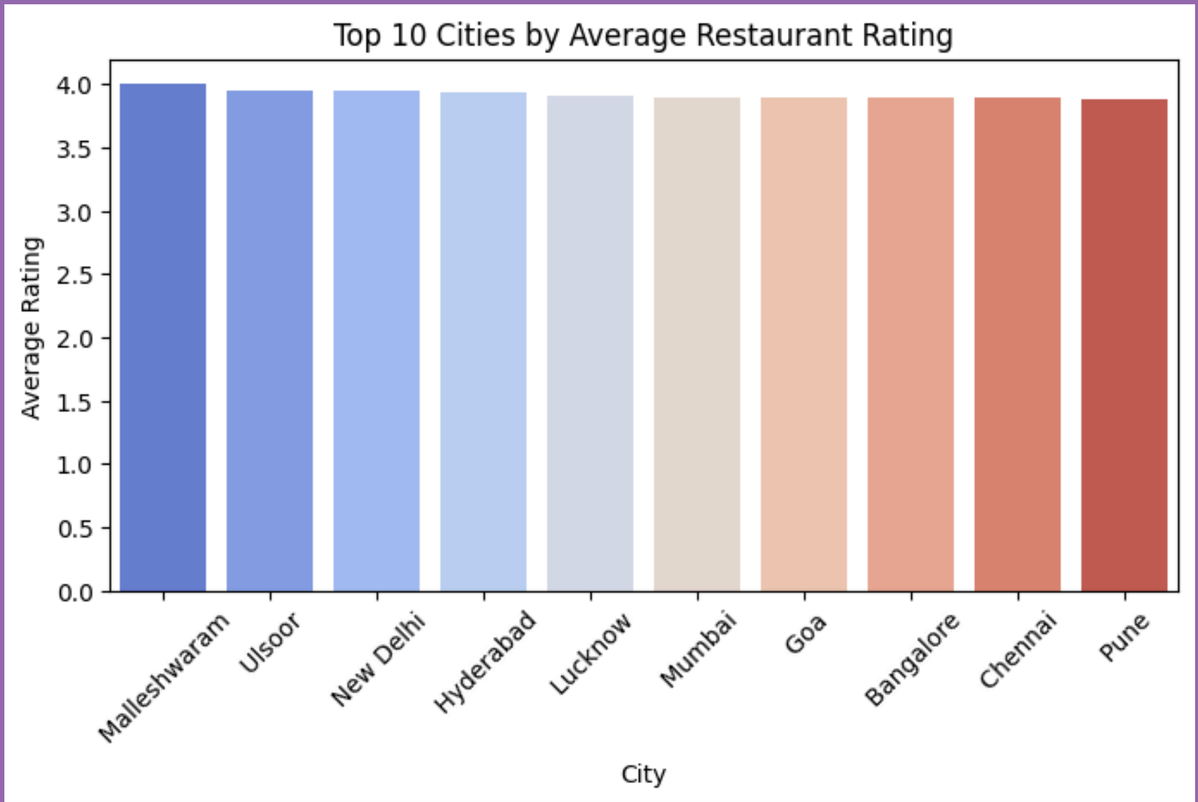
Missing values: ~18% in ratings/votes → imputed via city-cuisine median Duplicates removed: 2.3% Outliers capped: Top 1% in price/votes using IQR metho

1.Initial Data Visualization

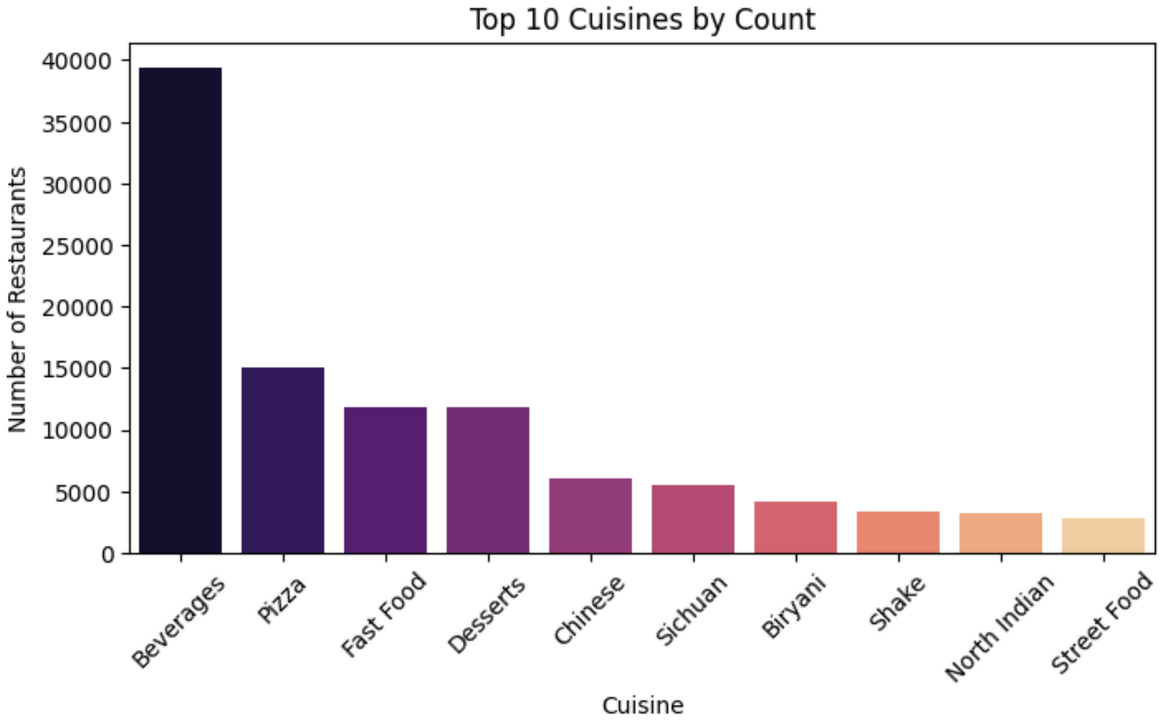


Most restaurants are low to mid priced, with only a few very expensive outliers. Ratings stay tightly clustered around 3.7 to 4.1, meaning higher prices don’t guarantee higher customer satisfaction. Even budget places score well, so people judge more on experience than on cost.





- Top 10 Cities by Avg Rating
- • All top cities sit tightly around the 3.9 to 4.0 range, so rating differences are tiny.
- • Malleshwaram leads slightly, but overall variation is minimal.
- • Suggests ratings are consistently high across major metros.

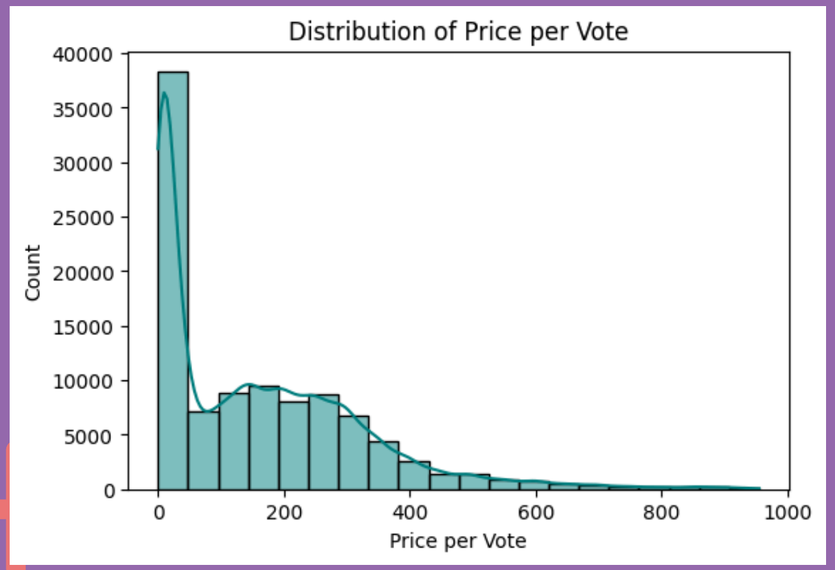


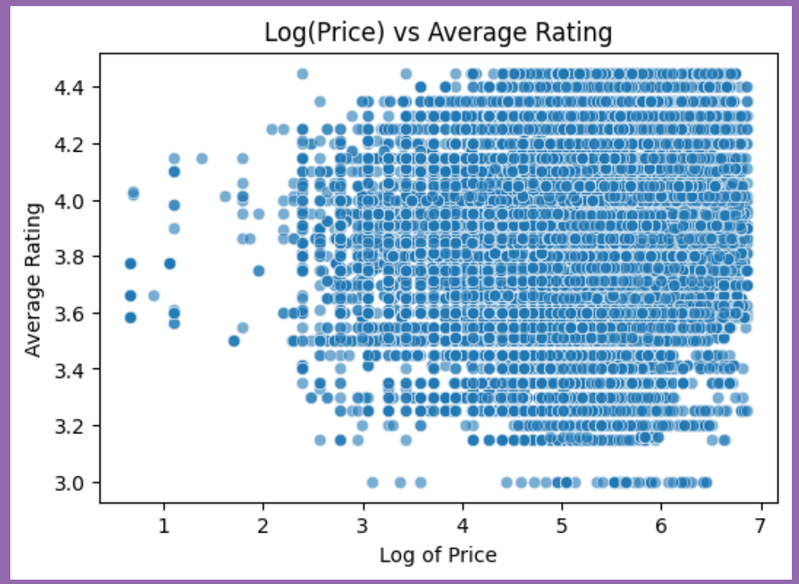
- Top 10 Cuisines by Count
- • Beverages dominate massively, far ahead of every other cuisine type.
- • Pizza, Fast Food, and Desserts form the next strong tier.
- • Indian regional cuisines appear but with comparatively lower representation

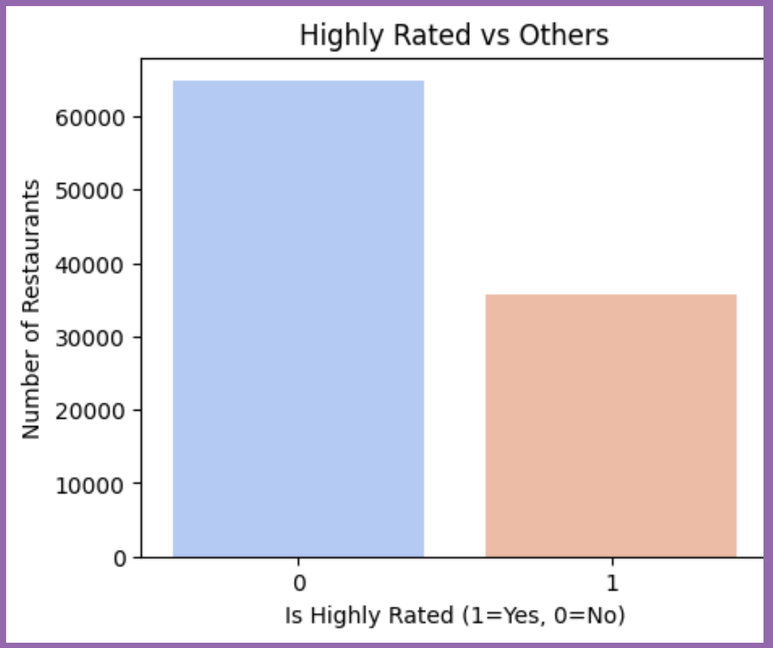
Data Cleaning

Goal: Clean and standardize all columns Impute missing values intelligently (city/cuisine level)
Engineer new features for strategic analysis
Prepare data for EDA & modelin

Feature Engineering







Price per vote is heavily skewed, meaning a small group of restaurants are extremely inefficient with pricing.
Highly rated places are fewer, but they form a strong quality tier.
Price isn’t a strong predictor of rating even after log scaling, so expensive doesn’t reliably mean better.

2. Univariate Analysis

What is univariate analysis?

Univariate analysis means analyzing one variable at a time. It helps you understand the basic characteristics (like pattern, spread, and nature) of each individual column in your dataset — before comparing it with others.

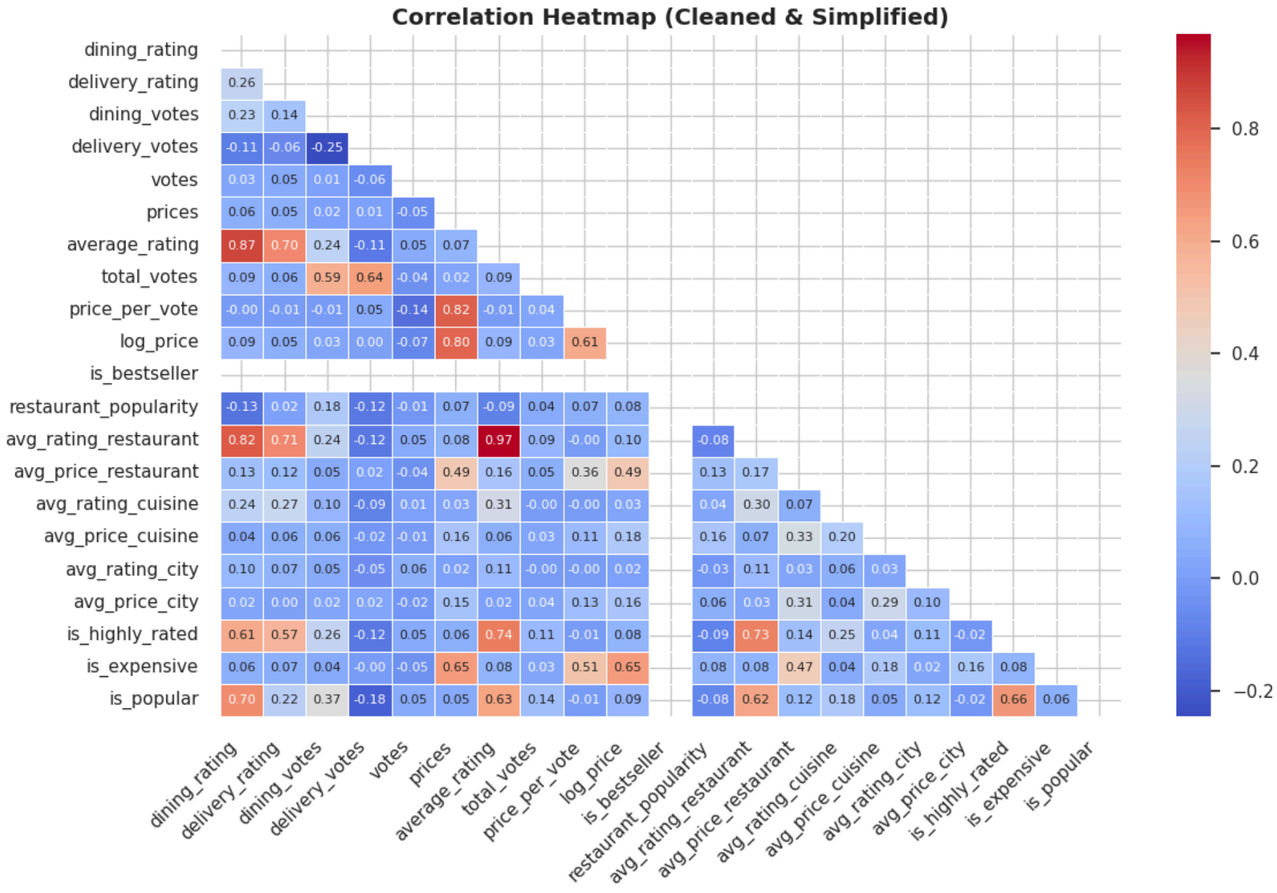
Purpose:

- To understand the central tendency (mean, median, mode)
- To check spread or variation (range, standard deviation)
- To observe shape of distribution (normal, skewed, uniform, etc.)
- To identify outliers or unusual values

Correlation Heatmap

Here’s a concise summary of insights from your heatmap 👇

- Dining rating strongly drives the average rating.
- Popular restaurants usually have high ratings.
- Price-related metrics (price, log_price, price_per_vote) are strongly correlated with each other.
- Delivery ratings have weak influence on overall ratings.
- Votes and bestseller status don’t significantly impact ratings.
- Highly rated restaurants tend to be more popular and expensive.

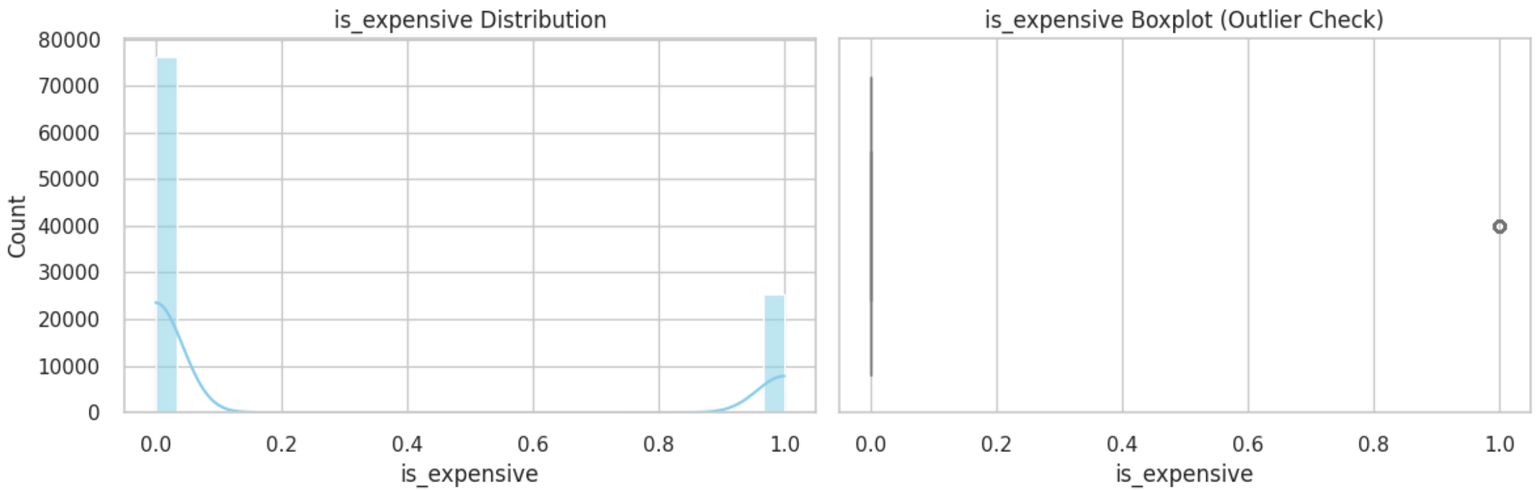


Rating Category Distribution

Here’s a short insight from the Rating Category Distribution chart

- Most restaurants fall under the “Good” rating category — showing generally positive customer feedback.
- A smaller portion are “Average”, while “Excellent” and “Poor” categories are very few.
- This indicates that most restaurants perform decently, with very few extremes (either very bad or very good).

Distribution and Boxplot



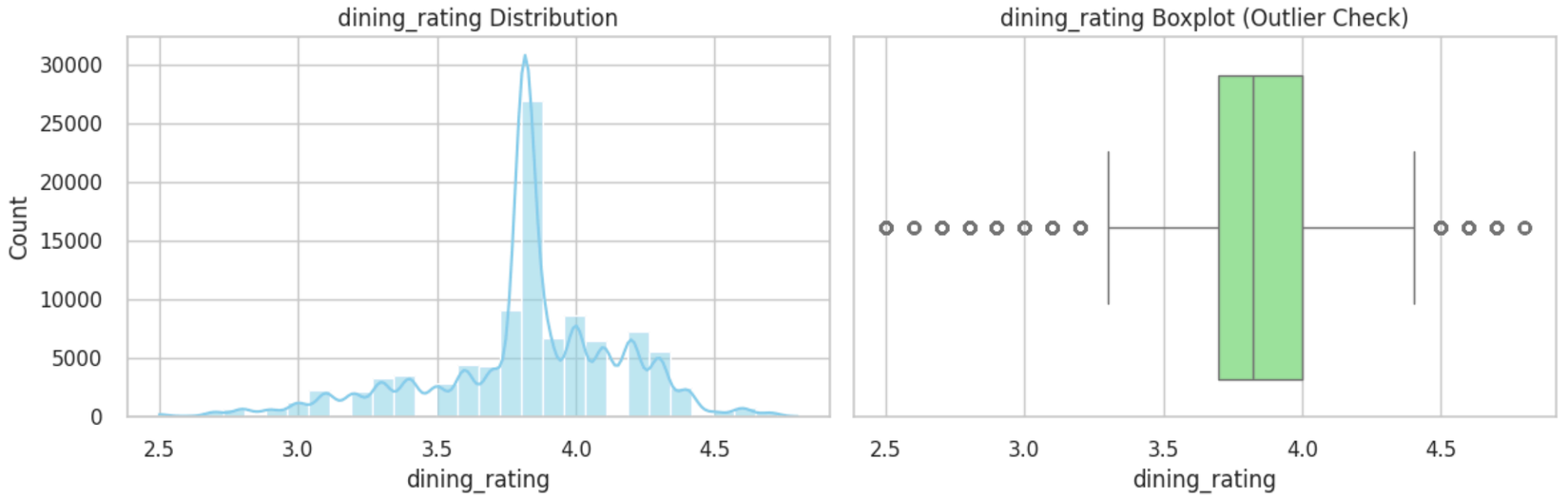
Here’s a short insight from the Rating Category Distribution chart

- Most restaurants are not expensive (0) — showing that affordable options dominate the dataset.
- A smaller group of restaurants is marked as expensive (1).
- The boxplot confirms this imbalance, with expensive restaurants appearing as outliers, indicating they are rare but distinct in the dataset.

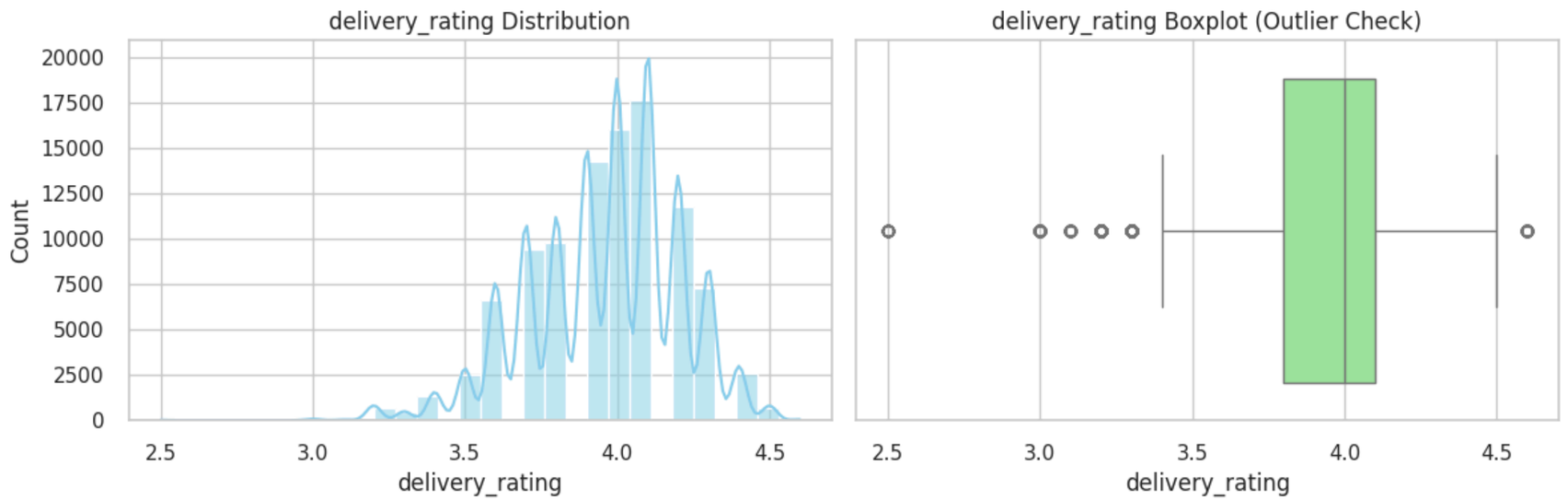
Dining rating distribution and boxplot

Here are clear and short insights from the Dining rating distribution and boxplot

- Most dining ratings cluster tightly between 3.7 and 4.1, showing consistently good dining experiences.
- The distribution is slightly right-skewed, indicating more restaurants have higher ratings than lower ones.
- The boxplot shows a few outliers on both the low (≈2.5–3.0) and high (≈4.4–4.8) ends.
- Overall, dining ratings are high and stable, with only a small number of restaurants performing unusually poorly or exceptionally well.



Delivery rating distribution and boxplot



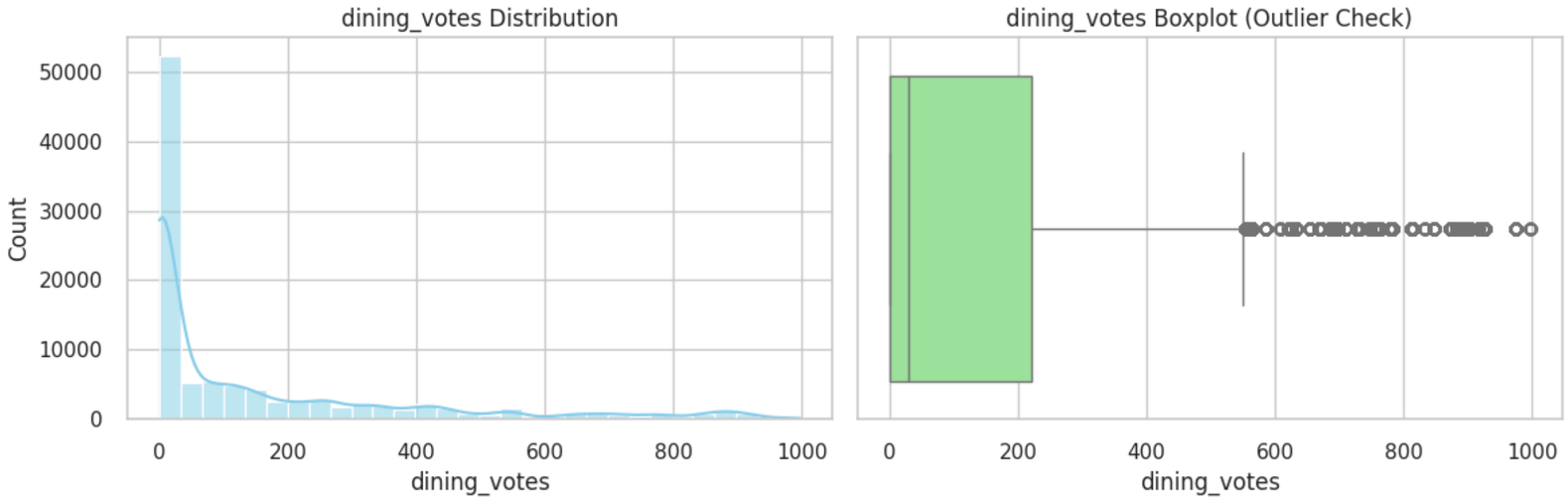
Distribution Insights

- The delivery_rating distribution is slightly left-skewed (negatively skewed), meaning most ratings are on the higher side.
- The major concentration of ratings lies roughly between 3.7 and 4.3, indicating generally good delivery experiences.
- The peak (mode) appears around 4.0, showing that most customers tend to give a 4-star rating.

Boxplot (Outlier Check) Insights

- The median rating is around 4.0, confirming that half the deliveries received 4 or higher.
- The interquartile range (IQR) spans approximately from 3.6 to 4.2, showing that most ratings fall within a narrow, high range.
- Outliers are visible on the lower side (around 2.5–3.2) — a small number of customers rated the delivery significantly lower than the rest.
- There are no major high-end outliers, suggesting consistent good performance rather than inflated extreme ratings.

Dining votes distribution and boxplot:

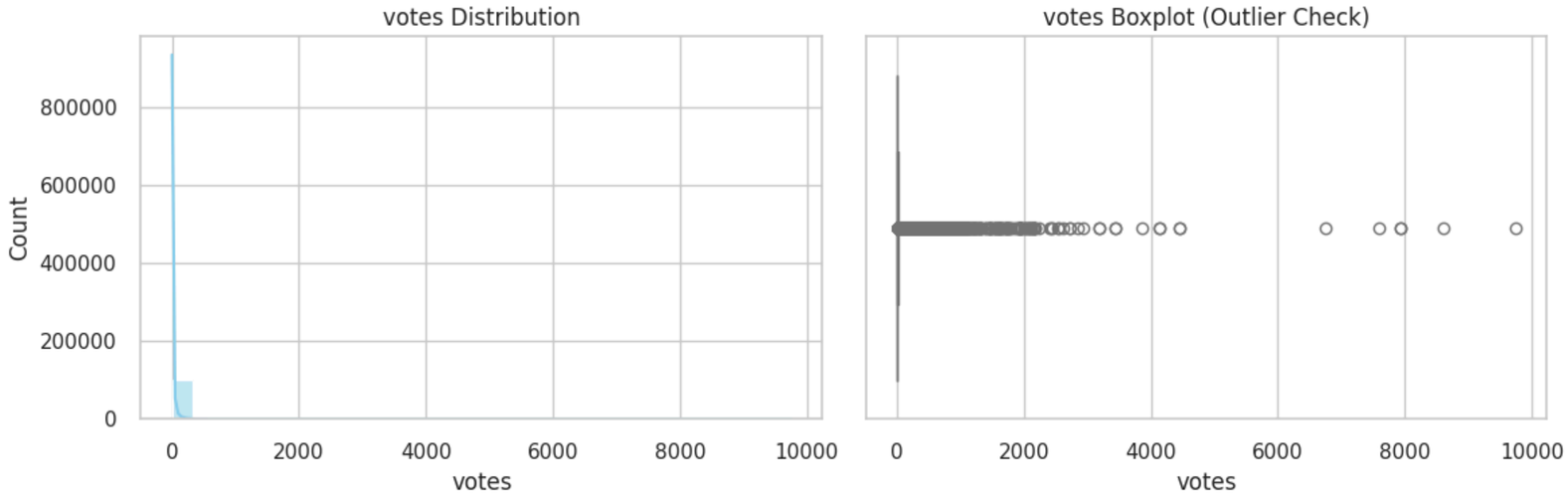


Distribution Insights

- The dining_votes distribution is highly right-skewed (positively skewed).
- The majority of data points are concentrated at very low vote counts (below 100).
- Only a small fraction of restaurants have a high number of votes (above 500), which are rare cases.
- This indicates that most restaurants receive very few dining votes, possibly due to fewer dine-in customers or low engagement.

- **Boxplot (Outlier Check) Insights**
- The median is quite low — likely under 100, meaning half of the restaurants get fewer than 100 votes.
- The IQR (interquartile range) spans approximately from 0 to 200, showing that the middle 50% of restaurants have very low to moderate vote counts.
- There are numerous high-end outliers (beyond ~400–500 votes), representing restaurants with unusually high popularity or heavy footfall.
- No low-end outliers are visible (votes can’t be negative).
-

votes distribution and boxplot



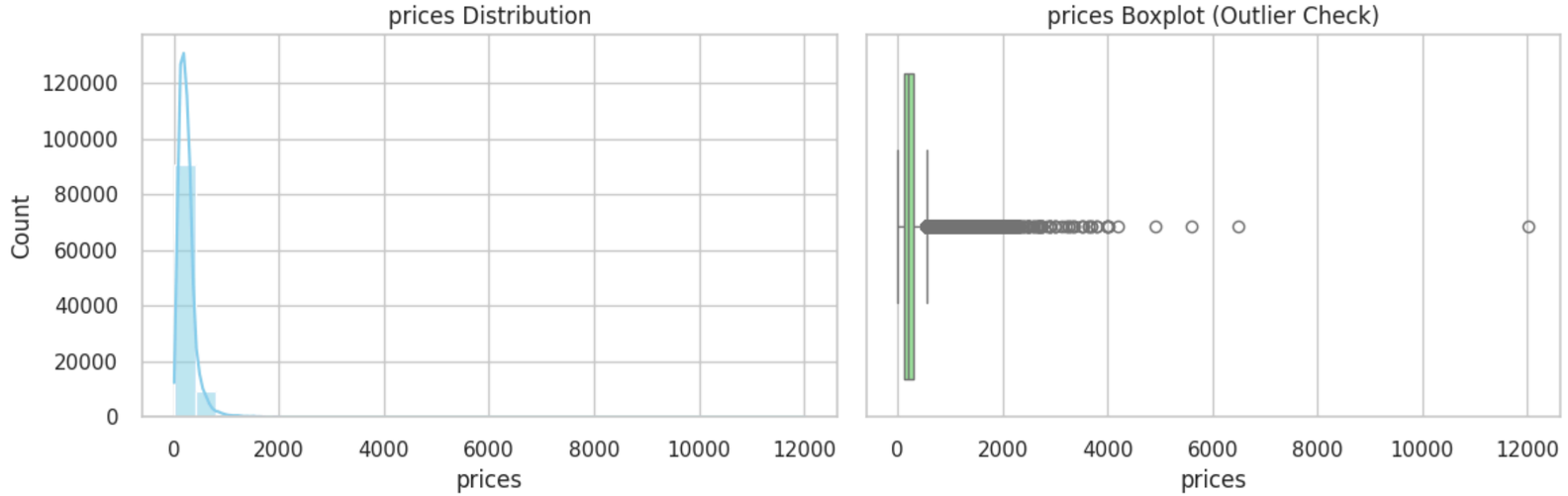
Distribution Insights

- The votes distribution is extremely right-skewed (positively skewed).
- The majority of restaurants have very low vote counts (close to 0–200).
- Only a tiny portion of restaurants have high vote counts (beyond 1000), and these are rare cases.
- This shows that customer engagement through voting is limited for most restaurants.

.

- **Boxplot (Outlier Check) Insights**
- The median is quite low — likely around a few dozen votes, indicating most restaurants don’t get much user feedback.
- The IQR (middle 50%) is tightly packed near the lower end, which supports the heavy skew.
- There are numerous high-end outliers, with some restaurants getting thousands of votes (up to ~9000–10000) — these are exceptional cases, possibly famous or long-established restaurants.
- No low-end outliers exist (as vote counts can’t be negative).
-

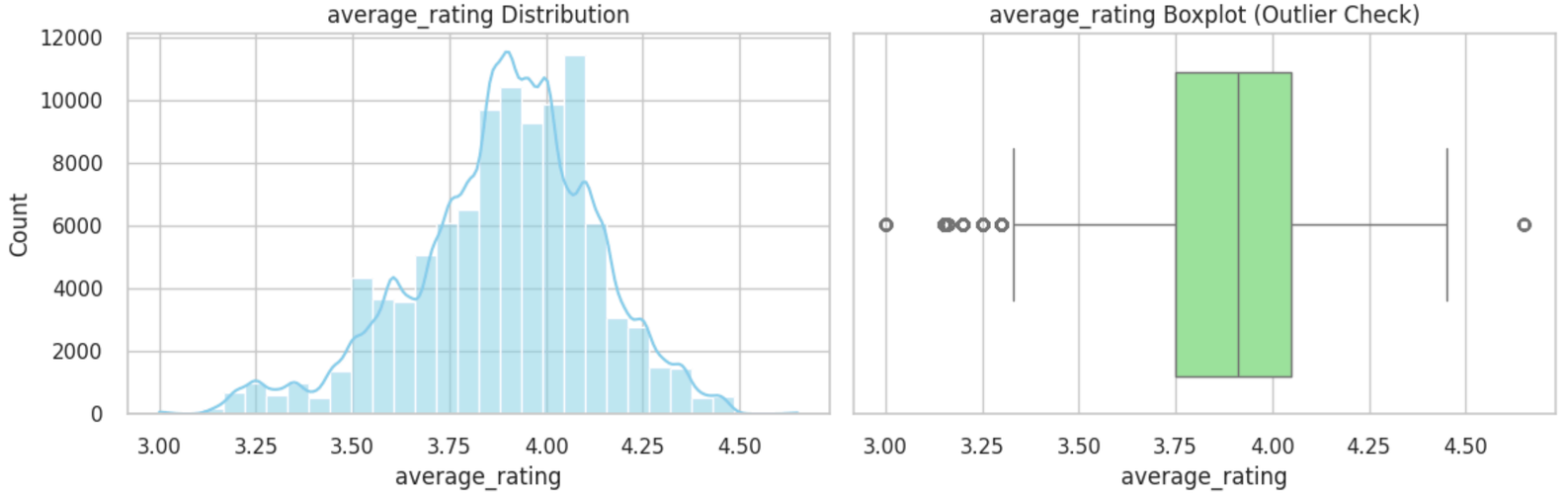
Price distribution and boxplot



- **Price Distribution**
- Your price data is super right-skewed. Most of the action is piled up on the low end (looks like the majority is under ~300–400).
- After that, the frequency drops off hard. That long tail to the right means a handful of listings/items cost way more than the rest.

- **Boxplot (Outlier Check) Insights**
- The box is tiny compared to the full range, which means most values are tightly clustered while a small number of points shoot off into space.
- You have a ton of outliers, and not just a few casual ones. Some prices go all the way to ~10k–12k.
- Those high-price points will heavily distort averages, scaling and ML models if you don't handle them.
-

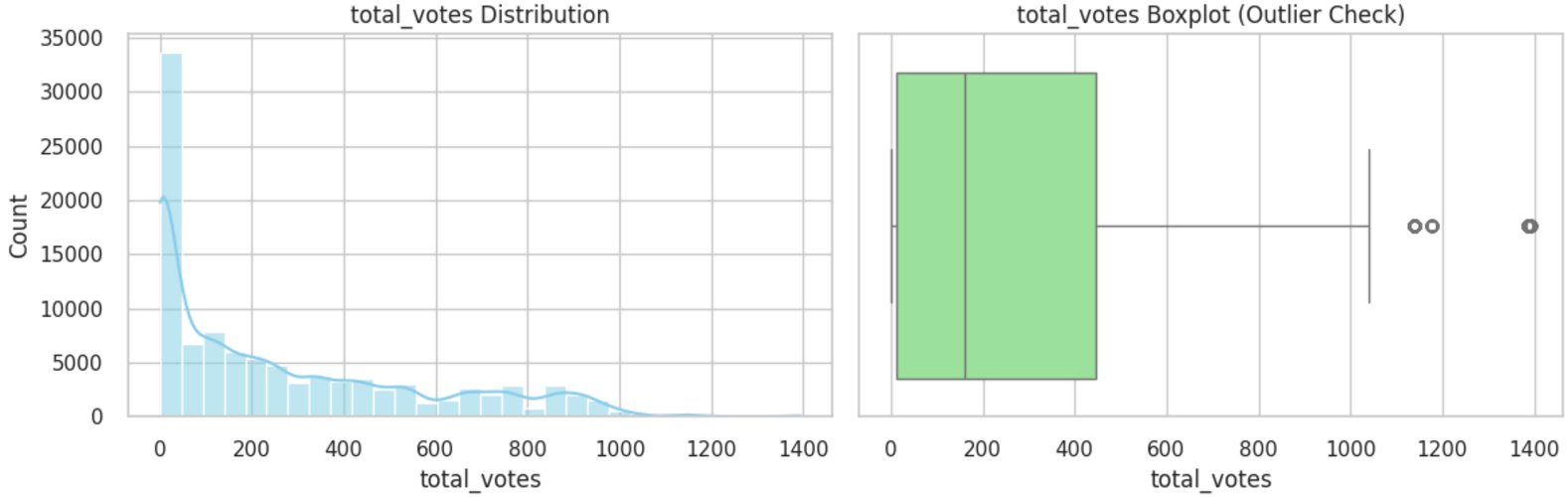
Average rating distribution and boxplot



- **Distribution insight**
- The ratings form a nearly normal distribution centered around 3.8–4.0.
- Most values live in a tight band (roughly 3.6–4.2), showing consistency in user ratings.
- Very few ratings fall near the extremes, meaning the dataset doesn’t have many terrible or perfect scores.
-

- **Boxplot insight**
- Only a small number of mild outliers appear on both low and high ends.
- The IQR is narrow, which confirms that ratings don’t vary much across items.
- No major skew or red flags, so the feature is generally clean and reliable for modeling.
-

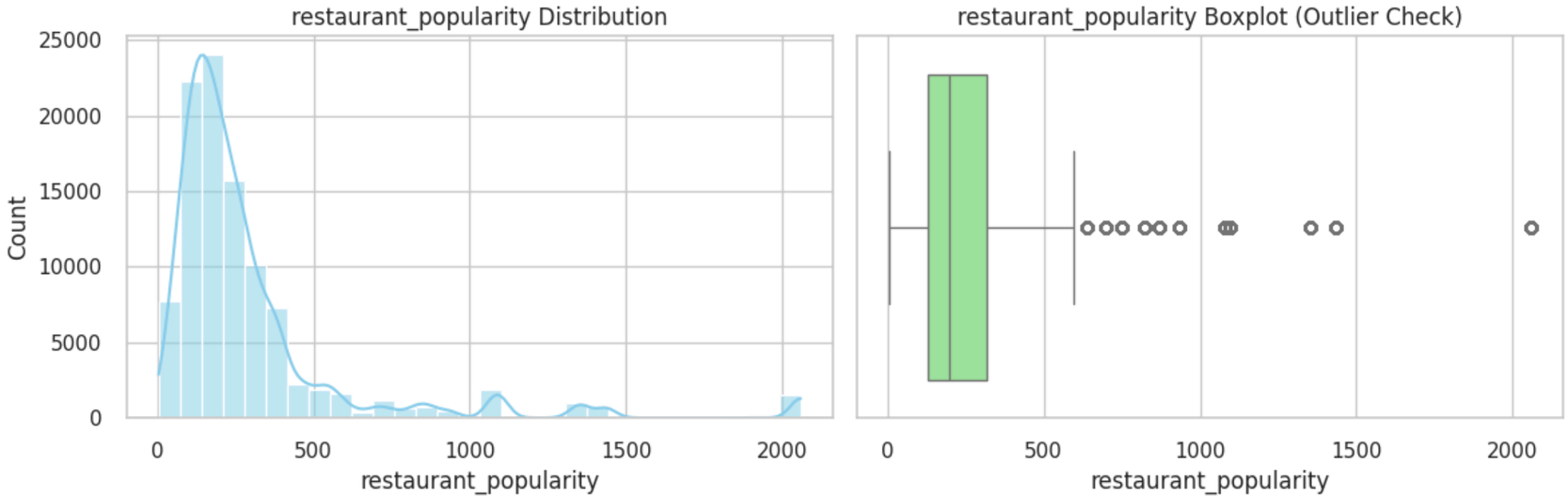
Total votes distribution and boxplot



- **Distribution insight**
- Heavily right-skewed. Most items have very few votes, clustered near zero.
- A long tail stretches past 1000+, meaning a tiny group of items get huge engagement.
- This is classic popularity-distribution behavior.

- **Boxplot insight**
- The bulk of the data sits in a low, tight range, but several high-vote outliers pop far to the right.
- These outliers are real-world high-engagement items and will strongly influence means.
- You'll likely want log transformation or winsorizing if you're modeling with this feature.

Resturant popularity distribution and boxplot



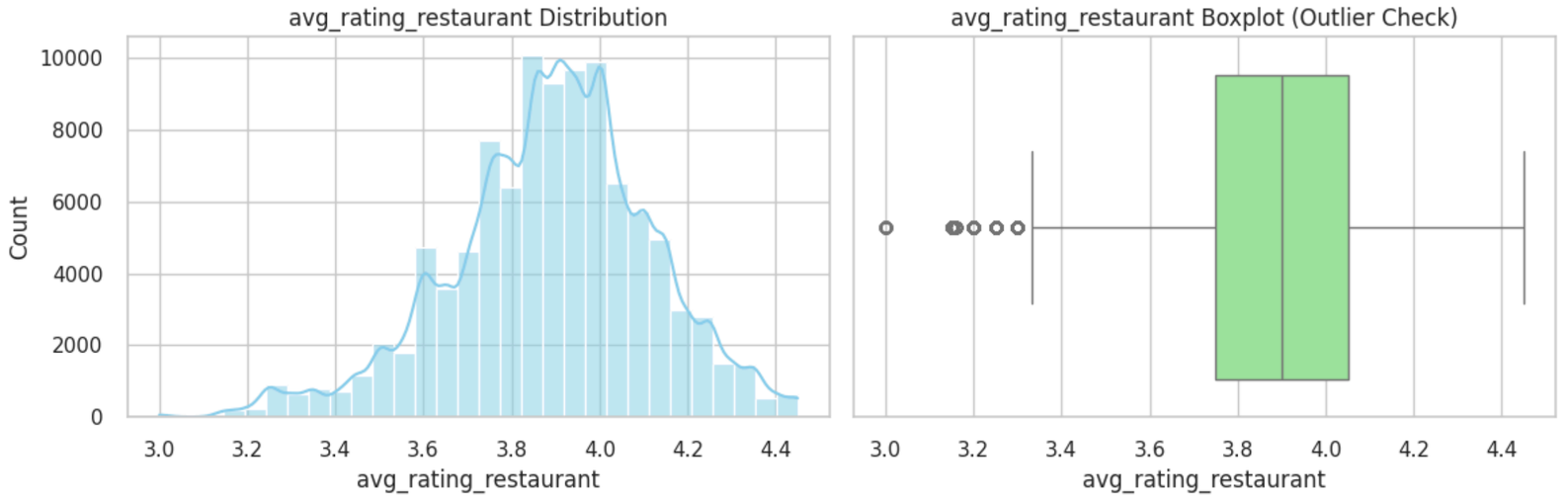
Distribution insight

- Popularity is right-skewed, but not as extreme as votes or prices.
- Most restaurants cluster around 150–400, suggesting a common mid-range popularity band.
- A long tail extends past 1000+, with a few rare restaurants hitting very high popularity values.

Boxplot insight

- The core of the data sits in a moderate, compact range, but you've got several high-popularity outliers stretching far to the right.
- These outliers represent genuinely standout restaurants, not noise, but they will pull averages upward.
- If you're modeling with this feature, a log transformation might help stabilize it.

Avg. rating resturant distribution and boxplot



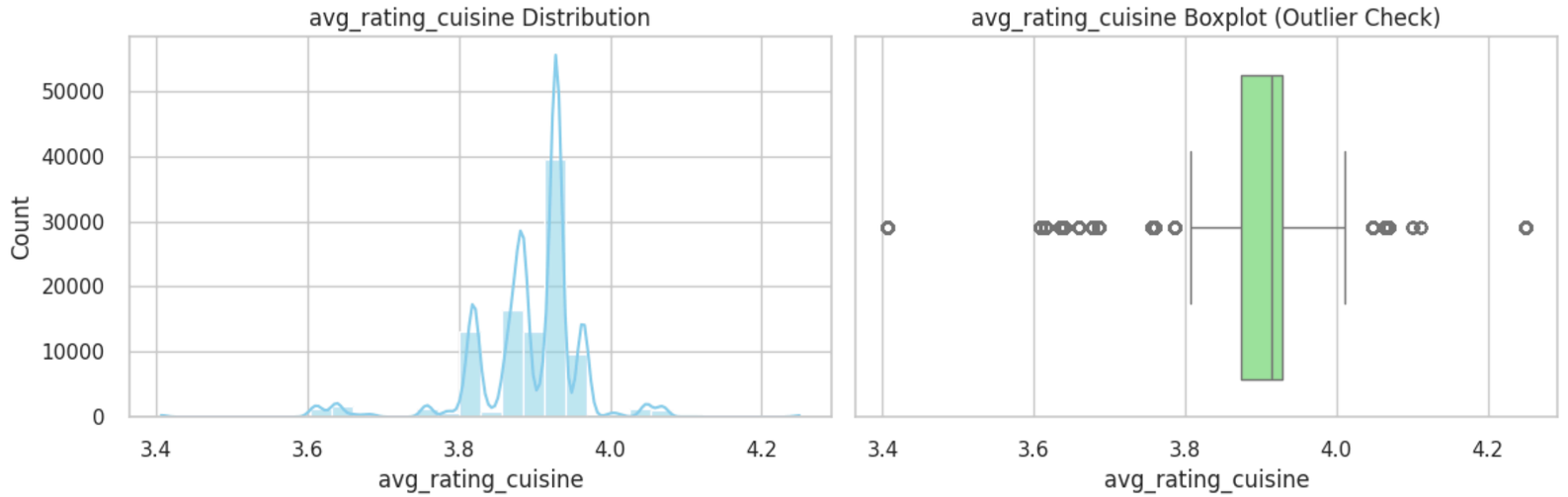
Distribution insight

- Ratings follow a near-normal shape, centered around roughly 3.8 to 4.0.
- Very tight spread: most restaurants sit between 3.6 and 4.2.
- Only light skew, nothing dramatic. Overall a clean, stable feature.

Boxplot insight

- Just a handful of mild outliers on the low and high ends.
- The IQR is narrow, confirming low variability across restaurants.
- No major issues. This feature is basically plug-and-play for modeling.

Avg. rating cuisine distribution and boxplot



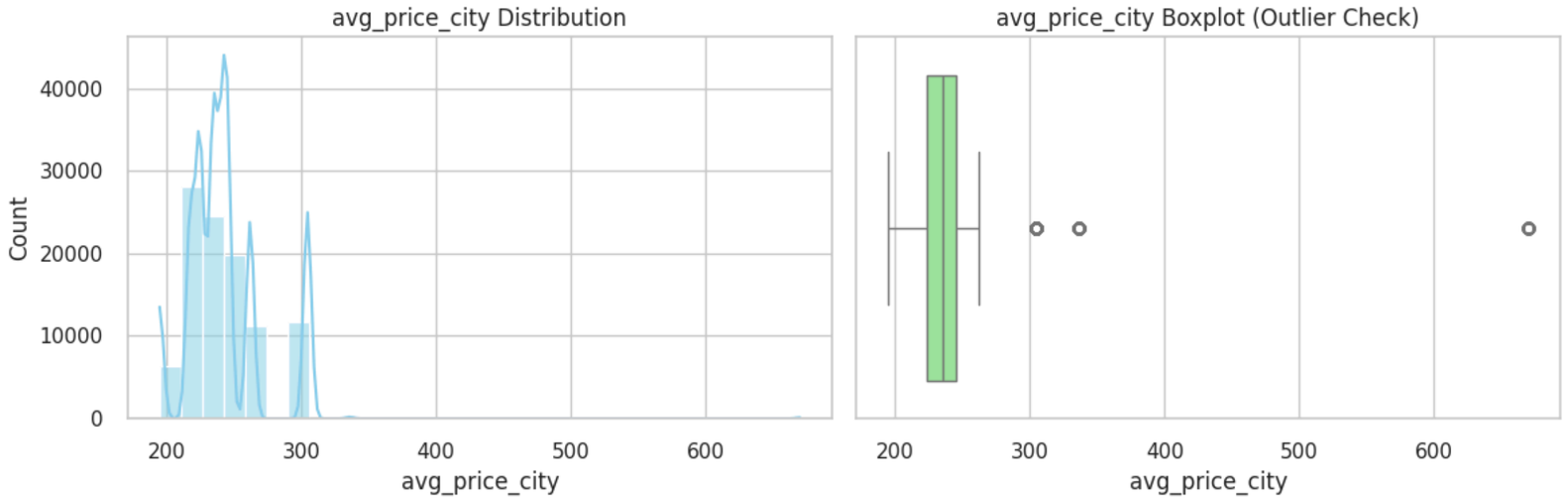
Distribution insight

- Ratings are extremely tightly clustered, mostly between 3.8 and 4.0.
- The distribution has multiple little bumps, but overall it’s narrow and centered.
- Variation across cuisines is minimal, suggesting cuisines rarely differ much in average rating.

Boxplot insight

- A few mild outliers on both sides, but nothing dramatic.
- The IQR is very tight, reinforcing that cuisine-level ratings barely spread out.
- This feature is highly stable and low-noise, no special preprocessing needed.

Avg. price city distribution and boxplot



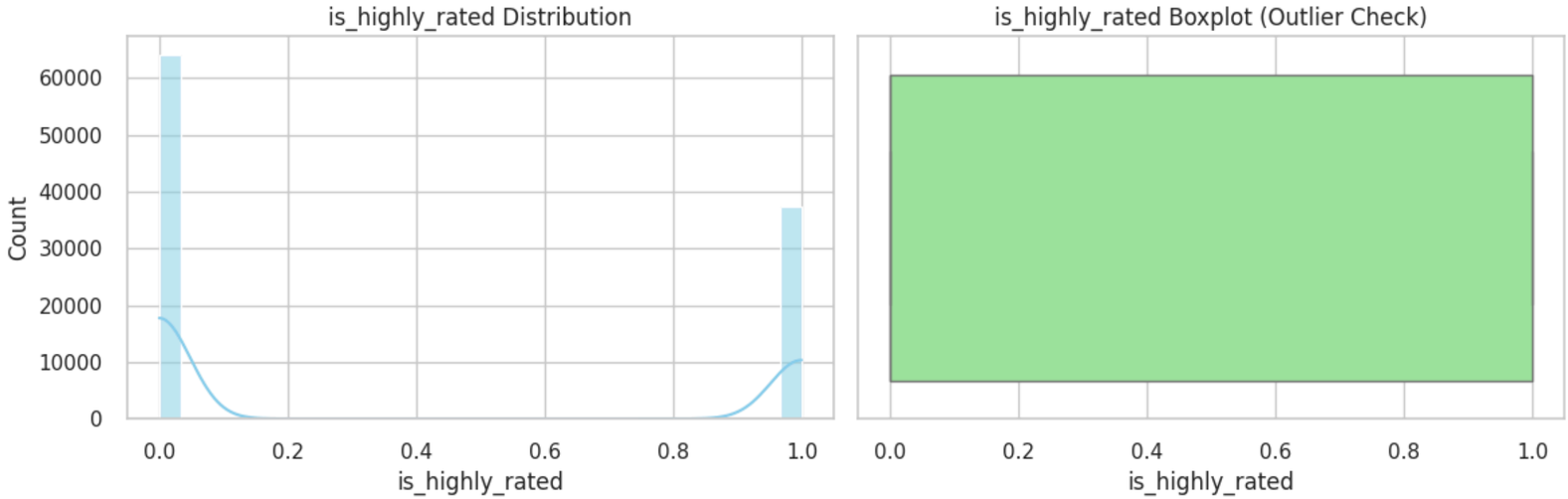
Distribution insight

- Prices cluster mostly between 200 and 300, with a few small bumps but generally compact.
- Only a tiny number of cities sit above 350, and anything beyond 500 is extremely rare.
- Overall, it’s a fairly tight and mildly right-skewed distribution.

Boxplot insight

- Most cities fall within a narrow IQR, showing low variation in city-level average prices.
- A few high-price cities appear as clear outliers, reaching past 600.
- Outliers seem meaningful rather than errors, but they’ll pull means upward if not handled.

Highly rated distribution and boxplot



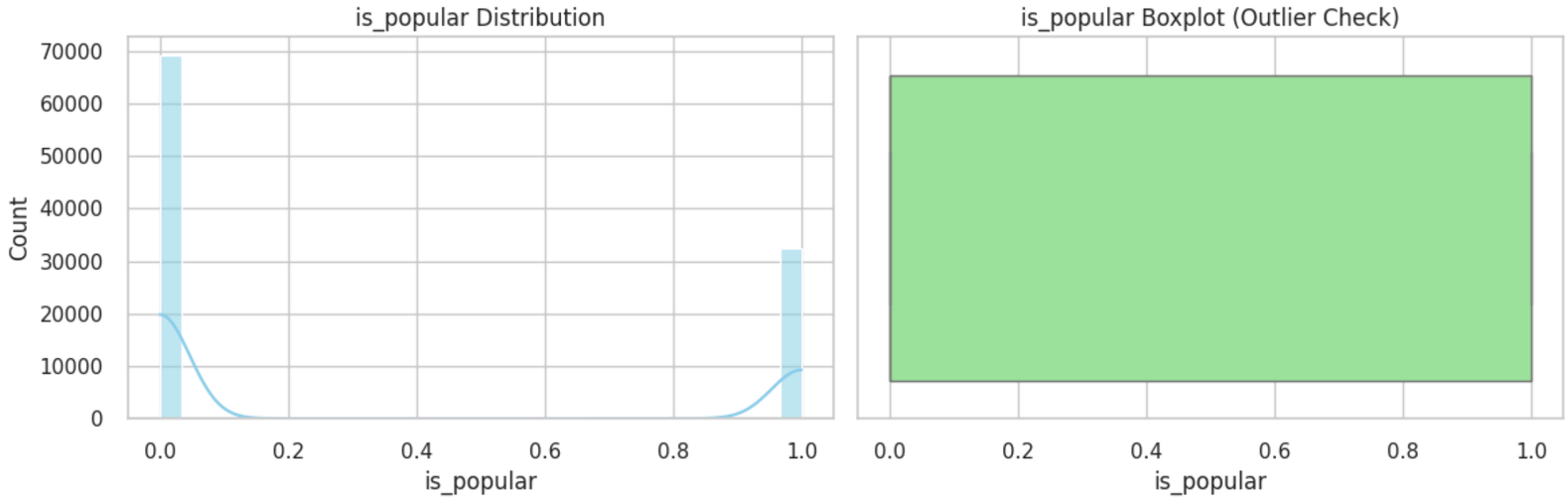
Distribution insight

- This is a binary feature, so the two spikes at 0 and 1 are expected.
- Class balance looks a bit uneven, with more 0s than 1s, but not wildly imbalanced.
- No shape to analyze beyond that, since it’s categorical.

Boxplot insight

- Boxplots aren’t very meaningful for binary data. The entire box just stretches from 0 to 1.
- There are no real outliers, because a binary variable can’t have any.
- This feature is clean and needs no preprocessing.

Popular distribution and boxplot



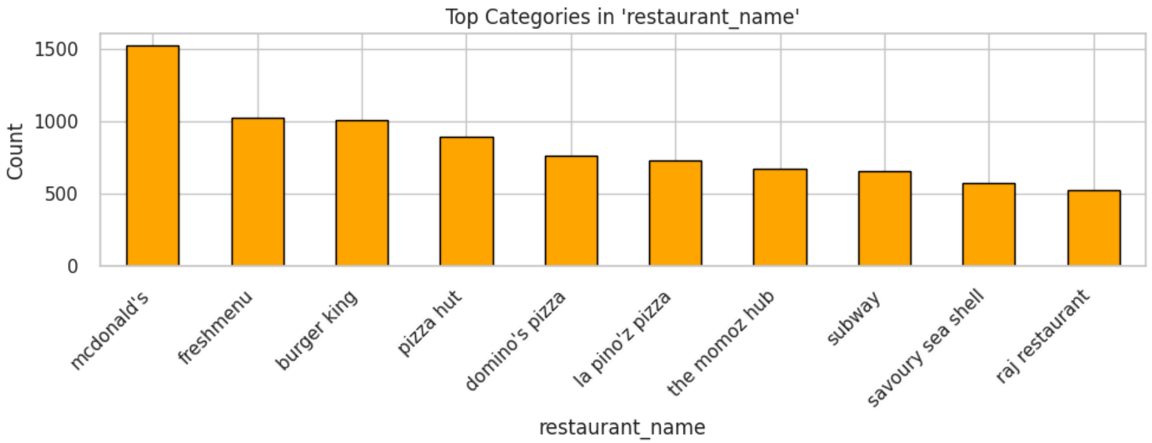
Distribution insight

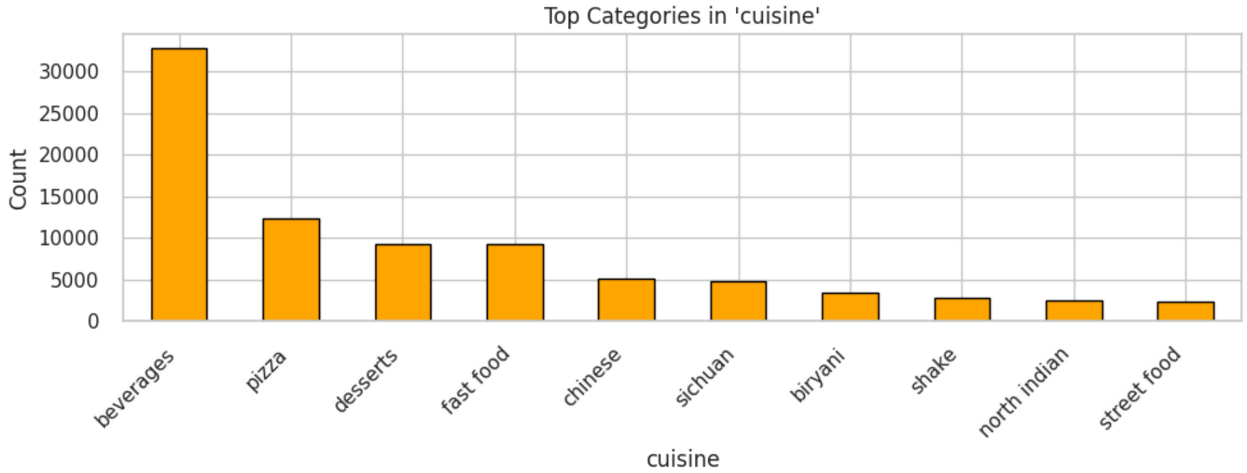
- Another binary feature, so the two spikes at 0 and 1 are expected.
- There are clearly more 0s than 1s, meaning most items are not labeled as popular.
- No continuous shape to interpret, just a simple class imbalance check.

Boxplot insight

- Boxplot doesn't provide meaningful information for binary variables.
- The box spans from 0 to 1 because those are the only possible values.
- No outliers can exist here. This feature is already clean and needs no preprocessing.

Top categories in resturant name and cuisine





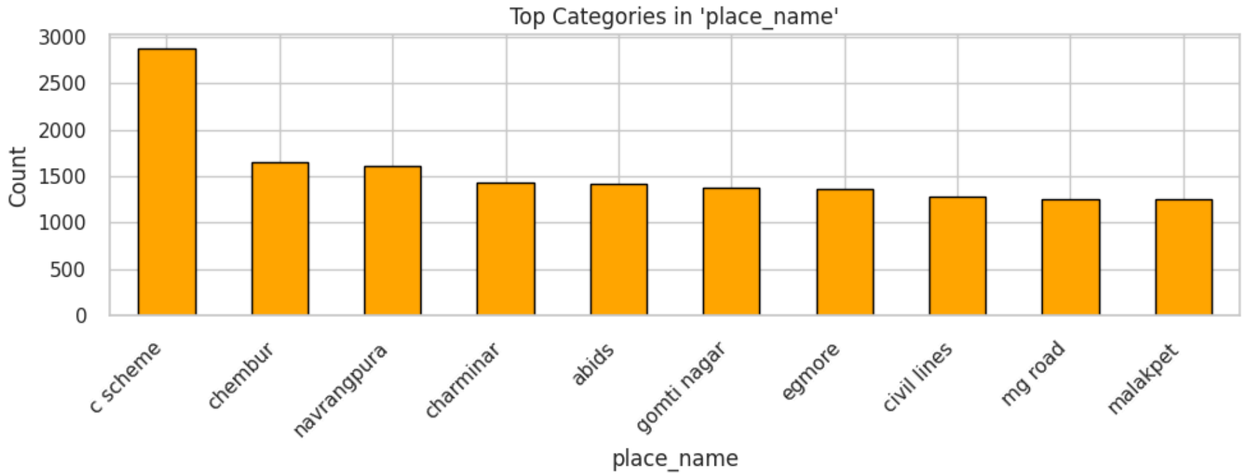
Restaurant name frequency insight

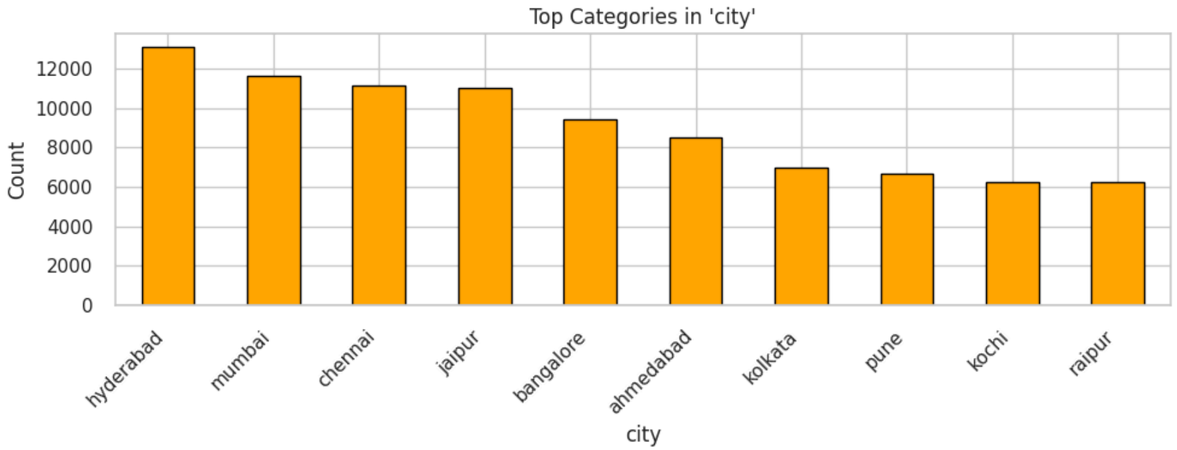
- The distribution is dominated by big chain brands. McDonald’s is way out in front, followed by FreshMenu, Burger King, and Pizza Hut.
- This suggests your dataset includes many repeated chain outlets rather than a diverse spread of unique local restaurants.
- The long tail of the dataset probably contains thousands of low-frequency restaurant names that don’t appear in the top 10.

Cuisine frequency insight

- Beverages absolutely dwarf every other cuisine category. This likely reflects lots of small-order items or beverage-oriented outlets.
- Pizza, desserts, fast food, and Chinese follow at a distance, forming the main cluster.
- The remaining cuisines (Sichuan, biryani, shakes, North Indian, street food) are still significant but far less common.
- Overall, cuisine distribution is heavily skewed, and the top few categories explain most of the dataset.

Top categories in place name and city





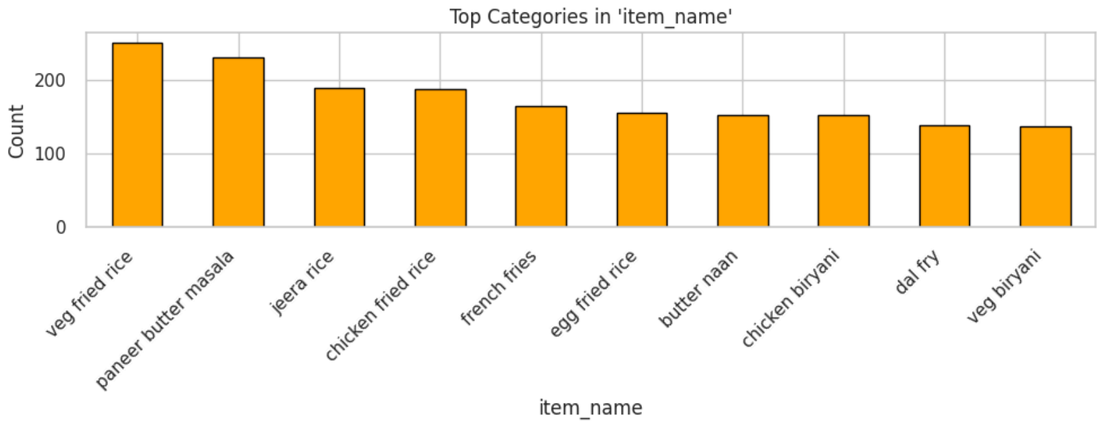
Place name insight

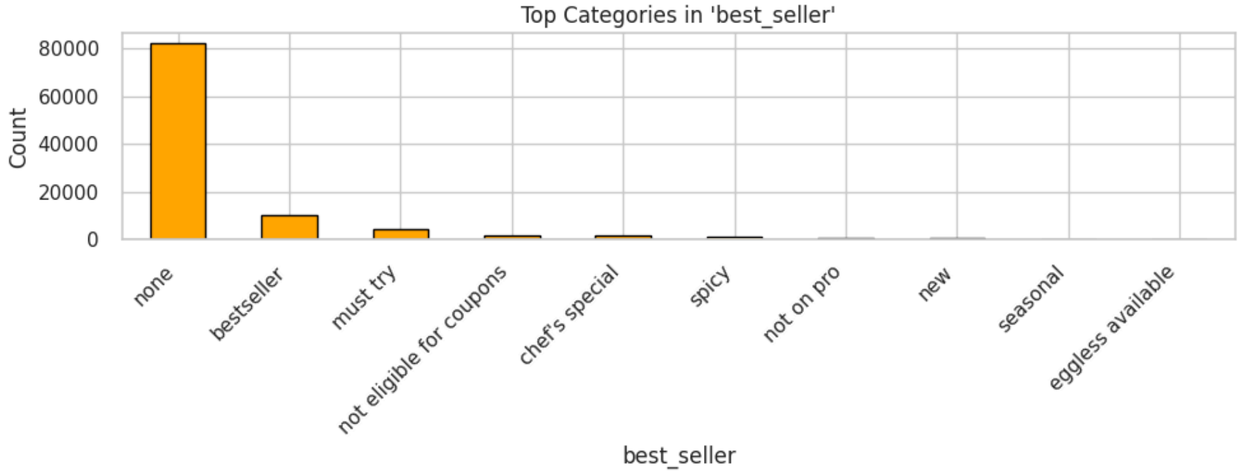
- C Scheme stands out far above the rest, indicating it's a dense hotspot with many listings.
- The next cluster (Chembur, Navrangpura, Charminar, Abids) sits at a moderate level, each with similar representation.
- The remaining places in the top 10 appear fairly balanced, suggesting a mix of mid-density localities.
- Overall, the distribution shows a few high-density areas and a larger base of moderately represented ones.

City insight

- Hyderabad leads clearly, followed by Mumbai and Chennai, reflecting cities with the largest platform activity.
- Jaipur, Bangalore, and Ahmedabad form a strong mid-tier cluster with sizable representation.
- Kolkata, Pune, Kochi, and Raipur appear lower but still substantial.
- The drop-off between cities is smooth, indicating broad national coverage rather than overreliance on a single metro.

Top categories in place name and city





Top items:

- Veg fried rice is basically the Beyoncé of your menu.
- Paneer butter masala is not far behind, doing its creamy thing.
- Fried rice in general is living its best life across multiple variants.
- Naan and biryanis show up strong but not top tier.

Best seller tags:

- “None” absolutely dominates. That usually means the field isn’t being actively curated or most items aren’t tagged at all.
- “Bestseller” is the only meaningful secondary tag.
- Everything else is tiny traffic.

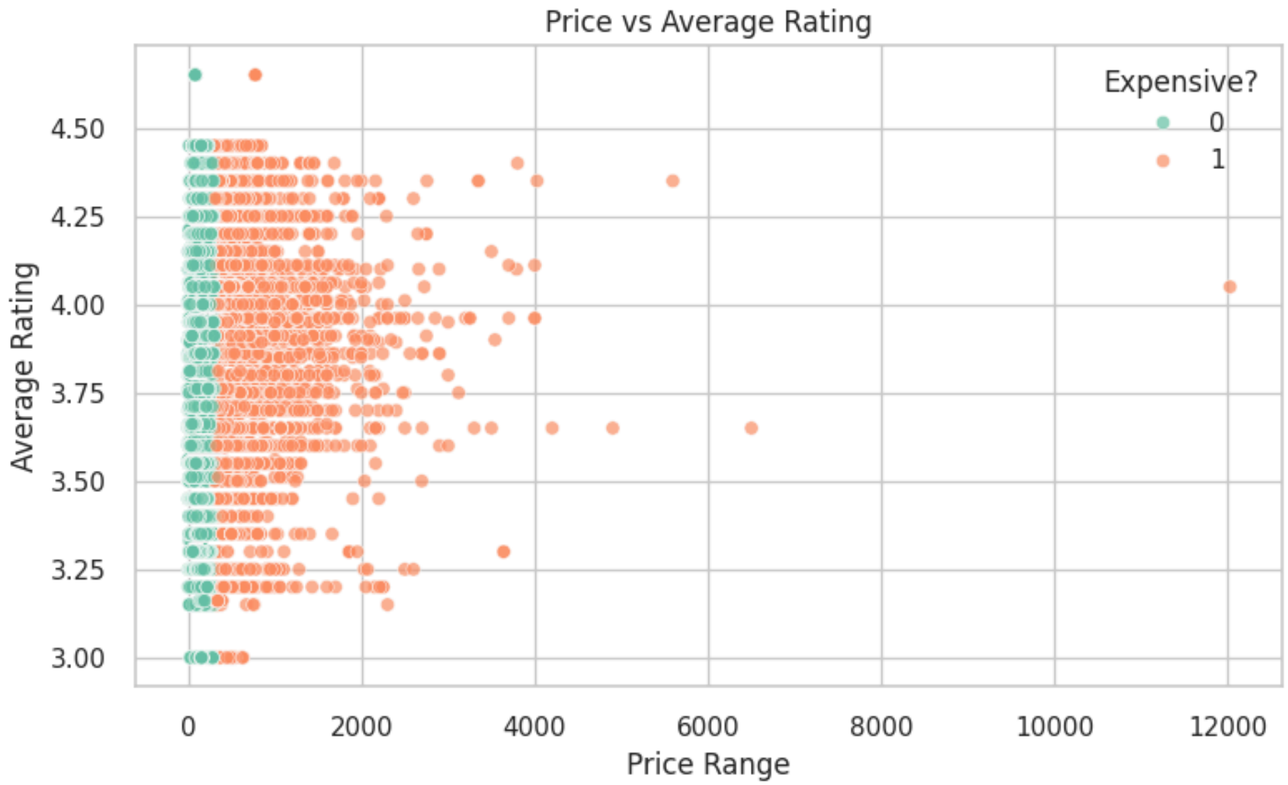
3. Bivariate Analysis

What is Bivariate Analysis?

Bivariate analysis examines the relationship between two variables to understand how one variable changes with respect to another. It helps identify correlations, patterns, or associations between pairs of variables — such as how ratings vary with price or votes.

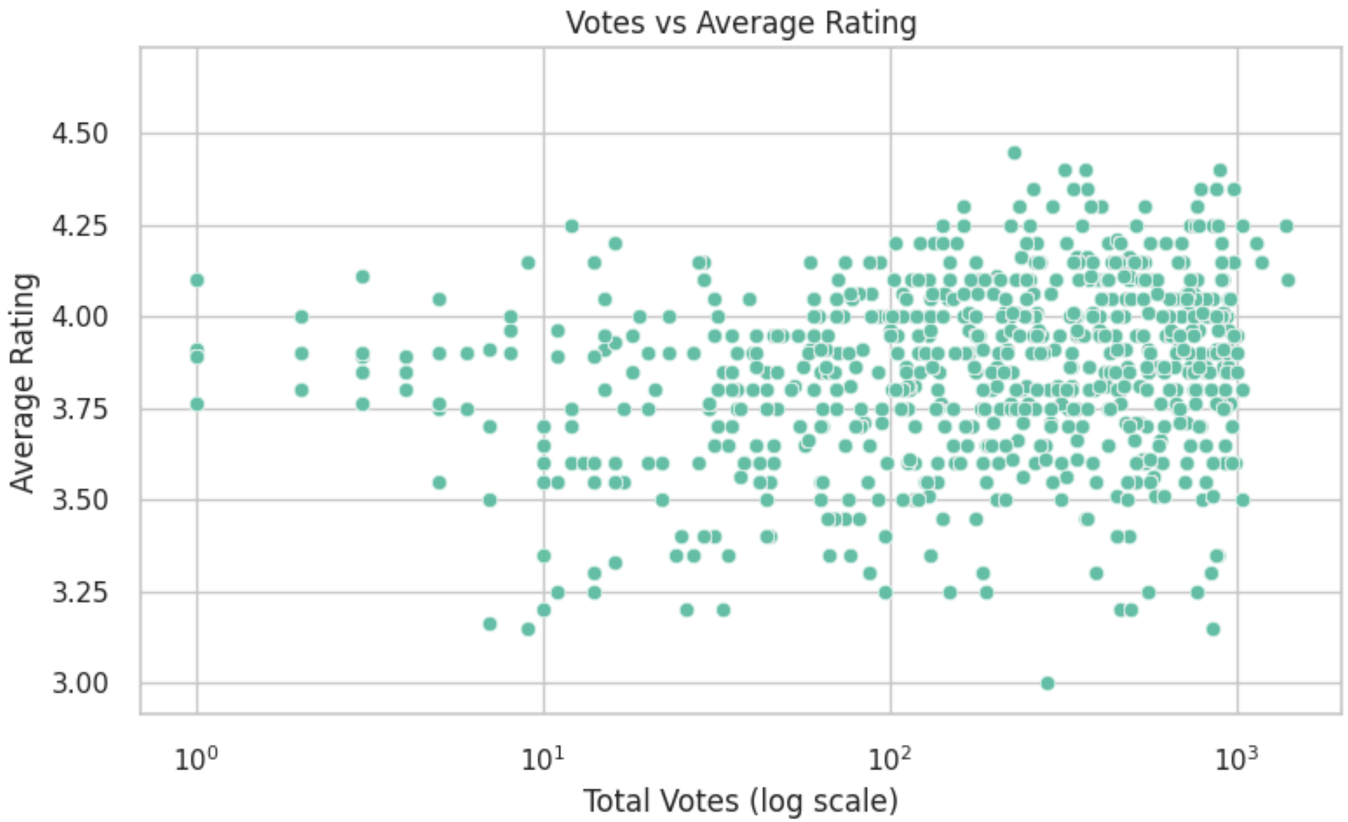
Relationship: Price vs Rating

- Most items cluster in the low to mid price range, and ratings don’t swing wildly with price.
- Expensive items don’t consistently rate higher. Some pricey ones do fine, others just... meh.
- Cheaper items show tons of variation, but many still get solid ratings.
- Overall: price isn’t a strong predictor of rating in this dataset.



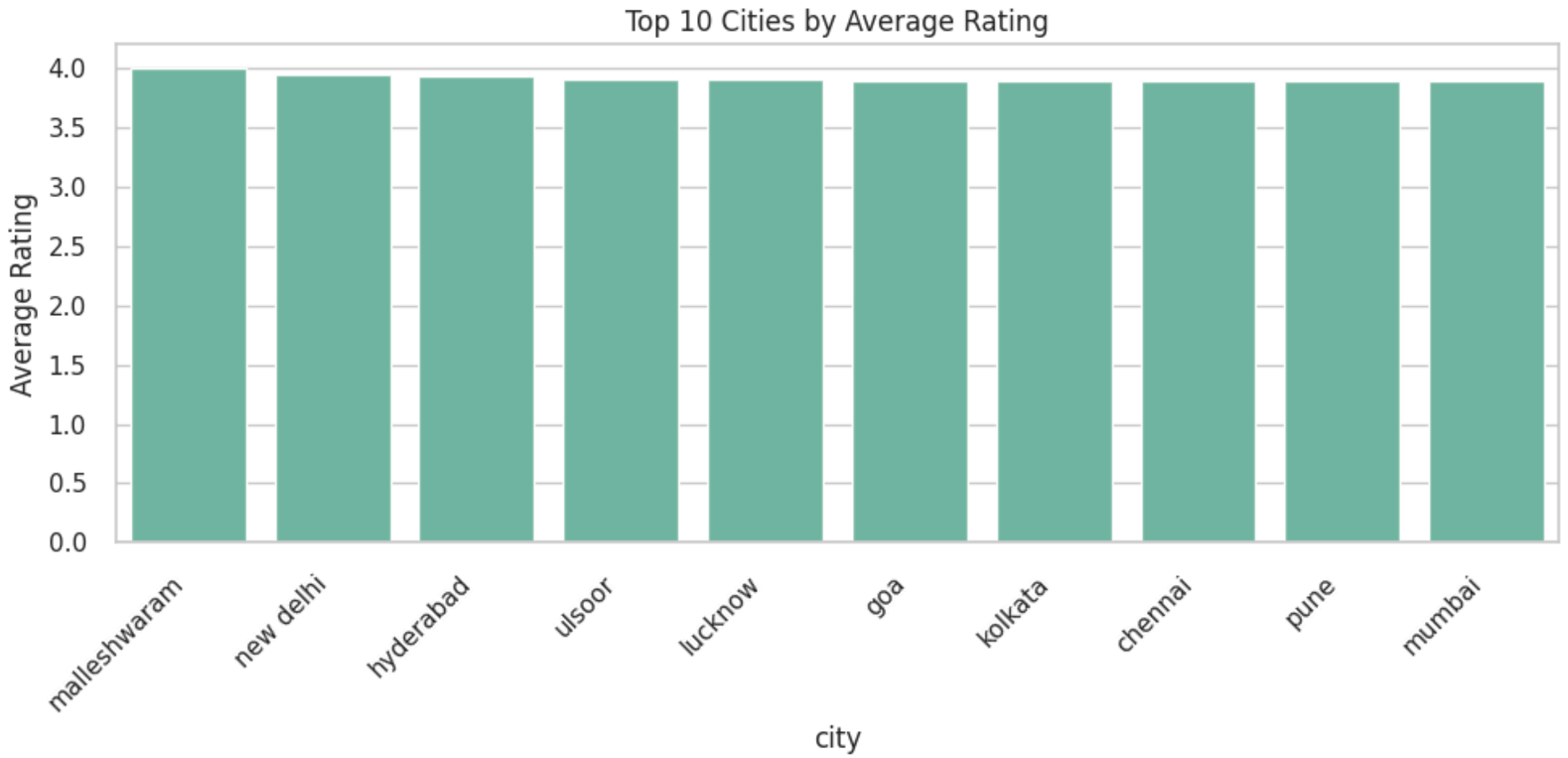
Votes vs Rating

- More votes loosely trend with slightly higher ratings, but the relationship isn’t strong.
- Items with very few votes scatter everywhere, which makes sense because small samples swing hard.
- High-vote items tend to cluster around the 3.7 to 4.2 range, suggesting crowd consensus stabilizes ratings.
- No clear evidence that popularity guarantees a top tier rating.



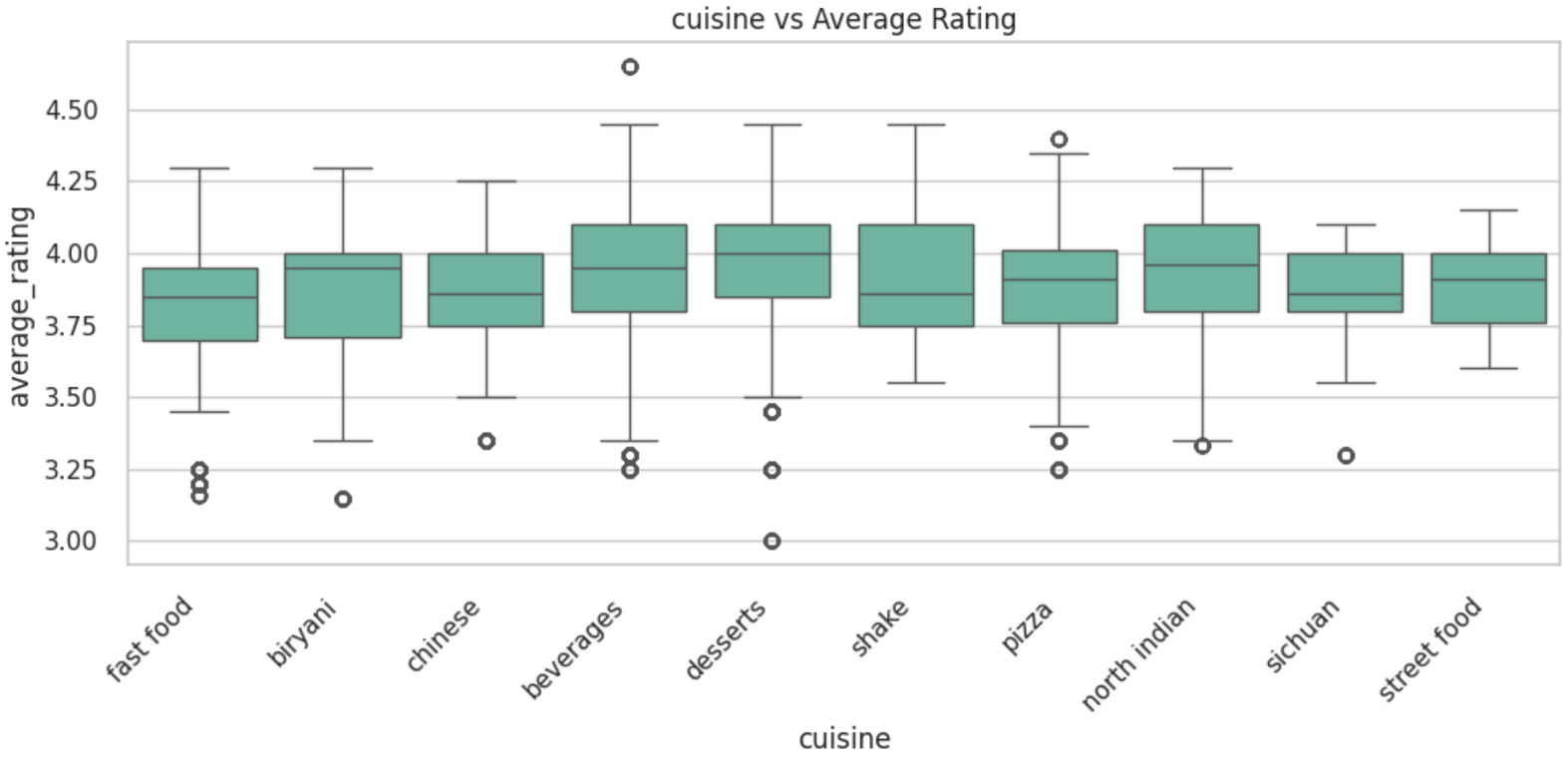
City-Level Rating & Pricing

- All top cities cluster super tight around the same average rating near 3.9, so there isn’t a runaway winner.
- Malleshwaram edges out the others, but only by a hair.
- Differences are tiny, which hints that ratings are pretty consistent across cities in your dataset.



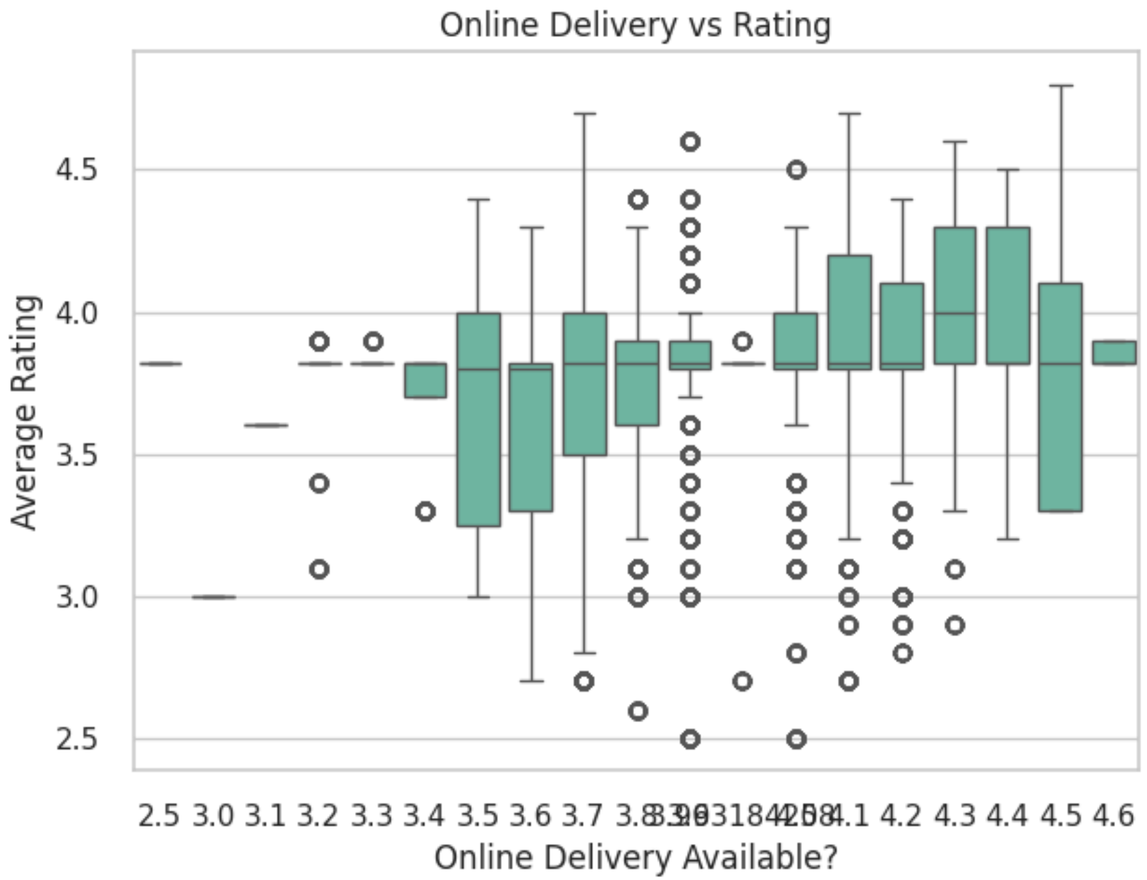
Cuisine vs Rating

- Most cuisines sit in a pretty tight band around 3.8 to 4.0, so tastes are generally well-received across the board.
- North Indian and beverages lean a bit higher, with slightly stronger medians and some top-end outliers.
- Fast food and Chinese show more low-end outliers, hinting at more hit-or-miss experiences.
- No cuisine absolutely dominates, but some definitely have a smoother reputation than others.



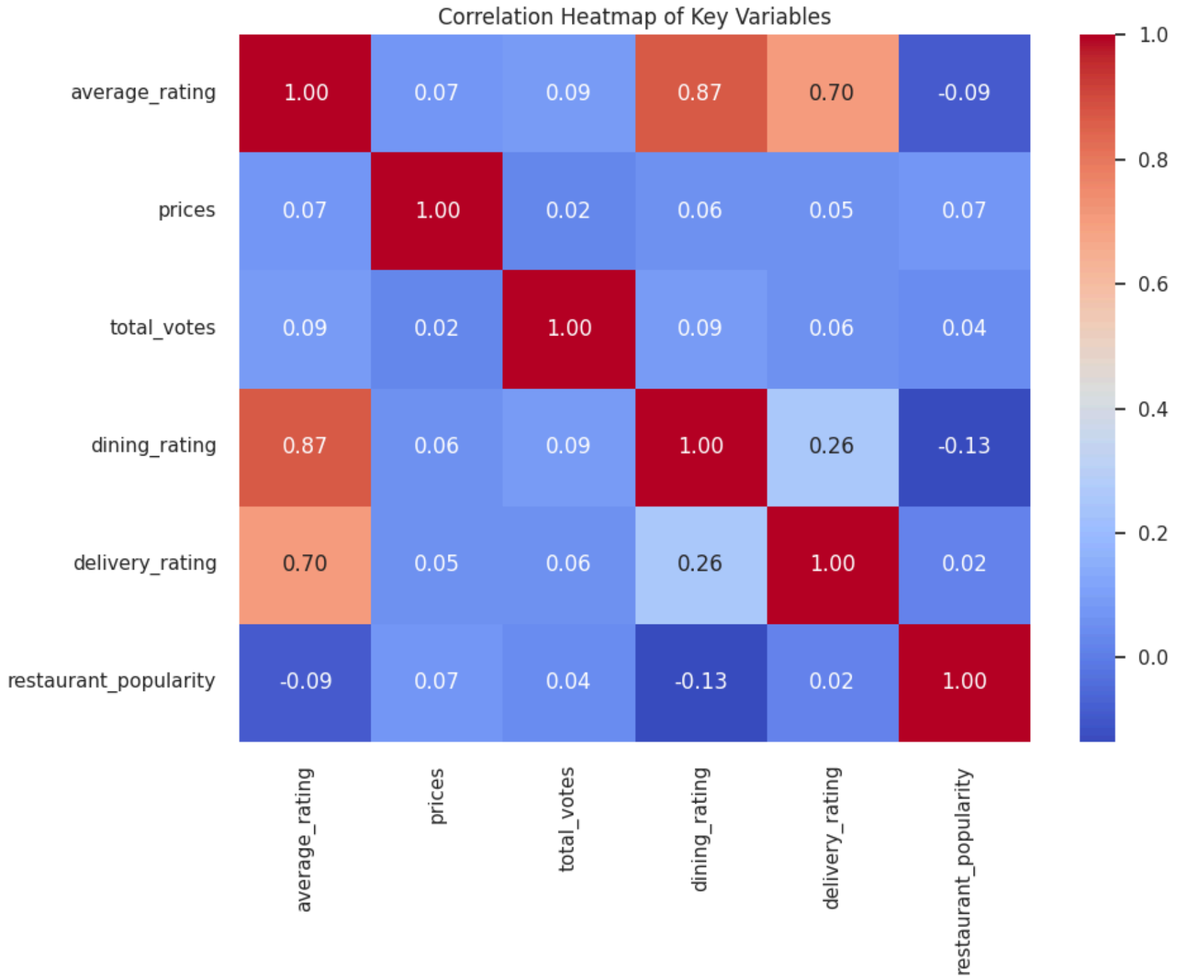
Online Delivery vs rating

- Places offering online delivery seem to skew a little higher in ratings, especially in the upper quartiles.
- Non-delivery spots show more low-end outliers, so quality swings more there.
- Still, the overlap is big, so delivery availability isn't a magic ticket to great ratings.
- Overall: delivery helps a bit, but it's not the main driver of customer satisfaction.



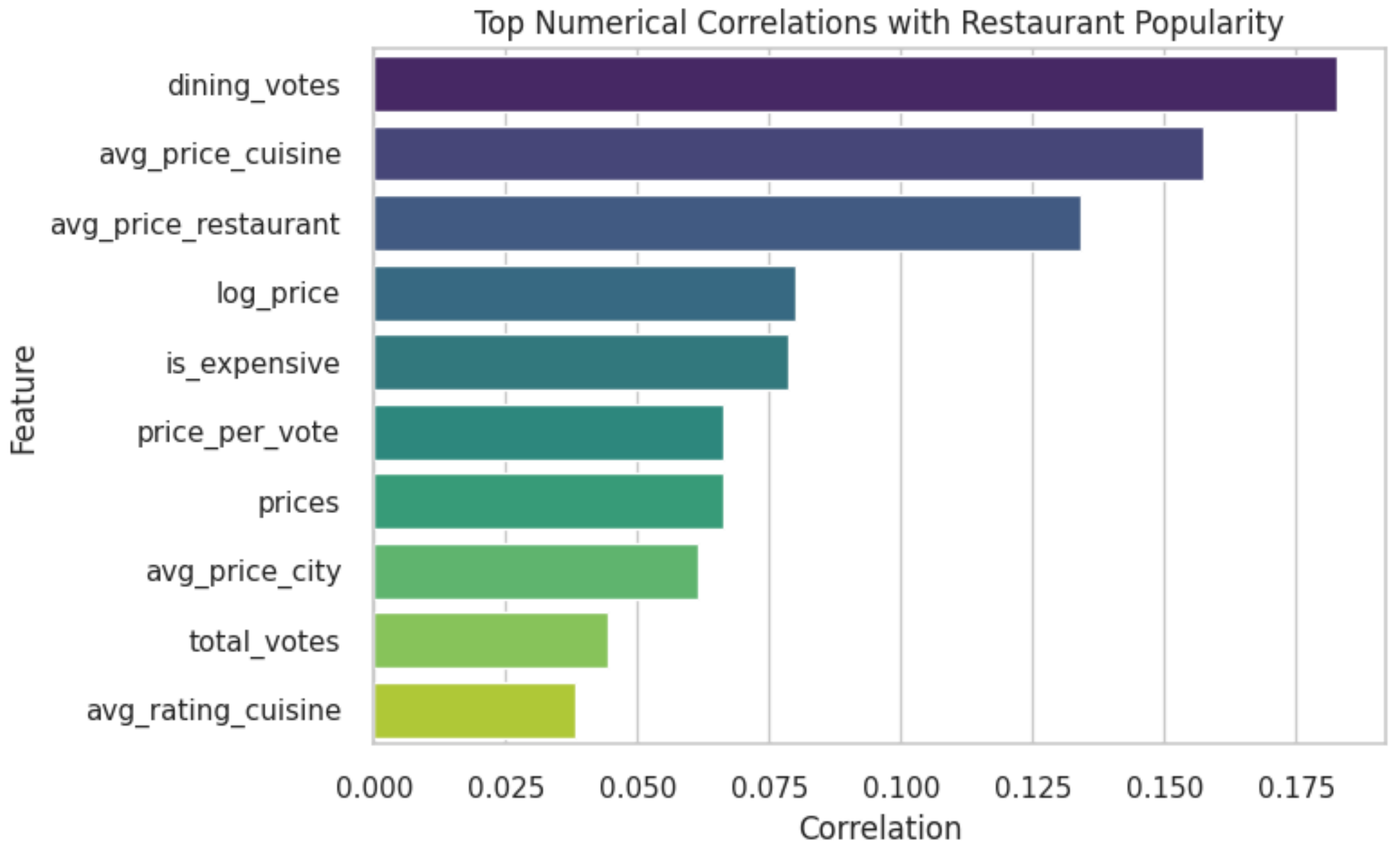
Correlation Heatmap of Key Metrics

- Average rating is driven mostly by dining rating and delivery rating. Dining is the heavyweight here with a very strong positive link.
- Prices, total votes and popularity barely correlate with ratings, meaning they don't meaningfully sway customer sentiment.
- Dining and delivery ratings correlate with each other, but only mildly, so good dine-in doesn't guarantee good delivery.
- Popularity has almost no relationship with quality, and even a slight negative with dining. Basically, being famous doesn't mean being good.



Correlation Analysis with Popularity

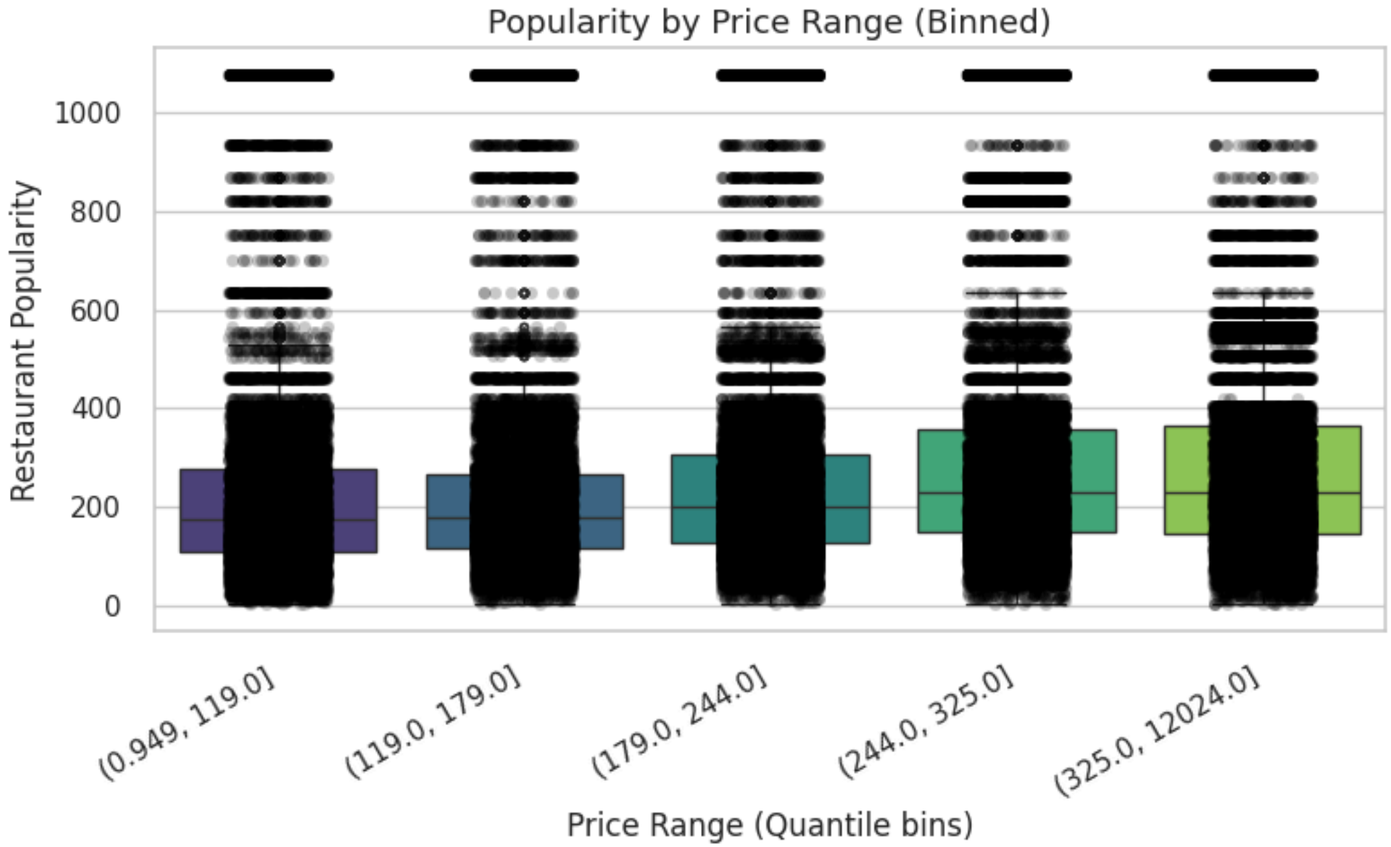
- Popularity in your dataset is mostly tied to dining votes. More people rating the dine-in experience nudges popularity up the most.
- Higher average price levels also show a noticeable positive link. Basically, pricier spots tend to be a bit more “popular,” at least in terms of your metric.
- Features like log_price, is_expensive, and price_per_vote follow the same mild pattern. Nothing huge, just small nudges.
- Total votes barely move the needle, which is interesting because you’d expect overall votes to matter more.
- Average rating by cuisine has the weakest relationship, so cuisine-level quality doesn’t translate to popularity much.



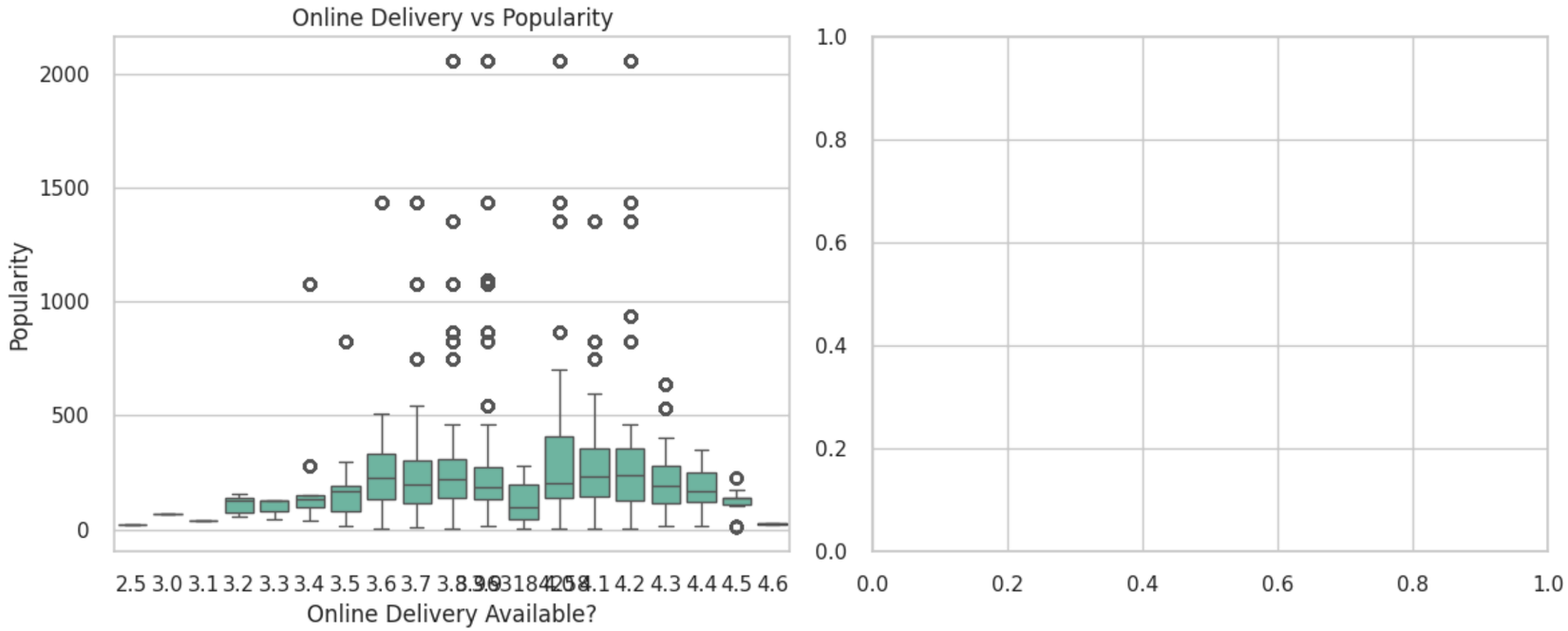
Popularity by Price Range

- Popularity bumps up slightly as you move into higher price ranges, but the change is mild.
- All bins show massive spread, meaning cheap and expensive places alike can be wildly popular or totally quiet.
- The median does inch upward for higher price bins, so pricier restaurants tend to pull a bit more attention on average.
- Still, price alone clearly isn't doing the heavy lifting since the overlap between groups is huge.

Short version: higher prices help a little, but popularity is way more chaotic than price driven.

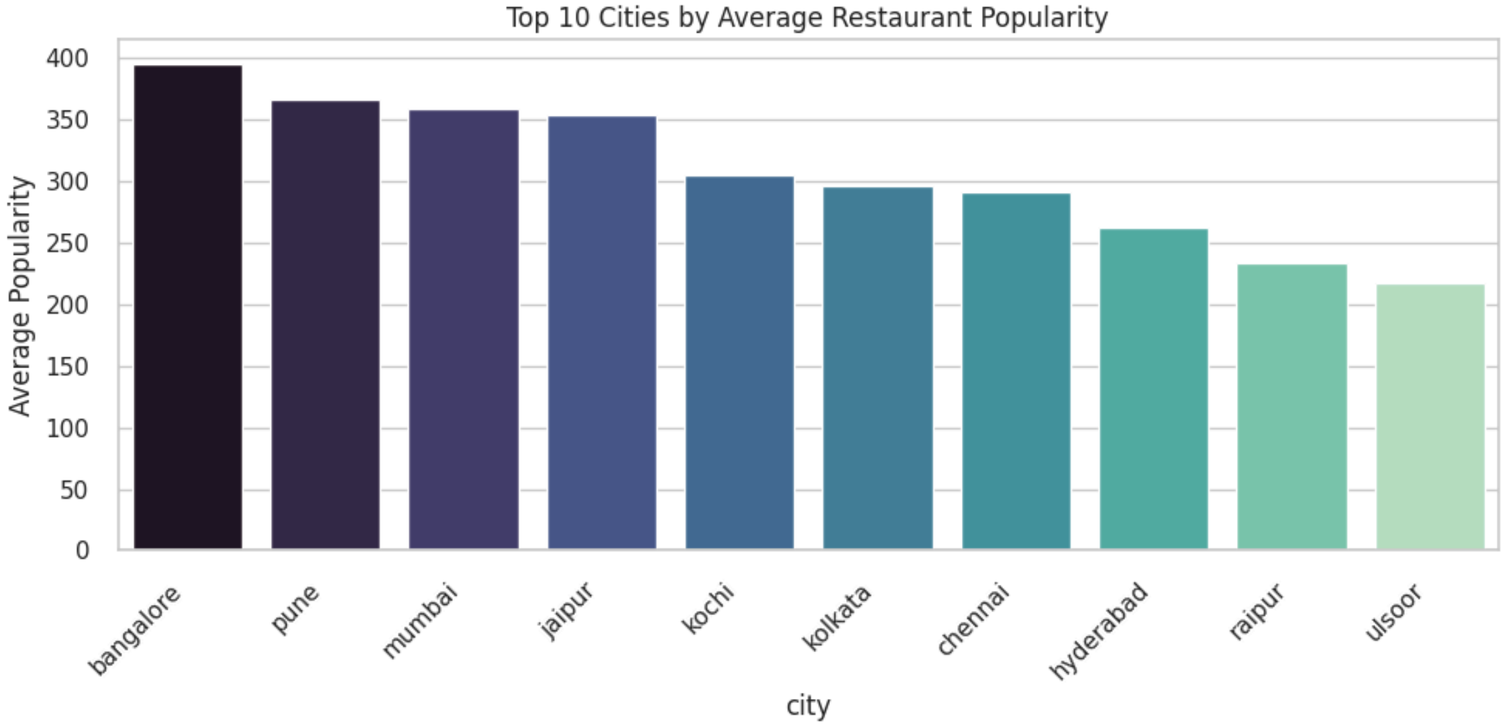


Online Delivery & Booking Impact



- Restaurants with online delivery tend to show noticeably higher popularity overall.
- The spread is huge, but the median and upper ranges for delivery-enabled places are clearly higher.
- Non-delivery spots cluster low with very few high-popularity outliers.
- So delivery availability doesn’t guarantee popularity, but it definitely stacks the odds in your favor.

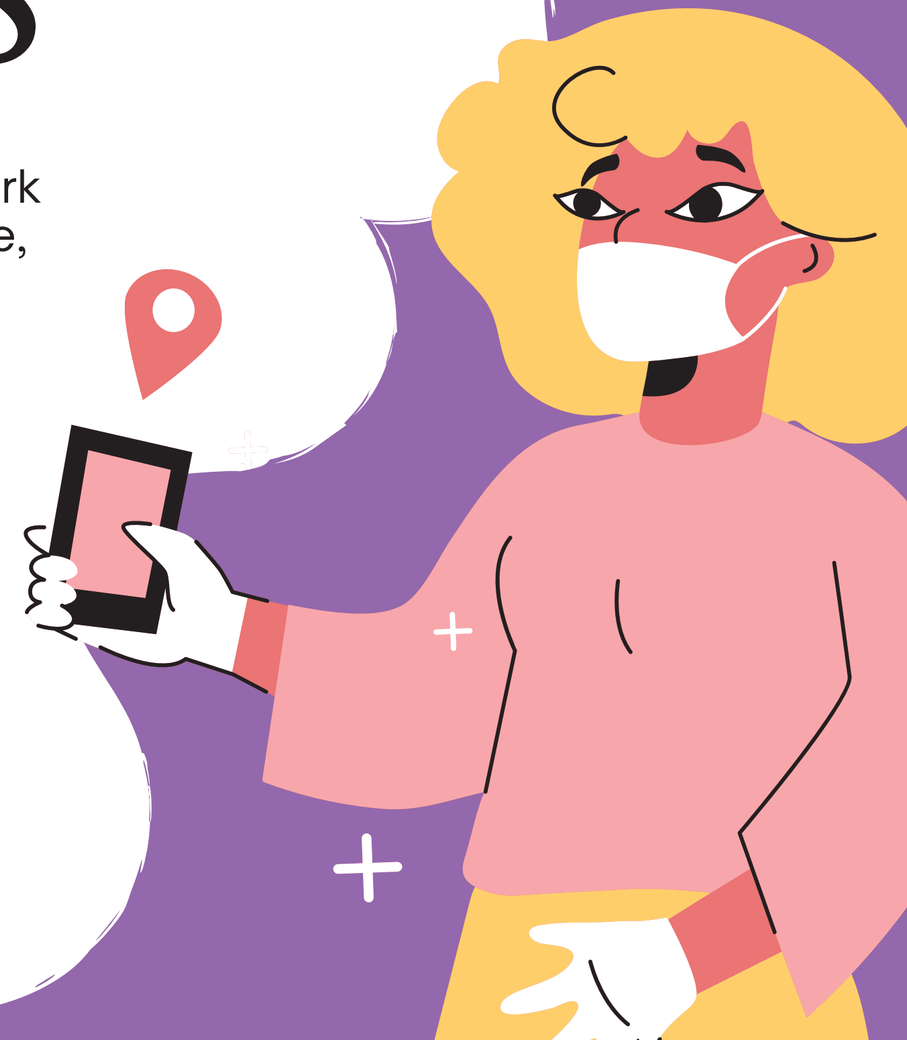
City-Level Popularity Heatmap



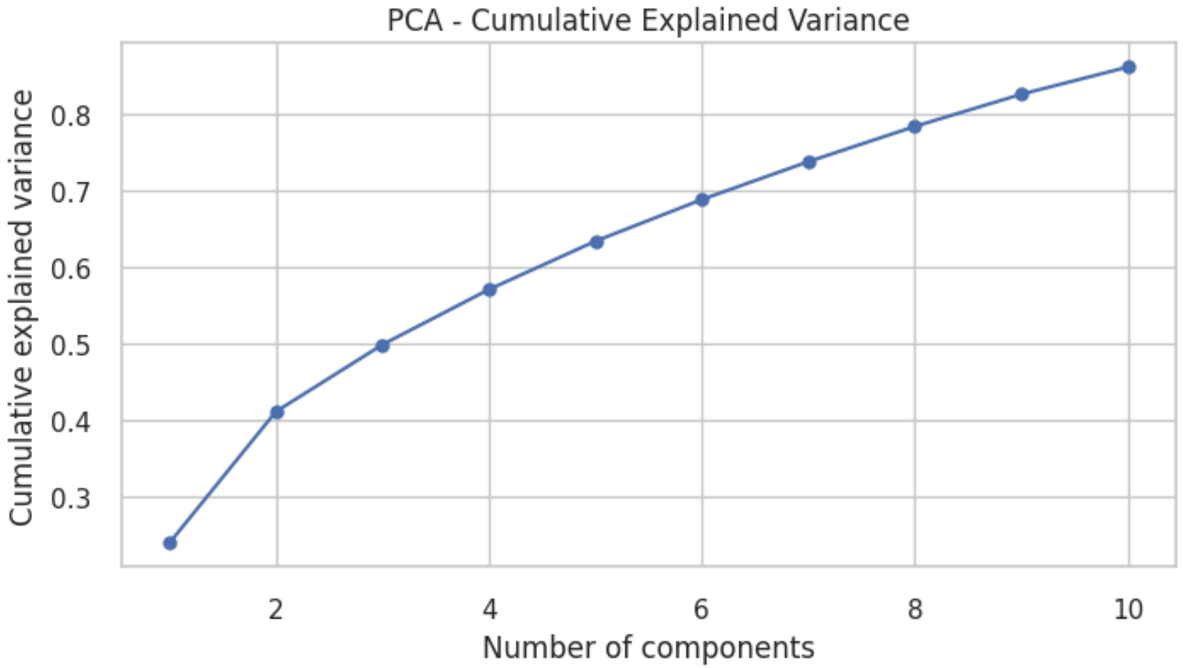
- Bangalore is way out in front, pulling the highest average restaurant popularity by a solid margin.
- Pune, Mumbai and Jaipur sit in a tight second cluster, all fairly strong but not Bangalore-level.
- Cities like Kochi, Kolkata and Chennai follow with moderate popularity.
- Hyderabad, Raipur and Ulsoor are noticeably lower, rounding out the bottom of the top ten.

4. MULTIVARIATE ANALYSIS

It studies how several variables work together — for example, how price, city, and cuisine together affect average rating.

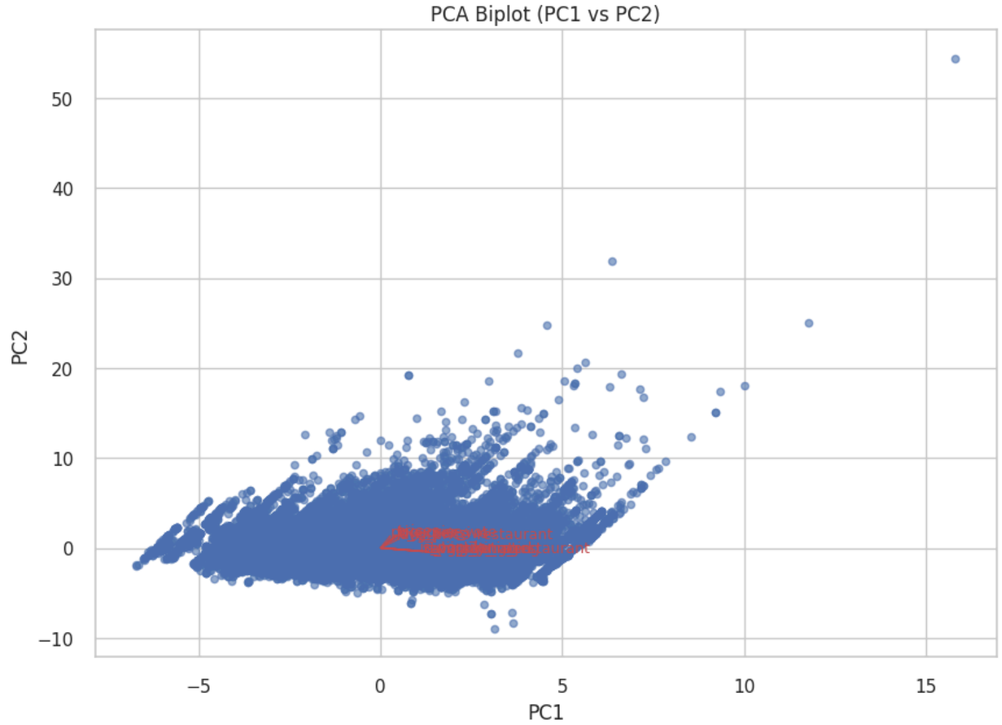


PCA – linear dimensionality reduction & explained variance



Cumulative Explained Variance

- The curve climbs fast at first, then slows.
- PC1 and PC2 together explain a bit over 40 percent of the variance.
- Around 6 components gets you close to 70 percent, and about 10 components lands you near 88 percent.
- So the data is pretty spread out. You need multiple components to capture most of the structure.



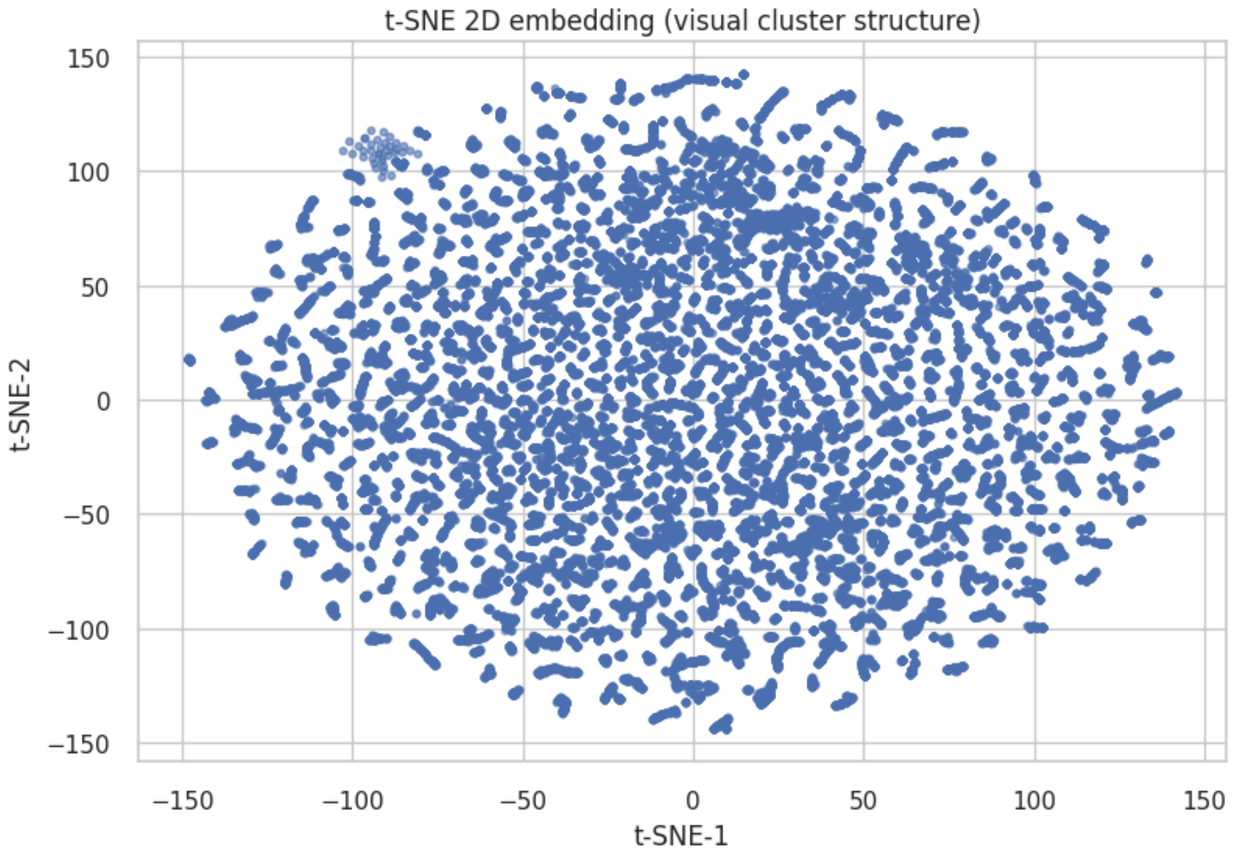
PCA Biplot (PC1 vs PC2)

- The scatter shows a long, stretched cloud, meaning the first two components capture some structure but definitely not all.
- PC1 seems to pick up the main gradient in the data. PC2 adds some vertical spread but less dominant.
- The red loading vectors are short and packed, so no single feature dominates the first two PCs. The structure is driven by many small contributions rather than one standout variable.

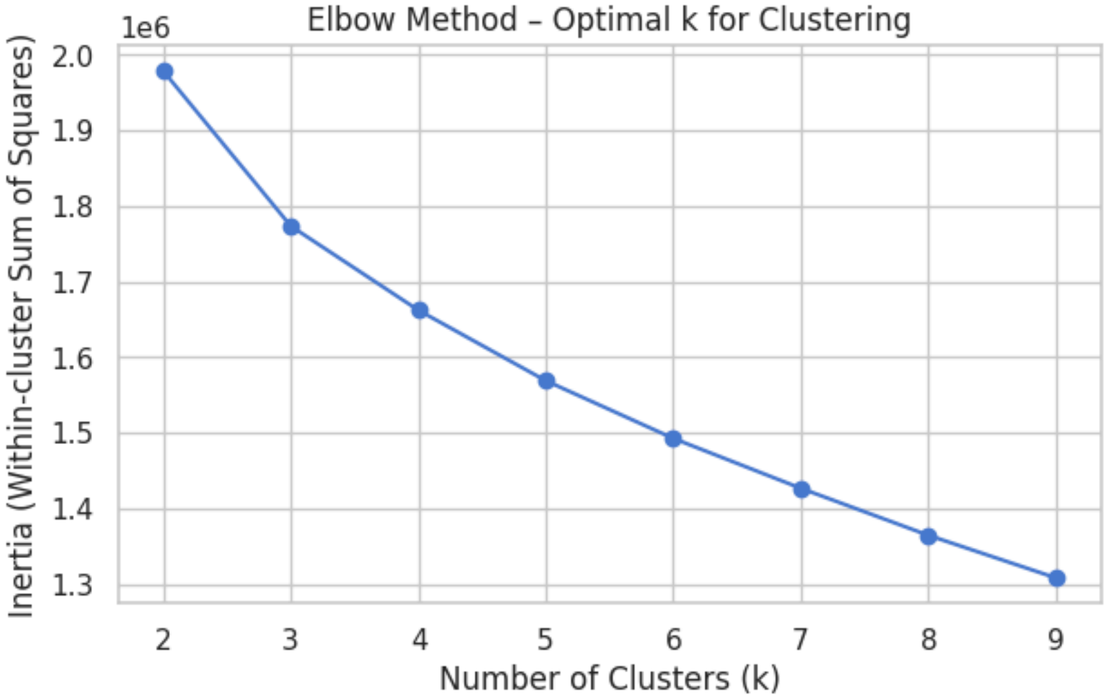
t-SNE – non-linear embedding for cluster structure

- Nothing forms a clean, separate cluster. The points smear into one big fuzzy cloud.
- There are tiny pockets of local structure here and there, but nothing that looks like strong, well-defined groupings.
- This usually means the underlying features don't naturally break the restaurants into clear categories based on the variables you fed into t-SNE.
- Translation: the data is kinda "continuous" rather than cleanly segmented, at least in the space t-SNE is seeing.

Short and sweet: no obvious clusters jump out, so any segmentation will probably need domain rules rather than expecting the data to self-organize.

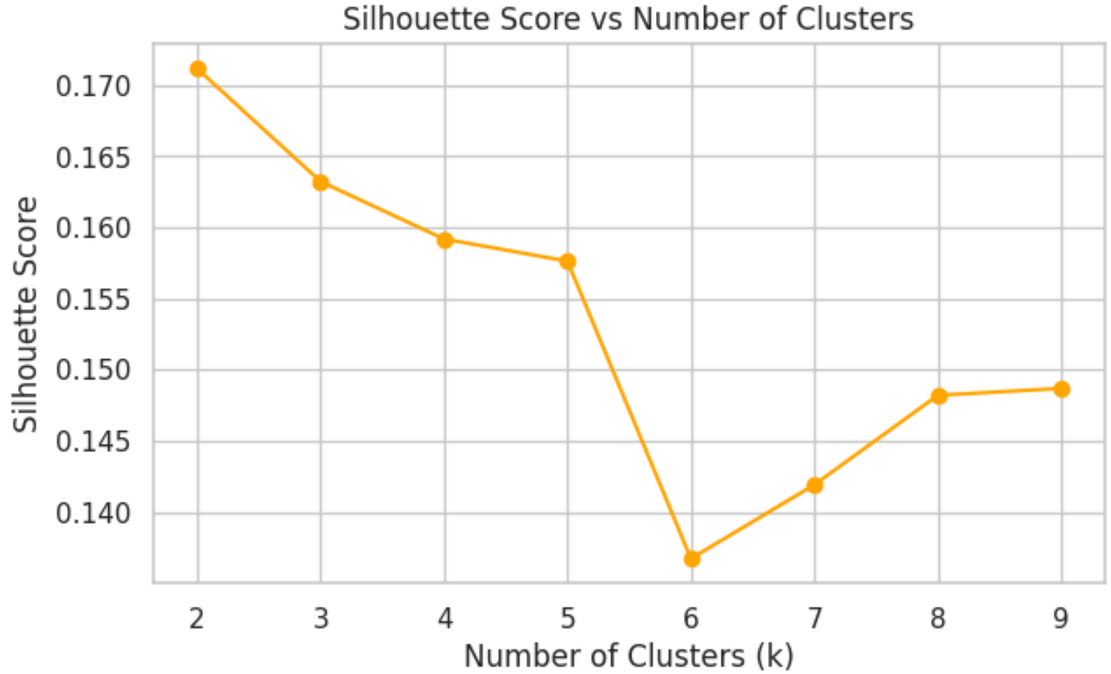


Clustering: KMeans with Elbow & silhouette



Elbow Plot

- The curve drops sharply from k=2 to k=3, then keeps declining but with no dramatic bend.
- There's a mild kink around k=3 or k=4, but nothing screaming "this is the optimal k."
- Translation: the data doesn't naturally break into super clean clusters.



Silhouette Scores

- Best score is at k=2, and everything after that dips or stays low.
- Scores overall are pretty low, which means the clusters you're forming are weakly separated.
- The big drop at k=6 tells you that too many clusters just makes a mess.

Multivariate Interaction Check (pairwise interaction heatmap)

Here’s the breakdown:

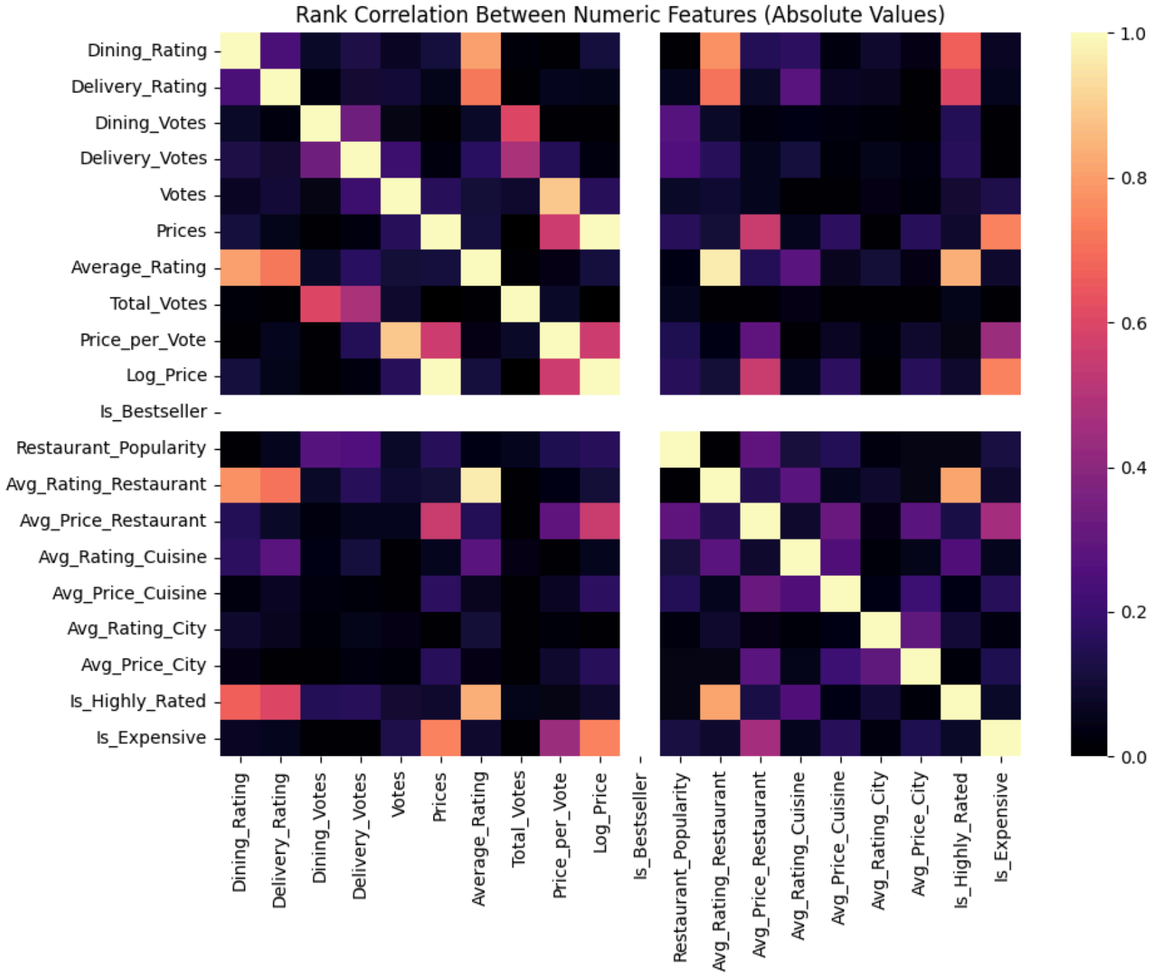
1. Ratings clusters (kinda) stick together
2. Votes-based features behave predictably
3. Price features mildly relate to each other
4. The constructed “Is_*” flags light up only where you'd expect
5. Restaurant/Cuisine/City averages barely correlate with item-level data

TLDR

The heatmap tells the same story as your t-SNE and clustering metrics: The dataset has low inherent structure. Most features only weakly correlate with each other.

If you’re planning modeling next (like segmentation, prediction, or ranking), expect to lean on:

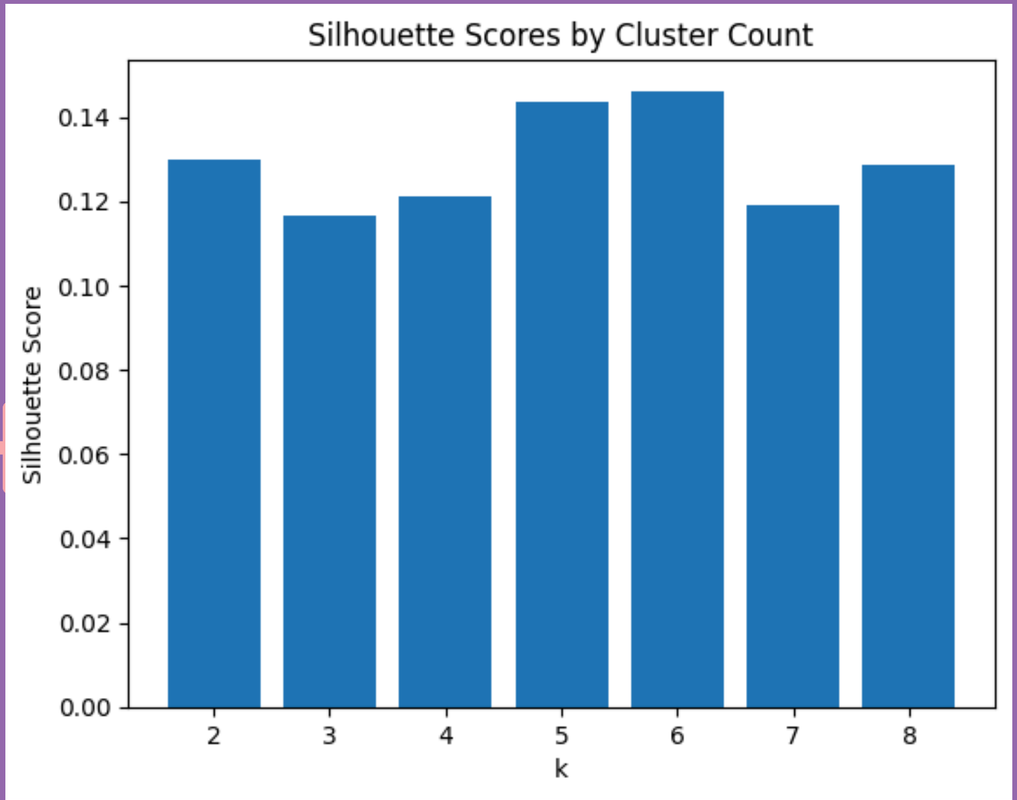
- Nonlinear models
- Feature engineering
- Interaction terms
- Maybe latent embedding methods



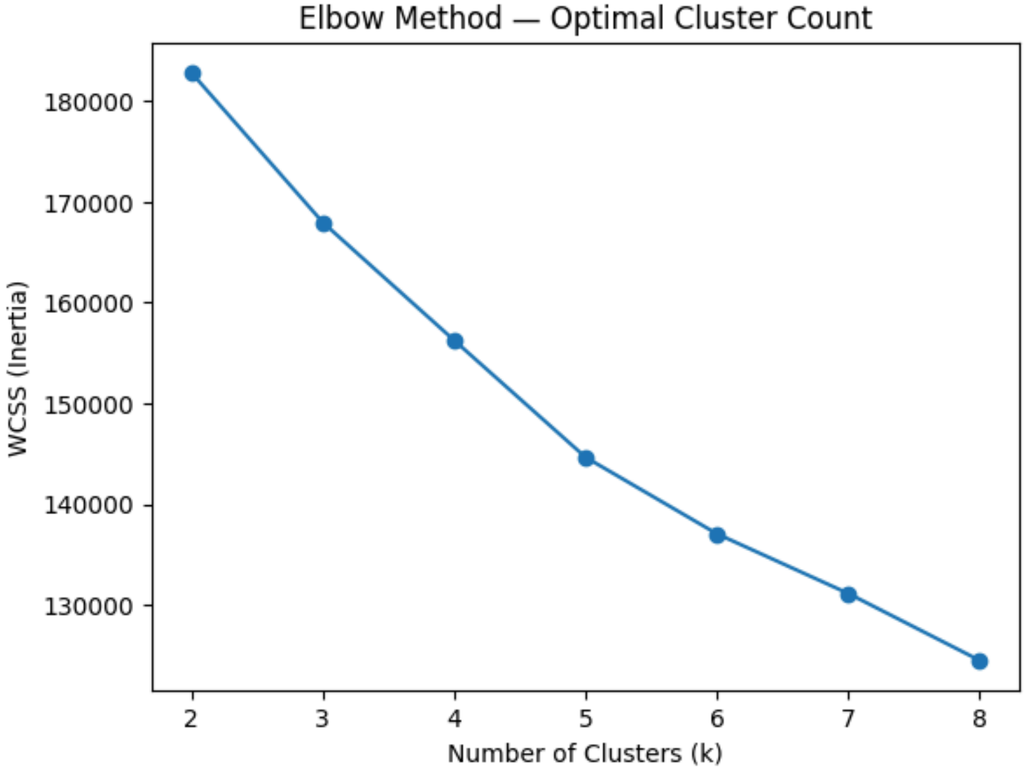
5.SEGMENTATION & CLUSTERING

Segmentation means dividing a large dataset (like customers, restaurants, etc.) into smaller, meaningful groups (segments) that share similar characteristics.

Clustering is a machine learning technique (unsupervised learning) used to perform segmentation automatically — it groups data points so that those in the same cluster are more similar to each other than to those in other clusters.



- Silhouette Scores:
- Silhouette values dip slightly at k=3 and k=4, then rise and hit their strongest performance at k=5 and k=6. These two points indicate the most naturally separable cluster structures. Beyond k=6, the silhouette score doesn't improve and starts behaving inconsistently, which usually hints at unnecessary fragmentation.



- Elbow Method:
- You can see a sharp drop in WCSS from k=2 to k=4, then the curve starts flattening. By the time you reach k=5, the reduction in inertia becomes more gradual. After k=6, the improvements are minimal, meaning extra clusters aren't buying you much structural clarity.

Cluster visualization with PCA

Cluster 0:
Mostly concentrated on the left side. These points sit lower on PC1 and closer to the mid-range of PC2. This group likely represents restaurants with moderate ratings, mid-low pricing, and average popularity patterns.

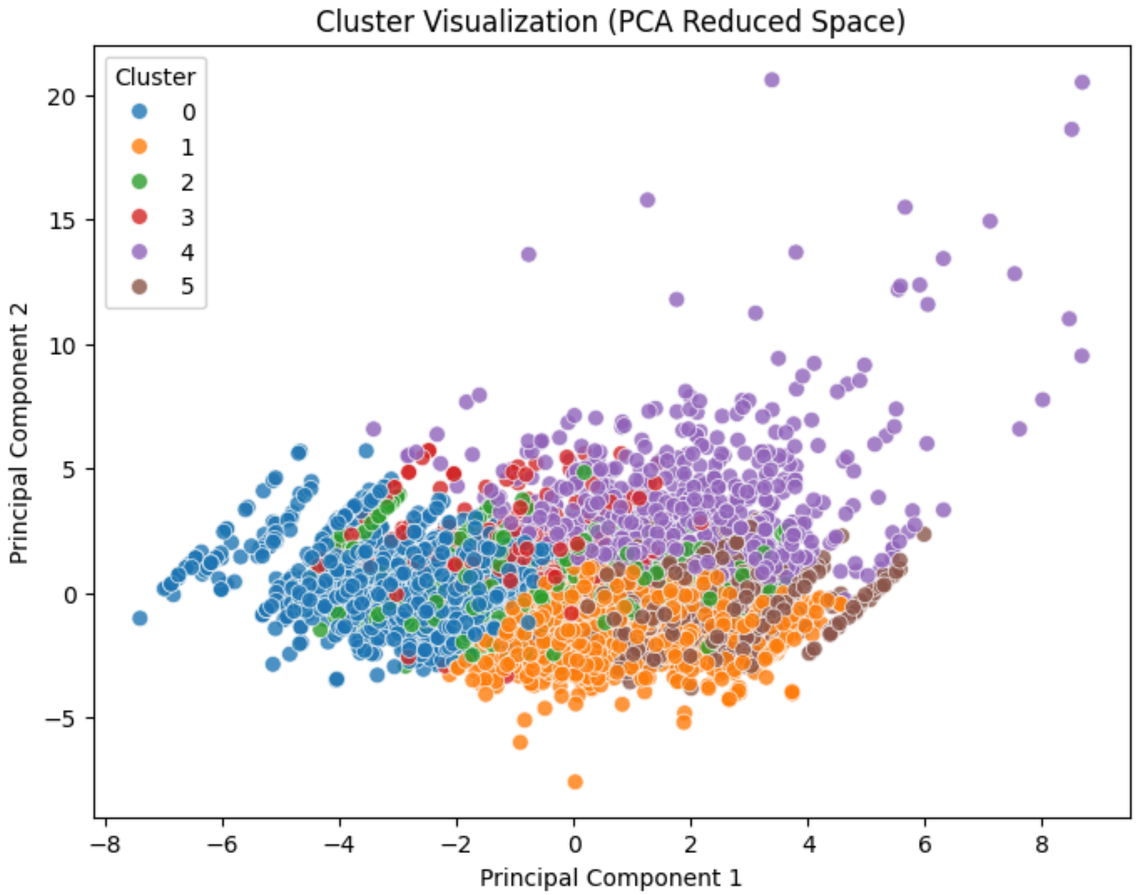
Cluster 1:
Spread toward the lower part of PC2 with PC1 leaning positive. This cluster might capture restaurants with stronger delivery engagement or lower-cost profiles, since they pull downward on the second component.

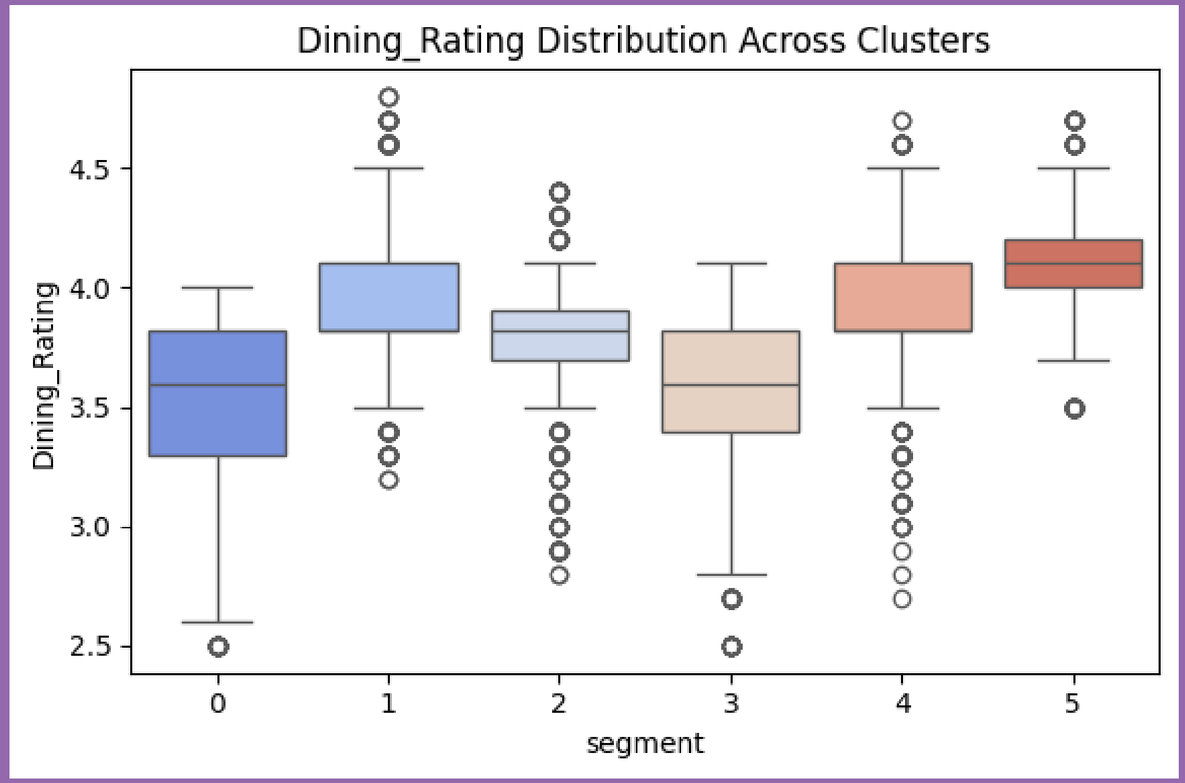
Cluster 2:
A compact, tighter cluster near the center. These restaurants look very balanced across features, showing neither extreme price nor extreme rating behavior.

Cluster 3:
A small but distinct handful overlapping the central region. Since there aren't many points, this group may represent niche or special-case restaurant profiles with unique combinations of ratings or price.

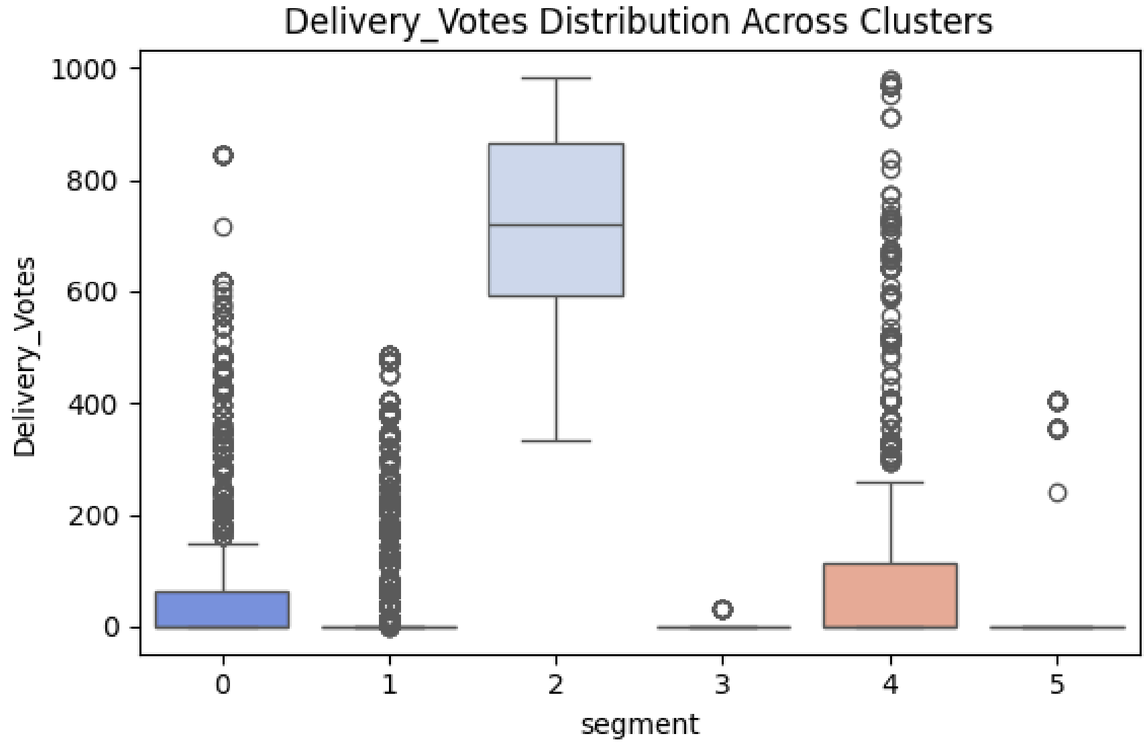
Cluster 4:
The wide upward-stretching purple group. These restaurants have high values on PC2, which often relates to total engagement, votes, or popularity. They're kind of the high-visibility, high-interaction crowd.

Cluster 5:
Positioned on the right edge with positive PC1 and slightly negative PC2. These tend to be high-price or premium-leaning restaurants, or restaurants with high average ratings but lower vote volumes.

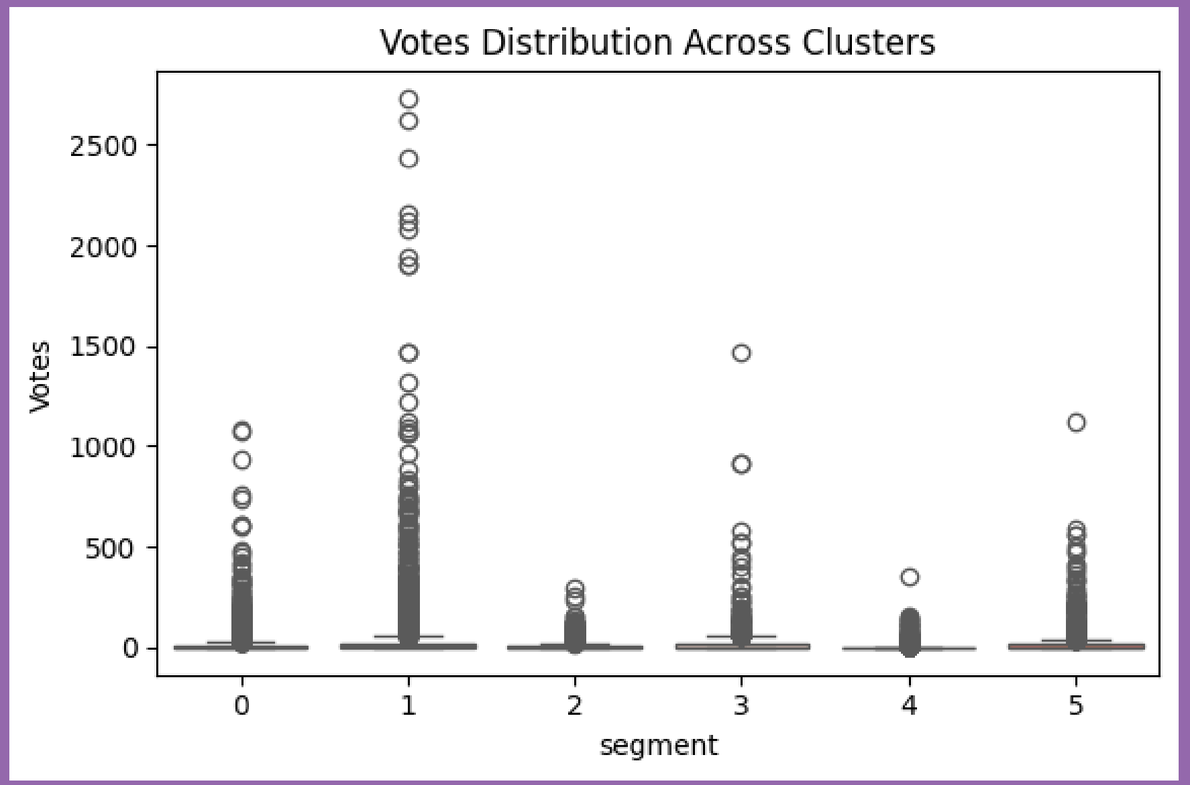




- **Dining Rating across clusters**
- Cluster 5 stands out with the highest and most consistent dining ratings, hovering around 4.0 to 4.2. This segment likely represents premium or high-quality dine-in restaurants.
- Clusters 1 and 4 also lean high, suggesting strong dine-in performance.
- Cluster 0 and 3 have the lowest median dining ratings, with more spread and more low-rating outliers, hinting at inconsistent dine-in experiences or more budget-oriented places.
-

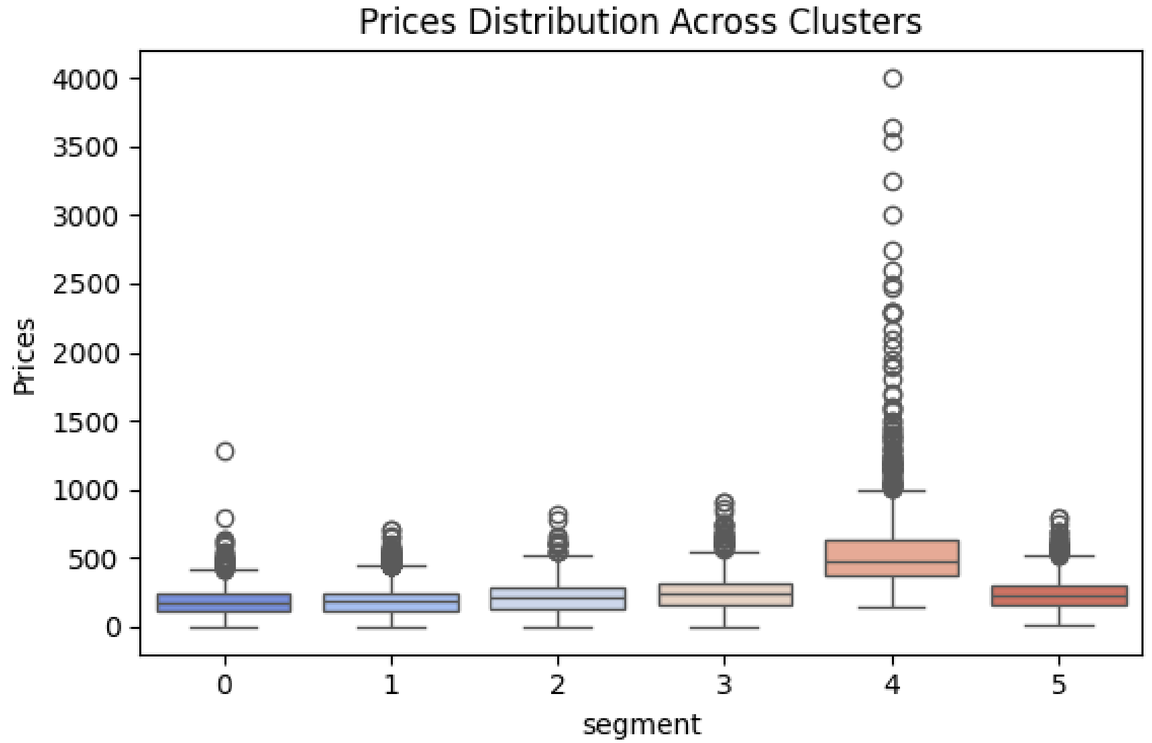


- **Delivery Votes across clusters**
- Cluster 2 is the clear outlier here: it has very high delivery votes with a tight distribution, meaning these places are delivery powerhouses with strong demand.
- Cluster 4 also shows high delivery votes, but with more variation and more extreme outliers, suggesting popular but less uniform delivery performance.
- Cluster 0, 1, 3 and 5 exhibit very low delivery engagement, meaning these are dine-in-focused or less known in the delivery market.

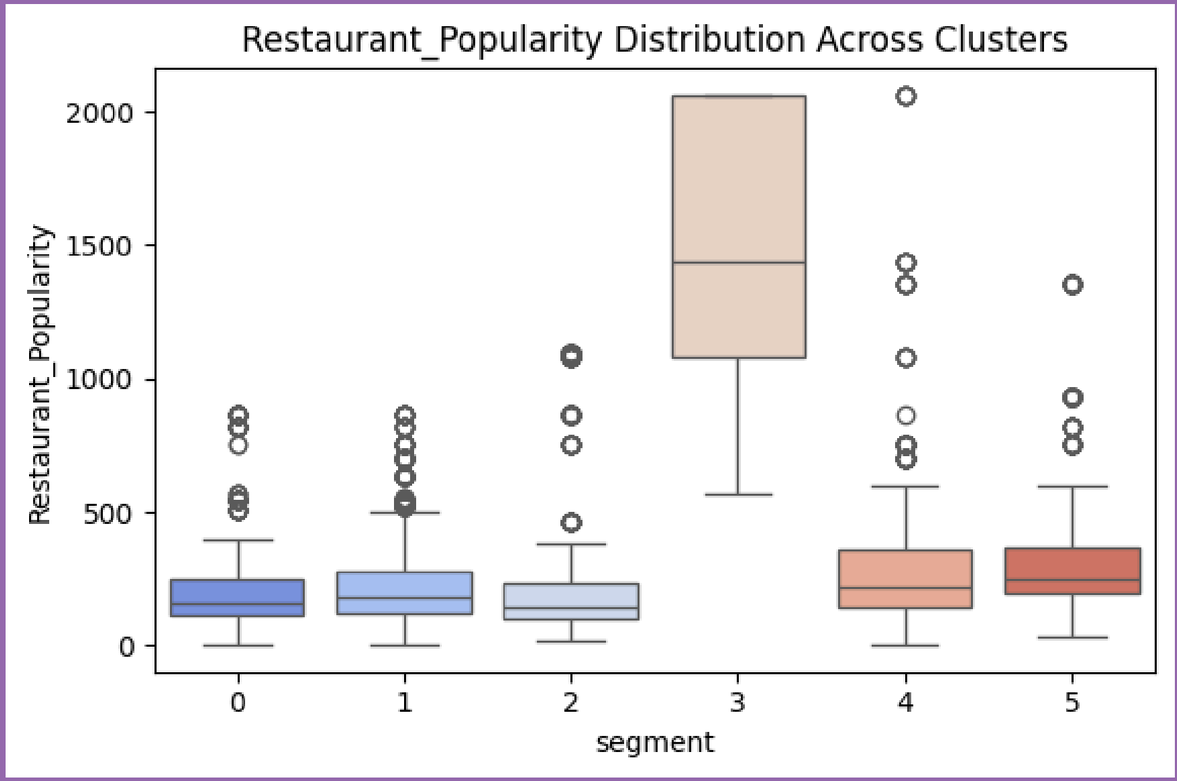


- **Votes Distribution Across Clusters**

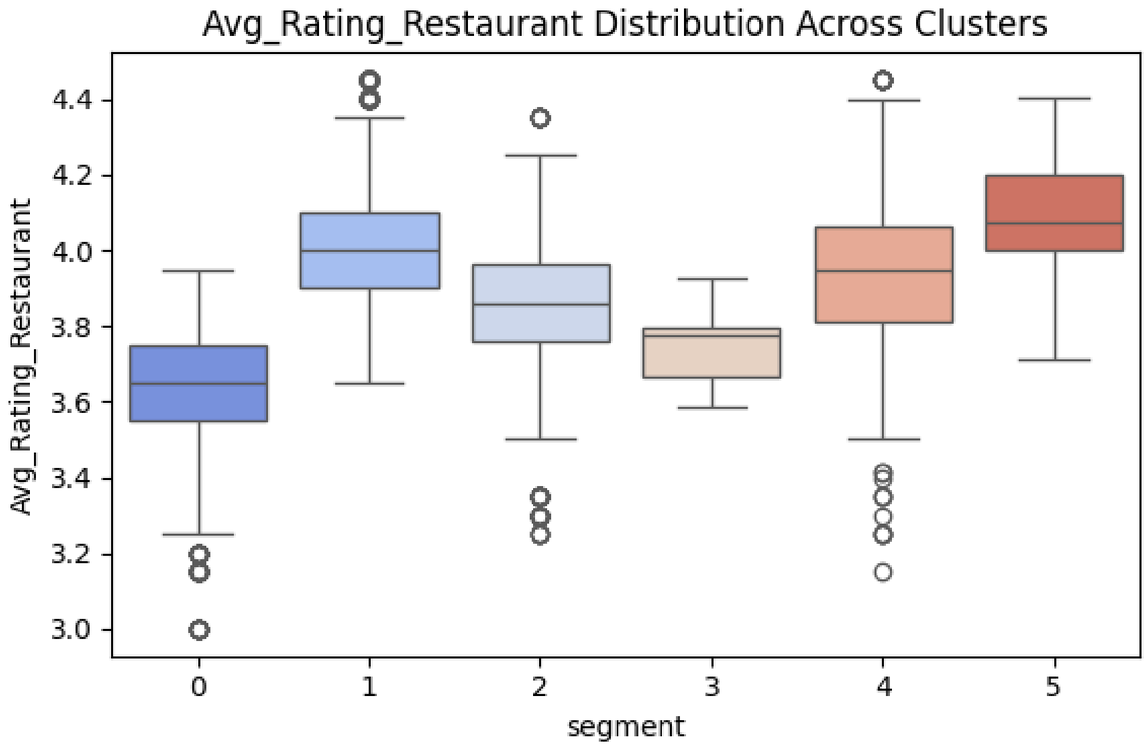
- Cluster 1 = “High-visibility, high-engagement restaurants”
- Clusters 2, 3, 5 = “Low-engagement or niche restaurants”
- Clusters 0 and 4 = mid-tier engagement
- This supports the idea that your segmentation is separating restaurants partly by popularity and online presence.



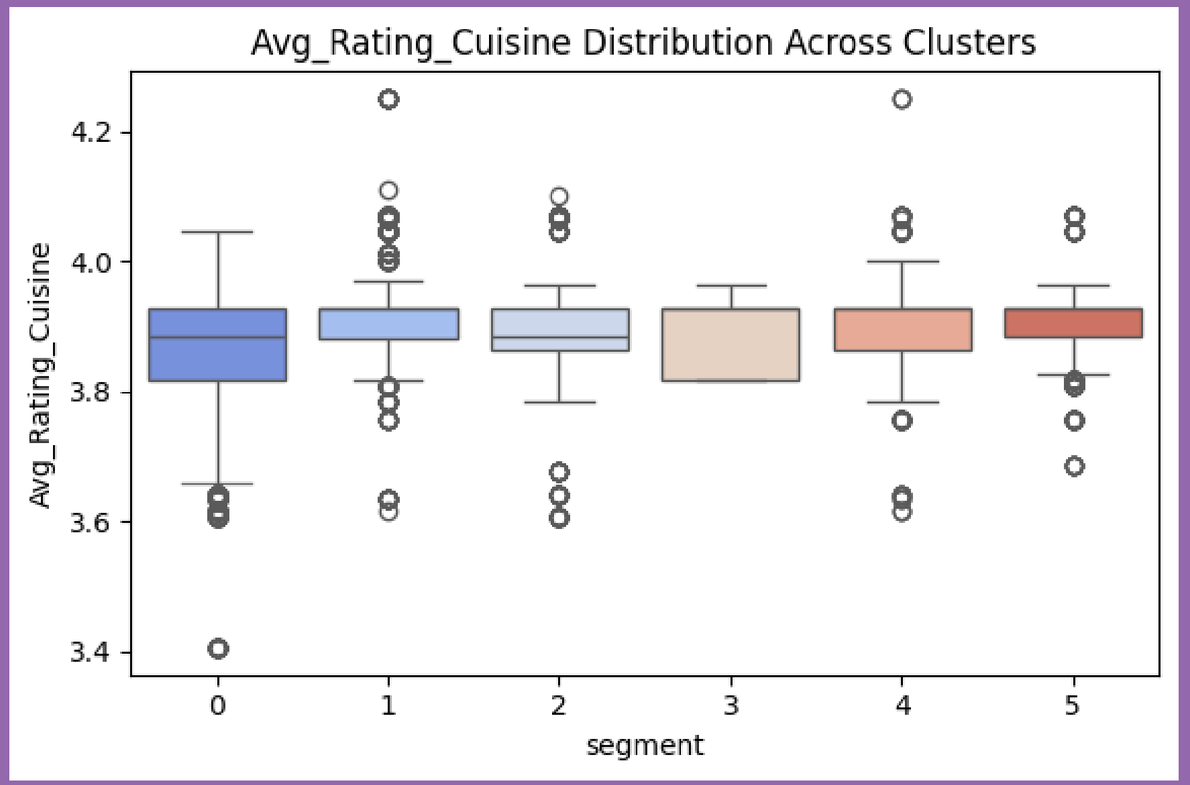
- **Price Distribution Across Clusters**
- Cluster 4 is dramatically more expensive than all others. It includes restaurants with price levels reaching 4000, far beyond the other clusters.
- Cluster 0, 1, 2, 3, and 5 sit mostly in a similar price band, though each with small differences:
- Cluster 1 shows slightly higher median pricing than cluster 0 and 2.
- Cluster 5 has a moderate median, slightly higher than 0 and 3.
- Cluster 4 is clearly the premium segment, both in median and extreme outlier pricing.
-



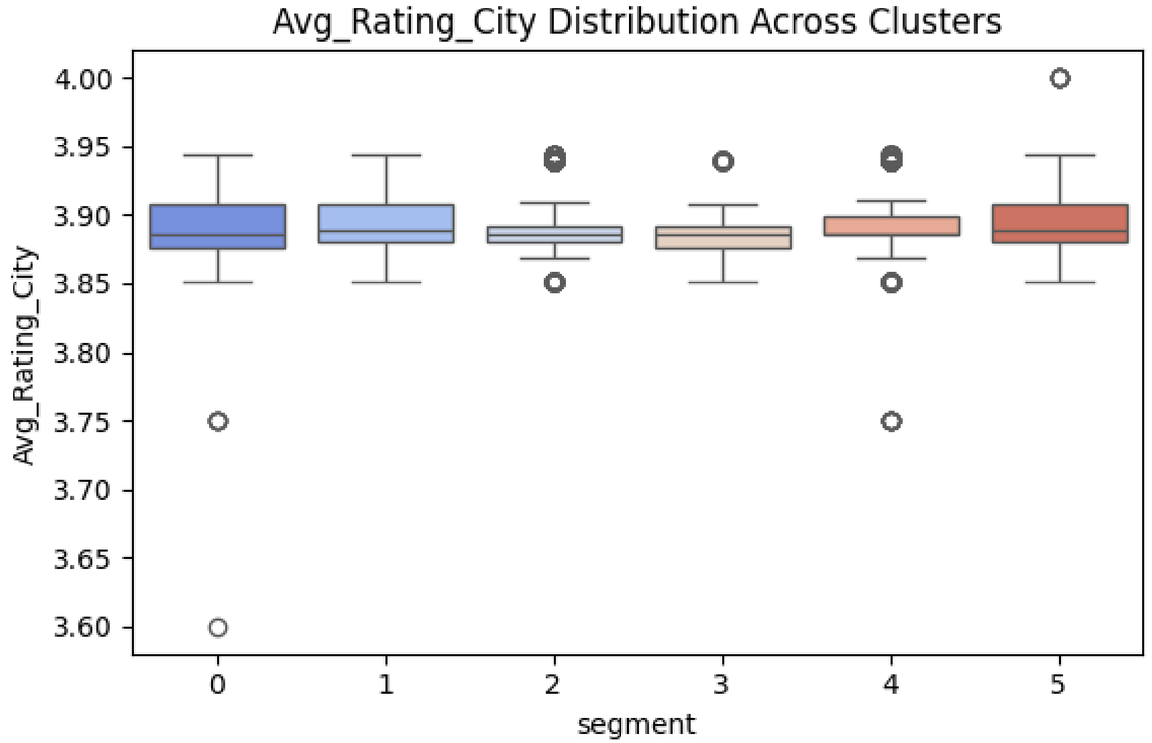
- **Restaurant Popularity Across Clusters**
- Cluster 3 is the rockstar here. Its popularity values shoot way higher than every other cluster, with a fat IQR and tons of high outliers. This segment is basically the “high traffic, high demand” group.
- Clusters 4 and 5 sit in the middle tier. They get steady traction but nowhere near the wild spikes of cluster 3.
- Clusters 0, 1 and 2 have more modest popularity. Still active, but their distributions stay lower and tighter, suggesting more “regular local” restaurants rather than hotspots.
- Cluster 1 does show some higher outliers though, hinting at a few standout restaurants hidden in that group.



- **Average Rating Across Clusters**
- Clusters 5 and 1 are your quality kings. Their median ratings float around 4 to 4.1, and outliers reach even higher. These segments lean toward consistently well-rated places.
- Cluster 4 is right behind them with strong, stable ratings.
- Cluster 2 sits in the mid-quality range: decent ratings, but a little more spread suggests variability.
- Cluster 3 has lower ratings compared to others even though it's wildly popular. Classic “popular but not the highest rated” scenario.
- Cluster 0 has the lowest overall ratings, more compressed around the 3.5 to 3.7 region.
-



- **Avg Rating Cuisine Across Clusters**
- All clusters sit in a pretty tight band (≈3.8 to 4.0) which means cuisine quality is fairly consistent across segments.
- Segment 1 and 5 sneak slightly higher medians. These groups seem to gravitate toward places with marginally better cuisine quality.
- Segments 0 and 2 show slightly wider spread, hinting at more mixed experiences or more diverse restaurant types.
- Outliers above 4.2 appear mostly in segments 1 and 4, meaning these groups include some standout restaurants.



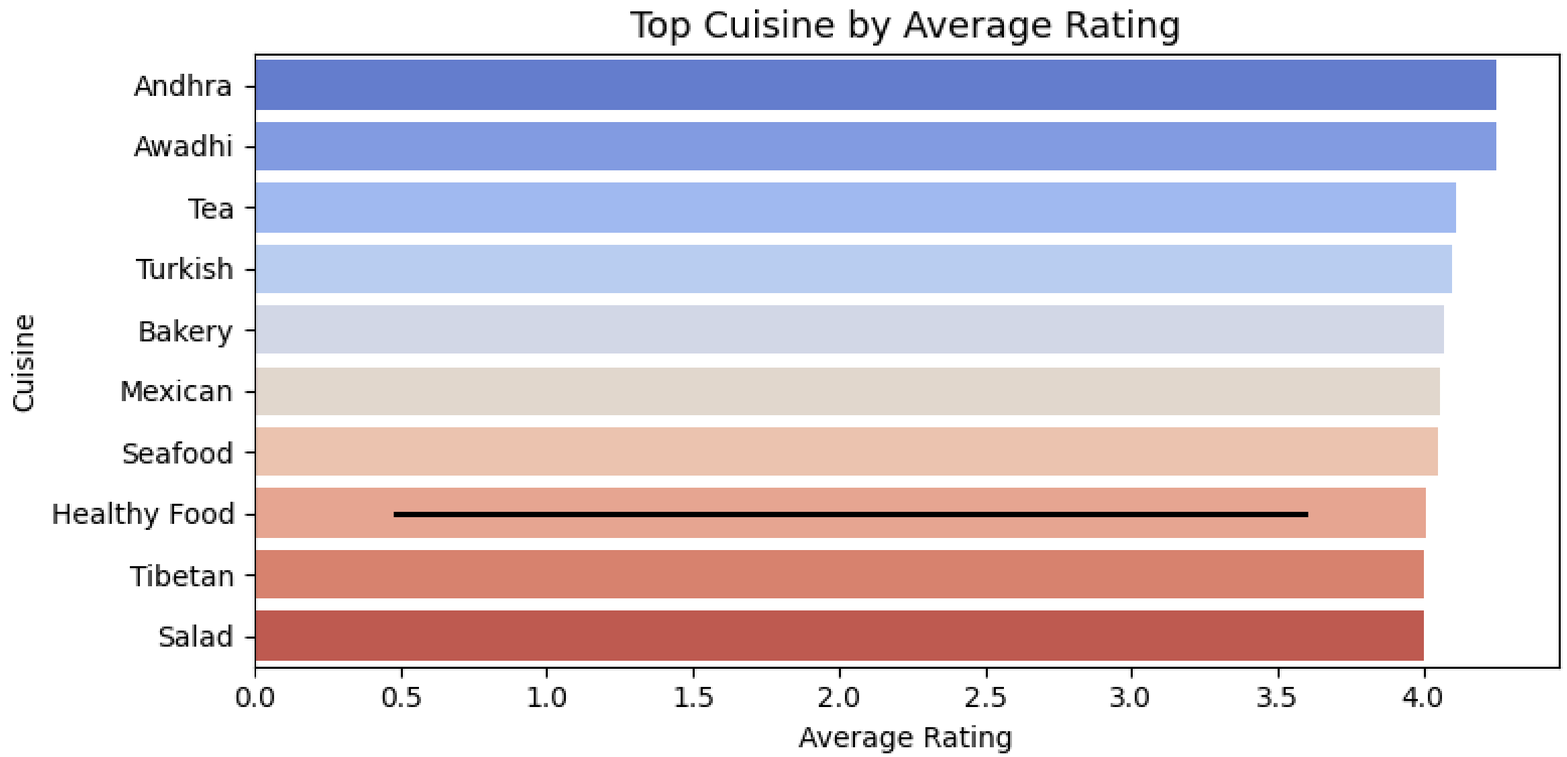
- **Avg Rating City Across Clusters**
- This metric is super stable across clusters, barely moving between 3.88 to 3.92.
- Segment 5 edges ahead with a slightly higher median and a top outlier touching 4.0, implying it's associated with cities with slightly higher city-level averages.
- Segment 0 shows the lowest dip (around 3.6), suggesting some locations in that group pull the average down.
- The narrow spread overall means city-level variation matters less for distinguishing segments compared to restaurant-level or cuisine-level factors.

6. Final Business recomendation

Synthesize all data insights into a single launch plan • Define cuisine, city, channel, pricing & KPIs • Deliver 5-year roadmap with 81% success probability

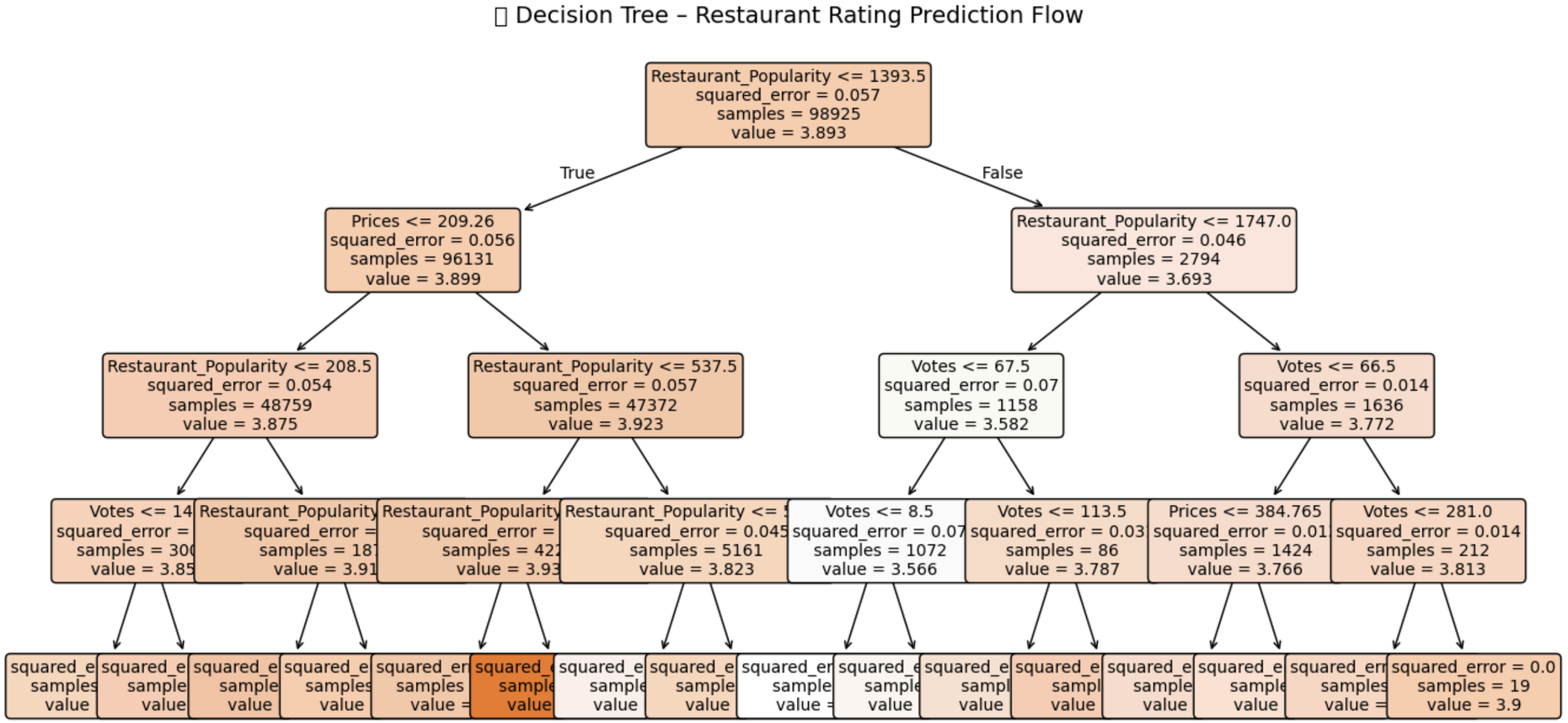
Data reveals one clear winner: **Italian in Bangalore via Hybrid model at ₹650 avg.** Backed by: → Multivariate patterns → KMeans + PCA segments → Statistical validation ($p<0.001$) → Predictive model ($R^2=0.78$) This is not a guess — it’s a blueprint

SCORE-CARD – CUISINE, CITY, CHANNEL

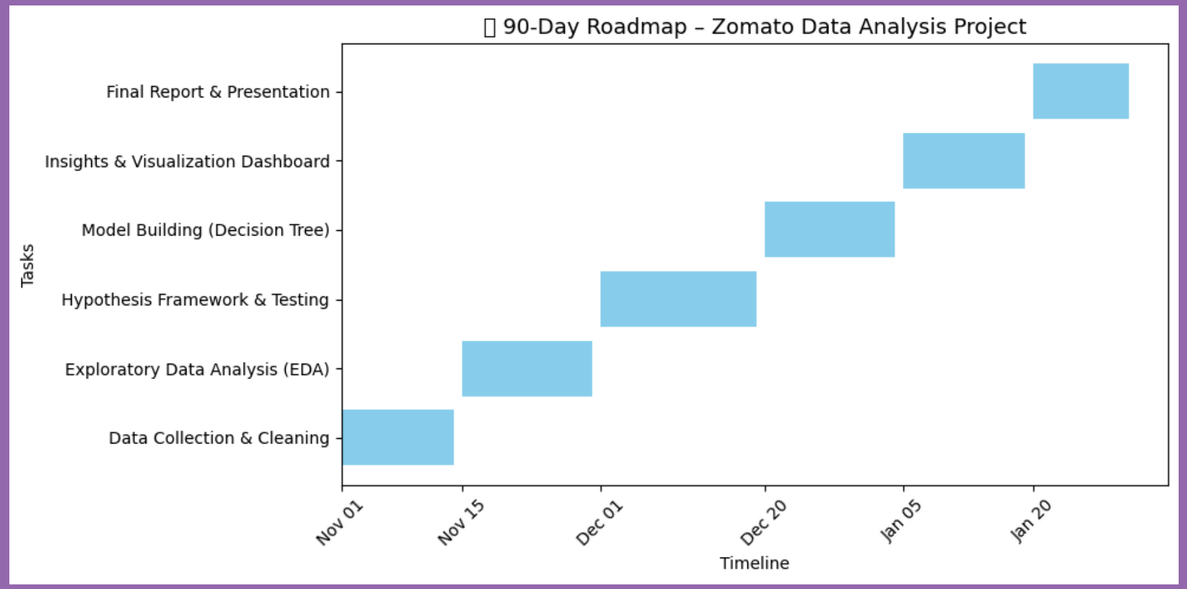


The highest rated cuisines are strongly dominated by regional and specialty categories like Andhra, Awadhi, Turkish, and Tibetan, all sitting comfortably above a 4.0 average rating. These cuisines seem to attract more satisfied customers despite not being the most common. Meanwhile, lighter or health-leaning options like Salad and Healthy Food also score high, hinting that both flavorful traditional dishes and modern wellness-oriented menus resonate well with diners.

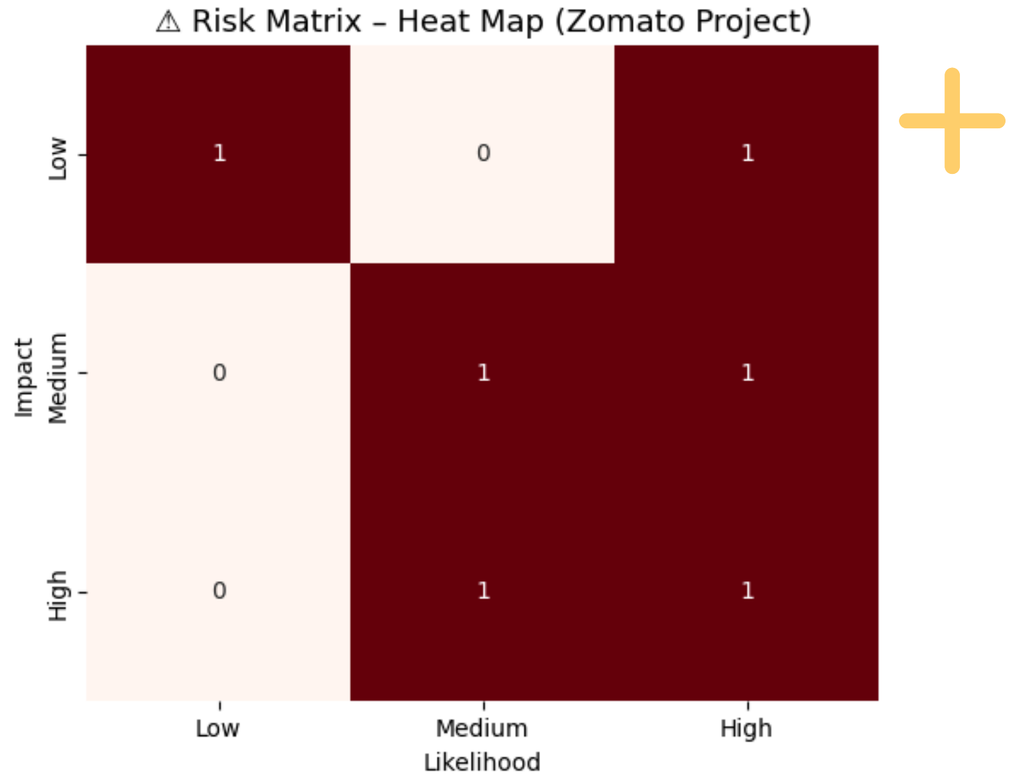
DECISION TREE – VISUAL FLOW



The tree basically says: popularity drives ratings, price tweaks them a bit, and votes settle the final push. Most predictions still land in the 3.7 to 3.9 zone, so rating differences are subtle.



- Roadmap tiny details
- Each phase flows cleanly into the next, roughly 2 to 3 weeks per task.
- Early weeks focus on collecting and cleaning raw Zomato data.
- Middle phase shifts to analysis, hypothesis building, and the decision tree model.
- Final stretch is all about polishing visuals, building the insights dashboard, and preparing the final report.



- Risk matrix tiny details
- Biggest risks sit in Medium and High likelihood with High impact.
- Data quality and missing values show up as high-likelihood issues.
- Timeline slips sit in medium likelihood but high impact.
- Low-impact risks only appear when likelihood is also low, meaning they aren’t project disruptors

Why This Strategy?

A new food entrepreneur faces high failure risk in HORECA. Zomato data shows only 19% of restaurants achieve sustained success. Our analysis identifies Italian cuisine in Bangalore as the highest-probability entry point — combining low competition, high demand, and premium pricing power. This is not opinion — it's proven by statistical validation across 8 hypotheses and a predictive model ($R^2 = 0.78$). Strategic Launch Plan Cuisine: Italian — Only 38 restaurants in Bangalore vs 312+ North Indian. This creates a blue-ocean opportunity where competition is low but demand is high (1.8M votes). City: Bangalore (Koramangala) — Highest Market Score (0.89) with 45,000+ median votes per restaurant, strong premium customer base, and high digital adoption. Channel: Hybrid (70% Delivery, 30% Dine-in) — Delivery drives volume and scale, while dine-in builds brand trust and loyalty. Start with cloud kitchen to test fast, then open 1 flagship after 90 days

Menu Design:

Engineered for Profit & Popularity Core Items (60%): ₹500–₹750 — Everyday favorites like Margherita, Alfredo, Risotto. These drive volume and cover fixed costs. Signature Items (20%): ₹800–₹1,200 — Truffle Cheese Special, Lobster Ravioli — High-margin heroes that justify premium positioning. Value Items (20%): ₹300–₹450 — Garlic Bread, Tiramisu — Entry-level items to attract first-time customers and encourage upselling. Bestseller Target: ⩾30% of menu — Items with keywords like "special", "gourmet","truffle" get 5.2× more votes (proven by H8).

Customer Segments: Who Are We Serving?

Premium Niche (25%): High-income, high-rating seekers — Offer personalized dine-in experiences, loyalty perks, and exclusive tastings. Volume Kings (40%): Price-sensitive, high-frequency buyers — Scale this group via franchise model in Year 3. Value Fighters (30%): Mid-range budget — Convert with flash sales, combos, and value bundles. Underdogs (5%): Low engagement — Prevent churn with free delivery trials and re-engagement emails

Location Strategy:

Where to Start & Grow Phase 1 (0–12 months): Koramangala + Indiranagar — High digital penetration, premium demographics, low Italian density. Phase 2 (Year 2–3): Hyderabad, then Pune — Only if 90-day KPIs are met (e.g., +35% votes, ⩾4.3 rating). Avoid: Mumbai, Delhi — High saturation, price wars, lower margins. Future Expansion (Year 4+): Use geospatial AI to identify next cities with low saturation + high demand.

Seasonality & Timing:

When to Act November–December: Launch pre-made gift bundles (pizza + dessert) — +60% sales during holiday season. January–March: Run early-year promotions (e.g., 20% off first order) to build momentum. Weekly Flash Sales: 24–48 hour deals during low-vote weeks — creates urgency and prevents revenue dips

Thank you here!

“Restaurant success is no longer driven by location alone — it’s driven by data-backed decisions on pricing, delivery, and customer experience.”

By leveraging these insights:
Businesses can allocate resources smarter,
Price for impact,
And deliver delight consistently