CoffeeKing

DATA BASED INSIGHTS AND RECOMMENDATIONS

Project

Using Yelp Data to Guide CoffeeKing’s Market Entry Strategy

Client

CoffeeKing Leadership Team

Goal

- Identify where to open
- How to operate
- Define what experience to offer
- Understand how to drive engagement & ratings

Initial Hypotheses

- H1: Location strongly impacts engagement and satisfaction
- H2: Higher engagement leads to higher ratings and visibility
- H3: Longer or more consistent hours increase customer interaction

Initial Key Questions

- Which cities and states show the strongest coffee performance?
- Does higher engagement (reviews, photos, tips, check-ins) lead to better ratings?
- What business attributes and service experiences differentiate top coffee shops?

Yelp Dataset

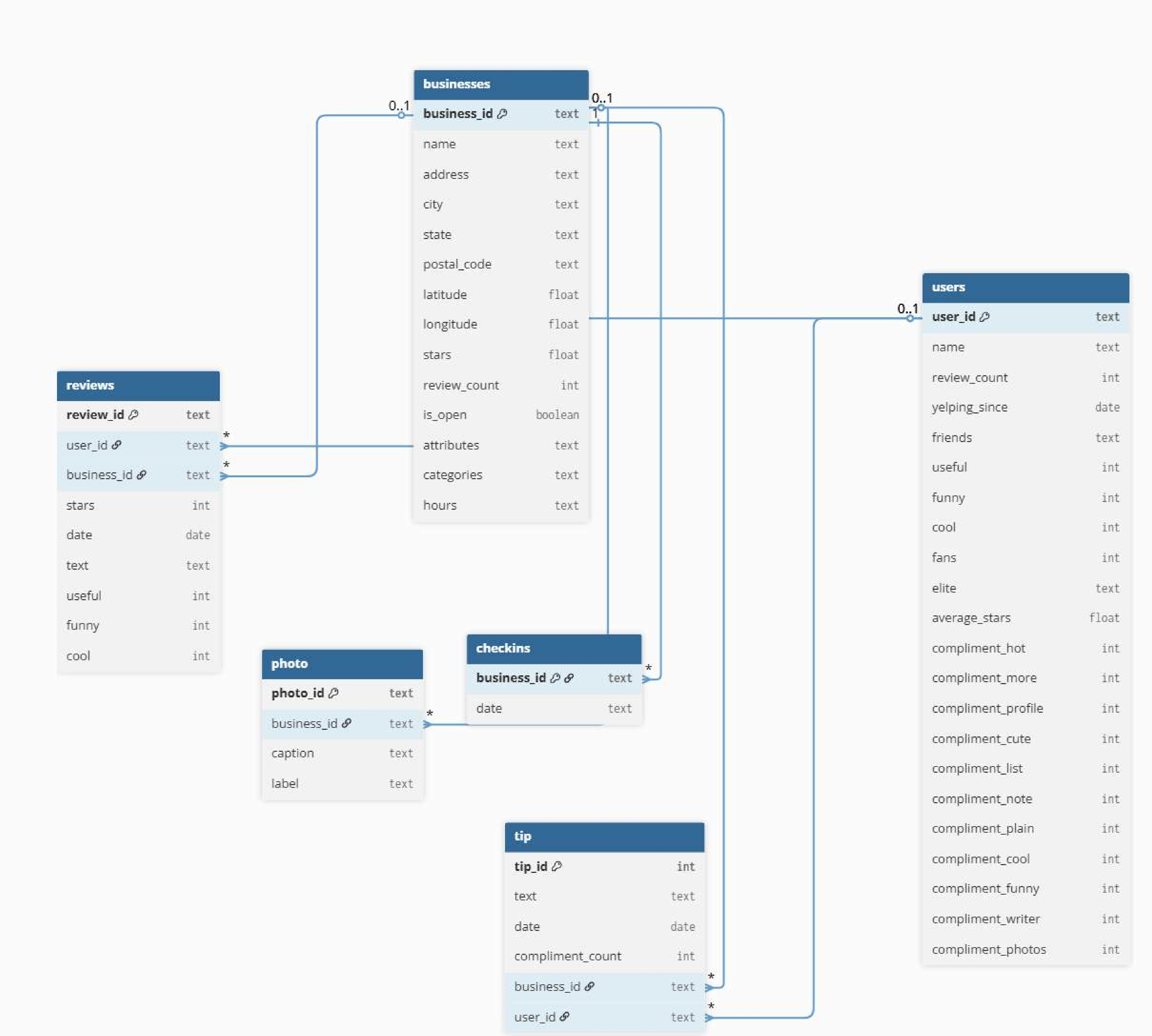
- Businesses (location, attributes, hours)
- Reviews (ratings, text)
- Users (activity, influence)
- Check-ins, tips, photos

Why Yelp?

- Real customer voice
- Location + engagement + sentiment in one ecosystem

Analytical Approach

- Descriptive statistics
- Correlation & regression
- TF-IDF text analysis
- Custom engagement metrics (SESS, OFI)



Coffee Business by City and State

What We Analyzed

- Business density
- Avg star rating
- Median review count
- Rating variability

These Outputs show

- Total coffee businesses by cities and states
- Average ratings, rating variability through standard deviation
- Median reviews

Key Insight

- Large cities → more competition, lower consistency
- Mid-sized cities → higher ratings & stability

Recommendation for CoffeeKing

- Target mid-sized, less saturated cities for first locations

```
coffee_state_stats = (
    coffee_df
    .groupby("state")
    .agg(
        avg_rating=("stars", "mean"),
        rating_std=("stars", "std"),
        median_reviews=("review_count", "median"),
        business_count=("business_id", "count")
    )
    .sort_values("business_count", ascending=False)
)

coffee_state_stats.head(5)
```

✓ 0.1s

state	avg_rating	rating_std	median_reviews	business_count
PA	3.459225	0.990807	22.0	1729
FL	3.537287	1.074643	27.0	1113
TN	3.557460	1.050361	32.0	496
LA	3.665612	0.967494	37.5	474
IN	3.543710	1.038113	30.0	469

```
n"),
"),
ount", "median"),
_id", "count")

, ascending=False)
```

city	avg_rating	rating_std	median_reviews	business_count
Philadelphia	3.547433	0.961106	31.0	896
Tampa	3.591940	1.020490	28.0	397
Edmonton	3.572650	0.855900	12.0	351
New Orleans	3.872642	0.805329	47.0	318
Tucson	3.515723	0.983975	32.0	318

Coffee Business by City and State

These Outputs show total coffee businesses by cities and states

- i.e. PA has 1729 coffee businesses
- i.e. Tampa has 397 coffee shops

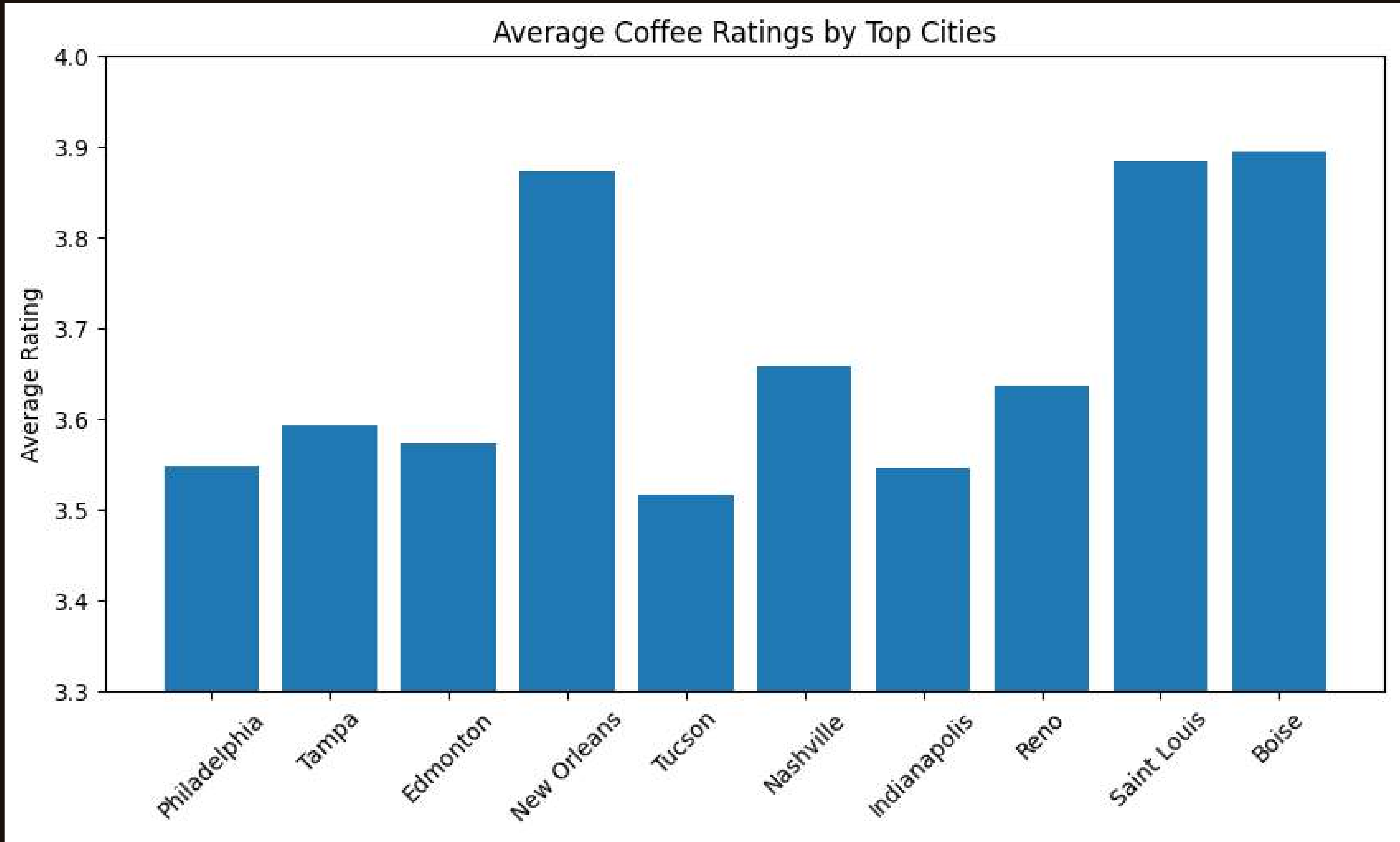
```
coffee_df["state"].value_counts().head(10)
✓ 0.0s

state
PA    1729
FL    1113
TN     496
LA     474
IN     469
MO     427
NJ     410
AB     380
AZ     352
NV     269
Name: count,
```

```
coffee_df["city"].value_counts().head(10)
✓ 0.6s

city
Philadelphia    896
Tampa           397
Edmonton        351
New Orleans     318
Tucson          318
Nashville       313
Indianapolis    302
Reno            216
Saint Louis     142
Boise           138
Name: count, dtype: int64
```

Hypothesis 1 Results (Location Matters)

City-Level Highlights

- Philadelphia, Tampa: high volume, moderate ratings, high variability
- Boise, St. Louis, New Orleans:
  - Higher avg ratings (~3.87–3.89)
  - Lower variability
  - Strong engagement despite smaller size



Customer Engagement vs Ratings

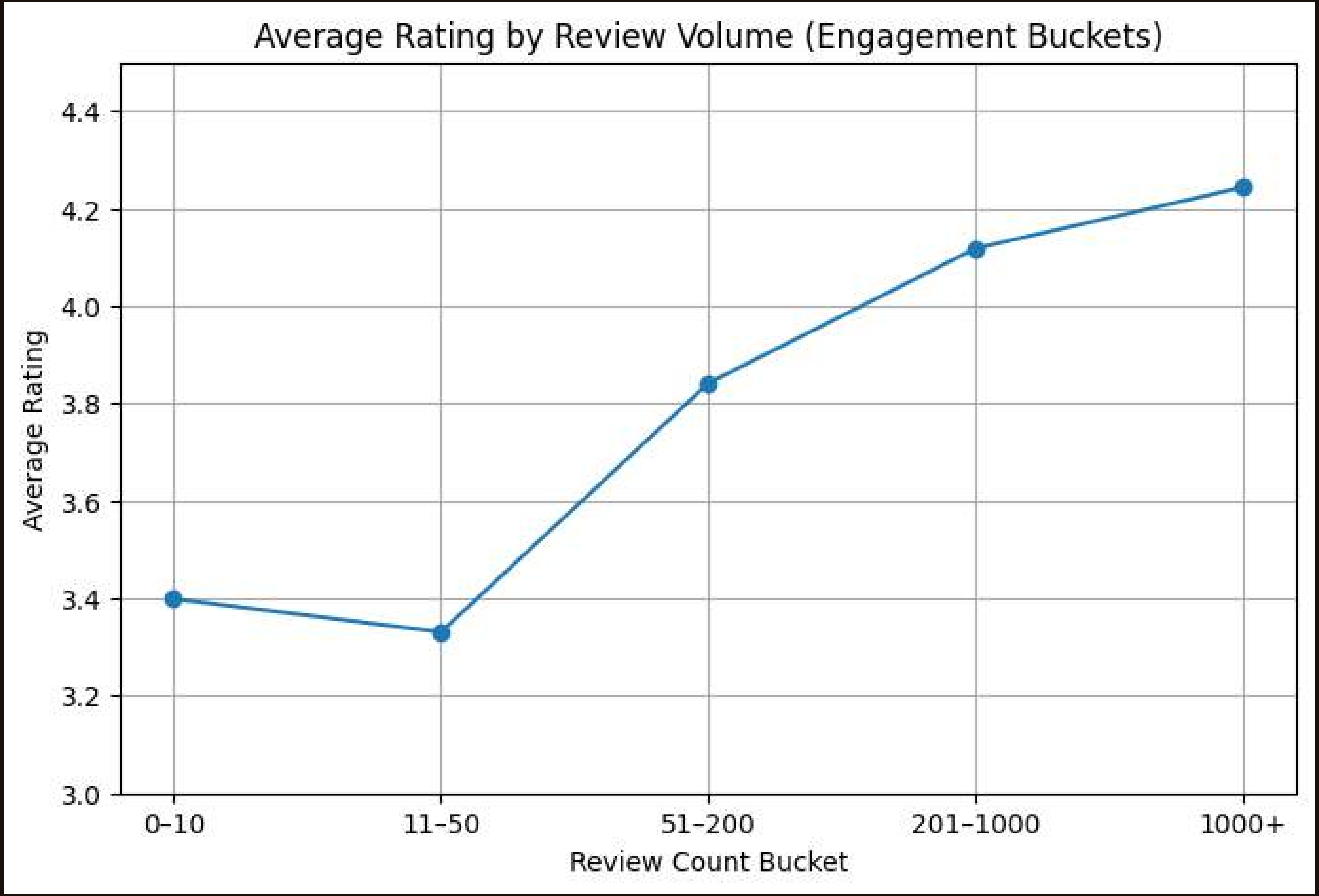
Engagement Signals Used

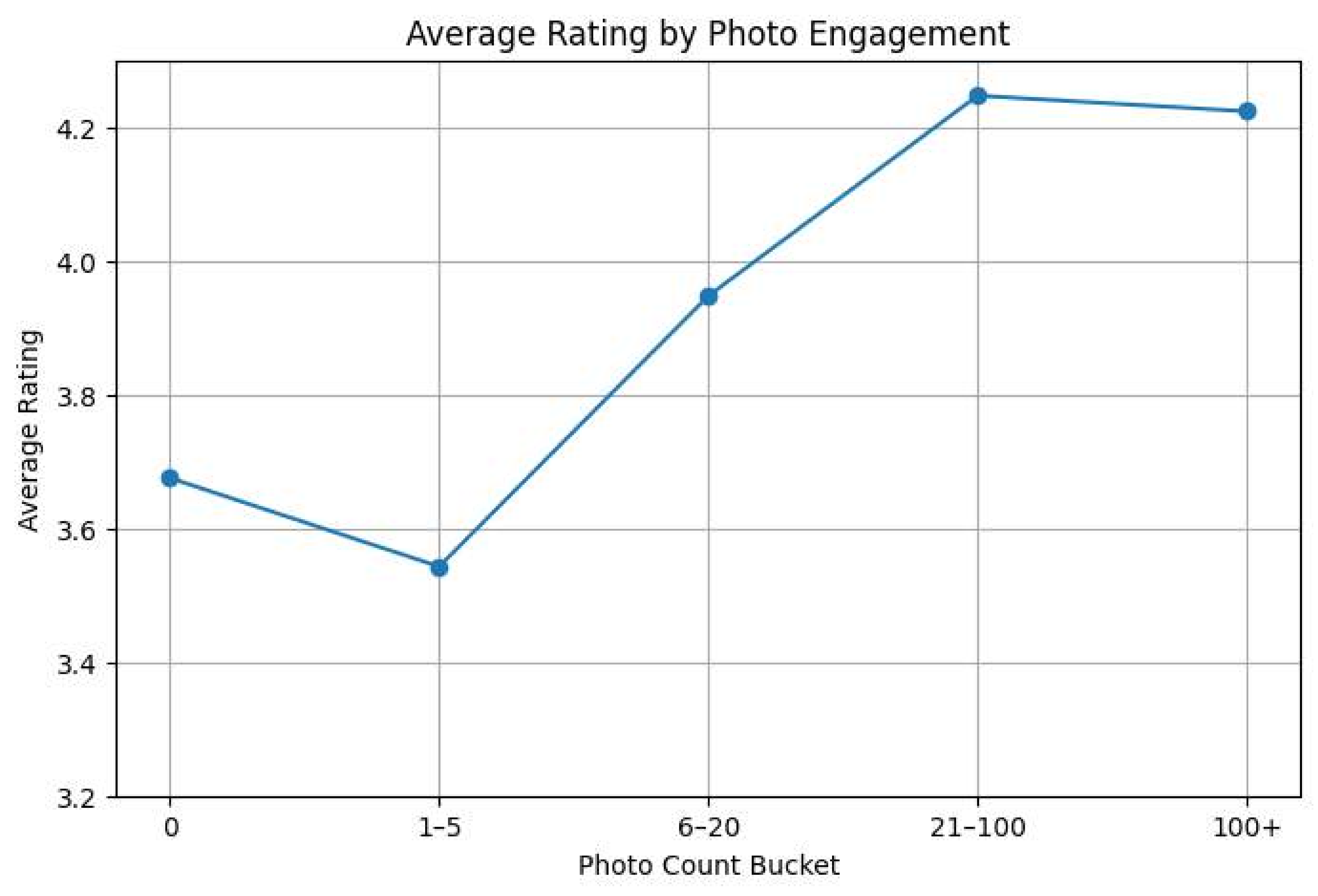
- Reviews
- Tips
- Photos
- Check-ins

```
coffee_engagement_summary = (
    coffee_review_stats
    .join(coffee_checkin_counts, how="left")
    .join(coffee_tip_counts, how="left")
    .join(coffee_photo_counts, how="left")
    .fillna(0)
)

coffee_engagement_summary.describe()
```

✓ 1.7s

	review_count	avg_rating	checkin_count	tip_count	photo_count
count	6728.000000	6728.000000	6728.000000	6728.000000	6728.000000
mean	66.095571	3.517584	0.990636	11.178062	3.517390
std	151.298295	0.999095	0.096320	39.047883	9.861686
min	5.000000	1.000000	0.000000	0.000000	0.000000
25%	11.000000	2.809066	1.000000	2.000000	0.000000
50%	26.000000	3.765069	1.000000	4.000000	1.000000
75%	61.000000	4.333333	1.000000	11.000000	4.000000
max	5778.000000	5.000000	1.000000	2571.000000	528.000000


Average Rating by Review Volume (Engagement Buckets)
Average Rating
4.4
4.2
4.0
3.8
3.6
3.4
3.2
3.0
0–10
11–50
51–200
201–1000
1000+
Review Count Bucket


Average Rating by Photo Engagement
Average Rating
4.2
4.0
3.8
3.6
3.4
3.2
0
1–5
6–20
21–100
100+
Photo Count Bucket

Hypothesis 2 Results (Engagement Drives Ratings)

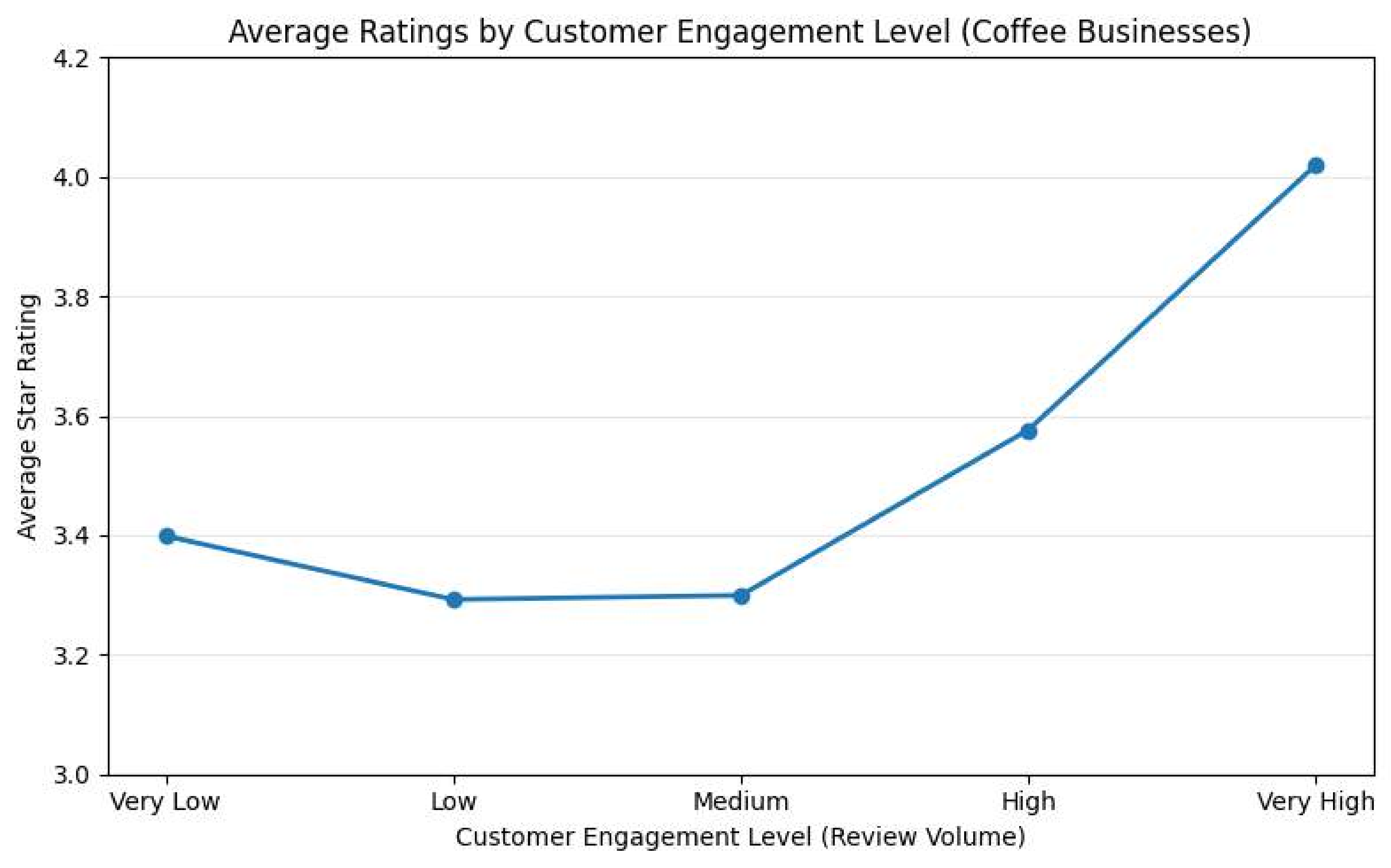
Key Pattern

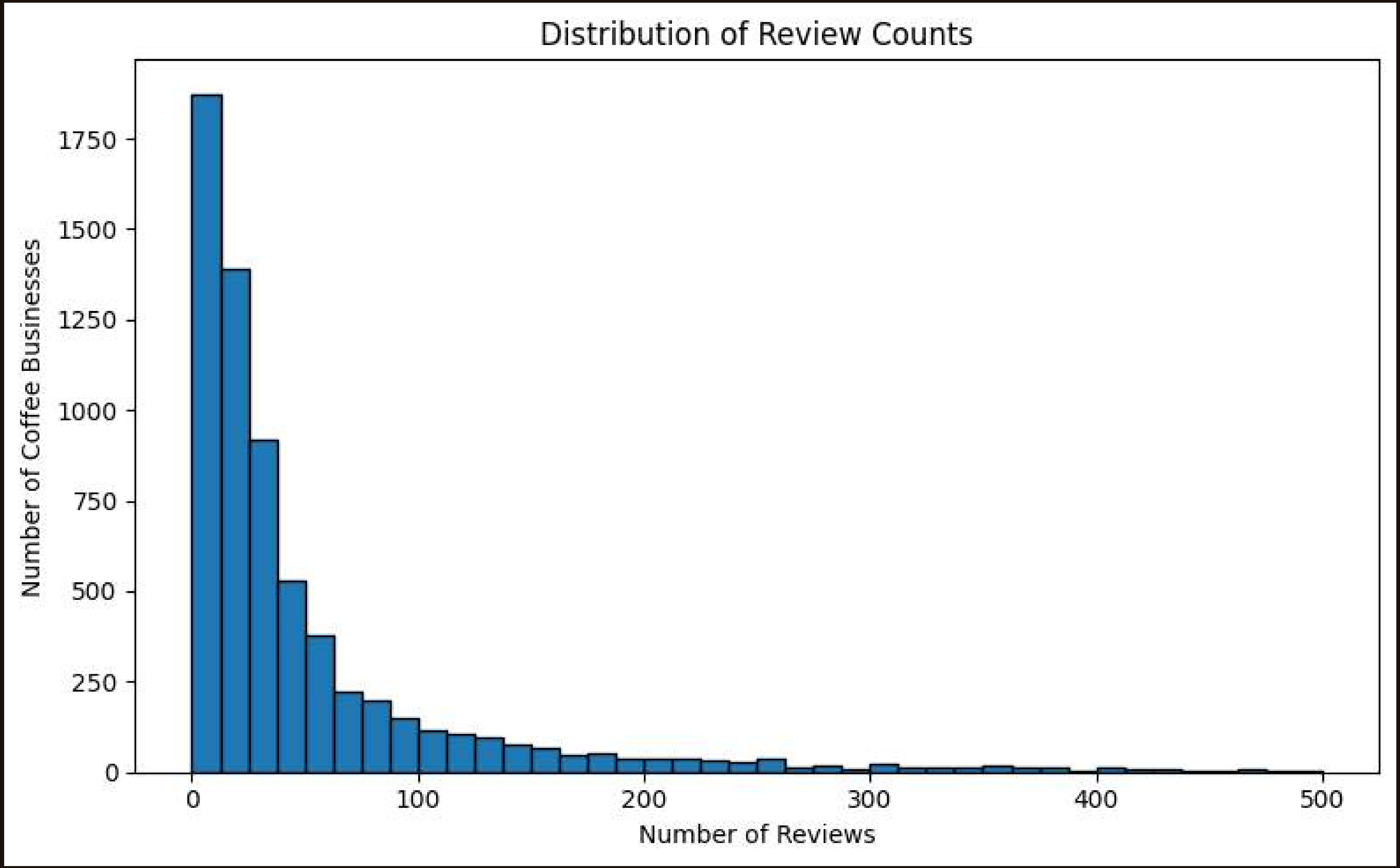
- High engagement = significantly higher ratings
- Medium engagement = exposure effect (more mixed reviews)

Conclusion

Hypothesis supported

Visibility + interaction reinforce reputation





Distribution of Review Counts
Number of Coffee Businesses
1750
1500
1250
1000
750
500
250
0
0
100
200
300
400
500
Number of Reviews

Relationship and Correlation Analysis

Holding tips constant, each additional photo associated with a coffee business is linked to ~3.2 additional reviews on average.

Holding photos constant, each additional tip is linked to ~1.9 additional reviews on average.

- Photos are a strong engagement amplifier
- Visual content encourages more people to interact and leave reviews
- Photos likely increase trust and visibility
- Tips reflect user effort, but are less powerful than photos
- Tips still contribute meaningfully to visibility

```
coef_df = pd.DataFrame({
    "feature": X.columns,
    "coefficient": model.coef_
})

coef_df
```

✓ 0.0s

	feature	coefficient
0	photo_count	3.207856
1	tip_count	1.897749

Model Score (R-squared)

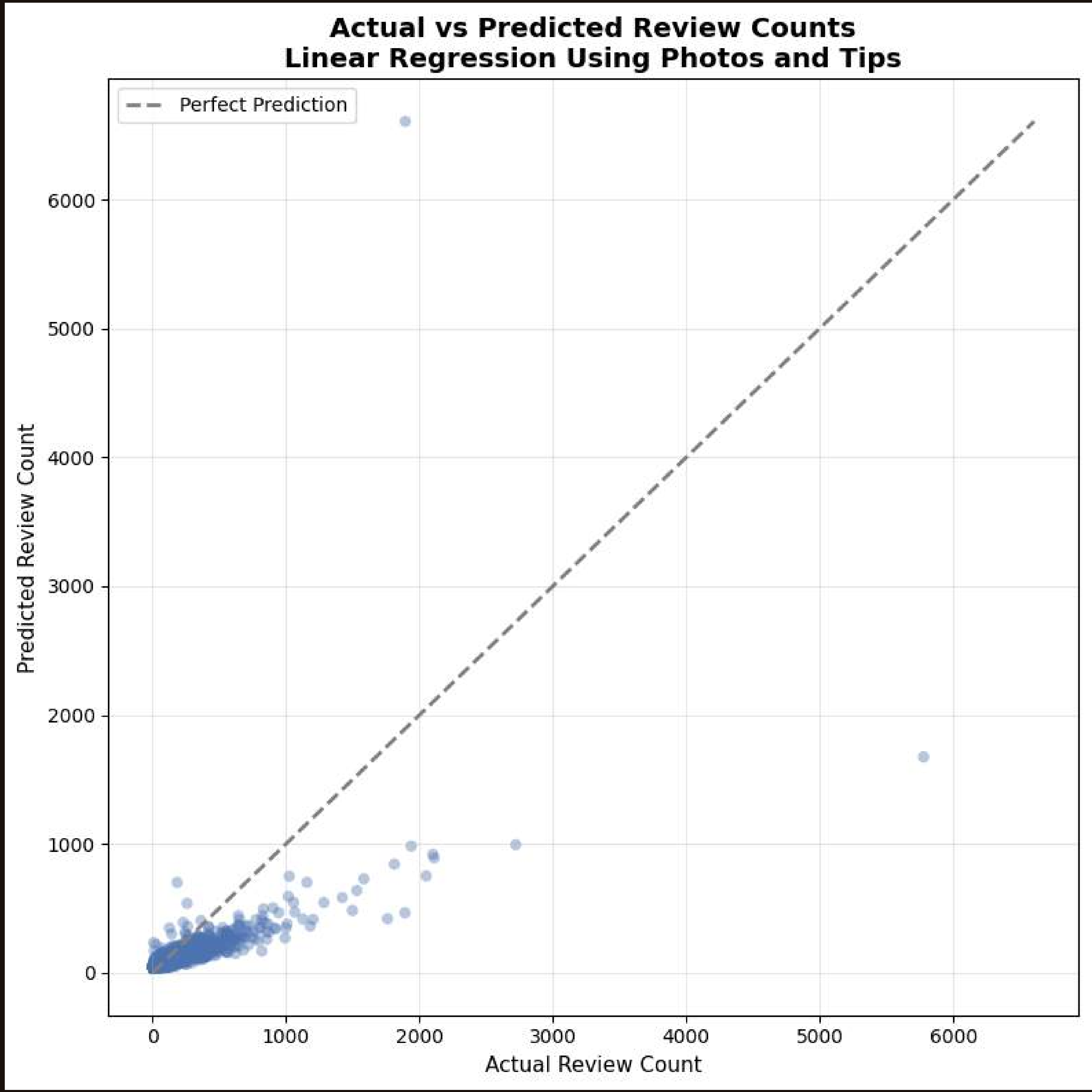
- About 43% of the variation in review counts across coffee businesses can be explained by photo count and tip count.
- In social / behavioral data (Yelp, reviews, human behavior):
- R^2 values of 0.2–0.5 are common
- Human decisions are noisy and influenced by many hidden factors
- A value of 0.43 is actually:
- Moderate to strong
- Strong enough to support insight

What the remaining ~57% represents

The unexplained variance likely comes from:

- Business age
- Location (foot traffic)
- Brand recognition
- Pricing
- Service quality
- Yelp algorithm effects
- Random user behavior

```
model.score(X, y)
✓ 0.0s
0.4341749091942011
```


Actual vs Predicted Review Counts
Linear Regression Using Photos and Tips
Perfect Prediction
Predicted Review Count
Actual Review Count
0
1000
2000
3000
4000
5000
6000

Operating hours and Engagement/Ratings

Metrics Created

- Weekly hours
- Hours consistency (variance)
- Review_count
- Check-ins_count
- stars

pearson correlation among these metrics

Correlation Findings

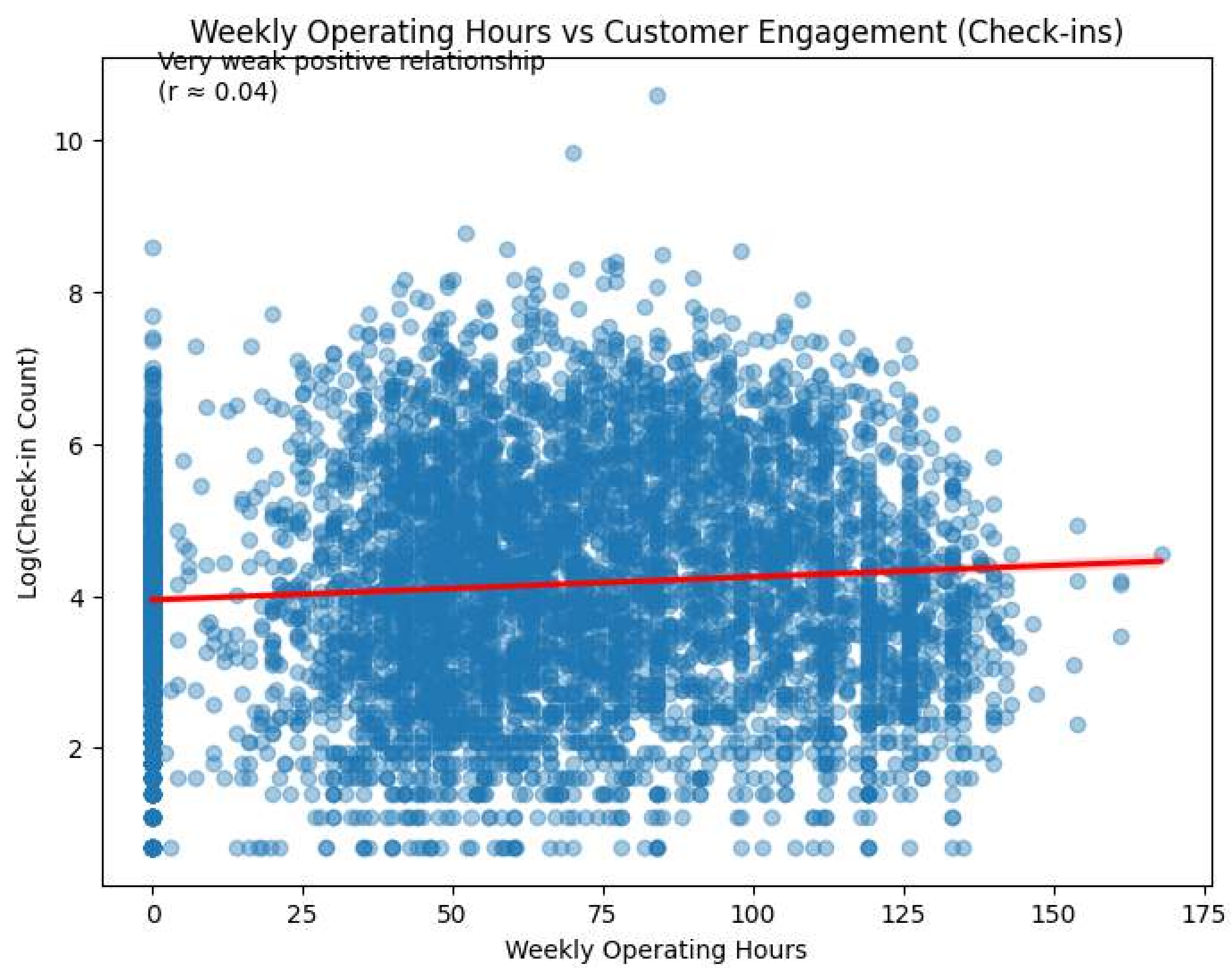
- Hours ↔ Check-ins: **very weak**
- Reviews ↔ Check-ins: **strong (0.67)**

Conclusion

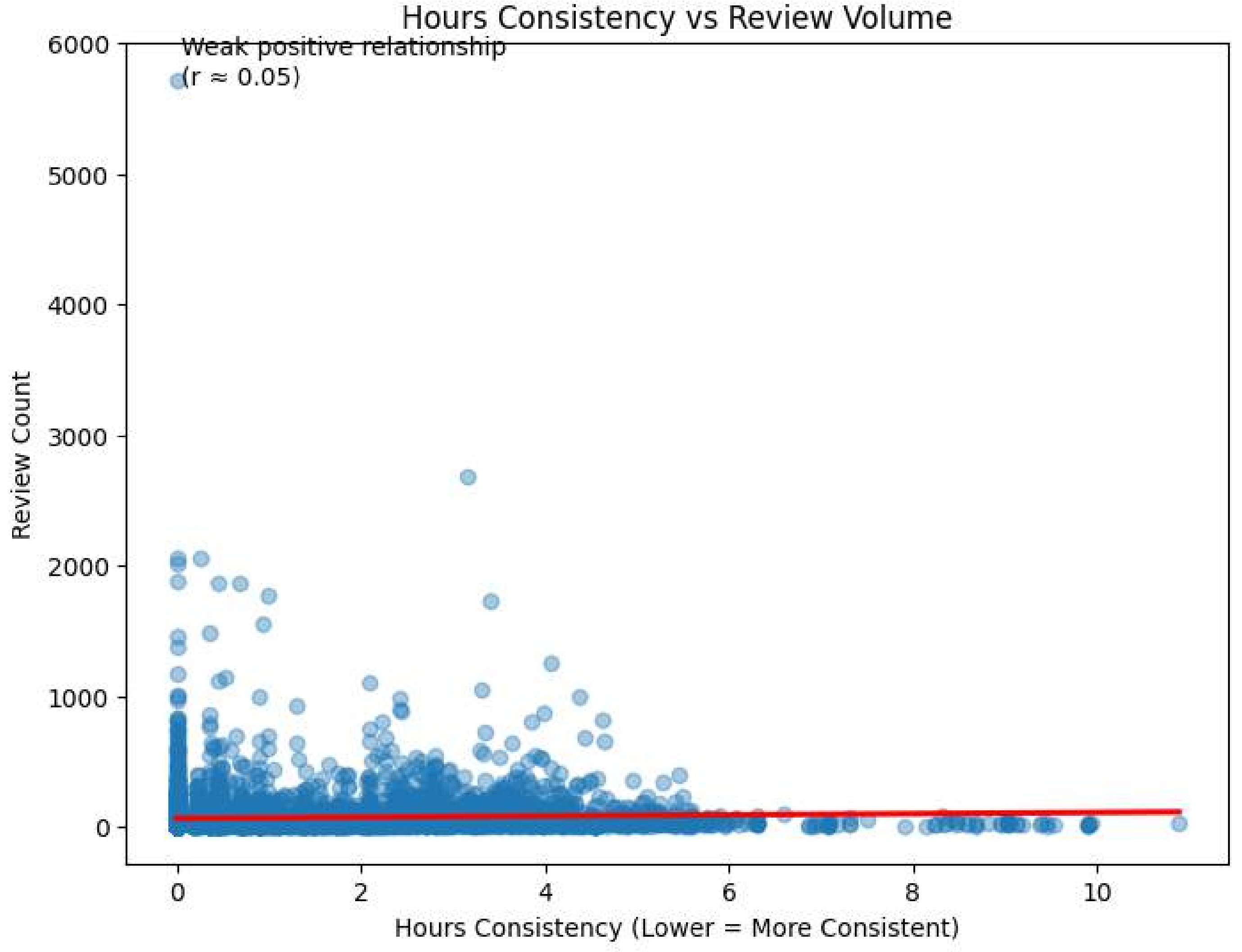
Longer hours alone do not drive traffic
Experience quality does

```
corr_cols = [
    "weekly_hours",
    "hours_consistency",
    "review_count",
    "stars",
    "checkins_count"
]

corr_matrix = coffee_hours_engagement[corr_cols].corr(method="pearson")
corr_matrix
```

	weekly_hours	hours_consistency	review_count	stars	checkins_count
weekly_hours	1.000000	0.081411	0.006689	-0.167771	0.042186
hours_consistency	0.081411	1.000000	0.049725	0.155604	0.047461
review_count	0.006689	0.049725	1.000000	0.164618	0.667739
stars	-0.167771	0.155604	0.164618	1.000000	0.072818
checkins_count	0.042186	0.047461	0.667739	0.072818	1.000000


Weekly Operating Hours vs Customer Engagement (Check-ins)
Very weak positive relationship
(r ≈ 0.04)
Log(Check-in Count)
Weekly Operating Hours


Hours Consistency vs Average Star Rating
Moderate positive relationship
(r ≈ 0.16)
Average Rating
Hours Consistency


Hours Consistency vs Review Volume
Weak positive relationship
(r ≈ 0.05)
Review Count
Hours Consistency (Lower = More Consistent)

Who are Coffee Customers?

Coffee Users vs General Yelp Users

- Write 3× more reviews
- Have more fans
- Slightly higher average ratings

Implication

Coffee customers are highly vocal & influential

CoffeeKing Action

Prioritize experiences worth talking about

```
users_df["is_coffee_user"] = users_df["user_id"].isin(coffee_user_ids)

user_stats = users_df.groupby("is_coffee_user").agg(
    avg_reviews=("review_count", "mean"),
    avg_rating=("average_stars", "mean"),
    avg_fans=("fans", "mean"),
    user_count=("user_id", "count")
)

user_stats
```

✓ 1.7s

is_coffee_user	avg_reviews	avg_rating	avg_fans	user_count
False	18.291572	3.605876	0.985625	1728100
True	57.337113	3.794250	4.659342	259797

Performace Categorization

Coffee business are categories in three perfromace group based on ratings and review volume

Top if review count >= 200 and stars >= 4
total top coffee businesses = 387

Low if review count < 100 and stars <= 3
total low coffee businesses = 2256

Businesses not in these two categories are in the mid range category

There is significant difference of average ratings and median reviews among three categories

```
coffee_df.assign(
    performance_group = coffee_df["business_id"].apply(
        lambda x: "Top" if x in top_coffee_ids
        else "Low" if x in low_coffee_ids
        else "Mid"
    )
).groupby("performance_group").agg(
    avg_rating=("stars", "mean"),
    median_reviews=("review_count", "median"),
    business_count=("business_id", "count")
)
```

0.2s

	avg_rating	median_reviews	business_count
performance_group			
Low	2.306959	17.0	2256
Mid	4.121420	29.0	4085
Top	4.264858	318.0	387

Coffee Business Attributes

Most common attributes among all the coffee businesses ragardless of performance

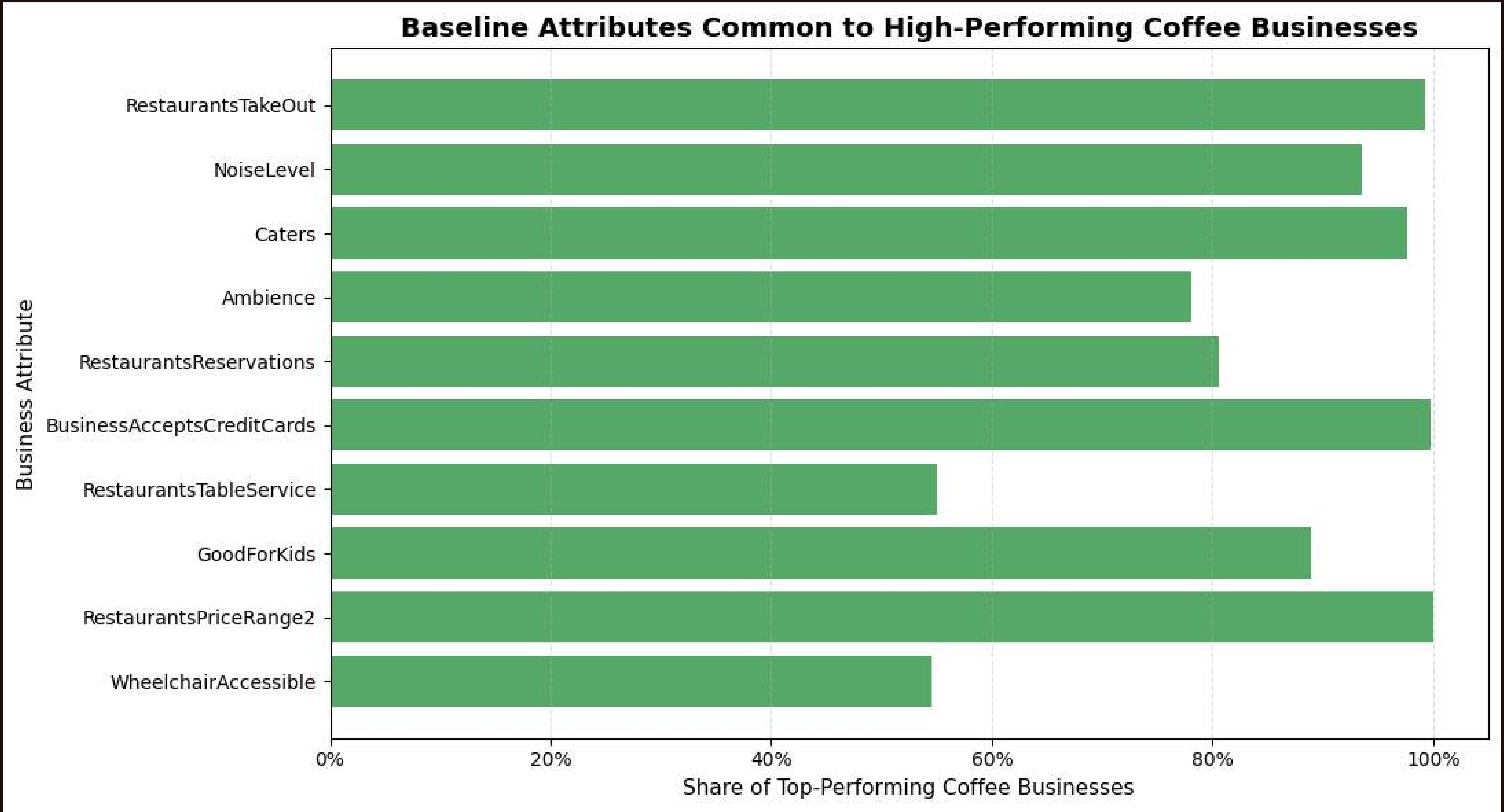
These are table stakes, not differentiators

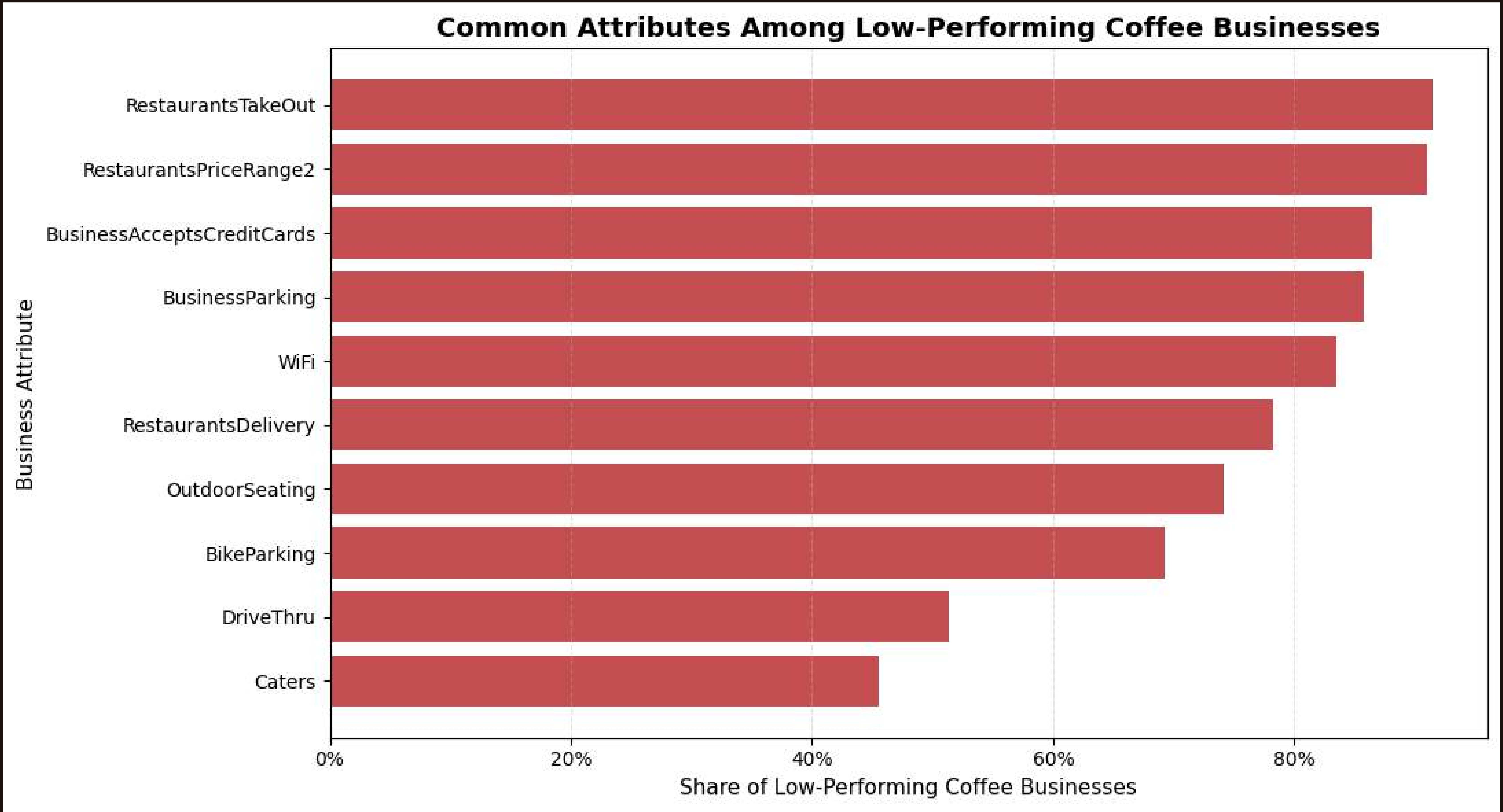
Following slides show attributes common among top performing coffee businesses, low performing coffee businesses, and the ones that is most differentiating among two performance groups

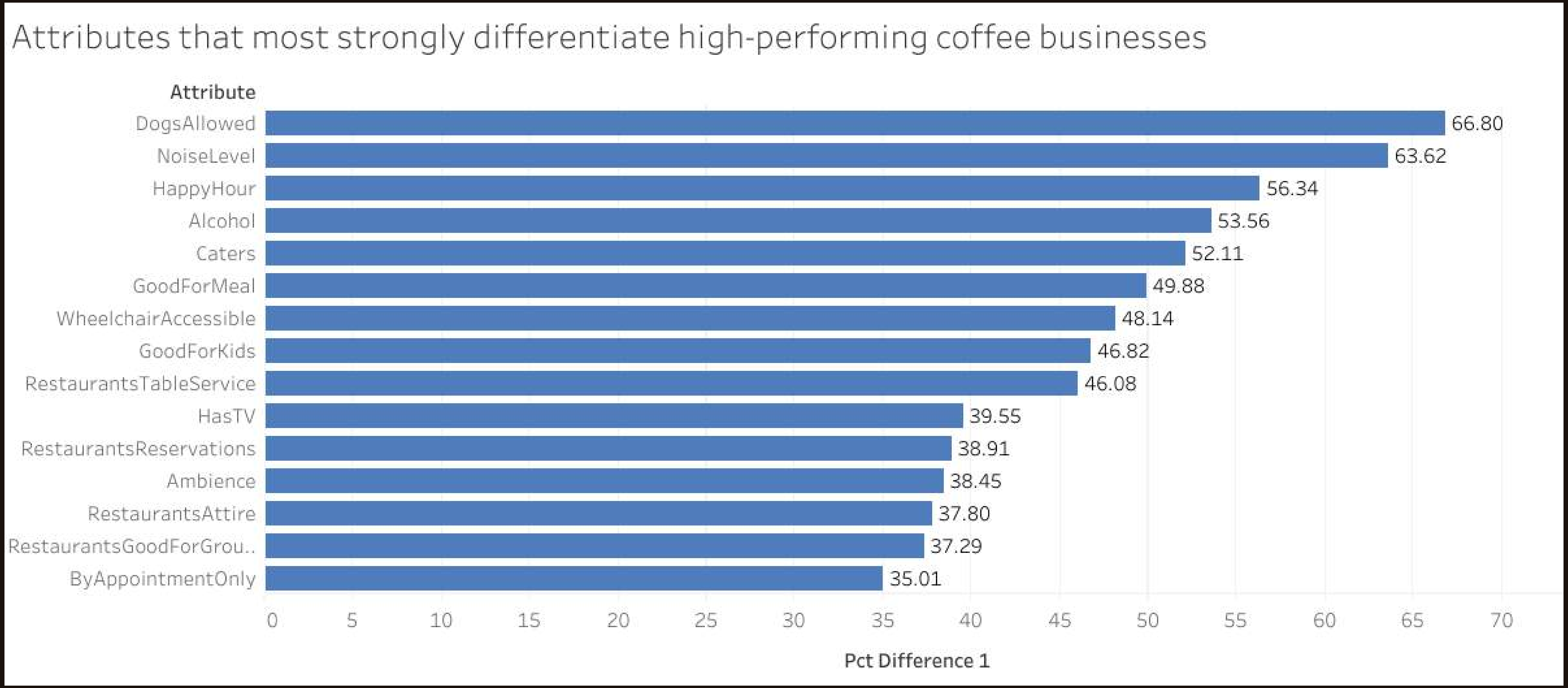
```
attr_freq_all = (
    pd.Series(all_attrs)
    .sort_values(ascending=False)
    .reset_index()
    .rename(columns={"index": "attribute", 0: "count"})
)

attr_freq_all.head(10)
```

✓ 0.1s

	attribute	count
0	BusinessParking	6047
1	RestaurantsTakeOut	5935
2	WiFi	5816
3	BusinessAcceptsCreditCards	5810
4	RestaurantsPriceRange2	5707
5	OutdoorSeating	5415
6	BikeParking	4937
7	RestaurantsDelivery	4927
8	Caters	3884
9	GoodForKids	2838


Baseline Attributes Common to High-Performing Coffee Businesses
RestaurantsTakeOut
NoiseLevel
Caters
Ambience
RestaurantsReservations
BusinessAcceptsCreditCards
RestaurantsTableService
GoodForKids
RestaurantsPriceRange2
WheelchairAccessible
Business Attribute
0%
20%
40%
60%
80%
100%
Share of Top-Performing Coffee Businesses


Common Attributes Among Low-Performing Coffee Businesses
RestaurantsTakeOut
RestaurantsPriceRange2
BusinessAcceptsCreditCards
BusinessParking
WiFi
RestaurantsDelivery
OutdoorSeating
BikeParking
DriveThru
Caters
Business Attribute
0%
20%
40%
60%
80%
Share of Low-Performing Coffee Businesses



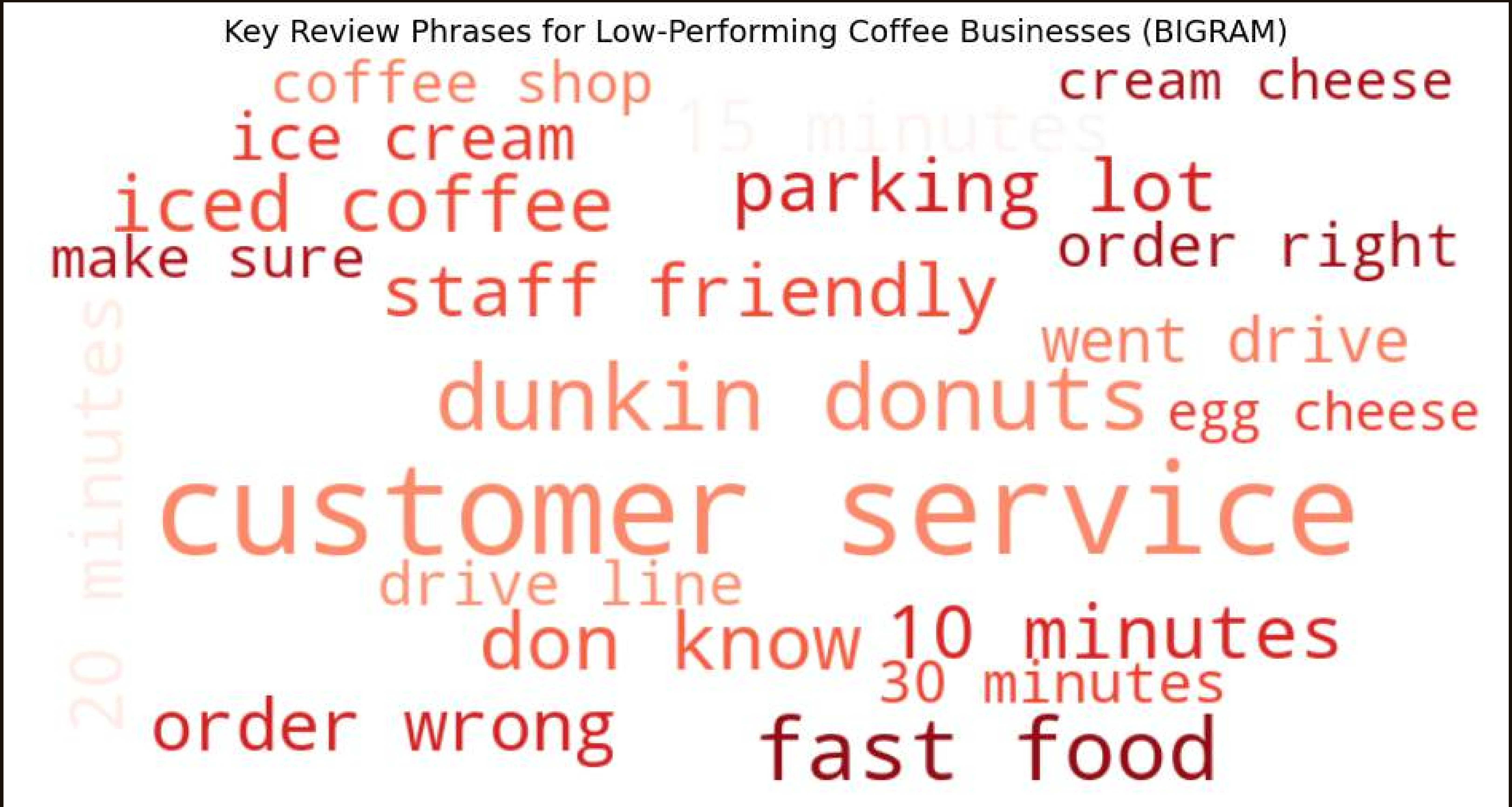
This graph shows percenatge difference of each attribute for top and low performing coffee businesses
How much more common this attribute is in top coffee businesses (top_pct – low_pct)
for example: 71% top performing coffee business while only 4.5% low performing coffee businesses has DogsAllowed attribute. and the Difference is 71%-4.5% = 66.80%

About 7 out of 10 top-performing coffee shops allow dogs, while fewer than 1 out of 20 low-performing shops do. This makes “Dogs Allowed” one of the strongest distinguishing features of successful coffee businesses in this dataset.

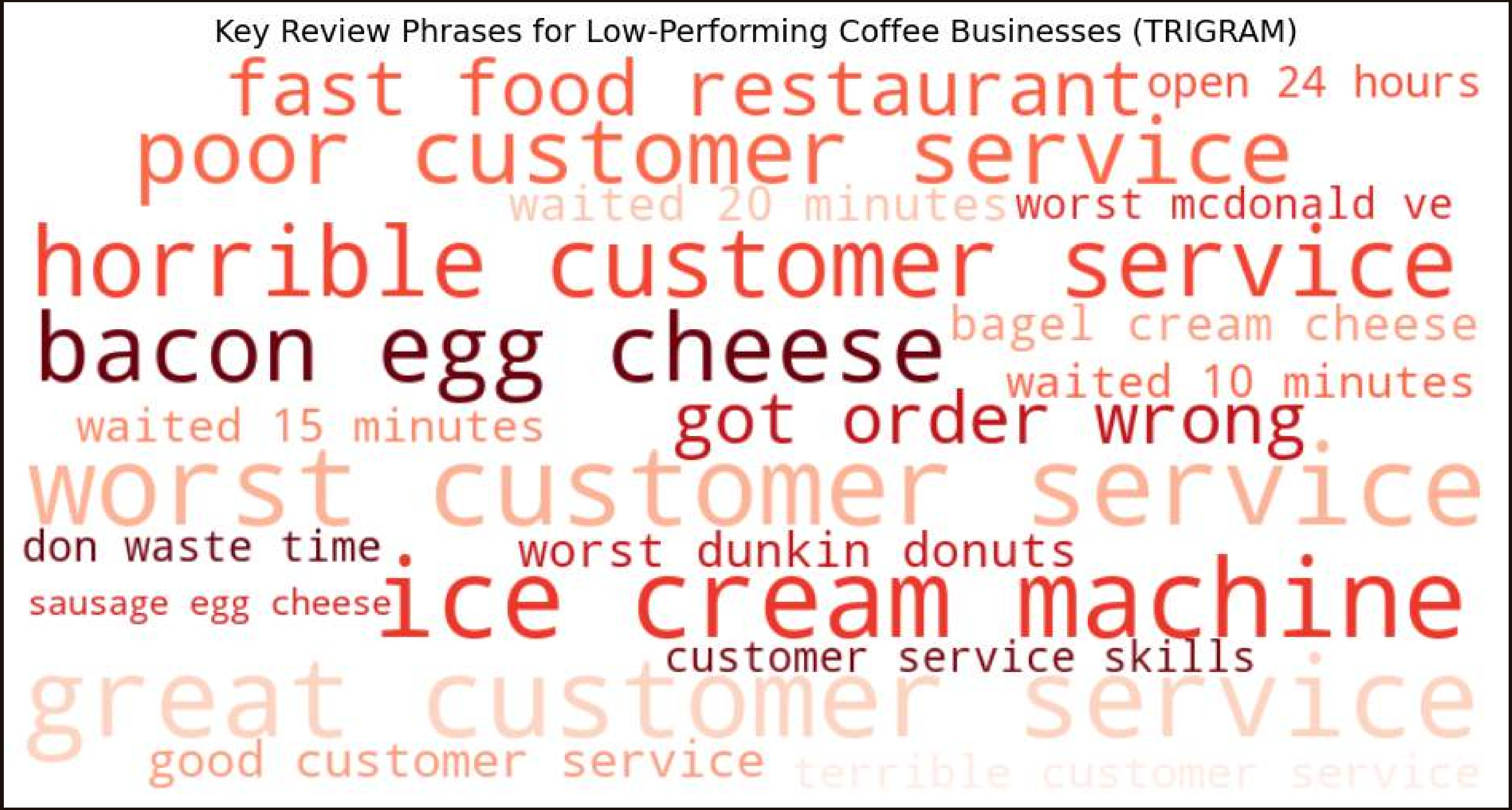
BIGRAM Review Text Analysis (TF-IDF)

TRIGRAM to see phrases

What customers complain about



BIGRAM and TRIGRAM



New Metric 1: SESS (Service Experience Signal Score)

What It Measures

SESS=(Positive Service Mentions−1.5×Negative Service Mentions)/Total Reviews

Why It Matters

- Captures how customers feel about service
- Strongly differentiates top vs low performers

Result

sess_top = 0.082

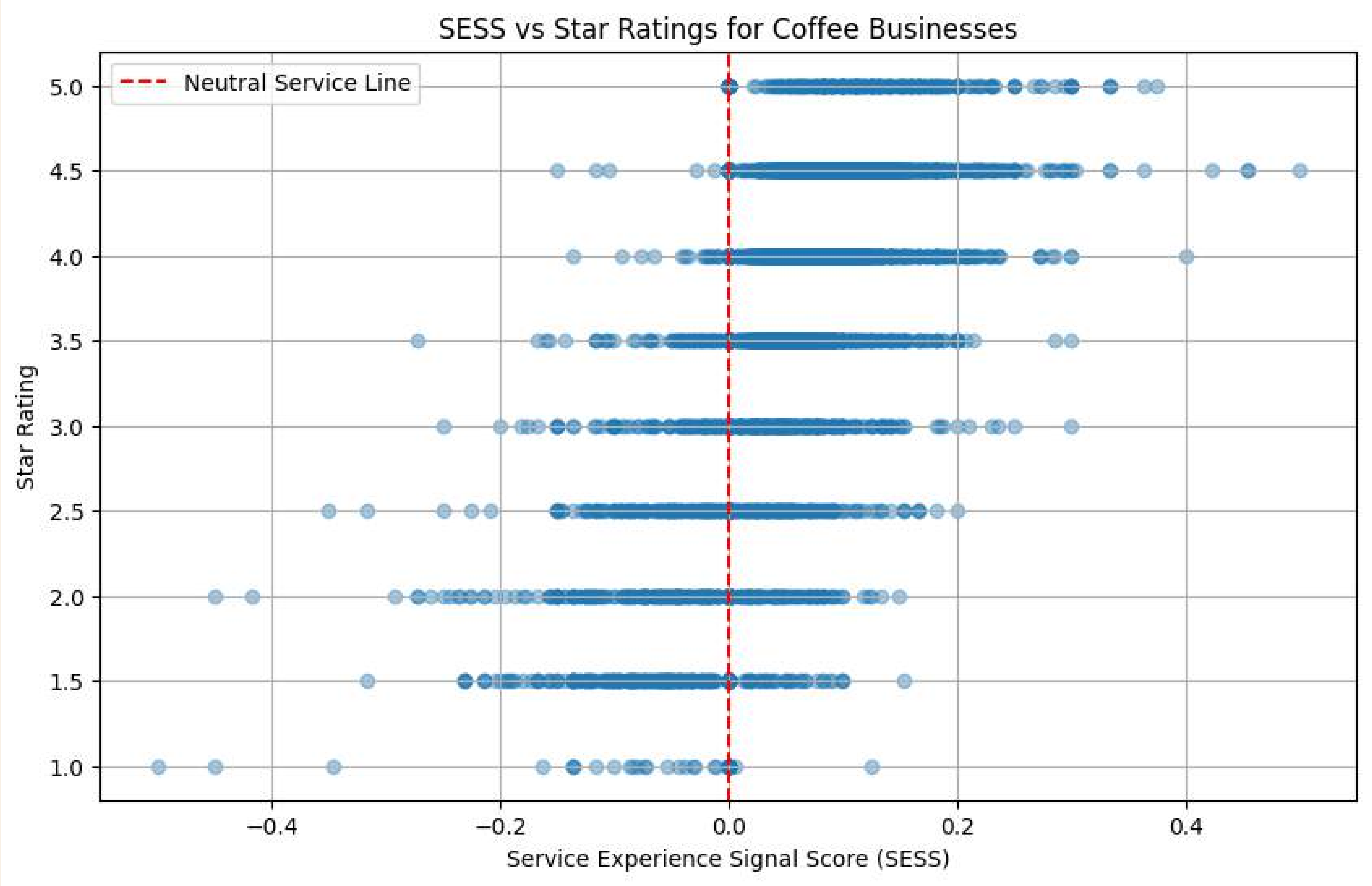
sess_low = -0.009

Key Insight

- Higher SESS aligns with higher star ratings
- low SESS aligns with low ratings

Recommendation for CoffeeKing

Track service quality early, before reviews accumulate



New Metric 2: OFI (Operational Friction Index)

What It Measures
Frequency of wait times, order errors, drive-thru issues
OFI=Operational Friction Mentions/ Total Reviews

Why It Matters
- Quantifies operational pain points
- Predicts negative sentiment before ratings drop

Result
- Low-performing coffee has 7× higher OFI

ofi_top = 0.036
ofi_low = 0.252

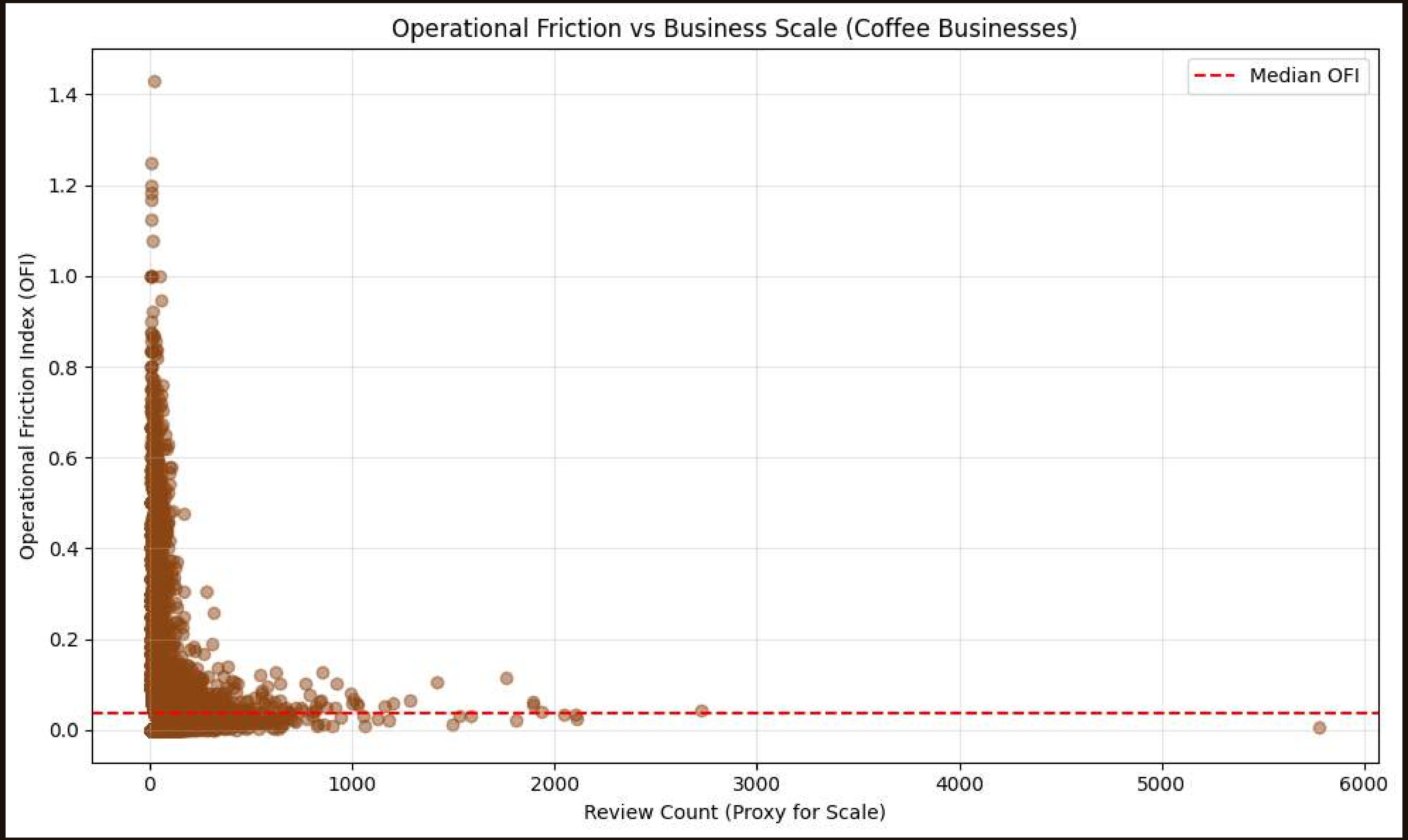
Operational Fraction vs Business Scale

What this shows

- As business scale increases, operational friction does not disappear — it often increases
- High-volume coffee businesses face execution challenges, not demand problems
- OFI helps distinguish:
  - Popular & well-run businesses
  - Popular but operationally stressed businesses
-

Why this matters for CoffeeKing

- Growth alone does not guarantee customer satisfaction
- Scaling locations without operational discipline increases:
  - Wait-time complaints
  - Order accuracy issues
  - Negative service sentiment in reviews



Operational RISK VS SERVICE EXPERIENCE

Quadrants

1. **Top-Left (Low OFI, High SESS)**
2. **Operationally smooth + great service**
3. **→ Best-in-class performers**
4. **Top-Right (High OFI, High SESS)**
5. **Good service but operational strain**
6. **→ Scaling risk (service goodwill may erode over time)**
7. **Bottom-Left (Low OFI, Low SESS)**
8. **Operationally fine but weak service**
9. **→ Training & culture opportunity**
10. **Bottom-Right (High OFI, Low SESS) ← High-Risk Underperformers**
11. **Operational breakdown + poor service perception**
12. **→ Urgent intervention required**

```python
star_threshold = 3.5
sess_threshold = 0.01
ofi_threshold = 0.15
metrics_df["risk_flag"] = (
    (metrics_df["stars"] < star_threshold) &
    (metrics_df["SESS"] < sess_threshold) &
    (metrics_df["OFI"] > ofi_threshold)
)

high_risk = metrics_df[metrics_df["risk_flag"]]
high_risk.head()
```

✓ 0.0s

	business_id	stars	SESS	OFI	review_count	risk_flag
7	-3dkEoYgH8AlUtBMZvzUfg	2.5	0.000000	0.333333	21	True
13	-7Rx5jVeQmlVoAU_oXrzew	1.0	-0.136364	0.454545	11	True
38	-QbbFXdiWQb2vKaDlky4Pw	2.5	0.000000	0.583333	12	True
44	-TboXPMTf45s24FPyD8OAA	2.0	0.000000	0.230769	52	True
48	-Wv0KRW7vv77bjOKLrxpXg	2.5	0.000000	0.153846	13	True

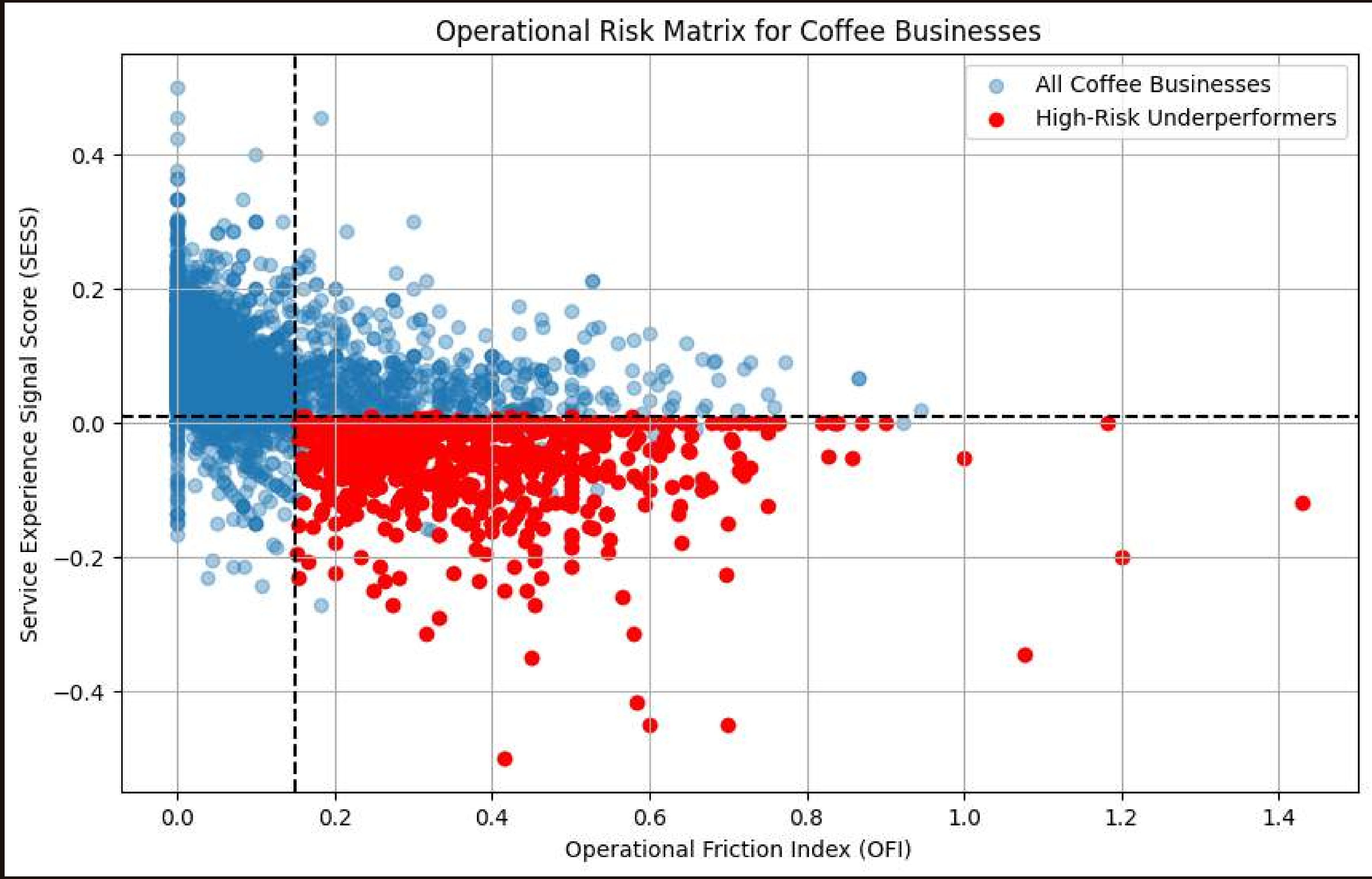
High-Risk Underperformer Quadrant

Quadrants
Top-Left (Low OFI, High SESS)
Operationally smooth + great service
→ Best-in-class performers

Top-Right (High OFI, High SESS)
Good service but operational strain
→ Scaling risk (service goodwill may erode over time)

Bottom-Left (Low OFI, Low SESS)
Operationally fine but weak service
→ Training & culture opportunity

Bottom-Right (High OFI, Low SESS) ← High-Risk Underperformers
Operational breakdown + poor service perception
→ Urgent intervention required



Executive Summary

What we found

- Location and market context strongly influence coffee business success on Yelp
- Customer engagement (reviews, photos, check-ins) is a key driver of higher ratings
- Operating longer or more consistent hours alone does not guarantee higher engagement
- Coffee businesses attract a small but highly influential user base
- Certain attributes and experiences consistently differentiate high-performing coffee shops

What this means for CoffeeKing

- Strategic location + engagement-first operations matter more than long hours alone
- CoffeeKing should design stores, hours, and offerings to maximize engagement quality, not just availability

Location Strategy

- Smaller and mid-sized cities often outperform large metro areas on ratings and consistency
- Expansion success depends on **market saturation**, not just population size

Actionable Recommendations

- Prioritize **mid-sized or under-saturated coffee markets**
- Evaluate cities using:
  - Average coffee ratings
  - Median review counts
  - Rating consistency (low variance)
- Avoid over-crowded markets unless CoffeeKing offers a **clear experiential differentiator**

Business Impact

- Higher chance of strong early reviews
- Faster reputation building
- Lower competitive pressure

Smarter Hours Not Longer Hours

Actionable Recommendations

- Optimize hours around peak demand, not maximum availability
- Focus on:
- Morning commuter rush
- Weekend high-traffic windows
- Maintain predictable schedules rather than extended hours

Business Impact

- Lower operational costs
- Higher staff efficiency
- Better customer experience during peak times

Positioning and Differentiation

Actionable Recommendations

- Position CoffeeKing as:
- An experience, not just a beverage provider
- Invest in:
- Signature drinks
- Store aesthetics
- Staff training
- Align branding with emotions customers express in reviews

Business Impact

- Higher emotional connection
- More positive review language
- Stronger brand recall

Engagement Strategy

Actionable Recommendations

- Design stores to encourage:
  - Photo-worthy interiors
  - Social moments (latte art, branded cups, merch)
- Actively prompt:
  - Reviews
  - Photos
  - Check-ins via signage or loyalty perks
- Treat Yelp engagement as a growth channel, not just feedback

Business Impact

- Stronger online presence
- Faster brand recognition
- Higher perceived quality

Final Recommendations

What CoffeeKing Should Do

1. Expand into strategically chosen mid-sized markets
2. Design engagement-first store experiences
3. Optimize hours for demand, not duration
4. Target influential coffee users
5. Invest in product + ambience differentiation
6. Monitor engagement quality as you scale

Next Steps

Immediate

- Identify 3–5 target cities using Yelp metrics
- Pilot engagement strategies in first locations

Mid-Term

- Track review sentiment and engagement growth
- Adjust hours and offerings based on data

Long-Term

- Build a repeatable, data-driven expansion model


Thank You