Capstone Project

Explaining Customer Ratings of Tesco and Sainsbury's

Zixiong Long
August 2020

Introduction

What Drives Customer Ratings?

- Grocery stores are everywhere and seemly identical

- Endogenous: Completeness of goods? Service? Manager?

- Exogenous: Location? Competitive environment? House price?

Focus of This Research

- London
- Tesco
- Sainsbury’s
- Exogenous factors

Data

Data Sources

- Foursquare API: Locations, venues and ratings
- Wikipedia: London locations
- The London Data Store: London house prices

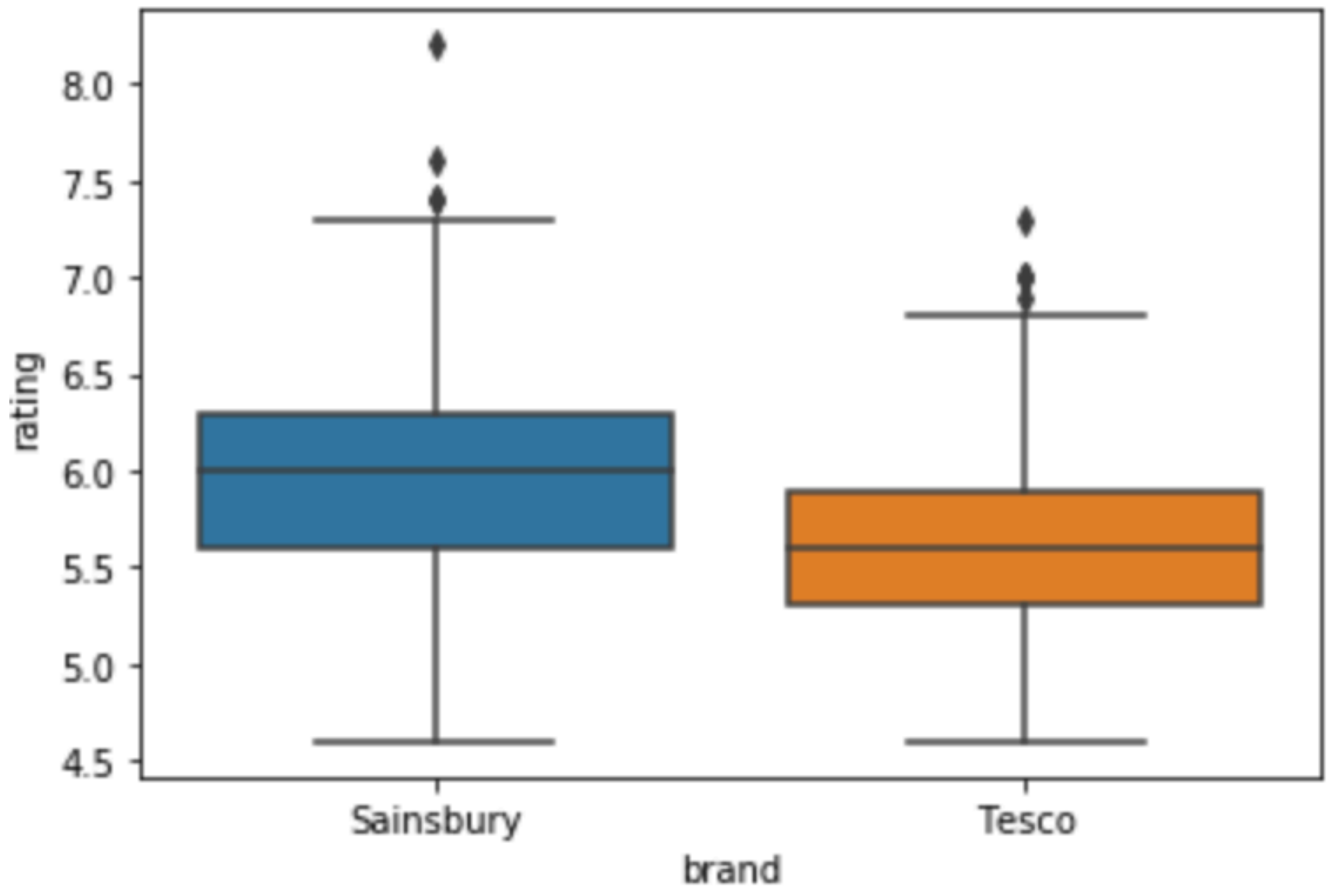
Data Collecting

- Raw London location data: 534 rows and 5 columns
- Raw grocery stores data: 2182 rows and 7 columns
- Raw house price data: 675 rows and 91 columns
- Raw neighbour venues data: 24631 rows and 2 columns
- Final cleaned data: 591 rows and 21 columns

Exploratory Analysis

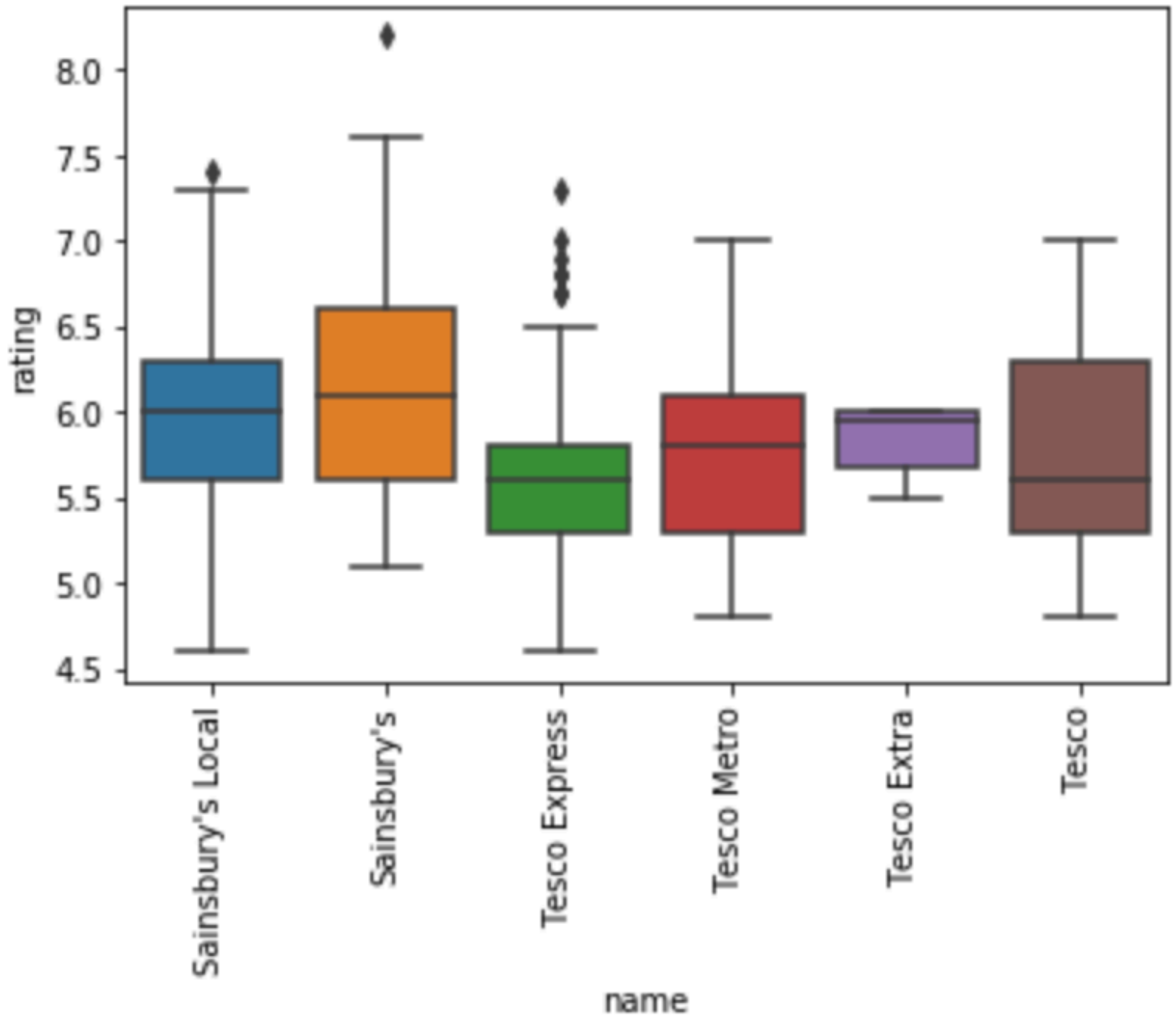
Rating vs Brand

- Tesco(5.6) on average has lower rating than Sainsbury(6.0)
- Tesco and Sainsbury have similar lower bound in rating, at 4.5
- Tesco(6.8) rating upper bound is lower than Sainsbury(7.4)
- Tesco(7.5) max rating is smaller than Sainsbury(8.5) max rating



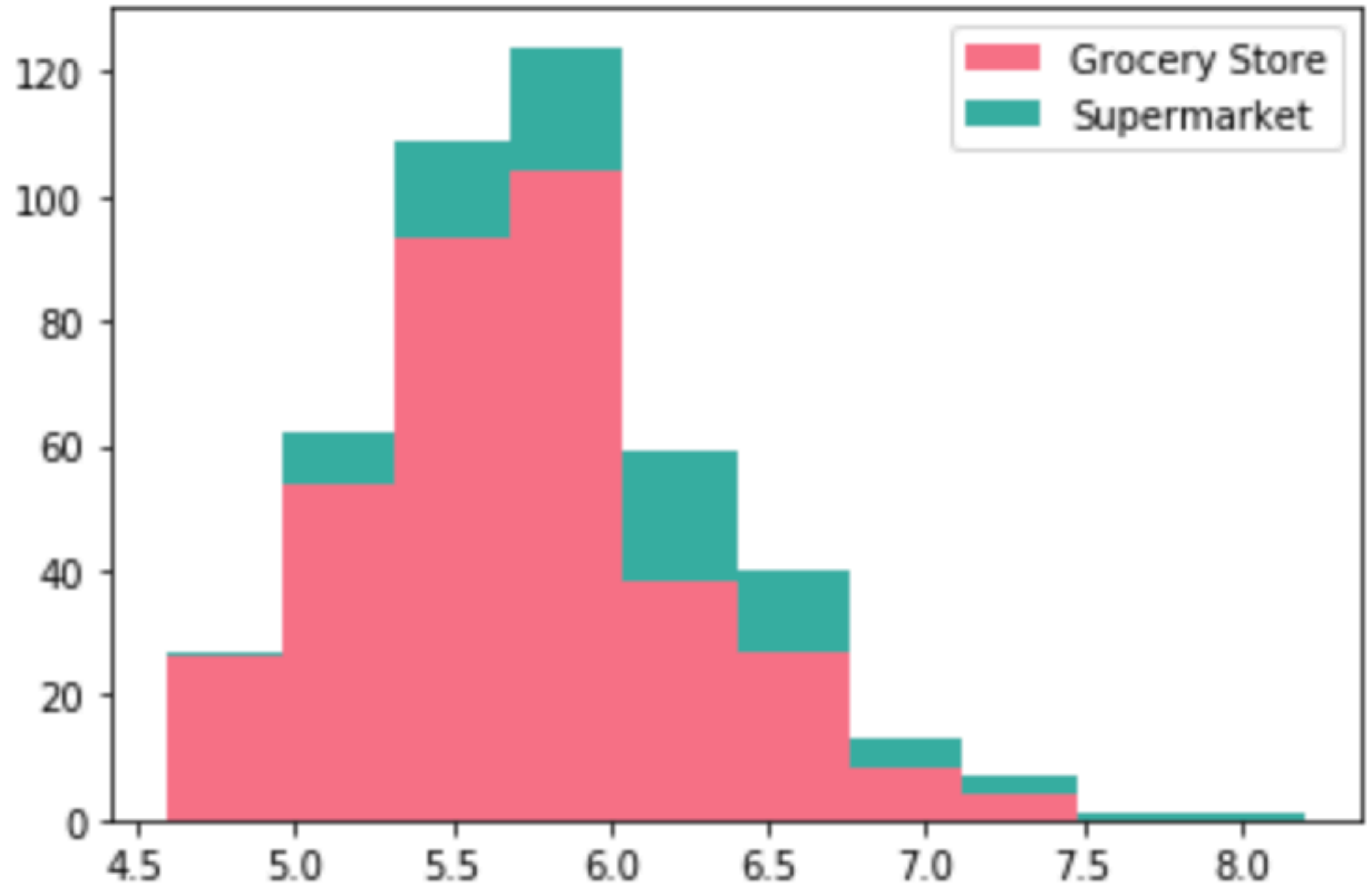
Rating vs Name

- No significant differences between sub-categories (names)
- Sainsbury's have a higher upper bound
- Within Tesco, Tesco Express and Tesco Extra make a big difference
- Within Sainsbury's, Sainsbury's is better than Sainsbury Local



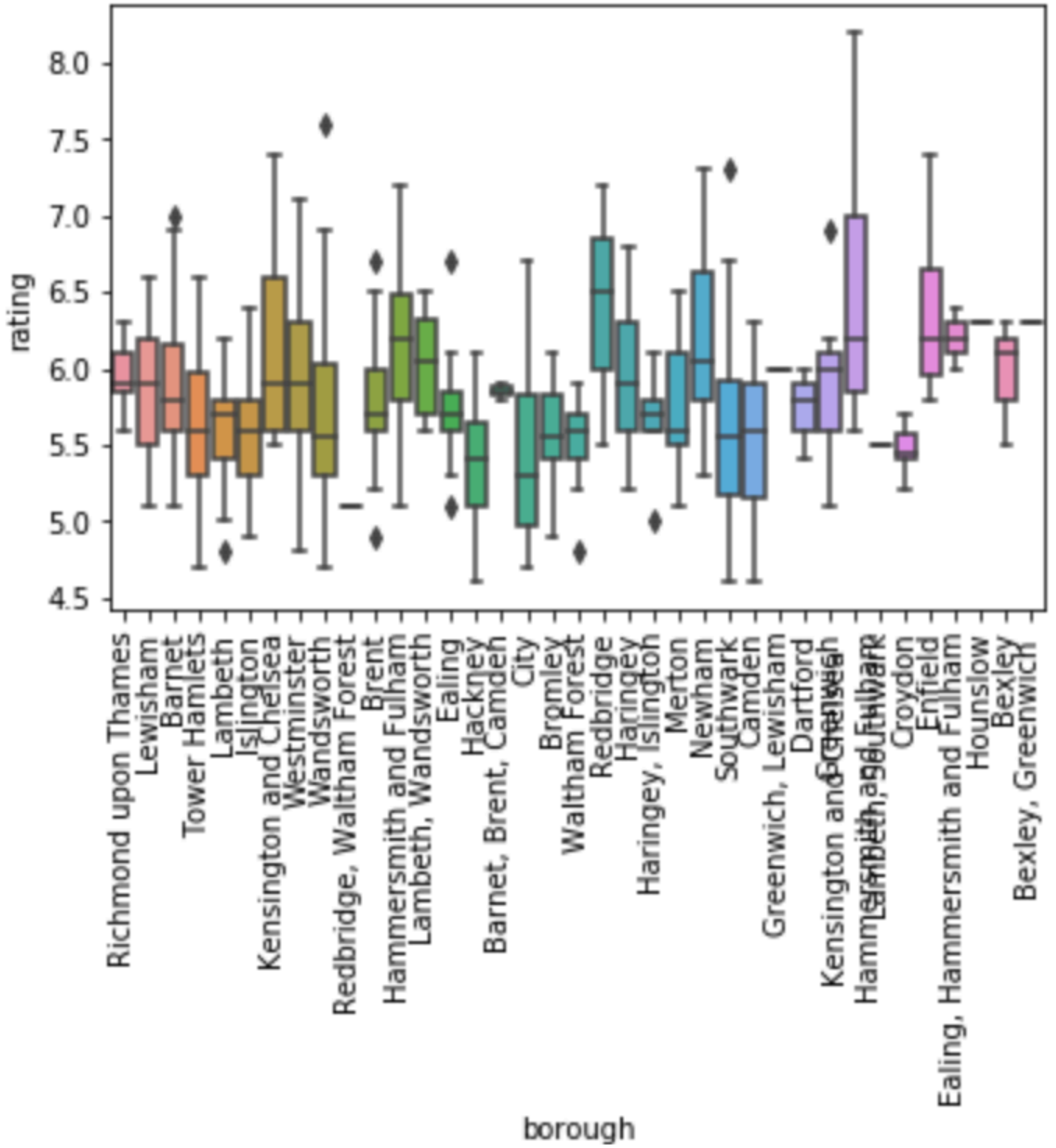
Rating vs Categories

- Supermarkets generally has a higher rating than grocery stores.
- It's a useful feature, however, the previous new name column already contains the information of classification of supermarket



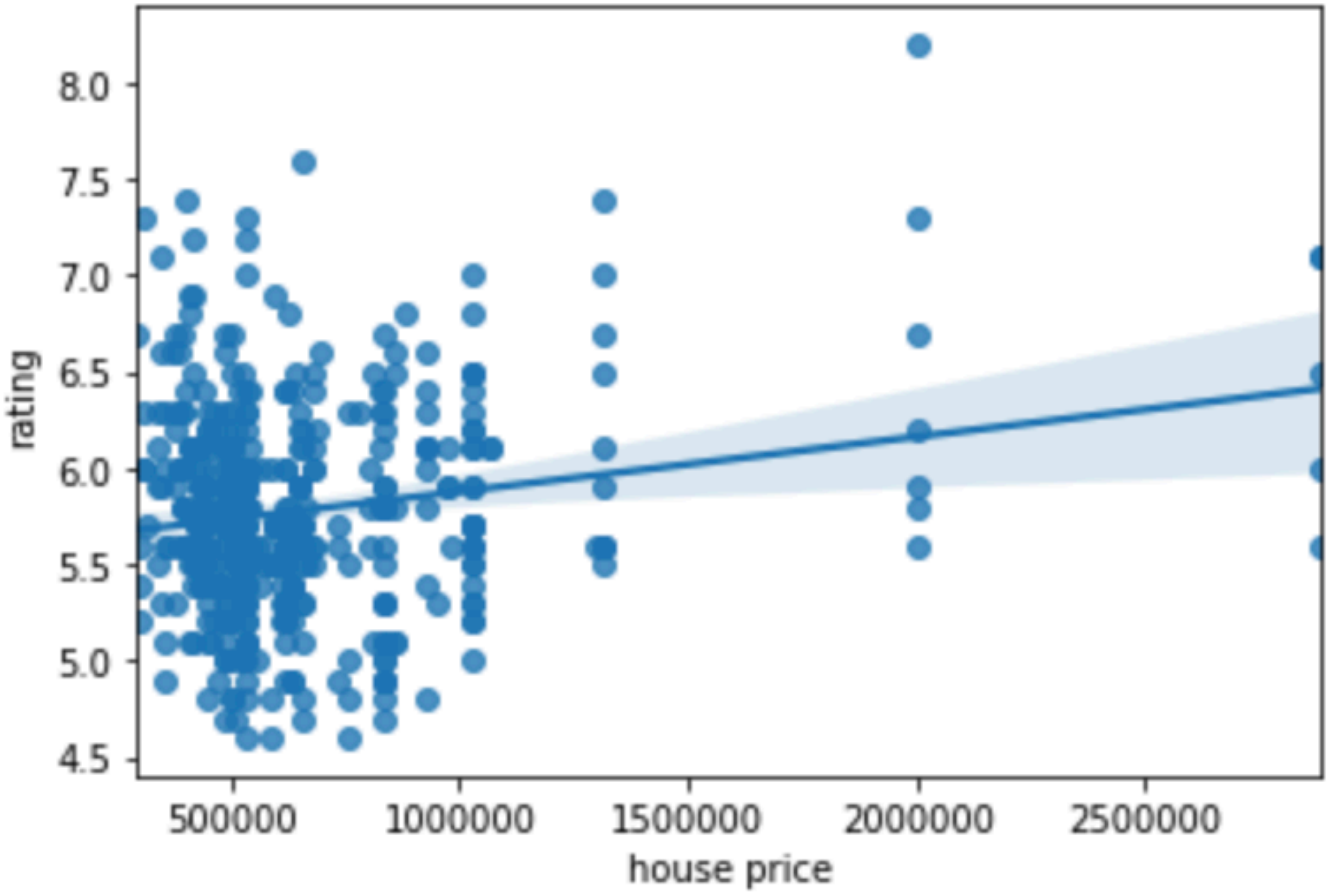
Ratings vs Borough

- The 10 boroughs — 'Richmond upon Thames' , ’Kensington and Chelsea’, 'Hammersmith and Fulham’, 'Lambeth, Wandsworth’, 'Barnet, Brent, Camden’, 'Redbridge', ’Newham', ’Kensington and Chelsea\nHammersmith and Fulham’, ‘Enfield', 'Ealing, Hammersmith and Fulham’ — exhibit relative high ratings than others.



Rating vs House Price

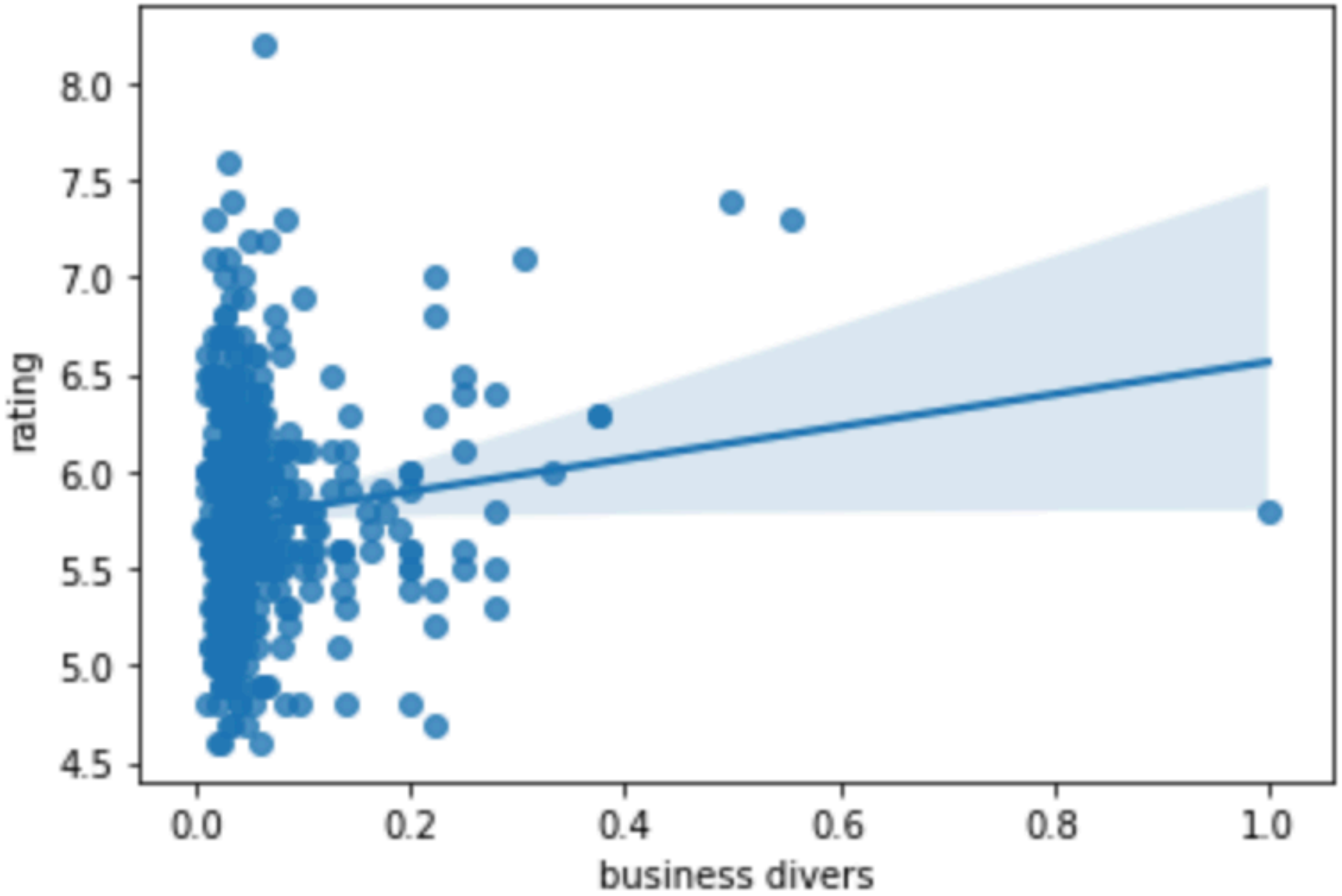
- There is weak positive relationship between house price and rating
- A significant difference can be found in low bounds between stores where house price above 1100000 and those below 11000000



Rating vs Business Diversification

- One way to measure business activity is to calculate the sum of squared percentages of each 5 popular venues to total numbers of venues

- A weak positive relationship exists for rating and business diversification



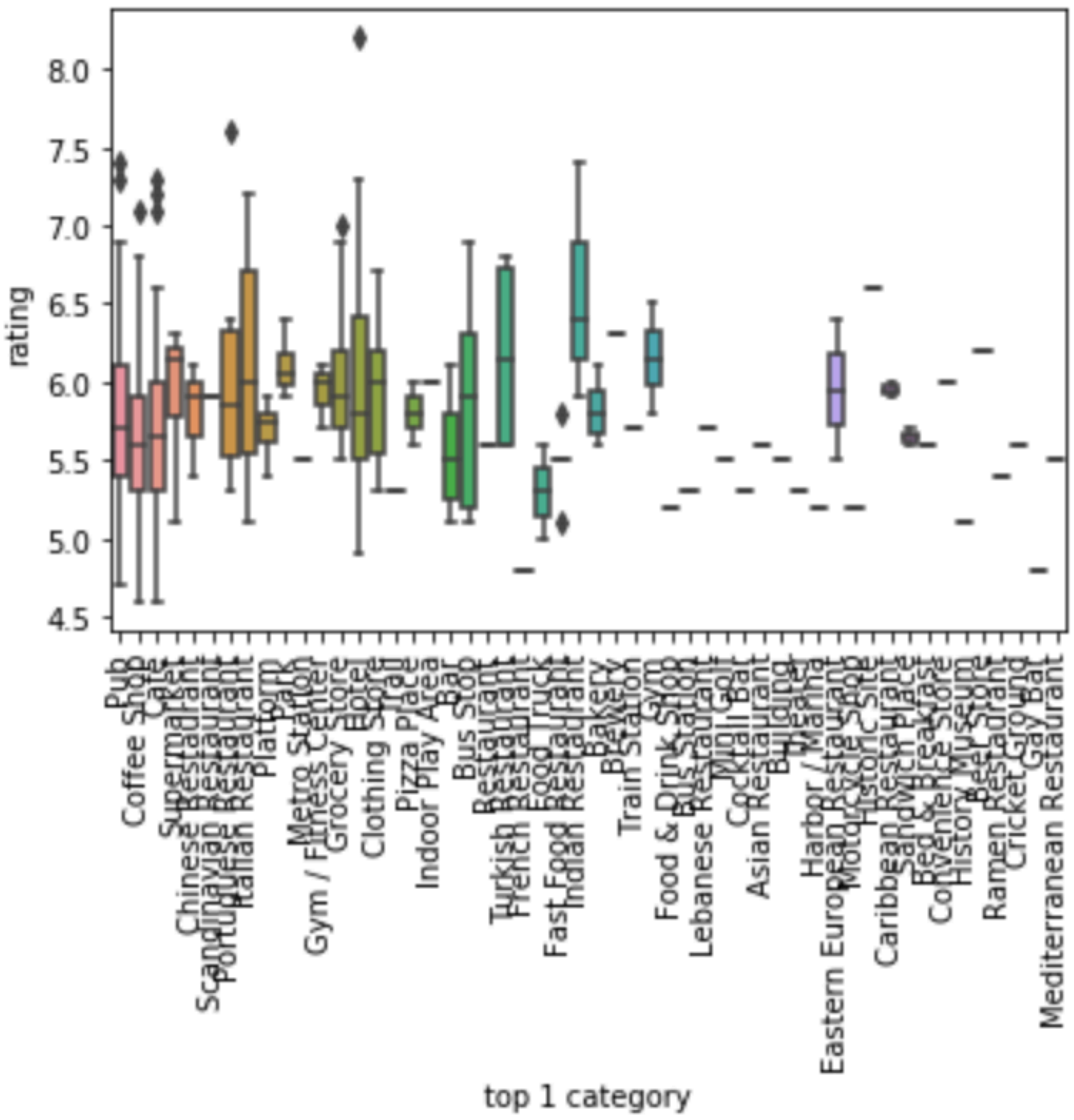
Rating vs Top 1 Category

- Three classes of venues can be identified by grouping ratings:

- high: 'Grocery Store’, 'Hotel' 'Italian Restaurant’, 'Clothing Store' ‘Supermarket', 'Portuguese Restaurant'

- low: 'Coffee Shop’, 'Café'

- median: ‘Pub', ‘Platform'



Models and Predictions

Machine Learning Models

- 2 baseline models: the first one randomly guessing labels with equal 33.3% probability and the second model randomly guessing labels with corresponding label frequencies in training data
- 5 fine-tuned machine learning models: Decision Tree, K Nearest Neighbours, Support Vector Machine, Random Forest and XGBoost
- 1 ensemble model of hard voting classifier with previous 5 fine-tuned models as the underlying

5-fold Cross Validations

- All machine learning models outperformed the baseline random guessing models. It means that input features and machine learning models do provide extra information to explain the store ratings

- Voting classifier beat all other models with smaller variance in prediction.

- The voting classifier is the final model to be used for prediction.

	mean accuracy	std accuracy
Voting	0.530516	0.0170639
XGBoost Tuned	0.514888	0.067094
Random Forest Tuned	0.50572	0.0284465
Random Forest Tuned 2	0.503447	0.0154755
XGBoost Tuned 2	0.501251	0.0548969
KNN Tuned 2	0.49216	0.0345727
KNN Tuned	0.487666	0.0459845
KNN	0.478652	0.0296194
Decision Tree Tuned	0.471859	0.0299297
Decision Tree Tuned 2	0.471859	0.0299297
SVC Tuned 2	0.469484	0.0315936
Decision Tree	0.465169	0.0485704
SVC	0.460444	0.0378738
SVC Tuned	0.458248	0.0208889
XGBoost	0.45383	0.0371145
Random Forest	0.451532	0.0481365
Base Model 2	0.336313	0.0389415
Base Model 1	0.329699	0.052785

Predictions

- Voting classifier, as the final model, still generates the best out-of-sample performance with accuracy close to cross validation performance
- Random forest and XGBoost are surprisingly beaten by other three methods, which indicates the overfitting of the two models in the parameter-tuning process
- All machine learning models outperform the baseline models

	accuracy
Voting	0.513514
KNN	0.493243
SVC	0.486486
Decision Tree	0.472973
Random Forest	0.452703
XGBoost	0.439189
Base Model 2	0.398649
Base Model 1	0.337838

Conclusions

Conclusions

- A grocery store located in a neighbourhood with high business diversity tends to have low ratings, or, surrounding commercial activity hurts grocery store ratings.
- Tesco Express has relatively low ratings comparing to other Tesco stores or Sainsbury's.
- A grocery store located in Richmond upon Thames, Kensington and Chelsea, Hammersmith and Fulham, Ealing and Redbridge tends to have high ratings.
- A grocery store that is close to its competitors, i.e. grocery stores and supermarkets, tends to have higher ratings.
- A neighbourhood with very high house prices tends to have positive impact on grocery store ratings.
- A voting classifier with decision tree, SVC, random forest, KNN and XGBoost can utilise the features well and produce reasonably good out-of-sample results.

Future Works

- Other exogenous factors such as ratings of neighbours, distance to nearest competitors, and a higher level classification of venues (e.g. group all kinds of restaurants as one class) could be potentially predictive for the ratings of grocery store.

- Endogenous factors such as completeness of goods of a store, service level of staff, and internal rating of store manager could also be powerful determinants of the ratings.

Thank You!