

Navigating New Terrains: Modeling Manhattan's Business Dynamics with Python & A Complete Guide to Azure Distributed Computing System Set-Up

Cihang Wang, Ruidi Mu, Jiajun Gu, Ella Shi, Yicheng Wang

Abstract

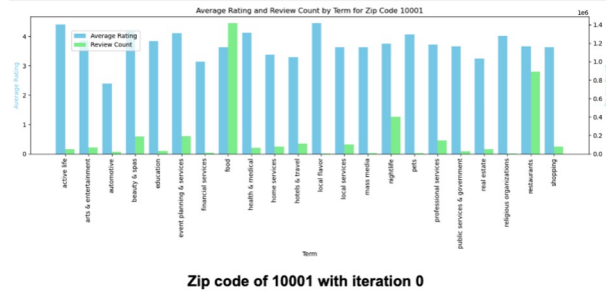
Forecasting changes in property prices stands as a paramount consideration for stakeholders in real estate, urban planning, and economic forecasting. Building on prior semester's work and on research linking Yelp data to gentrification indicators, we explore the predictive capacity of Yelp business data by carefully studying the correlation between Yelp features such as review count, rating, price level of businesses, as well as business operating hours and the median home sales price of the neighborhood.

Query Efficiency and Data Quality Improvements

We streamlined our query code, improving efficiency from 80% to 97%, and reduced API call waste by 80%. Key enhancements include handling the "is_open" field correctly and adopting accurate geographic filters, ensuring a reliable and comprehensive dataset for Manhattan's business landscape.

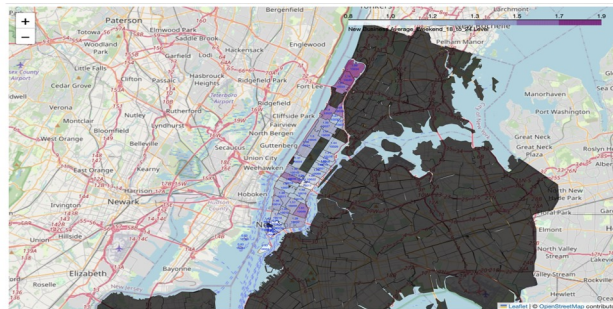
Comprehensive Exploratory Static Data Analysis

We manually input missing official zip codes for data from 2024-02-20 to 2024-03-29 and analyzed iterations 0, 1, and 2. Grouping data by zip code and term, we examined average ratings and review counts. Results show consistent trends in food, nightlife, and restaurant categories, with zip code 10001 displaying similar patterns across all iterations.



Dynamic Data Analysis and Visualization

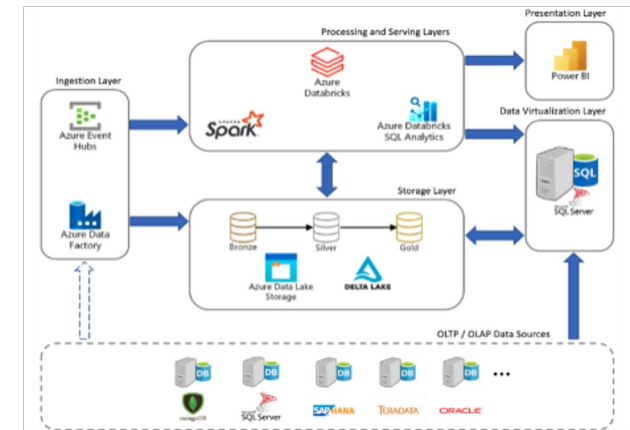
We created a comprehensive dataset of NYC business information using Pandas, filtered by official zip codes, and engineered new features for activity periods. Using GeoPandas and Folium, we conducted geospatial analysis and visualized data. Maps showed established food businesses had higher ratings downtown, new services excelled in Battery Park, and review counts were high in the Lower East Side and Chelsea. Fort George and Inwood had the most food businesses open late on weekends.



Cloud Environment Set Up

Azure and Databricks

- Scalability and Cost-Effectiveness: Quickly scale based on needs and pay only for used resources, reducing costs compared to physical servers.
- Security and Compliance: Robust security measures ensure data protection and regulatory compliance.
- Global Availability: Deploy and manage applications across multiple locations, enhancing performance and accessibility.



Conclusion

Our study demonstrates the power of integrating real-time data from digital platforms with advanced analytics and cloud computing to understand urban business dynamics. Key findings reveal patterns and trends in Manhattan, offering valuable insights for urban planners and policymakers. The project provides a robust framework for predicting neighborhood transformations and supports informed decision-making in urban planning and economic development.