Yelp Open Dataset

Design Document & Proposal

Catherine Liu

CS 181DV

Table of Contents

1. Project Definition and Scope

Problem Statement and Motivation

The goal of this project is to leverage Yelp's open dataset to derive insights into characteristics of businesses and potential factors of business success. These insights aim to guide small business owners toward strategic success and help consumers discover quality dining experiences.

The project will develop an interactive visualization system including geographic heat maps, bar graphs, a scatter plot, and a pie chart. The visualizations will use Yelp's data regarding businesses including name, geographic information, ratings, categories the business belongs to, and additional attributes. It will also use a reviews dataset containing information about the business it is for, the user, the rating, and the review text.

Project Definition and Scope

Target users and use cases + Expected insights

1. Small Business Owners

Goal: Help owners understand where and how to grow or compete.

- Insight: Correlation between geographic location and review metrics
    - Method: Heatmap of review density by location
- Insight: Identify “Top Cities” for highly rated businesses
    - Method: Bar chart ranking cities considering average ratings and total reviews
- Insight: Sentiment analysis for reviews and rating distribution
    - Method: Sentiment analysis score, pie chart of review sentiment distribution, bar chart of rating distribution

2. Consumers (Local and Foreign)

Goal: Help consumers find great food and explore regional cuisine trends.

- Insight: Popular cuisines and their geographic distribution
    - Method: Map with geographic cuisine clustering
- Insight: Top cuisines in each city
    - Visualization: Bar chart of top cuisines for large metropolitan areas
- Insight: Hidden Gem Businesses (high rating, low visibility)
    - Method: Scatter plot (review count vs. rating)

Project Definition and Scope

Expected outcomes

- **Interactive visualizations** that make Yelp data more **accessible** and **interpretable** for business owners and consumers
- Derive general **insights** and **trends** from Yelp's data about its businesses
- Better understanding for business owners about **customer sentiment** and **ratings**
- **Geographic heatmaps** that indicate regional reviews and **spatial understanding** of consumer engagement
- Show **buisness performance** across regions
- **Geographic insights** into specialization of **cuisines** by area
- **Hidden gem discovery** to promote and help customers find businesses which are high-quality, but lesser-known
- **Actionable insights** for small business decisions and better consumer choice

Project Definition and Scope

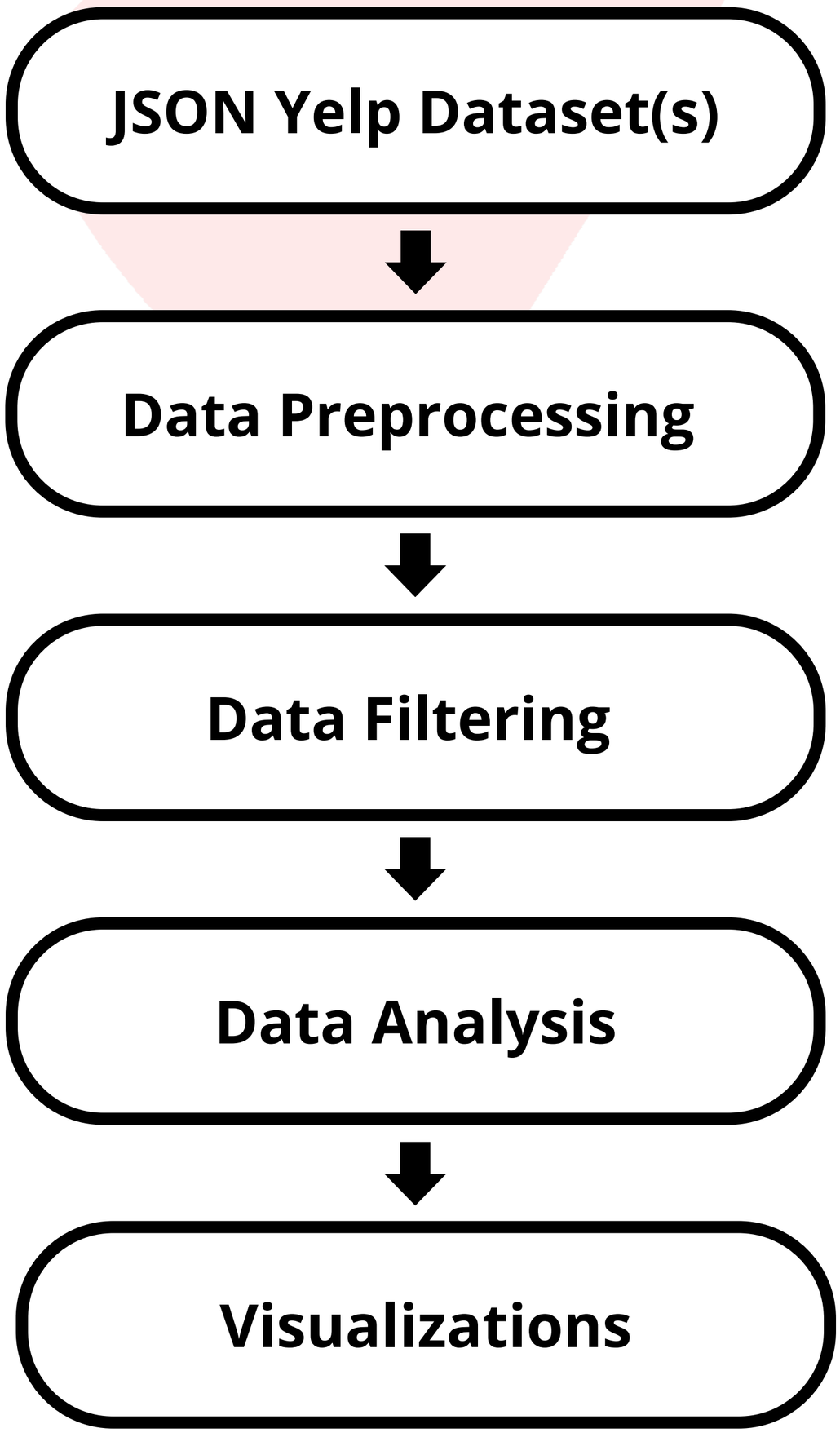
Project boundaries and constraints

1. **Scope:** The data is centered around specific cities and areas. It is not comprehensive.
2. **Data Quality:** There are inconsistencies in the data and categorization may be subjective.
3. **Bias:** The analysis is done specifically for business on Yelp, excluding data from other platforms and word of mouth.

Project Definition and Scope

2. Technical Design

Data flow and processing pipeline

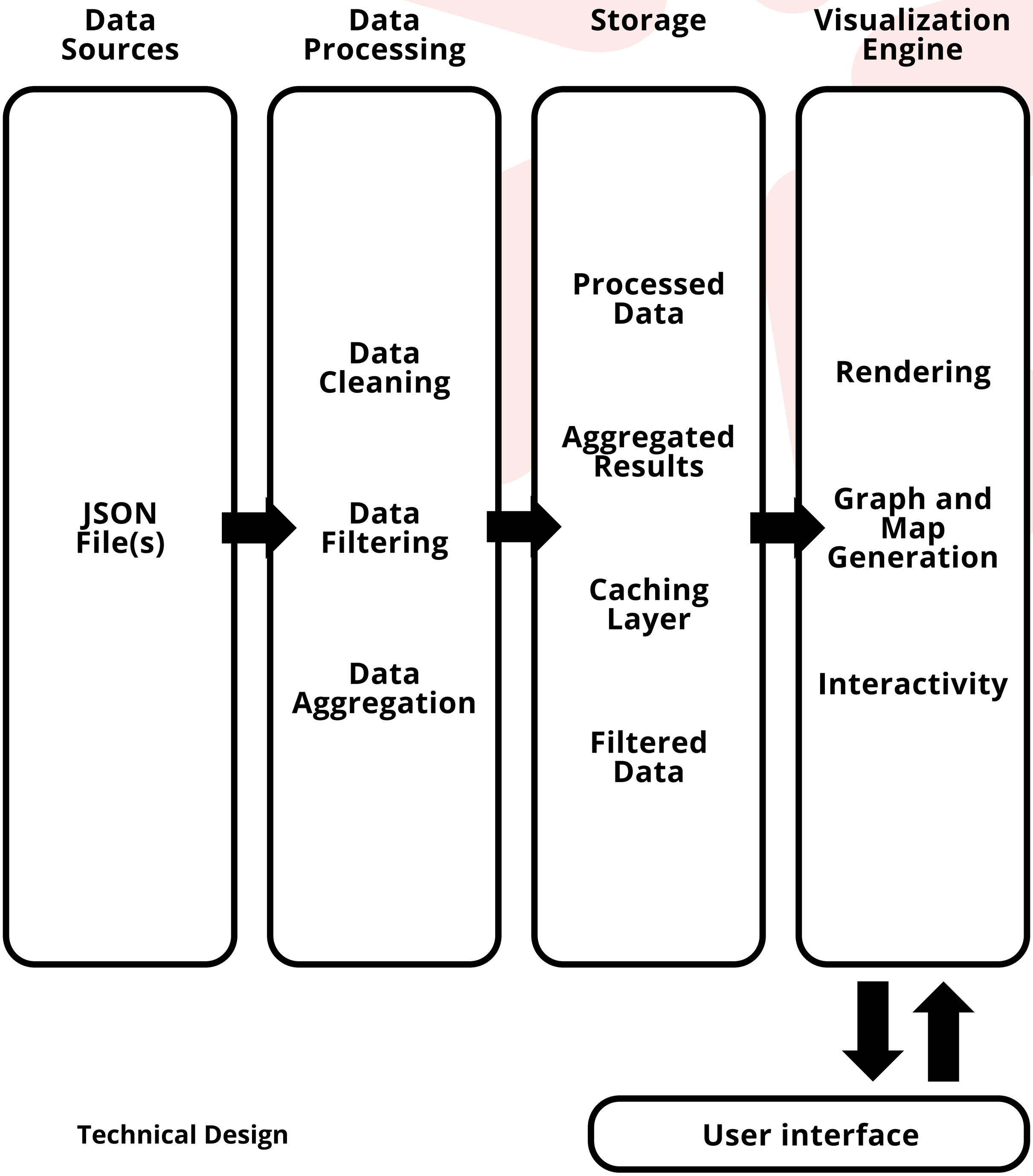


System architecture diagram #1

Data Source	Yelp Open Dataset (JSON)
Data Preprocessing	Python (pandas, textblob, pyarrow)
Data Storage	Local data files
Visualization	Plotly, Mapbox, Steamlit
Frontend	Steamlit
Local Hosting	Steamlit

Technical Design

System architecture diagram #2



Technology Stack Justification

1. **Data Source:** relevent, real world information with variety of information to derive insights from
2. **Data Preprocessing:** efficient, flexible, and easy to use
3. **Data Storage:** simplicity
4. **Visualization:** highly interactive, variety of possible data representations
5. **Frontend:** ease of use, integrates with Plotly
6. **Local Hosting:** ease of launch

(Refer to system architecture diagram #1)

Performance Considerations

1. **Large Files:** the Yelp datasets are very larges, so it would be a good idea to preprocess and save only the data that is applicable to use, use filtering and chunked processing
2. **High Memory Use:** process in chunks
3. **Plotting Speed:** use of aggregated data
4. **Browser Responsiveness:** lazy-loading

Technical Design

3. Data Strategy

Select and justify your dataset from recommended sources

For this project, I decided to use the Yelp Open Dataset as it provides relevant and diverse data that can be used to derive a variety of insights. Some factors I considered include:

1. **Real-World Relevance:** Yelp is a widely used platform for both consumers and business owners. It represents real-world behavior and preferences
2. **Ethical Use:** The dataset is made public by Yelp for academic and research purposes.
3. **Personal Interest:** I use Yelp often and I love trying different restaurants and cuisines.

Data Strategy

Data cleaning and preprocessing approach

The dataset to use:

yelp_academic_dataset_business.json: includes general characteristics of a business

yelp_academic_dataset_review.json: includes individual reviews and the context for them

1. Load JSON file using pandas
2. Handle missing values and invalid data, convert necessary columns to proper types
3. Filter data
    a. by businesses status (is_open == 1)
    b. by valid categories (must be a restaurant)
    c. remove businesses which lack necessary information for analysis
4. Feature Engineering through binned ratings and cuisine extraction
5. Chunked loading and sentiment analysis for review processing

Data Strategy

Storage and retrieval strategy

1. **Local storage**
    a. Raw data:
        yelp_academic_dataset_business.json,
        yelp_academic_dataset_review.json
    b. Post-processed data: store as parquet file
2. **Data loading**
    a. Load filtered dataset
    b. Use indexing on relevant columns
3. **Caching for speed**

Data Strategy

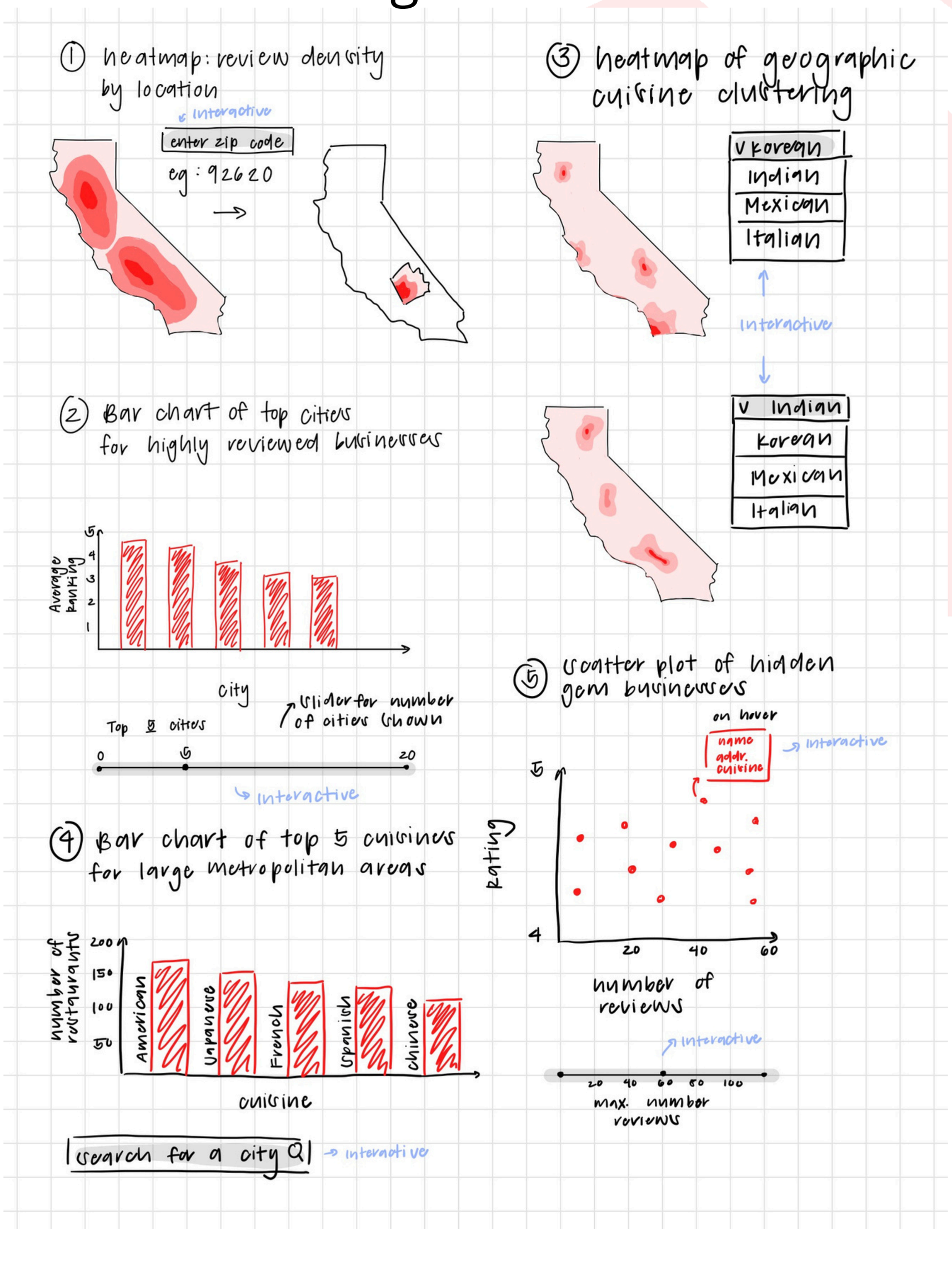
Data maintenance plan

1. Keep track of dataset versions: raw data, cleaned data, filtered data
2. Organize data in a directory
3. Use names which consistent and self-descriptive
4. Implement comprehensive error logging
5. Keep data on Github or Kaggle

Data Strategy

4. Visualization Design

Initial Plan/Design



Visualization Design

Created Visualizations

Business Metrics (review density): Business Owner > Business Metrics

Visualizations and Areas of Interactivity

- Geagraphic heatmaps
  - Zoom, pan, hover
  - Dropdown and type to search select city
- Bar graphs
  - Slider Top Number of Businesses
  - Zoom, pan, hover
  - Dropdown and type to search select city

Select City

All Cities

Number of Businesses

9963

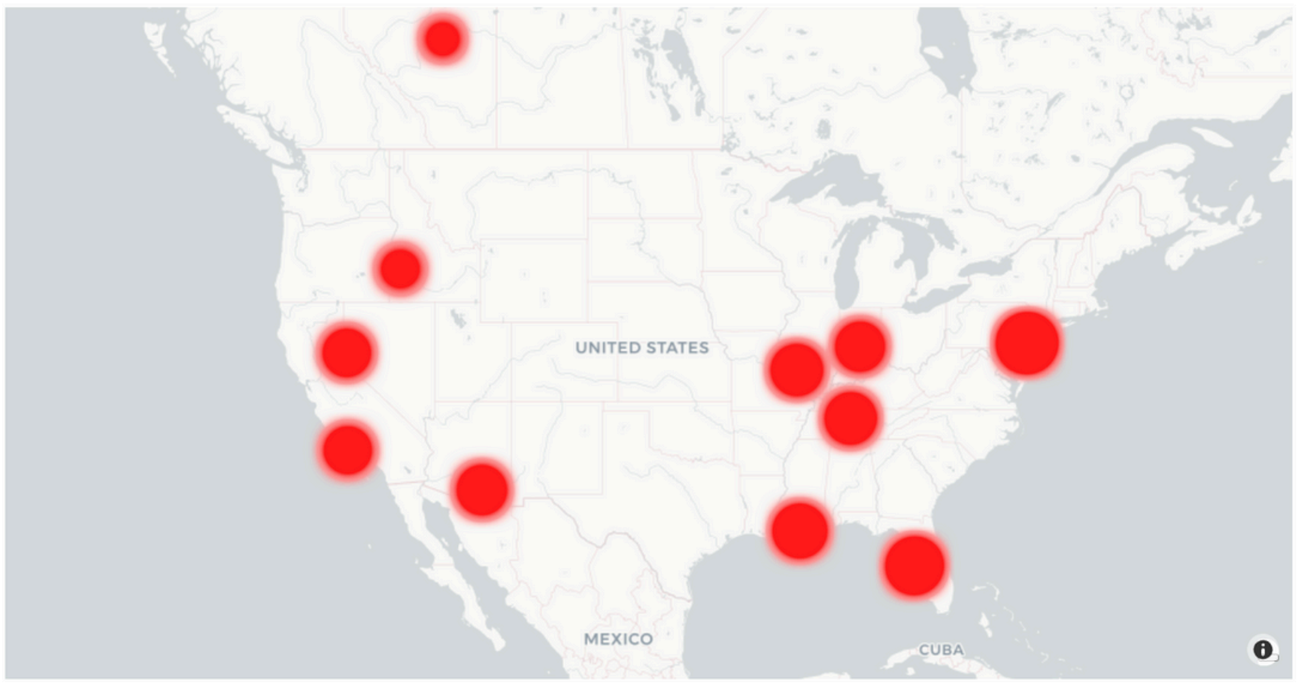
Average Rating

3.5

Average Reviews Per Restaurant

91

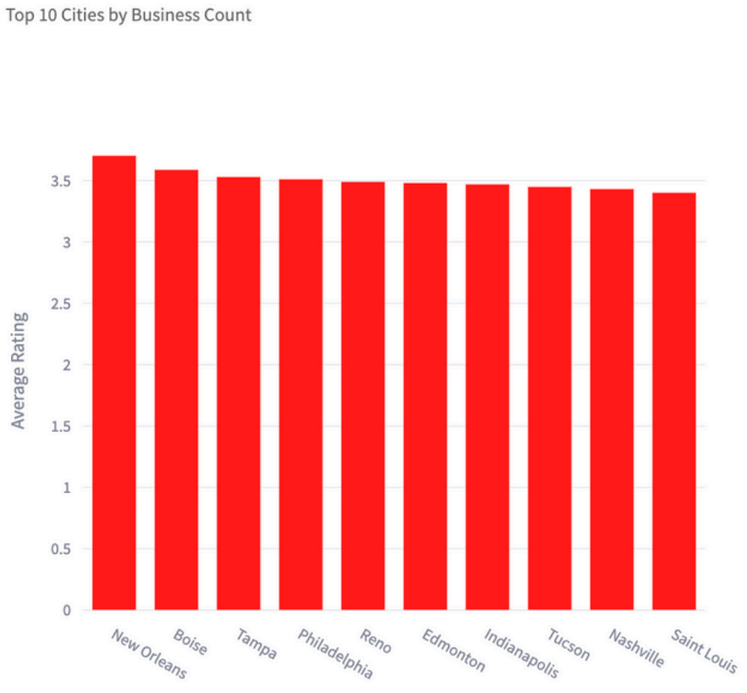
Review Density Map



Review density for all cities

Rating Distribution

Top Number of Businesses

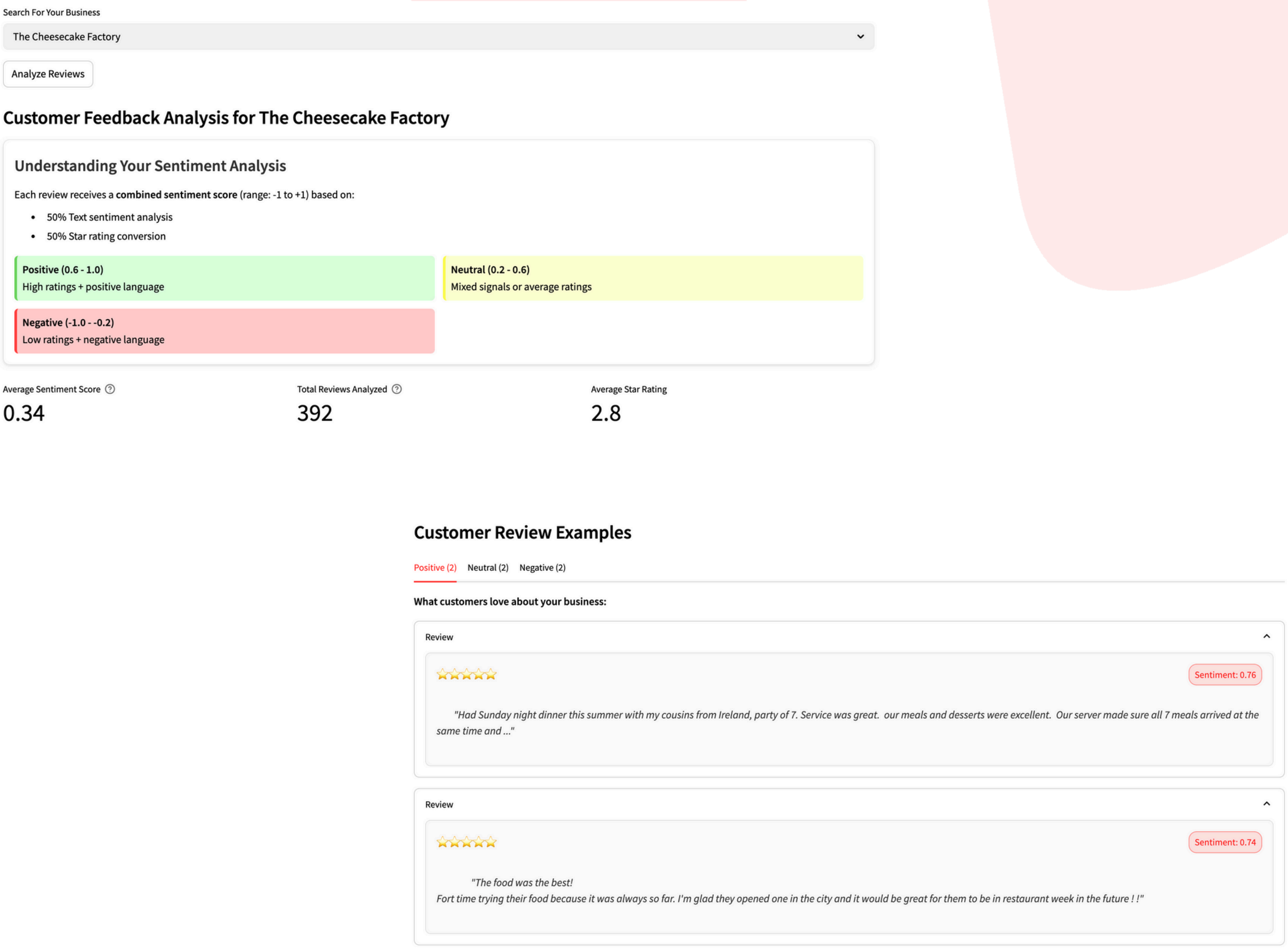


Visualization Design

Customer Reviews Analysis: Business Owner > Customer Reviews Analysis

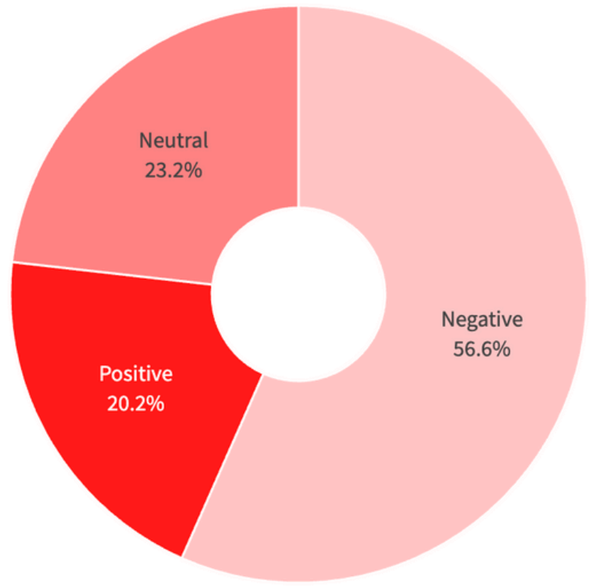
Visualizations and Areas of Interactivity

- Business search
- Sentiment distribution pie chart
    - Hover
- Sample Reviews
    - Positive, neutral, negative
- Rating distribution bar graph
    - Zoom, pan, hover

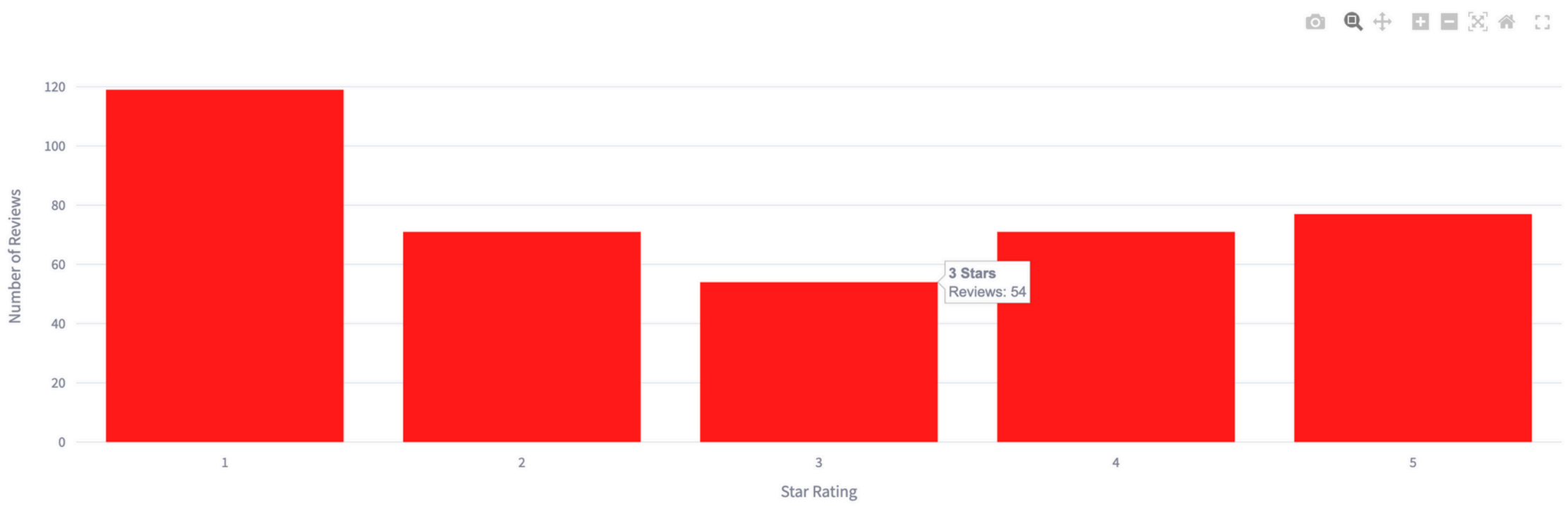


Visualization Design

Sentiment Distribution



Star Rating Distribution

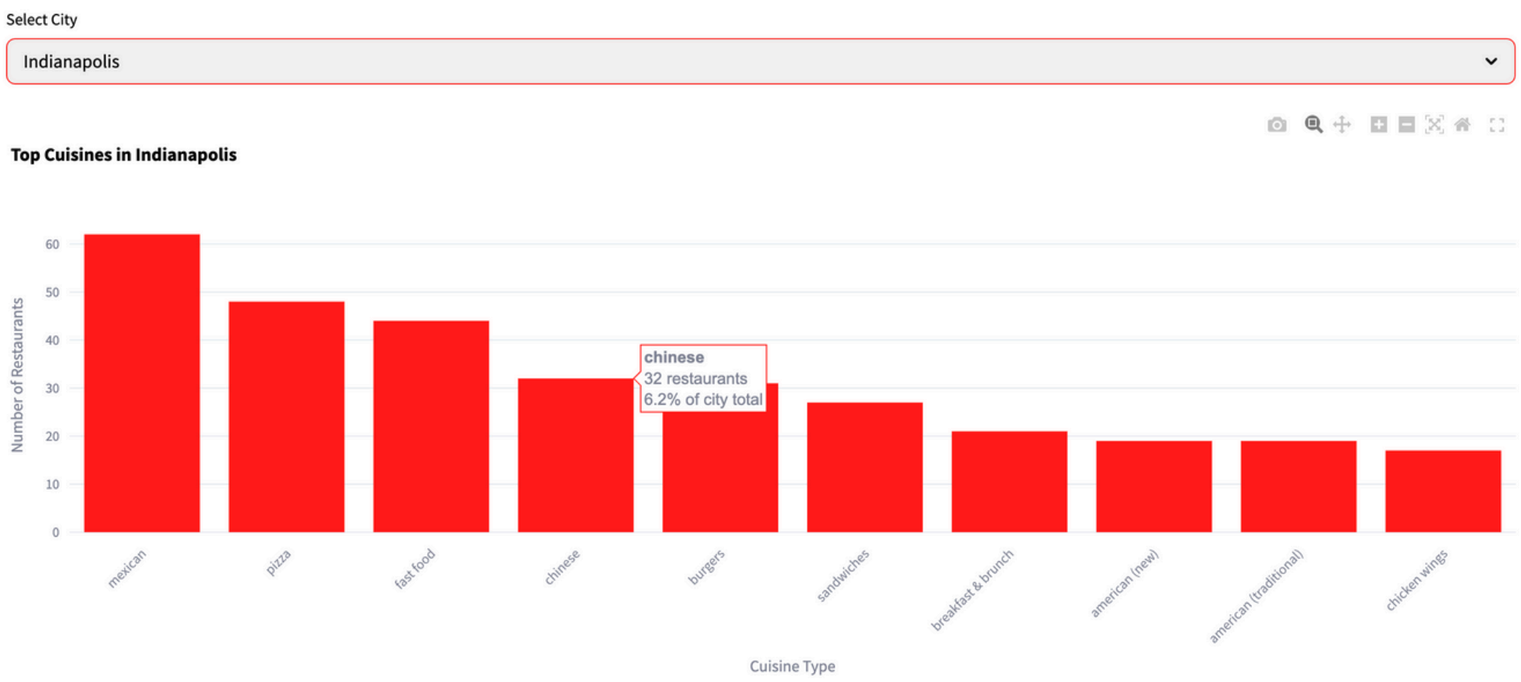


Visualization Design

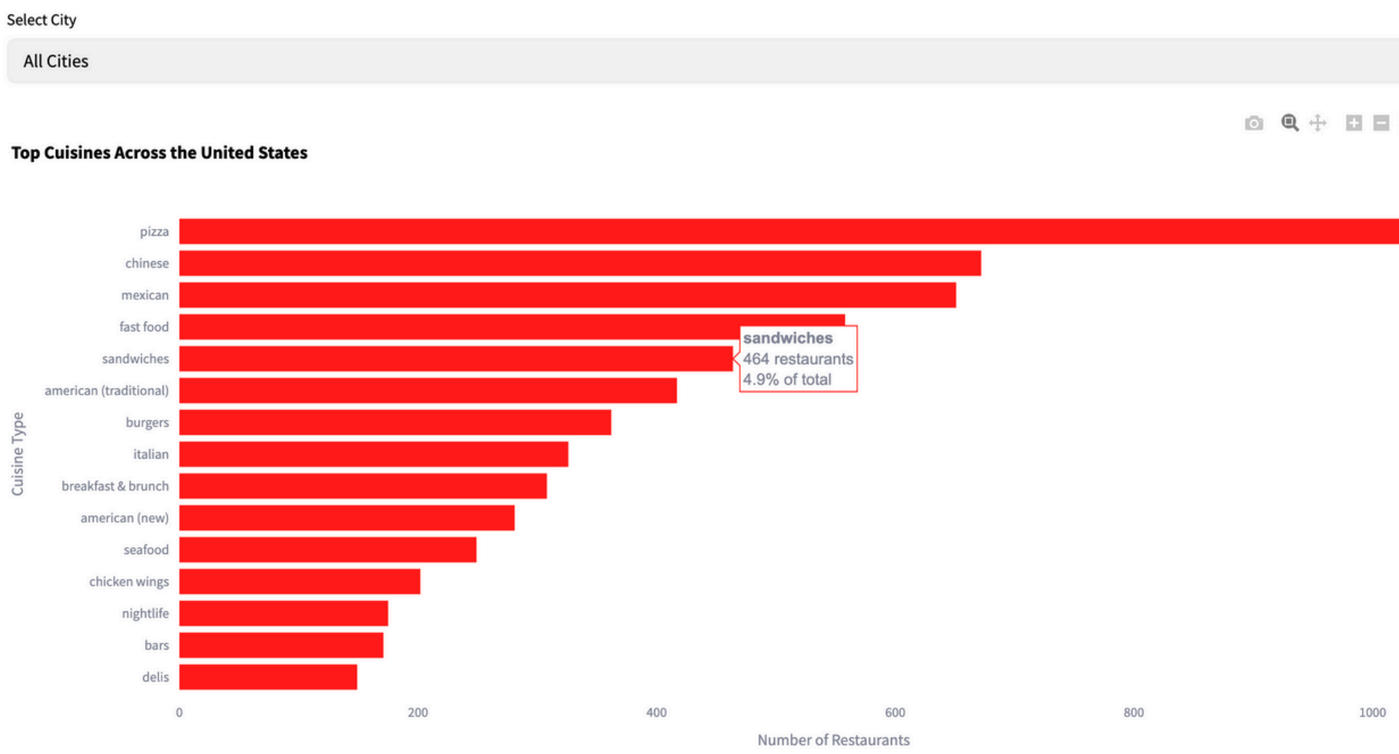
Top Cuisines: Consumer > Top Cuisines

Visualizations and Areas of Interactivity

- City dropdown choice and type to search
- Top cuisines bar graph
    - Zoom, pan, hover



Top cuisines in Indianapolis

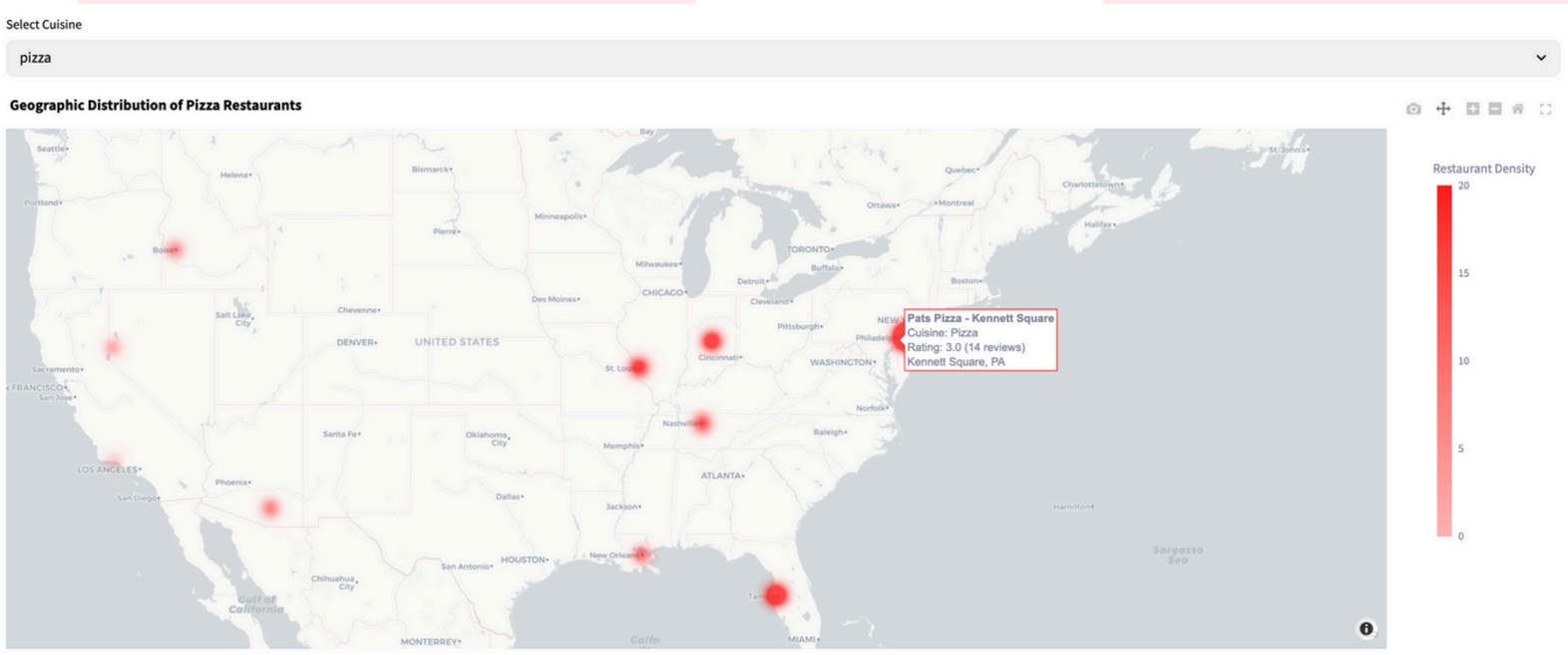


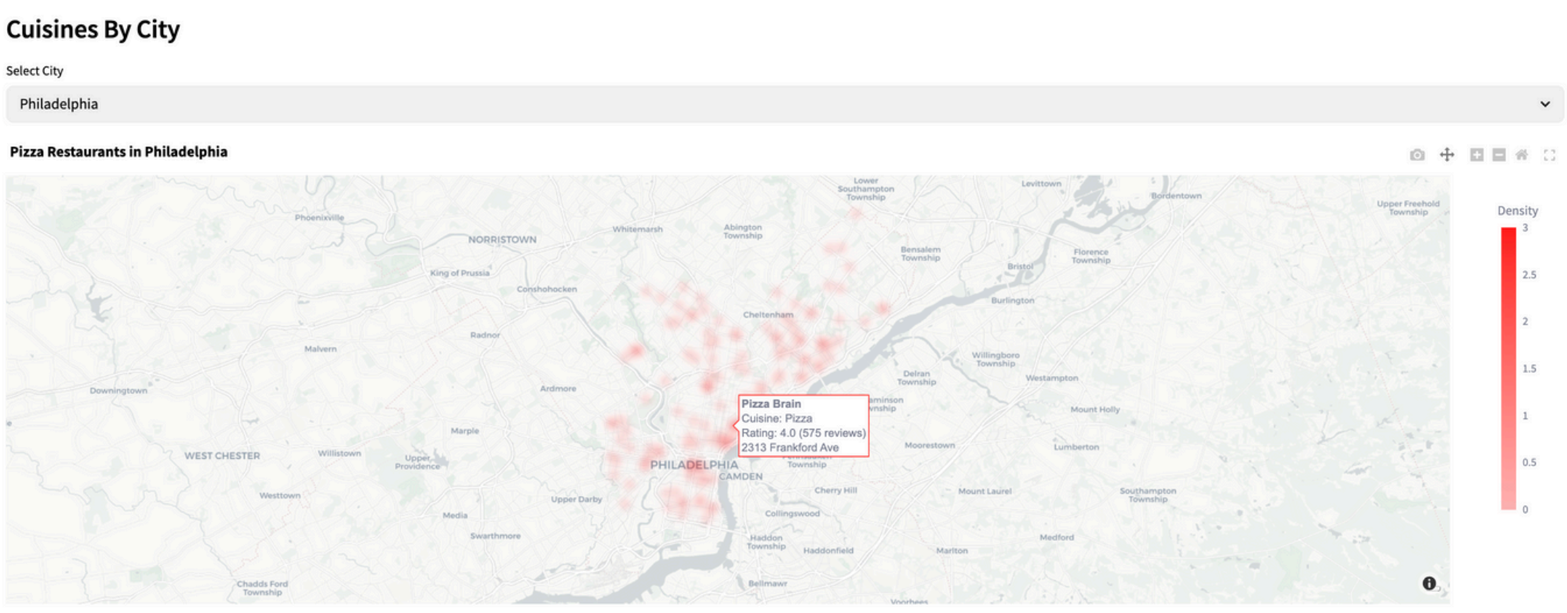
Visualization Design

Cuisine Distribution: Consumer > Cuisine Distribution

Visualizations and Areas of Interactivity

- Cuisine dropdown choice and type to search
- Geographic heatmap on number of restaurants for a given cuisine
    - Zoom, pan, hover
- City dropdown choice and type to search
- Geographic heatmap on number of restaurants for a given cuisine in a specific city
    - Zoom, pan, hover



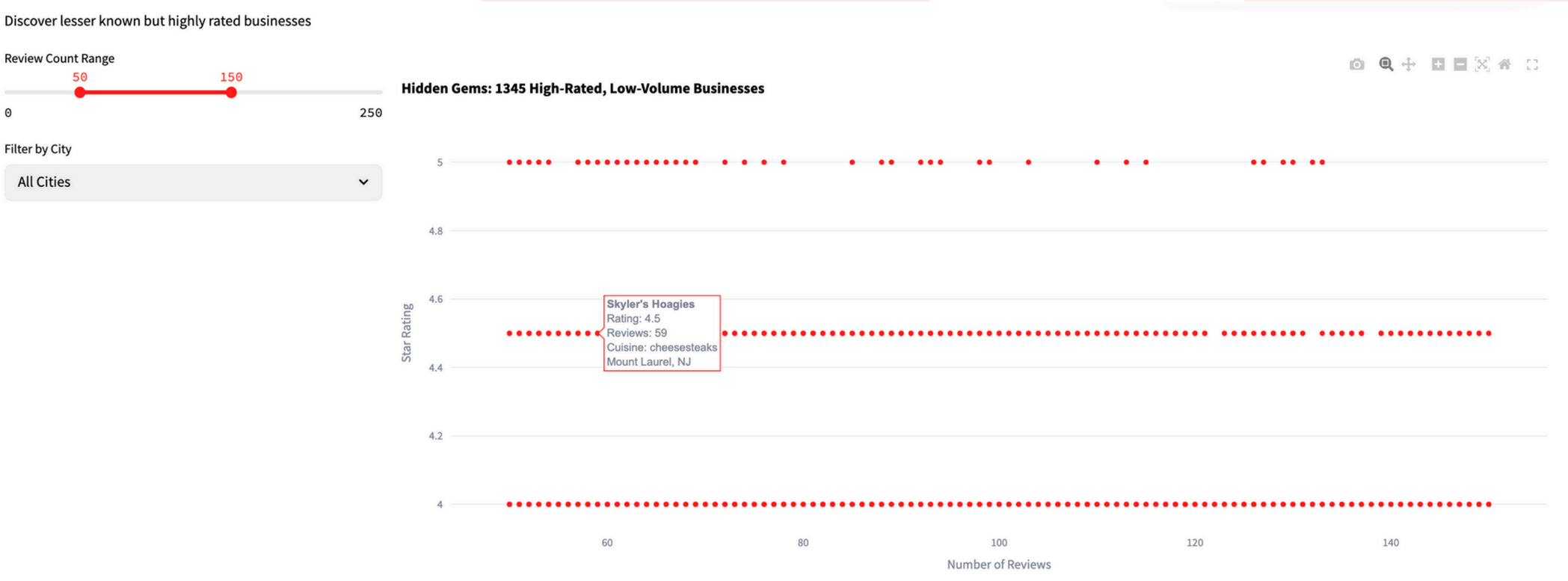


Visualization Design

Hidden Gems: Consumer > Hidden Gems

Visualizations and Areas of Interactivity

- Slider for minimum and maximum reviews
- Dropdown and type to search city
- Scatter plot of restaurants based on filers
    - Zoom, pan, hover



Found 1345 hidden gems:

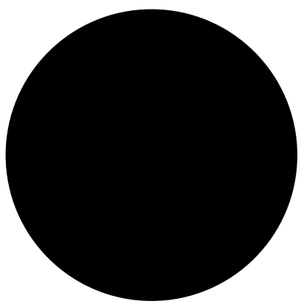
- Average rating: 4.2
- Average reviews: 90

Visualization Design

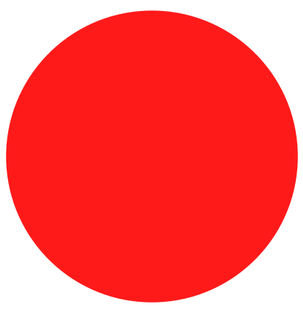
Design choices and accesibility

Colors:

Black



Red



- Brand recognition and alignment, makes project cohesive and clear
- High contrast colors make visualizations easy to see
- **Accessibility:** high contrast ratio, use of icons, shapes, labels in addition to color and text, analogous colors for visual hierarchy and visual comfortability

Visual encodings

- **Position:** height of bars in bar graph or scatter plot point positioning
- **Color saturation:** in geographic maps bright red represents high density and lower saturated reds represent lower density

Visualization Design