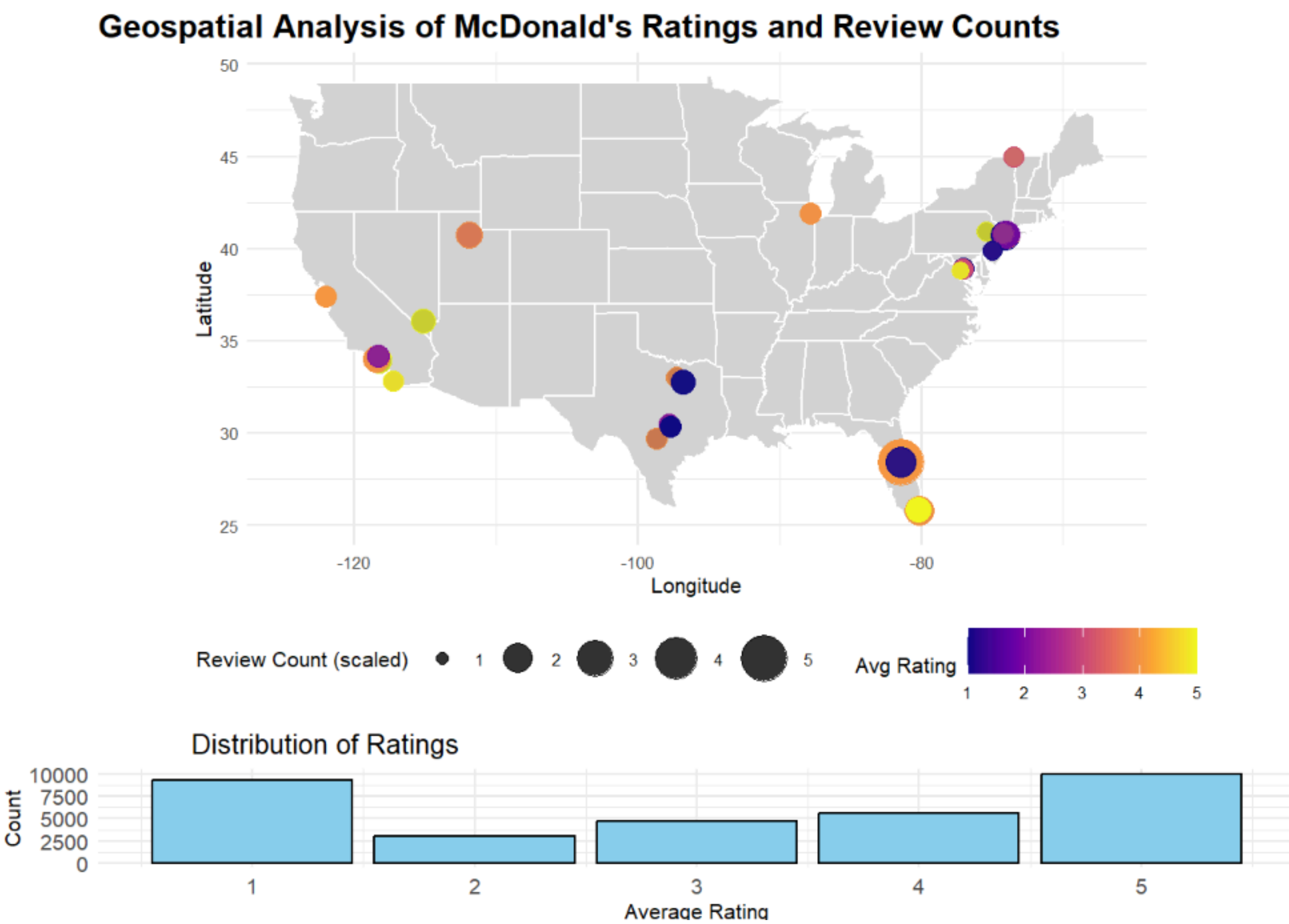
Mapping McDonald’s Customer Satisfaction: Geospatial Insights into Ratings and Reviews

Author: Aaron Yu

Visualization 1:



1. What is the essential question that your visualization is supposed to inform?

1. How are McDonald’s store ratings and review counts distributed geographically?

2. Are extreme ratings (1 or 5 stars) more common than moderate ratings (2, 3, or 4 stars)?
3. How do geographical patterns relate to the overall rating distribution?

---

2. How do aspects of your design support exploration of the essential question?

- **Design Choices:**
  - The **map** uses colors to represent average ratings and point sizes for review counts, helping visualize geographical trends.
  - The **bar chart** summarizes the overall rating distribution, showing a higher frequency of 1-star and 5-star ratings.
- **Trade-offs:**
  - The map focuses on ratings and review counts, omitting other metrics like sentiment score to keep it clear.
  - The bar chart doesn't differentiate regions, prioritizing a simple overview of rating distribution.

---

3. What are your key findings? How do they relate to your prior understanding?

- **Key Findings:**
  - High-rating stores (yellow) cluster in specific regions, while low ratings (purple) are dispersed, indicating regional performance differences.
  - Extreme ratings (1 and 5 stars) dominate, while moderate ratings are rare, showing polarized customer feedback.
  - Combining the map and bar chart highlights the interplay between regional patterns and overall trends.
- **Relation to Prior Understanding:**
  These findings align with the expectation of polarized reviews but reveal new insights, like regional clustering of high ratings.

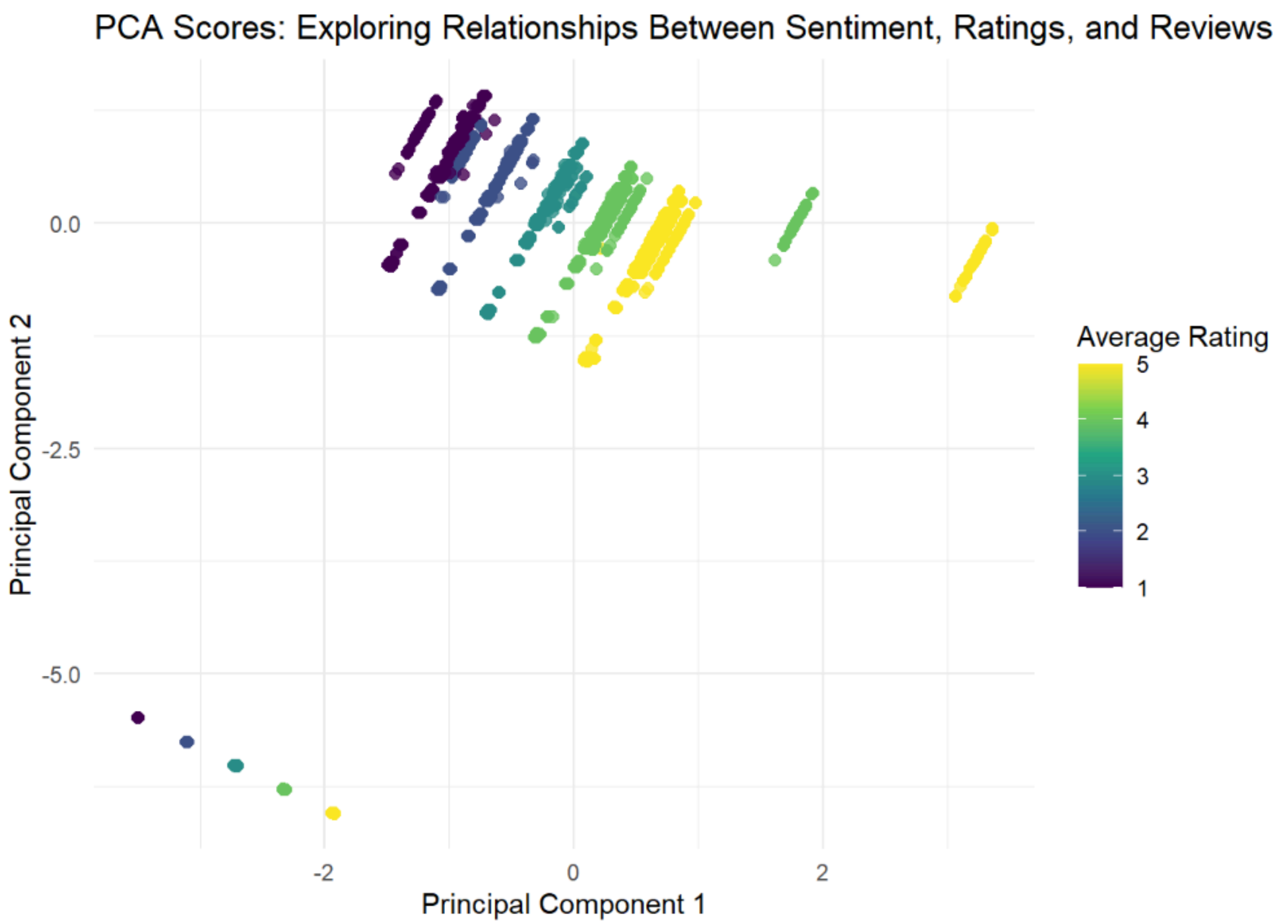
---

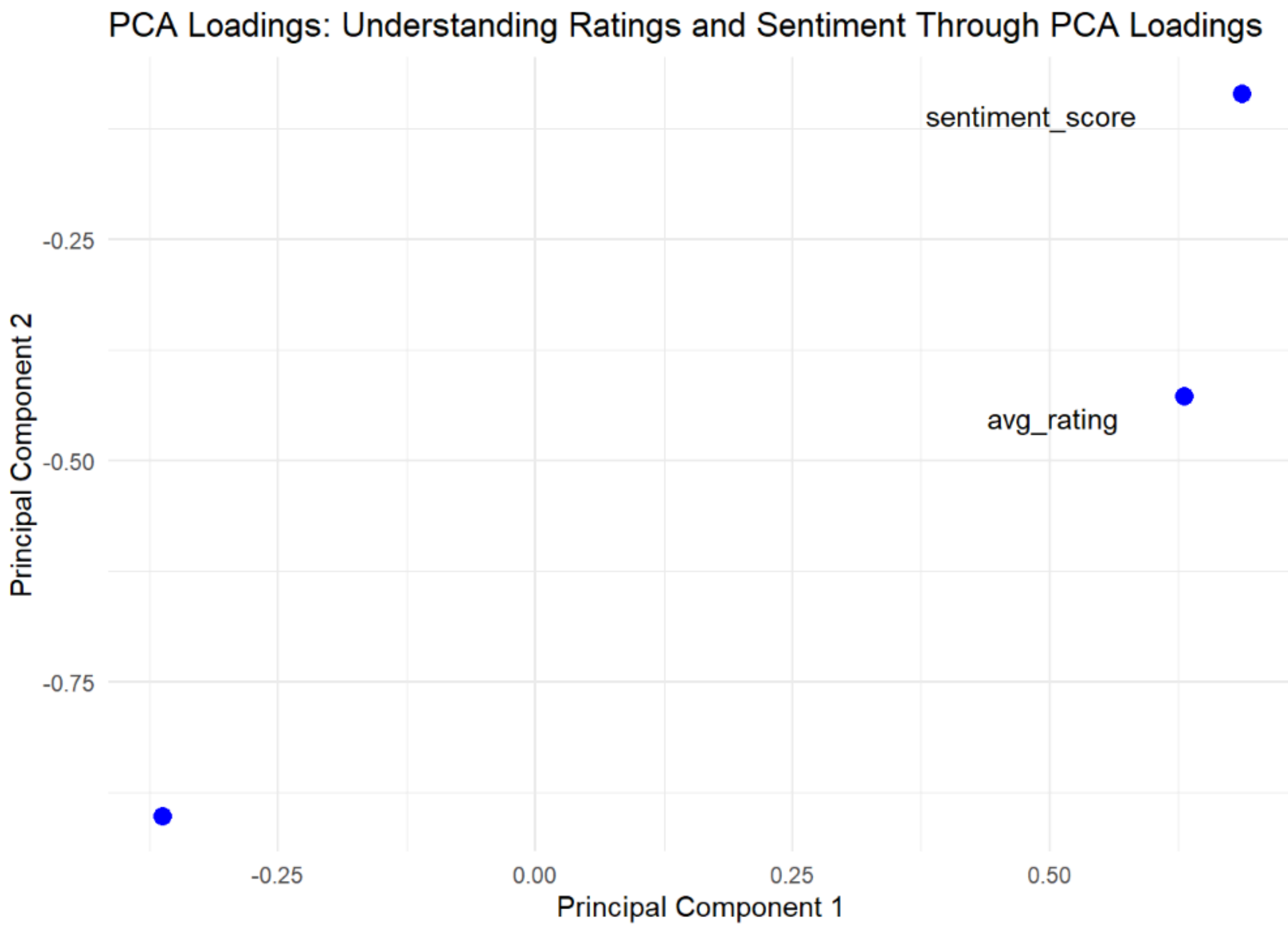
4. How did you create the visualizations? Were there any data preparation steps?

1. **Data Preparation:**
   - Converted ratings to numeric values and dropped missing geographic data.
   - Scaled review counts for proportional visualization and created an `sf` object for mapping.
2. **Visualization:**
   - The **map** displays geographic trends using `ggplot2`, with `avg_rating` mapped to color and scaled `rating_count` to size.
   - The **bar chart** summarizes rating frequencies.

- Both visualizations were combined using the `patchwork` package.

Visualization 2:






```
## Importance of components:
##                           PC1    PC2    PC3
## Standard deviation     1.1846 0.9799 0.7978
## Proportion of Variance 0.4678 0.3201 0.2122
## Cumulative Proportion  0.4678 0.7879 1.0000
```

1. What is the essential question that your visualization is supposed to inform?

- **PCA Scores Plot:** What are the patterns in the relationship between ratings, sentiment scores, and review counts? Are there clusters or trends?
- **PCA Loadings Plot:** How do ratings, sentiment, and review counts contribute to the principal components?
- **Variance Table:** How much variance is captured by the principal components, and is focusing on the first two sufficient?

---

2. How do aspects of your design support exploration of the essential question?

- **PCA Scores Plot:**
  - **Design:** Color encodes `avg_rating`, and PC1/PC2 axes show relationships across 78.79% of variance.
  - **Trade-offs:** Higher components are ignored, and review counts are indirectly represented.
- **PCA Loadings Plot:**
  - **Design:** Variable positions and labels show contributions to PC1/PC2.
  - **Trade-offs:** Focuses only on the first two components, excluding others.
- **Variance Table:**
  - **Design:** Summarizes explained variance to justify dimensionality reduction.
  - **Trade-offs:** A cumulative variance plot could better visualize trends.

---

3. What are your key findings? How do they relate to your prior understanding?

1. **PCA Scores Plot:**
   - PC1 (46.78%) is strongly linked to `avg_rating`, while PC2 (32.01%) captures `sentiment_score`.
   - High ratings cluster, while low ratings are more dispersed.
2. **PCA Loadings Plot:**
   - `avg_rating` heavily contributes to PC1; `sentiment_score` drives PC2.
   - `rating_count` has minimal influence on both components.
3. **Variance Table:**
   - The first two components explain 78.79% of the variance, validating the focus on them.

Relation to Prior Understanding: These findings confirm that ratings and sentiment are dominant factors but reveal new nuances, such as variability in low ratings.

---

4. How did you create the visualizations? Were there any data preparation steps?

1. **Data Preparation:**
   - Cleaned text reviews, calculated sentiment scores using the Bing lexicon, and replaced missing values with 0.
   - Selected `avg_rating`, `rating_count`, and `sentiment_score` for PCA and standardized them using `scale()`.
2. **Visualization Creation:**
   - **PCA Scores Plot:** Scatter plot of PC1 vs. PC2, colored by `avg_rating` using `ggplot2`.
   - **PCA Loadings Plot:** Loadings matrix visualized with variable labels to explain PC contributions.
   - **Variance Table:** Summarized the variance captured by components.

The PCA analysis uncovers relationships between ratings, sentiment, and review counts. PCA Scores Plot shows trends and clustering, while PCA Loadings Plot reveals variable contributions. Together, they effectively highlight the dataset's primary patterns, confirming that ratings and sentiment dominate customer feedback analysis.