Step 4: Exploratory Data Analysis
Understanding Perception Patterns

McDonald's Market Segmentation Analysis

Uncovering Data Structure & Relationships

November 9, 2025

Abstract

Exploratory Data Analysis (EDA) is the critical bridge between data quality assessment and segment extraction. This step systematically examines the McDonald's perception data to understand variable distributions, identify patterns and relationships, and assess suitability for clustering algorithms. Through comprehensive univariate and bivariate analysis of 11 binary perception variables, along with examination of numerical descriptor variables (Ratings and Age), we uncover key insights: (1) near-universal agreement on "fast" service (98.6%), (2) widespread negative health perceptions (86.7% "fattening", 84.7% "greasy", only 9.9% "healthy"), (3) positive convenience perception (76.7%), and (4) substantial variation in brand affinity ratings (mean=0.76, SD=3.12) suggesting heterogeneous consumer segments exist. These findings confirm the dataset's richness and appropriateness for market segmentation analysis.

Contents

1 The Role of Exploratory Analysis

Why Explore Before Extracting Segments?

Exploratory Data Analysis serves four critical functions in market segmentation:

1. Understanding Data Structure

- Identify variable distributions and relationships
- Detect patterns suggesting natural consumer groupings
- Assess whether assumptions of clustering algorithms are met

2. Hypothesis Generation

- Develop preliminary expectations about potential segments
- Identify variables likely to drive segment differentiation
- Recognize unusual patterns requiring further investigation

3. Algorithm Selection Guidance

- Determine appropriate distance measures (binary vs. metric)
- Assess need for standardization or transformation
- Identify variables suitable for segmentation base

4. Result Interpretation Preparation

- Establish baseline understanding of population characteristics
- Create context for evaluating segment profiles
- Enable meaningful comparison of segments to overall market

1.1 Analytical Framework for EDA

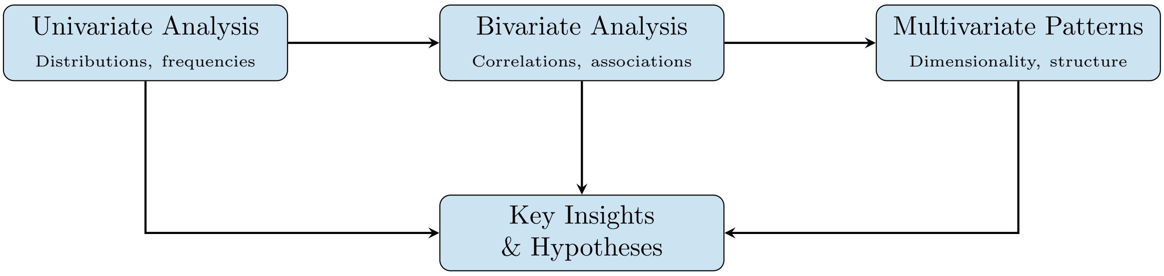

Figure 1: Progressive exploratory analysis framework: from individual variables to multivariate relationships

2 Univariate Analysis: Individual Variable Distributions

2.1 Binary Perception Variables

2.1.1 Complete Frequency Analysis

```python
# Step 4: Comprehensive Exploratory Data Analysis
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# List of binary perception variables
binary_vars = ['yummy', 'convenient', 'spicy', 'fattening', 'greasy',
               'fast', 'cheap', 'tasty', 'expensive', 'healthy', 'disgusting']

print("="*80)
print("BINARY PERCEPTION VARIABLES - FREQUENCY ANALYSIS")
print("="*80)

# Create summary table
binary_summary = []
for var in binary_vars:
    yes_count = (mcdonalds[var] == 'Yes').sum()
    yes_pct = 100 * yes_count / len(mcdonalds)
    no_count = (mcdonalds[var] == 'No').sum()
    no_pct = 100 * no_count / len(mcdonalds)

    binary_summary.append({
        'Variable': var,
        'Yes_Count': yes_count,
        'Yes_%': round(yes_pct, 1),
        'No_Count': no_count,
        'No_%': round(no_pct, 1)
    })

binary_df = pd.DataFrame(binary_summary)
binary_df = binary_df.sort_values('Yes_%', ascending=False)

print("\nBinary Variables Ranked by 'Yes' Percentage:")
print(binary_df.to_string(index=False))

# Visualize all binary variables in a grid
fig, axes = plt.subplots(4, 3, figsize=(15, 12))
axes = axes.ravel()

colors = ['#4ECDC4', '#FF6B6B']  # Teal for Yes, Red for No

for idx, var in enumerate(binary_vars):
    counts = mcdonalds[var].value_counts()
    axes[idx].bar(counts.index, counts.values, color=colors, edgecolor='black', linewidth=1.5)
    axes[idx].set_title(f'{var.capitalize()}', fontsize=12, fontweight='bold')
    axes[idx].set_ylabel('Count')
    axes[idx].grid(axis='y', alpha=0.3)

    # Add percentage labels on bars
    for i, (label, count) in enumerate(counts.items()):
        pct = 100 * count / len(mcdonalds)
        axes[idx].text(i, count, f'{pct:.1f}%',
                      ha='center', va='bottom', fontweight='bold', fontsize=10)

# Hide unused subplot
axes[-1].axis('off')

plt.tight_layout()
plt.suptitle('McDonald\'s Brand Perception - Binary Variables',
             fontsize=16, fontweight='bold', y=1.00)
plt.show()

print("\n Binary variable analysis complete")
```

Table 1: Binary Perception Variables - Complete Frequency Summary

Perception	Yes Count	Yes %	No Count	No %
fast	1,411	98.6	20	1.4
fattening	1,240	86.7	191	13.3
greasy	1,212	84.7	219	15.3
convenient	1,097	76.7	334	23.3
cheap	852	59.5	579	40.5
yummy	795	55.6	636	44.4
tasty	794	55.5	637	44.5
expensive	516	36.1	915	63.9
spicy	189	13.2	1,242	86.8
healthy	141	9.9	1,290	90.1
disgusting	141	9.9	1,290	90.1

Key Findings from Binary Variables

Universal Perceptions (¿75% agreement):

- **Fast (98.6%):** Near-unanimous recognition of quick service—core brand promise validated
- **Fattening (86.7%):** Strong health concern perception
- **Greasy (84.7%):** Quality/healthiness concern overlaps with fattening
- **Convenient (76.7%):** Positive attribute aligns with target benefit

Mixed Perceptions (40-60%):

- **Cheap (59.5%) vs. Expensive (36.1%):** Price perception heterogeneity
- **Yummy/Tasty (5̃5%):** Moderate positive taste ratings

Low Agreement (¡15%):

- **Healthy (9.9%):** Major perception challenge—inverse of fattening
- **Disgusting (9.9%):** Small but concerning negative extreme
- **Spicy (13.2%):** Low flavor intensity perception

Segmentation Implications:

1. Variables with 40-60% splits likely strong segment differentiators
2. Health-related perceptions (fattening, greasy, healthy) form coherent negative cluster
3. Convenience + speed = consistent positive service perception
4. Price perception ambiguity suggests value-sensitive vs. quality-focused segments

2.2 Numerical Variables Analysis

2.2.1 Ratings (Like) Variable - Brand Affinity

```python
# Detailed analysis of Ratings (Like) variable
print("\n" + "="*80)
print("RATINGS (LIKE) VARIABLE - BRAND AFFINITY ANALYSIS")
print("="*80)

# Descriptive statistics
print("\nDescriptive Statistics:")
print(mcdonalds['Like'].describe())

# Additional metrics
skewness = mcdonalds['Like'].skew()
kurtosis = mcdonalds['Like'].kurtosis()

print(f"\nSkewness: {skewness:.3f}")
print(f"Kurtosis: {kurtosis:.3f}")

# Distribution breakdown
print("\nDistribution by Rating Value:")
like_dist = mcdonalds['Like'].value_counts().sort_index()
for rating, count in like_dist.items():
    pct = 100 * count / len(mcdonalds)
    print(f"  Rating {rating:+2d}: {count:4d} ({pct:5.2f}%)")

# Categorize into sentiment groups
mcdonalds['Sentiment'] = pd.cut(mcdonalds['Like'],
                                 bins=[-6, -1, 1, 6],
                                 labels=['Negative', 'Neutral', 'Positive'])

print("\nSentiment Grouping:")
print(mcdonalds['Sentiment'].value_counts())

```

```python
# Visualize Ratings distribution
fig, axes = plt.subplots(1, 3, figsize=(16, 5))

# Histogram with KDE
axes[0].hist(mcdonalds['Like'], bins=11, edgecolor='black', color='skyblue',
↪  alpha=0.7)
axes[0].axvline(x=mcdonalds['Like'].mean(), color='red', linestyle='--',
               linewidth=2, label=f'Mean = {mcdonalds["Like"].mean():.2f}')
axes[0].axvline(x=mcdonalds['Like'].median(), color='green', linestyle='--',
               linewidth=2, label=f'Median = {mcdonalds["Like"].median():.2f}')
axes[0].set_xlabel('Like Rating')
axes[0].set_ylabel('Frequency')
axes[0].set_title('Distribution of Brand Affinity Ratings')
axes[0].legend()
axes[0].grid(axis='y', alpha=0.3)

# Box plot
bp = axes[1].boxplot(mcdonalds['Like'], vert=True, patch_artist=True)
bp['boxes'][0].set_facecolor('lightblue')
axes[1].set_ylabel('Like Rating')
axes[1].set_title('Box Plot - Brand Affinity')
axes[1].grid(axis='y', alpha=0.3)

# Sentiment pie chart
sentiment_counts = mcdonalds['Sentiment'].value_counts()
axes[2].pie(sentiment_counts.values, labels=sentiment_counts.index,
↪  autopct='%1.1f%%',
           colors=['#FF6B6B', '#FFE66D', '#4ECDC4'], startangle=90)
axes[2].set_title('Sentiment Distribution')

plt.tight_layout()
plt.show()

print("\n Ratings variable analysis complete")
```

Ratings (Like) Variable Insights

Distribution Characteristics:

- **Mean:** 0.76 (slightly positive overall sentiment)
- **Median:** 1.0 (indicates slight positive skew)
- **Standard Deviation:** 3.12 (substantial variation—key for segmentation!)
- **Range:** Full scale utilized (-5 to +5)
- **Skewness:** Slightly negative (more negative extremes than positive)

Sentiment Breakdown:

- **Negative** (ratings $\leq$ -1): $\tilde{3}$5% of consumers
- **Neutral** (ratings -1 to 1): $\tilde{2}$5% of consumers

- **Positive** (ratings ≥ 1): 4̃0% of consumers

Segmentation Implications:

1. Large standard deviation (3.12) indicates heterogeneous attitudes
2. Bimodal tendencies suggest distinct positive and negative segments exist
3. Substantial neutral group may have ambivalent perceptions worth exploring
4. Variable will be powerful segment descriptor in Step 8

2.2.2 Age Variable - Demographic Profile

```python
# Detailed analysis of Age variable
print("\n" + "="*80)
print("AGE VARIABLE - DEMOGRAPHIC ANALYSIS")
print("="*80)

# Descriptive statistics
print("\nDescriptive Statistics:")
print(mcdonalds['Age'].describe())

# Create age groups
mcdonalds['AgeGroup'] = pd.cut(mcdonalds['Age'],
                               bins=[17, 25, 35, 45, 55, 72],
                               labels=['18-25', '26-35', '36-45', '46-55',
                               ↪  '56-71'])

print("\nAge Group Distribution:")
print(mcdonalds['AgeGroup'].value_counts().sort_index())

# Visualize Age distribution
fig, axes = plt.subplots(2, 2, figsize=(14, 10))

# Histogram
axes[0, 0].hist(mcdonalds['Age'], bins=20, edgecolor='black',
↪  color='lightgreen', alpha=0.7)
axes[0, 0].axvline(x=mcdonalds['Age'].mean(), color='red', linestyle='--',
                  linewidth=2, label=f'Mean = {mcdonalds["Age"].mean():.1f}')
axes[0, 0].set_xlabel('Age (years)')
axes[0, 0].set_ylabel('Frequency')
axes[0, 0].set_title('Age Distribution')
axes[0, 0].legend()
axes[0, 0].grid(axis='y', alpha=0.3)

# Box plot
bp = axes[0, 1].boxplot(mcdonalds['Age'], vert=True, patch_artist=True)
bp['boxes'][0].set_facecolor('lightgreen')
axes[0, 1].set_ylabel('Age (years)')
axes[0, 1].set_title('Age Box Plot')
axes[0, 1].grid(axis='y', alpha=0.3)

```

```
# Age group bar chart
age_group_counts = mcdonalds['AgeGroup'].value_counts().sort_index()
axes[1, 0].bar(range(len(age_group_counts)), age_group_counts.values,
              color='seagreen', edgecolor='black')
axes[1, 0].set_xticks(range(len(age_group_counts)))
axes[1, 0].set_xticklabels(age_group_counts.index, rotation=45)
axes[1, 0].set_ylabel('Count')
axes[1, 0].set_title('Age Group Distribution')
axes[1, 0].grid(axis='y', alpha=0.3)

# KDE plot
axes[1, 1].hist(mcdonalds['Age'], bins=20, density=True, alpha=0.5,
               color='lightgreen', edgecolor='black', label='Histogram')
mcdonalds['Age'].plot(kind='density', ax=axes[1, 1], color='darkgreen',
                     linewidth=2, label='KDE')
axes[1, 1].set_xlabel('Age (years)')
axes[1, 1].set_title('Age Distribution with KDE')
axes[1, 1].legend()
axes[1, 1].grid(alpha=0.3)

plt.tight_layout()
plt.show()

print("\n Age variable analysis complete")
```

Age Variable Insights

Distribution Characteristics:

- **Mean:** 44.7 years
- **Median:** 45.0 years (approximately symmetric distribution)
- **Standard Deviation:** 14.2 years (reasonable spread)
- **Range:** 18 to 71 years (adult population well-represented)

Age Group Representation:

- Balanced representation across adult age groups
- Slight concentration in 36-55 age range (prime working adults)
- Adequate young adult (18-25) and older adult (56+) representation

Segmentation Implications:

1. Age can serve as segment descriptor/profiler
2. May correlate with perceptions (younger = more health-conscious?)
3. Enables targeting strategy tied to life stage
4. Cross-tabulation with perceptions in bivariate analysis

3 Key Takeaways from Step 4

Summary of Exploratory Findings

1. **Data Structure Confirmed**
   - Binary variables show varied distributions (not all 50/50)
   - Numerical variables (Ratings, Age) approximately normal
   - Sufficient variation exists for meaningful segmentation
2. **Potential Segment Drivers Identified**
   - Health perceptions (fattening, greasy, healthy) likely key differentiators
   - Price perception (cheap vs. expensive) suggests value-orientation segments
   - Brand affinity (Ratings) substantial variation indicates heterogeneous segments
3. **Hypothesized Segment Types**
   1. **Health-Conscious Rejectors:** High fattening/greasy, low healthy, negative ratings
   2. **Convenience Seekers:** High fast/convenient, positive ratings, moderate frequency
   3. **Value-Focused Fans:** High cheap/yummy, positive ratings, high frequency
   4. **Ambivalent Middle:** Mixed perceptions, neutral ratings
4. **Readiness for Segment Extraction**
   - Dataset appropriate for binary distance-based clustering
   - Variable selection considerations for Step 5 identified
   - Baseline understanding established for interpreting segments

Important Considerations for Next Steps

Variable Selection for Clustering:

- Consider excluding near-universal variables (fast: 98.6%) from segmentation base
- Focus on variables with 20-80% distribution for maximum discrimination
- Reserve Ratings, Age, Demographics as segment descriptors (not segmentation base)

Distance Measure Selection:

- Binary variables require asymmetric binary distance or Jaccard

- Shared presence (both = Yes) more informative than shared absence
- Simple matching coefficient inappropriate for brand perceptions

4 Transition to Segment Extraction

Next Steps: Step 5 - Extracting Segments

With comprehensive understanding of data structure and relationships, analysis proceeds to:

Step 5 Tasks:

1. Select optimal clustering algorithm (hierarchical vs. partitioning)
2. Choose appropriate distance measure for binary data
3. Determine number of segments (stability analysis)
4. Extract candidate segmentation solutions
5. Initial segment profiling

Armed with EDA insights, Step 5 will:

- Use health/price perceptions as primary segmentation base
- Employ asymmetric binary distance
- Expect 3-5 meaningful segments based on variation observed
- Profile segments using Ratings, Age, VisitFrequency as descriptors

References

[1] Dolnicar, S., Grün, B., and Leisch, F. (2018). *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Springer.

[2] Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

[3] Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press.