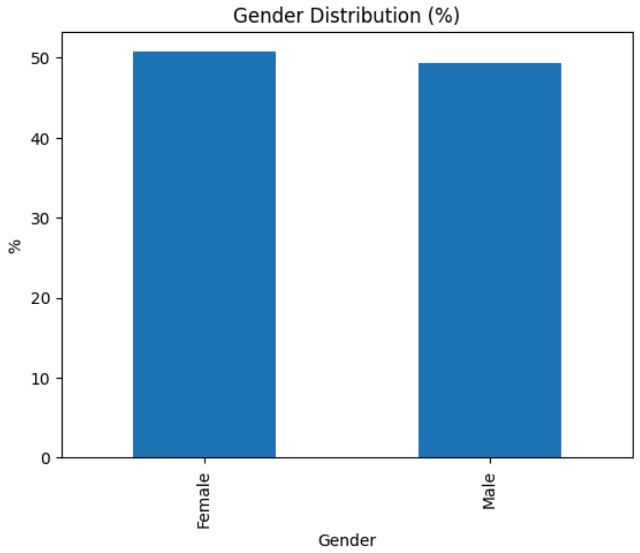
Demographic Profiling (Descriptive Only)

1. Gender Distribution

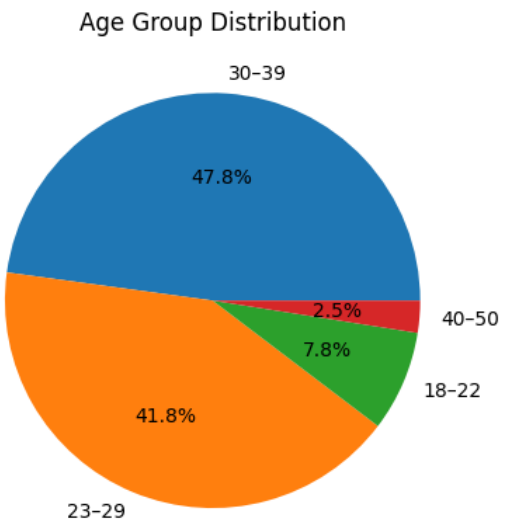
Description:
This chart shows the percentage distribution of male and female participants in the total sample. It helps us understand the gender balance in our dataset and verify alignment with the defined sample quota.



Intrepretation :

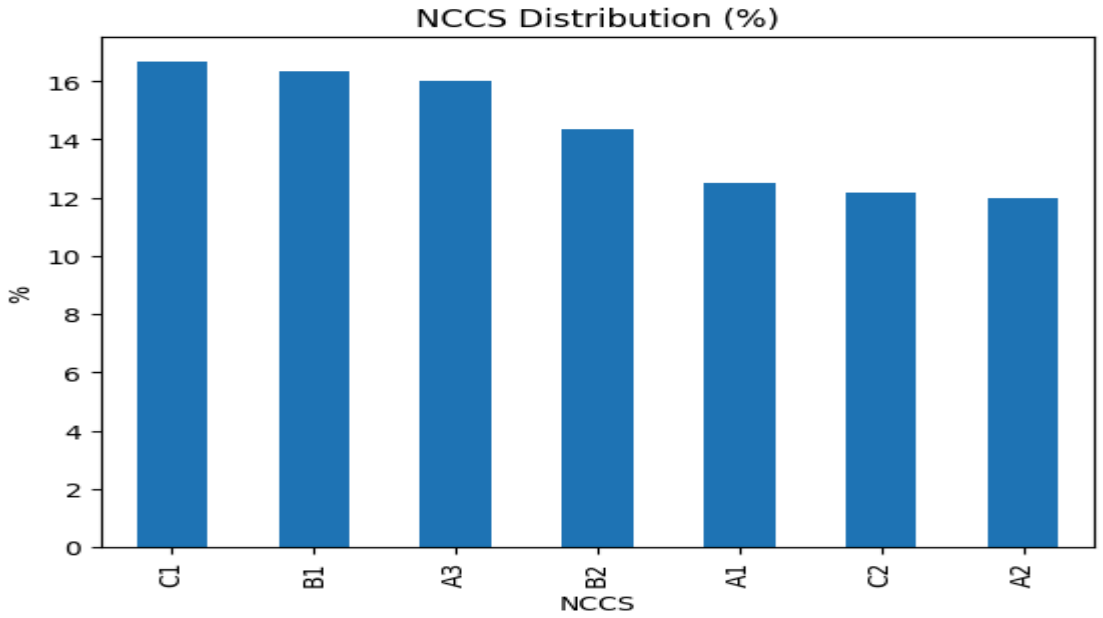
2. Age Group Distribution

Description:
The age distribution is segmented into four predefined bands (18–22, 23–29, 30–39, 40–50). This view provides clarity on the dominant age cohorts within the sample, which can be relevant for segmentation and targeting.



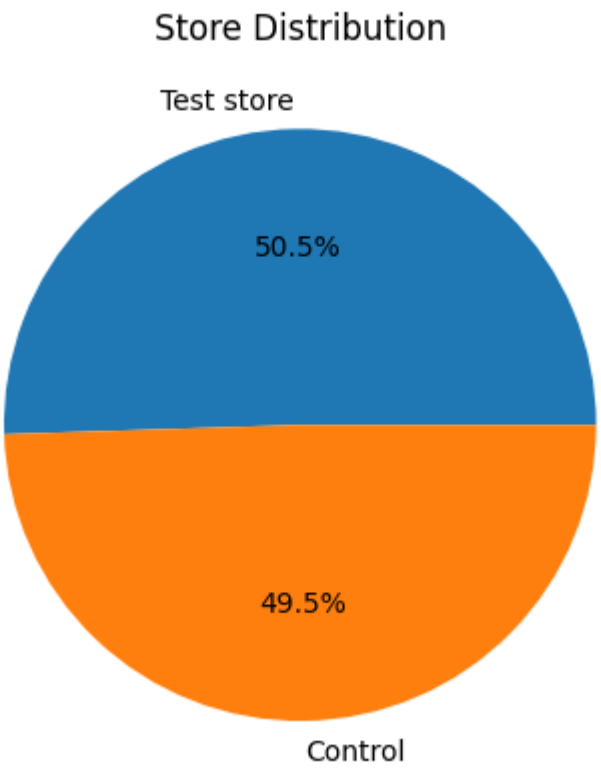
3. NCCS (Socioeconomic Classification)

Description:
This bar chart represents the distribution of participants by their NCCS classification. It helps assess the socioeconomic representation of the sample and its alignment with the research objectives.



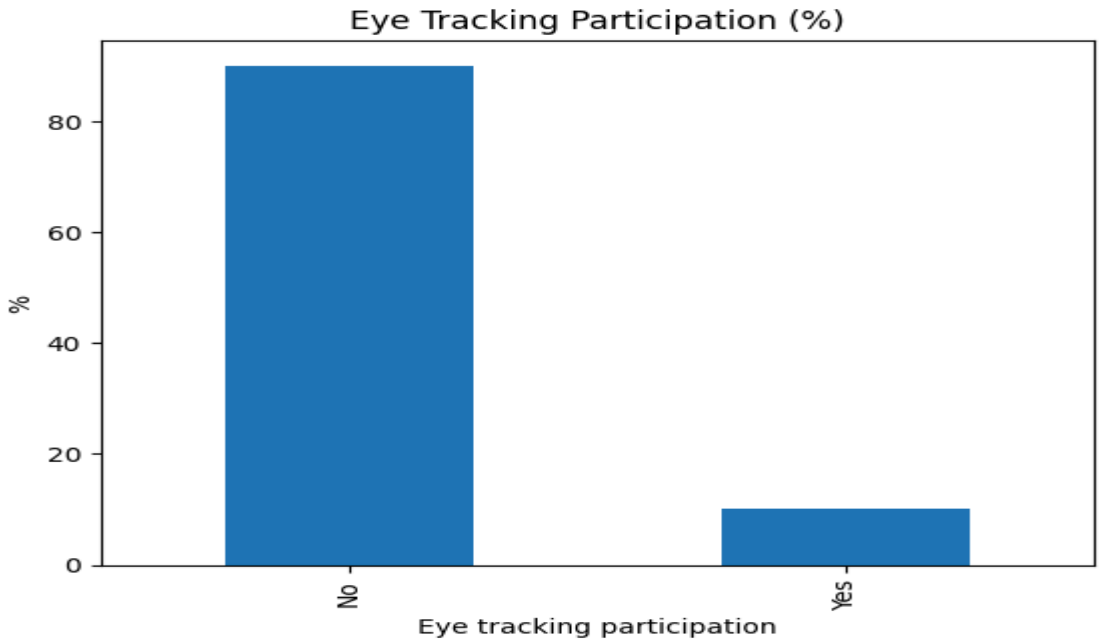
4. Store Type Distribution (Test vs Control)

Description:
This pie chart shows how the sample is distributed between the two store environments – Test (new packaging) and Control (existing packaging). It verifies sample balancing between test conditions.



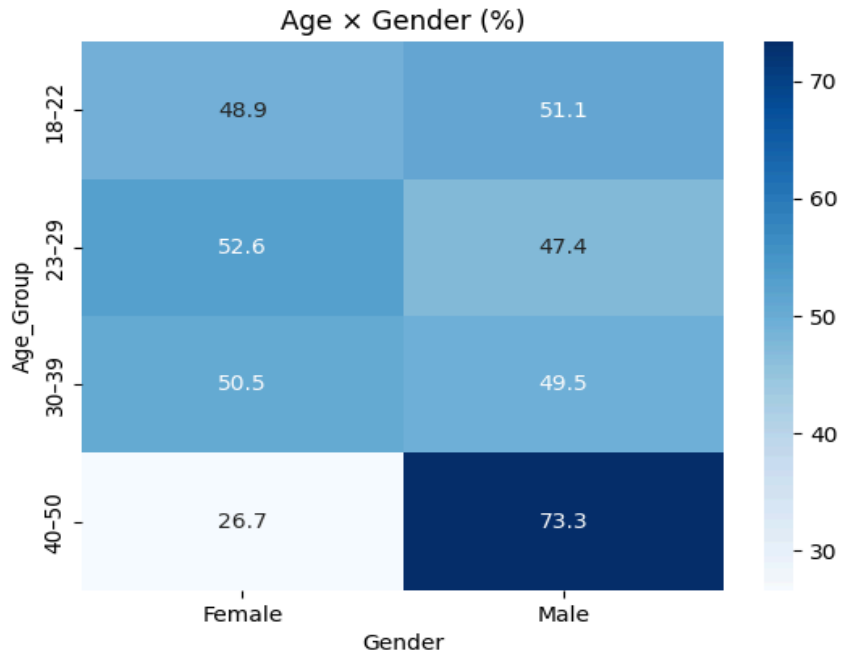
5. Eye-Tracking Participation

Description:
This visual highlights how many respondents participated in the eye-tracking module. It helps identify the size of the subset for deep packaging attention analysis.



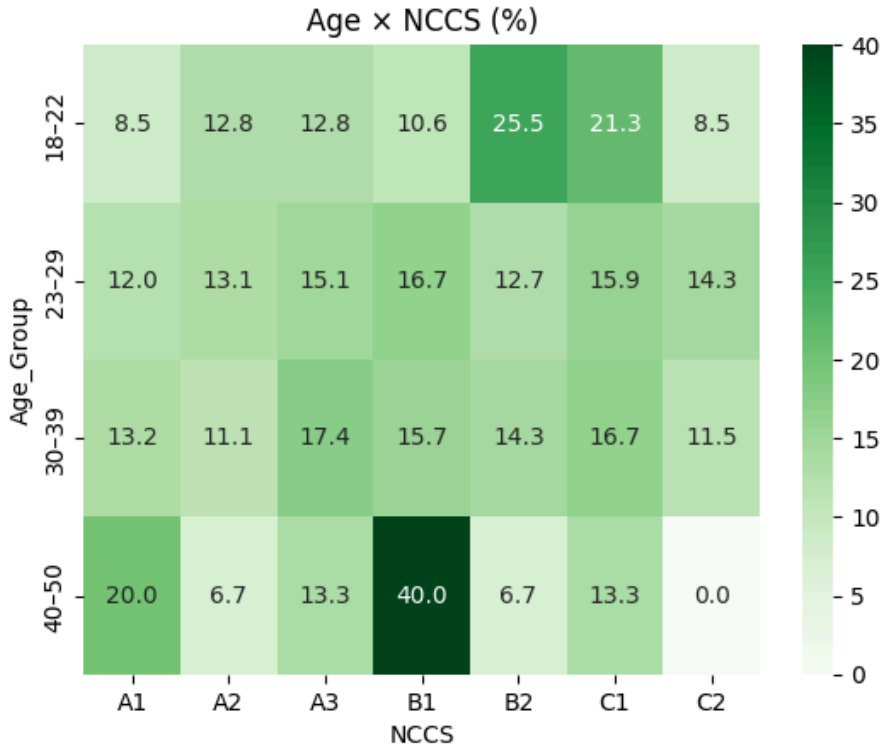
6. Age × Gender Crosstab

Description:
This heatmap shows how different age groups are distributed across gender. It provides insights into the demographic structure and potential bias or skew across age-gender combinations.



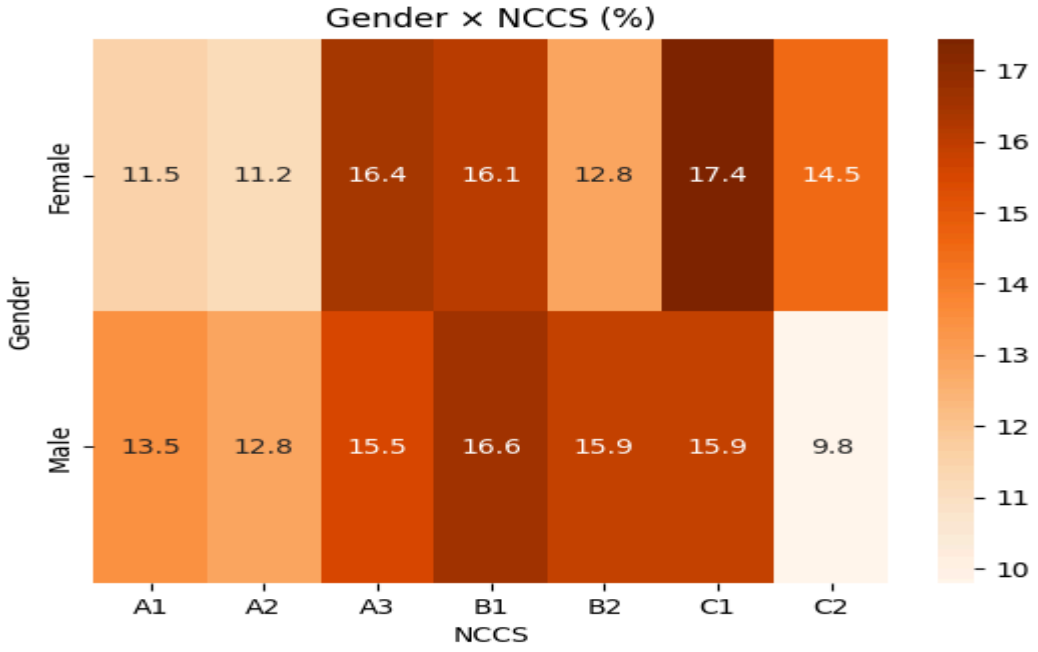
7. Age × NCCS Crosstab

Description:
This visual cross-tabulates age groups with NCCS classification. It helps in analyzing whether younger or older participants skew towards a specific socioeconomic category.



8. Gender × NCCS Crosstab

Description:
This heatmap presents the NCCS distribution by gender, helping identify any potential overrepresentation of a specific gender within certain socioeconomic classes.



Data Quality Checks Summary

Description

This section provides an overview of the data hygiene metrics conducted prior to analysis. The checks include validation for:

- Missing values
- Duplicate respondents
- Quota distribution skews
- Demographic balance
- Basic statistical validation

These steps ensure the dataset is clean, balanced, and statistically reliable.

---

1. Missing Data Check

Variable	Missing Values
Age	0
Gender	0
NCCS	0
Store	0

Interpretation:
There are no missing data entries across key demographic variables. This indicates a fully complete dataset, ready for analysis without imputation.

---

2. Duplicate Respondent Check

- **Duplicates Found:** `0`

Interpretation:
No duplicate records were identified. This confirms that all entries are unique and participant integrity is maintained.

---

3. Quota Distribution

Gender Distribution

Gender	Proportion
Female	50.67%
Male	49.33%

Interpretation:
The gender distribution is well-balanced, with near-equal representation from male and female respondents.

NCCS Distribution

NCCS	Proportion
C1	16.67%
B1	16.33%
A3	16.00%
B2	14.33%
A1	12.50%
C2	12.17%
A2	12.00%

Interpretation:
The NCCS (socio-economic classification) categories are fairly distributed across the sample. No segment is disproportionately represented, supporting diverse consumer profiling.

Store Distribution

Store Type	Proportion

Test Store	50.5%
Control Store	49.5%

Interpretation:
Store distribution is nearly evenly split, strengthening the validity of A/B testing and ensuring unbiased comparison between the two experimental conditions.

---

4. Basic Statistical Validations

- **Correlation (Brand Recall vs. Age):** `0.03`
  Interpretation: No meaningful linear correlation observed between respondent age and brand recall.

- **T-test: Purchase Intent by Gender**

  - **p-value:** `0.7087`
    Interpretation: No statistically significant difference in purchase intent between male and female respondents ($p > 0.05$). Gender does not influence intent to purchase in this sample.

Summary of Data Quality Status

Metric	Status
Missing Data	None
Duplicate Responses	None
Gender Balance	Balanced
NCCS Quota Coverage	Adequate
Store Group Split	Balanced
Age vs. Recall Correlation	Weak (0.03)
Gender Purchase Intent Bias	Not Significant

Description:

Why This Matters

Clean, reliable data is essential for any analysis. Without proper hygiene checks:

- Results can be misleading due to missing or duplicate data.
- Quotas might be off-target, invalidating segmentation insights.
- Demographic imbalances can bias conclusions.
- Statistical assumptions may not hold, reducing credibility of findings.

This step protects the **validity and credibility** of the entire research project.

How to Do It

✅ **Step 1: Check for Missing Values**
python
CopyEdit

```
# Identify missing values by column
missing_values =
df.isnull().sum().sort_values(ascending=False)
missing_percent = (df.isnull().mean() *
100).sort_values(ascending=False)
print(pd.DataFrame({'Missing Count': missing_values, 'Missing
%': missing_percent}))
```

✅ **Step 2: Detect Duplicate Respondents**
python
CopyEdit

```
# Assuming 'Respondent ID' or similar unique key exists
duplicates = df.duplicated(subset='Respondent ID', keep=False)
duplicate_count = duplicates.sum()
print(f"Duplicate respondents: {duplicate_count}")
```

✅ **Step 3: Validate Quota Distribution**

python
CopyEdit

```
# Replace 'Age Group' or 'Gender' with your actual quota
variables
quota_check = df['Age Group'].value_counts(normalize=True) *
100
print("Quota distribution (%):")
print(quota_check)
```

✅ **Step 4: Check Demographic Balance**

python
CopyEdit

```
# Cross-tab example for Age vs Gender
demo_balance = pd.crosstab(df['Age Group'], df['Gender'],
normalize='index') * 100
print("Demographic cross-tab (%):")
print(demo_balance)
```

✅ **Step 5: Basic Statistical Validation**

python
CopyEdit

```
# Descriptive stats for numeric fields (e.g., age, ratings)
print(df.describe())

# Outlier detection (example: if there's a scale of 1–5)
for col in ['Satisfaction Rating', 'Product Score']:  #
replace with your fields
    if col in df.columns:
        outliers = df[(df[col] < 1) | (df[col] > 5)]
        print(f"Outliers in {col}: {len(outliers)}")
```

Tools/Modules

- **Python**: `pandas`, `numpy`
- **Excel / Google Sheets**: Quick inspection for small datasets

- **Power BI / Tableau** *(optional)*: For visual quota and demographic heatmaps

---

Output

- **Missing Values Table**: By column with % missing
- **Duplicate Count**: Number of repeated respondent entries
- **Quota Distribution Check**: Category-wise distribution %
- **Demographic Cross-tab**: Age × Gender matrix (or other relevant pairs)
- **Basic Stats Summary**: Descriptive stats + outlier flags