Chapter 4: Data Analysis & Findings

4.1 Clustering Performance

To assess the effectiveness of clustering on scalp severity scores, KMeans was applied with 4 clusters, followed by evaluation using the Silhouette Score:

- **Silhouette Score:** `0.10`

A low silhouette score (close to 0) suggests that the clusters are not well-separated, indicating considerable overlap or ambiguity in cluster boundaries.

4.2 Demographic Associations via Chi-Square Test

A Chi-Square test was conducted to identify whether the scalp severity segments were significantly associated with demographic variables.

Variable	Chi-Square p-value	Interpretation
Gender	1.00	No significant association
NCCS	1.00	No significant association
Age Group	1.00	No significant association

> **Conclusion:** There is no statistically significant relationship between scalp severity segments and any of the tested demographic variables.

4.3 Advanced Statistical Testing

Further tests were conducted to evaluate relationships between behavioral or attitudinal metrics and demographic factors. Results are summarized below:

Test	Statistic	p-value	Interpretation
Pearson Correlation (Age vs Brand Recall Score)	`0.03`	`0.838`	Insufficient data / missing values

ANOVA (Purchase Intent × Age Group)	0.21	0.798	Data not available / failed assumptions
T-Test (Purchase Intent by Gender)	`-0.38`	`0.7087`	Data not available / failed assumptions

> **Note:** These results could not be computed due to missing or invalid data. Consider preprocessing or imputing missing values in future runs.

4.4 Persona Summary – Severe Scalp Issues

A detailed profile was generated for participants with the most severe scalp conditions:

Segment: Severe

Attribute	Value/Distribution
Mean Age	29.4 years
Gender Split	Female: 50.7% Male: 49.3%
NCCS Distribution	Top 3: C1: 16.7% B1: 16.3% A3: 16.0%
Mean Purchase Intent	3.03 (on a 1–5 scale)
Mean Brand Recall Score	6.43 (on a 0–10 scale)

> **Insight:** The “Severe” segment is young (average age ~29), evenly split by gender, and spans across mid to upper NCCS tiers. They show moderate purchase intent and strong brand recall, making them an attractive target for specialized scalp-care products.

Summary of Key Takeaways

- **Clustering quality** is weak; segmentation may benefit from improved input features.
- **Demographics** do not significantly predict scalp issue severity.
- **Behavioral correlations** could not be established due to missing values.

- The **"Severe" segment** shows potential with decent brand recall and purchase intent, and skews younger, warranting targeted marketing.

Attachments (Optional for Appendices or Appendices)

- Final Data Sheet with Clusters and Segments
- Visualizations: power bi

Description:

Why This Matters

A high Silhouette Score indicates well-defined, tight, and separated clusters — essential for reliable segmentation or consumer profiling.

🛠️ **How to Do It**

1. Select only numerical columns.
2. Normalize the data using `StandardScaler`.
3. Fit a `KMeans` model.
4. Predict clusters.
5. Compute the Silhouette Score.

📦 **Tools/Modules**

python

CopyEdit

```
from sklearn.cluster import KMeans

from sklearn.metrics import silhouette_score
```

```
from sklearn.preprocessing import StandardScaler
```

🧪 **Output**

python

CopyEdit

```
X = df_cluster.select_dtypes(include=['int64', 'float64'])
X_scaled = StandardScaler().fit_transform(X)

kmeans = KMeans(n_clusters=3, random_state=42)
labels = kmeans.fit_predict(X_scaled)

score = silhouette_score(X_scaled, labels)
print(f'Silhouette Score: {score:.3f}')
```

Example Output:

yaml

CopyEdit

```
Silhouette Score: 0.652
```

🔹 4.2 Demographic Associations via Chi-Square Test

✅ **Task Description**

Determine if there's a statistically significant relationship between two categorical variables (e.g., Gender and HairFall Concern).

📌 **Why This Matters**

Understanding such associations helps validate demographic relevance in behavioral or preference patterns.

🛠️ **How to Do It**

1. Create a contingency table using `pd.crosstab`.

2. Apply `chi2_contingency` to get the test statistic and p-value.

📦 **Tools/Modules**

python

CopyEdit

```
import pandas as pd

from scipy.stats import chi2_contingency
```

🧪 **Output**

python

CopyEdit

```
ct = pd.crosstab(df['Gender'], df['Q2'])  # Replace Q2 with actual column

chi2, p, dof, expected = chi2_contingency(ct)


print(f"Chi2 Statistic: {chi2:.3f}")

print(f"p-value: {p:.4f}")
```

Example Output:

yaml

CopyEdit

```
Chi2 Statistic: 8.934

p-value: 0.0031
```

🔹 4.3 Advanced Statistical Testing

🔸 A. T-Test (Independent Samples)

✅ Task Description

Compare the means of a continuous variable across two independent groups (e.g., Male vs Female scores).

📌 Why This Matters

Reveals if gender-based differences in a key metric (e.g., satisfaction) are statistically significant.

🛠️ How to Do It

1. Subset data into two groups.
2. Run `ttest_ind` to compare group means.

📦 Tools/Modules

python

CopyEdit

```python
from scipy.stats import ttest_ind
```

🧪 Output

python

CopyEdit

```python
group1 = df[df['Gender'] == 'Male']['Score']
```

```
group2 = df[df['Gender'] == 'Female']['Score']


t_stat, p_val = ttest_ind(group1, group2, nan_policy='omit')
print(f"T-stat: {t_stat:.3f}, p-value: {p_val:.4f}")
```

Example Output:

makefile

CopyEdit

```
T-stat: 2.134
p-value: 0.0350
```

🔸 **B. ANOVA (Analysis of Variance)**

✅ **Task Description**

Compare the means of a numerical variable across more than two groups (e.g., different AgeGroups).

📌 **Why This Matters**

Shows whether there are meaningful differences across age-based segments in terms of preferences or satisfaction.

🛠️ **How to Do It**

1. Segment data by group.
2. Apply `f_oneway` for ANOVA.

📦 **Tools/Modules**

python

CopyEdit

```python
from scipy.stats import f_oneway
```

🧪 **Output**

python

CopyEdit

```python
f_stat, p_val = f_oneway(
    df[df['AgeGroup'] == '18-25']['Score'],
    df[df['AgeGroup'] == '26-35']['Score'],
    df[df['AgeGroup'] == '36-45']['Score']
)
print(f"F-statistic: {f_stat:.3f}, p-value: {p_val:.4f}")
```

Example Output:

makefile

CopyEdit

```
F-statistic: 4.872
p-value: 0.0110
```

🔸 **C. Pearson Correlation**

✅ **Task Description**

Measure the strength and direction of the linear relationship between two numerical variables (e.g., DandruffScore and HairFallScore).

📌 **Why This Matters**

Understanding correlations helps in identifying patterns that can drive predictive models or key consumer insights.

🛠️ **How to Do It**

1. Identify numeric columns.
2. Run `pearsonr` to get correlation coefficient and significance.

📦 **Tools/Modules**

python

CopyEdit

```
from scipy.stats import pearsonr
```

🧪 **Output**

python

CopyEdit

```
corr, p_val = pearsonr(df['DandruffScore'],
df['HairFallScore'])

print(f"Pearson Correlation: {corr:.2f}, p-value:
{p_val:.4f}")
```

Example Output:

yaml

CopyEdit

```
Pearson Correlation: 0.61

p-value: 0.0000
```