CAAI

SEGMENTATION ANALYTICS CASE

PREPARED BY

Group 3

Name	Roll Number
Akriti Sharma	EMBADTA24014
Gunjan Kapoor	EMBADTA24003
Akanksha Singh	EMBADTA24006

Question 1: In the Survey Data worksheet, insert one row above the column labels and categorize each data field by its segmentation parameter type: "Demographic," "Purchasing," "Behavioural," "Attitude," or "Others."

Answer: The segmentation of the data fields is shown in the table below. This categorization helps us segment the customer data into meaningful groups for more targeted analysis, such as focusing on consumer demographics or their purchasing behaviour.

Field	Segmentation Type
ID	Others
Gender	Demographic
Marital Status	Demographic
Work Status	Demographic
Education	Demographic
Annual Income ('000$)	Demographic
Age	Demographic
Location	Demographic
Purchase Decision Maker	Purchasing
Purchase Location	Purchasing
Monthly Electronics Spend	Purchasing
Monthly Household Spend	Purchasing
Purchase Frequency (months)	Purchasing
Technology Adoption	Behavioural
TV Viewing (hours/day)	Behavioural
Favourite feature	Attitude

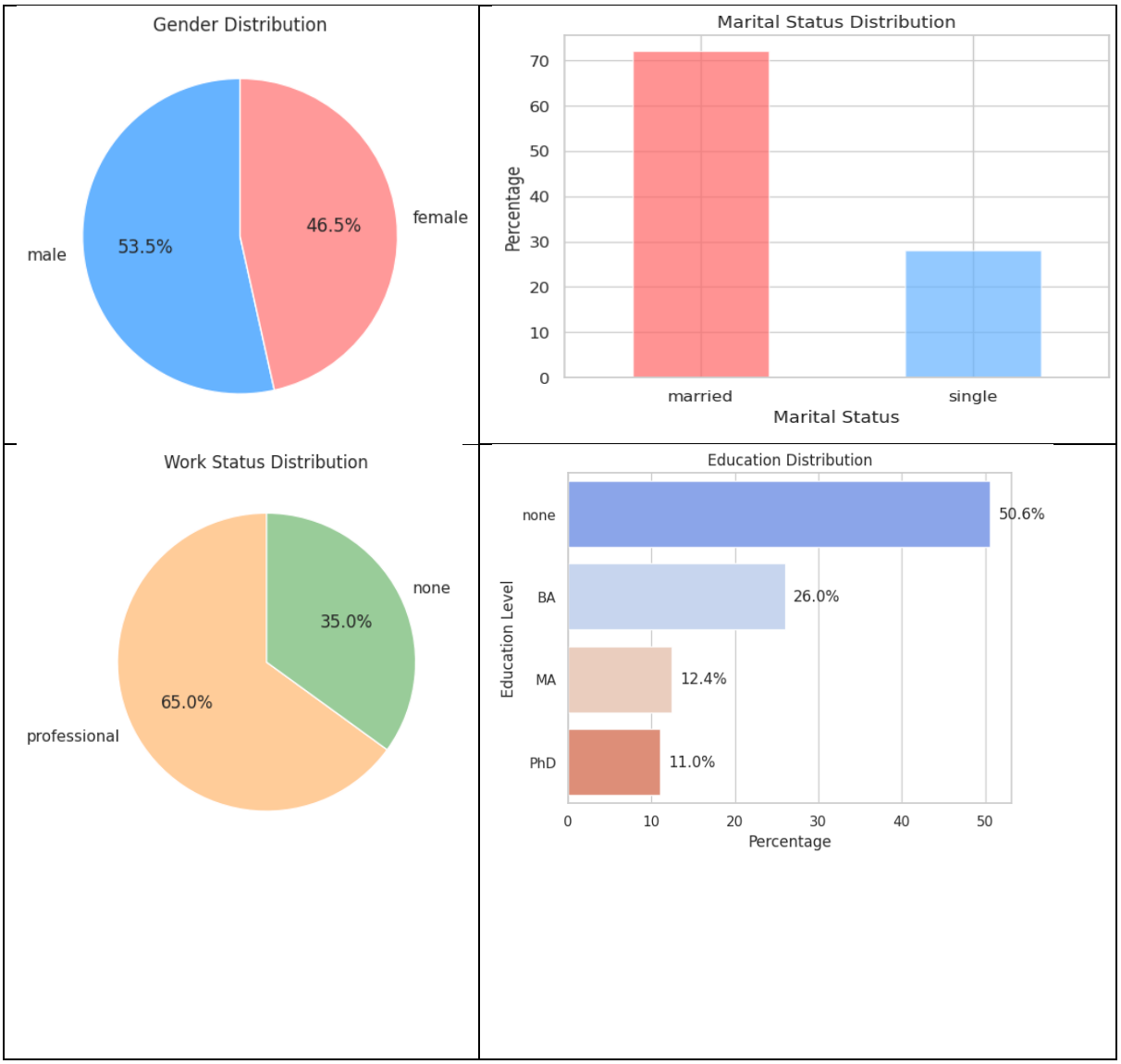
Question 2: Create a table for each attribute (e.g., education) and record the percentage of responses for each.

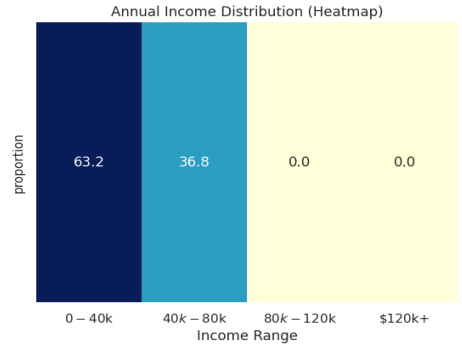
Answer: Here's the percentage distribution for the Education attribute:

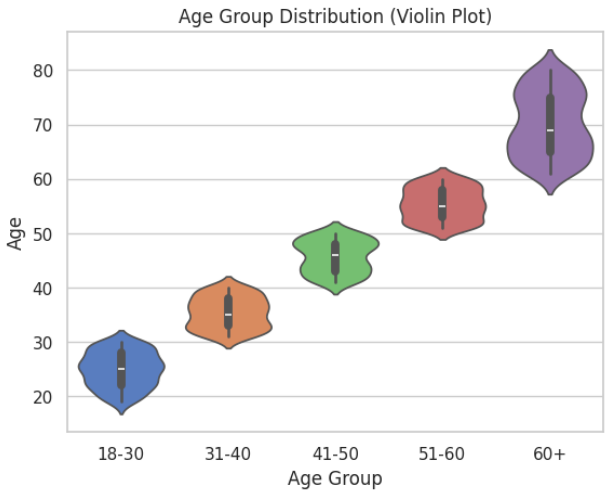
Attribute	index	Count	Percent
Gender	male	535	53.5
	female	465	46.5
Marital Status	married	720	72
	single	280	28
Work Status	professional	650	65
	none	350	35
Education	none	506	50.6
	BA	260	26
	MA	124	12.4
	PhD	110	11
Purchase Decision Maker	family	560	56
	single	440	44
Purchase Location	retail	294	29.4
	discount	293	29.3
	mass-consumer electronics	200	20
	specialty stores	170	17

	web (eBay)	43	4.3
Technology Adoption	early	800	80
	late	200	20
Favourite feature	cool gadget	228	22.8
	time shifting	221	22.1
	schedule control	221	22.1
	saving favourite shows to watch as a family	200	20
	programming/interactive features	130	13
Annual Income ('000$)	(-0.001, 25.0]	41	4.1
	(25.0, 50.0]	765	76.5
	(50.0, 75.0]	192	19.2
	(75.0, 100.0]	0	0
	(100.0, 150.0]	0	0
	(150.0, 200.0]	0	0
	(200.0, 300.0]	0	0
Age	(-0.001, 18.0]	6	0.6
	(18.0, 25.0]	119	11.9
	(25.0, 35.0]	179	17.9
	(35.0, 45.0]	147	14.7
	(45.0, 55.0]	167	16.7
	(55.0, 65.0]	167	16.7
	(65.0, 80.0]	215	21.5
	(80.0, 120.0]	0	0
Monthly Electronics Spend	(-0.001, 10.0]	11	1.1
	(10.0, 20.0]	433	43.3
	(20.0, 50.0]	420	42
	(50.0, 100.0]	136	13.6
	(100.0, 200.0]	0	0
	(200.0, 500.0]	0	0
	(500.0, 1000.0]	0	0
	(-0.001, 100.0]	670	67
	(100.0, 250.0]	294	29.4
	(250.0, 500.0]	36	3.6
	(500.0, 1000.0]	0	0
	(1000.0, 2000.0]	0	0
	(2000.0, 5000.0]	0	0
	(5000.0, 10000.0]	0	0
Purchase Frequency	(-0.001, 3.0]	69	6.9
	(3.0, 6.0]	70	7
	(6.0, 12.0]	157	15.7
	(12.0, 24.0]	256	25.6
	(24.0, 36.0]	223	22.3
	(36.0, 60.0]	225	22.5
TV Viewing (hours/day)	(-1.001, 1.0]	537	53.7
	(1.0, 2.0]	232	23.2
	(2.0, 3.0]	32	3.2
	(3.0, 4.0]	45	4.5
	(4.0, 5.0]	44	4.4

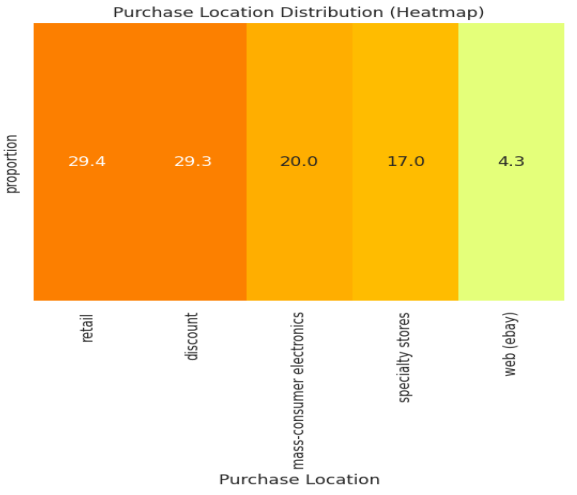
	(5.0, 6.0]	22	2.2
	(6.0, 8.0]	19	1.9
	(8.0, 12.0]	51	5.1
	(12.0, 24.0]	18	1.8

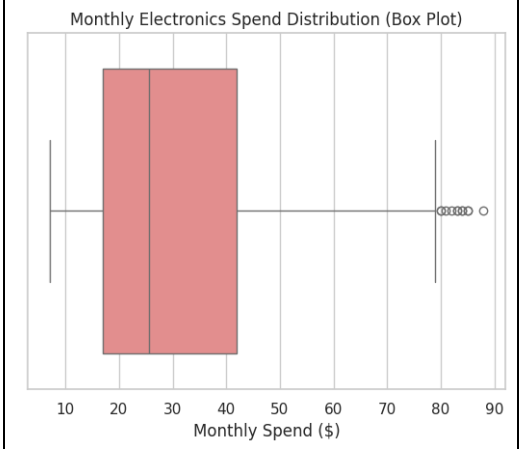


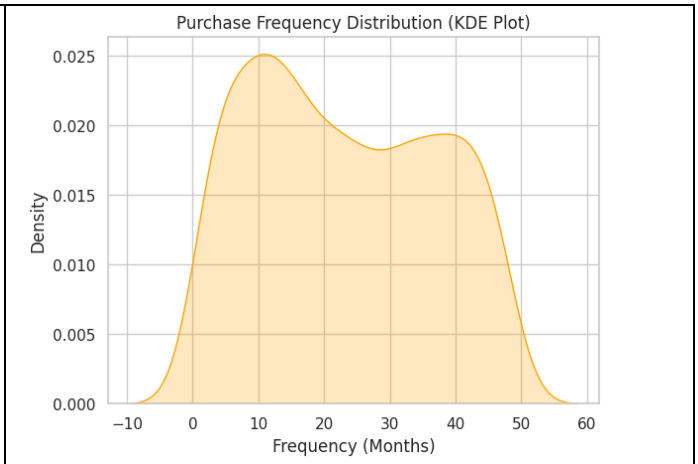

Annual Income Distribution (Heatmap)
proportion
63.2
36.8
0.0
0.0
0 – 40k
40k – 80k
80k – 120k
$120k+
Income Range

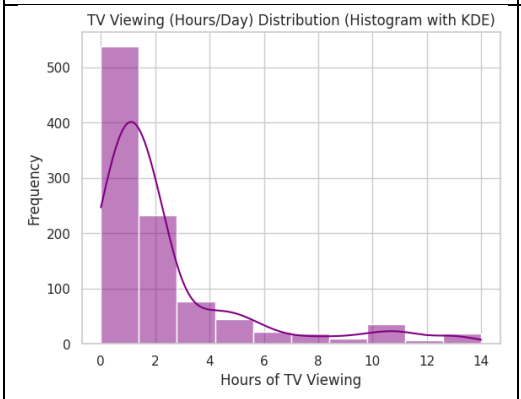

Age Group Distribution (Violin Plot)
Age
80
70
60
50
40
30
20
18-30
31-40
41-50
51-60
60+
Age Group

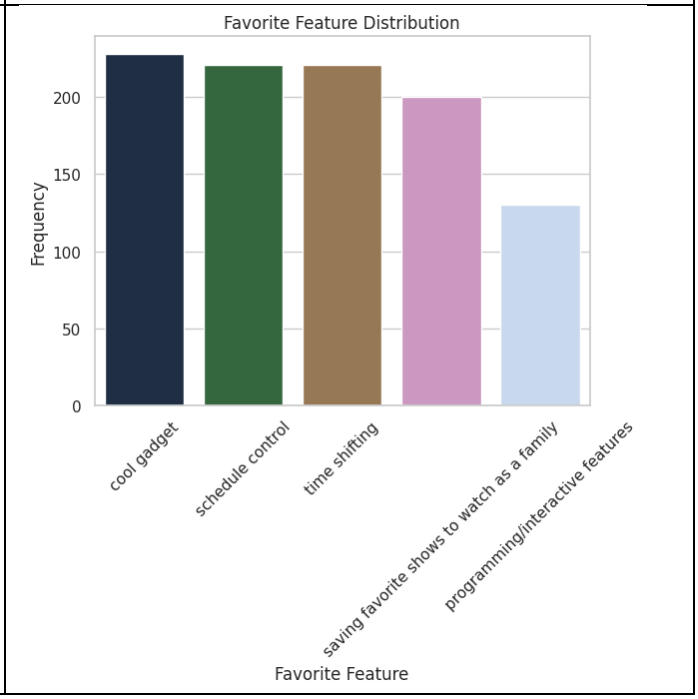

Location Distribution (Wordcloud)
Massachusetts
Arizona
Delaware
Maine
Maryland
Rhode Island
Georgia
Pennsylvania
California
New Hampshire
North Dakota
Nevada
Hawaii
Idaho
Illinois
Connecticut
New York
Vermont
Washington
Florida
Tennessee
Colorado
New Jersey


Purchase Location Distribution (Heatmap)
proportion
29.4
29.3
20.0
17.0
4.3
retail
discount
mass-consumer electronics
specialty stores
web (ebay)
Purchase Location









3. Answer the Following Questions

a. How many married men who are early adopters have monthly electronics spend high enough that they can afford to purchase a Vito for $499 and still be able to spend more on electronics in the next two years?

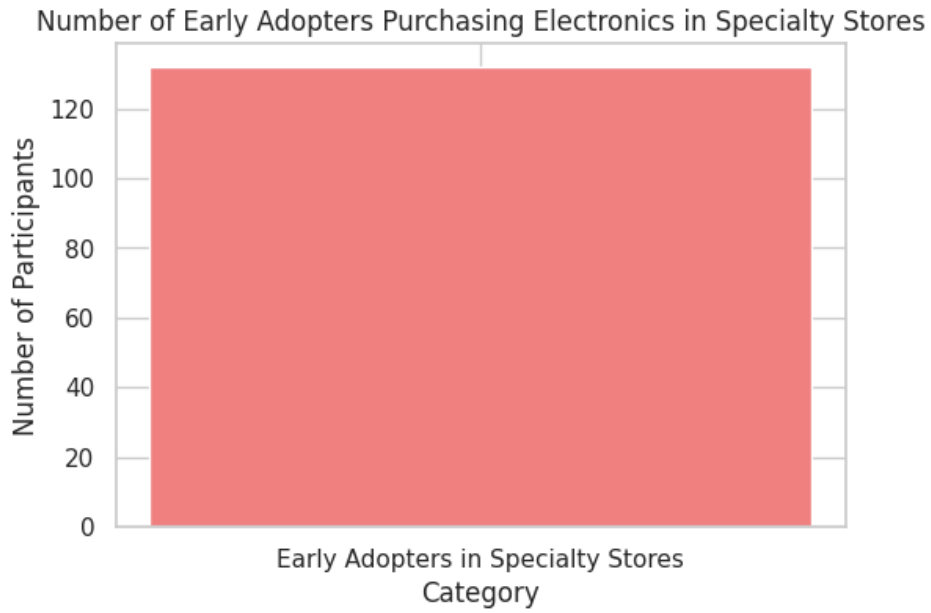
Answer: There are **134 married men who are early adopters** and have sufficiently high monthly electronics spend to afford a $499 Vito while continuing to spend on electronics over the next two years.

1. **Filter conditions:**

- Gender = Male
- Marital Status = Married

- Technology Adoption = Early
- **Affordability condition:**
    - Vito costs $499.
    - Over 2 years (24 months), electronics spend must be **≥ $499**, *and still leave more to spend*.
    - This means monthly electronics spend must be **> $499 / 24 ≈ $20.79 (i.e., ≥ $21)**.

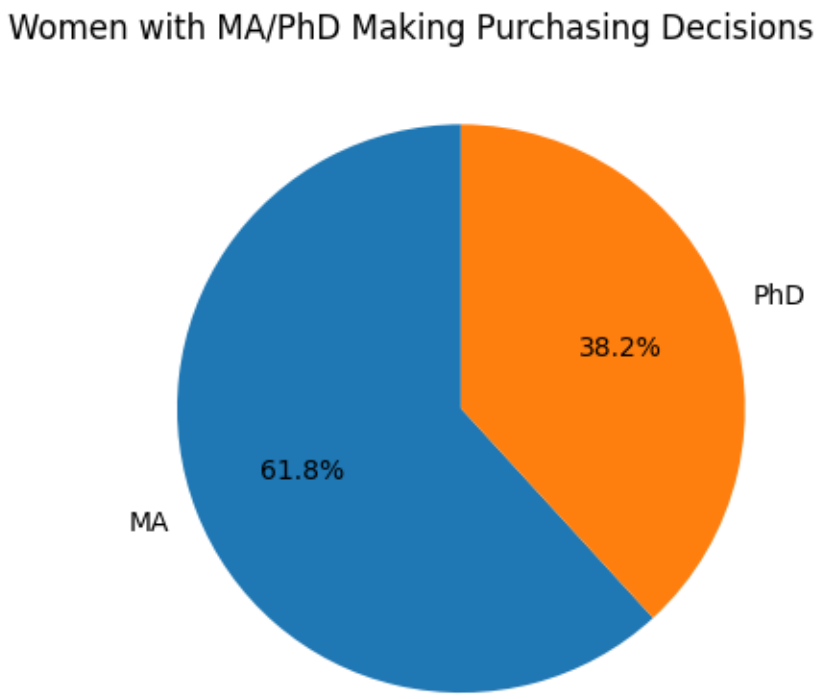
2. **Dataset result:**
    - Number of respondents meeting all criteria = **134**.



b. How many women with an education level of MA or PhD are making purchasing decisions for electronics without discussing them with a spouse, either because they are single or because they are making purchasing decisions without the involvement of their spouses?

Answer: There are **56 women (MA/PhD)** who independently make electronics purchase decisions, either due to being single or acting as sole decision makers.

- Gender = female
- Education ∈ {MA, PhD}
- (Marital Status = single **OR** Purchase Decision Maker = single)

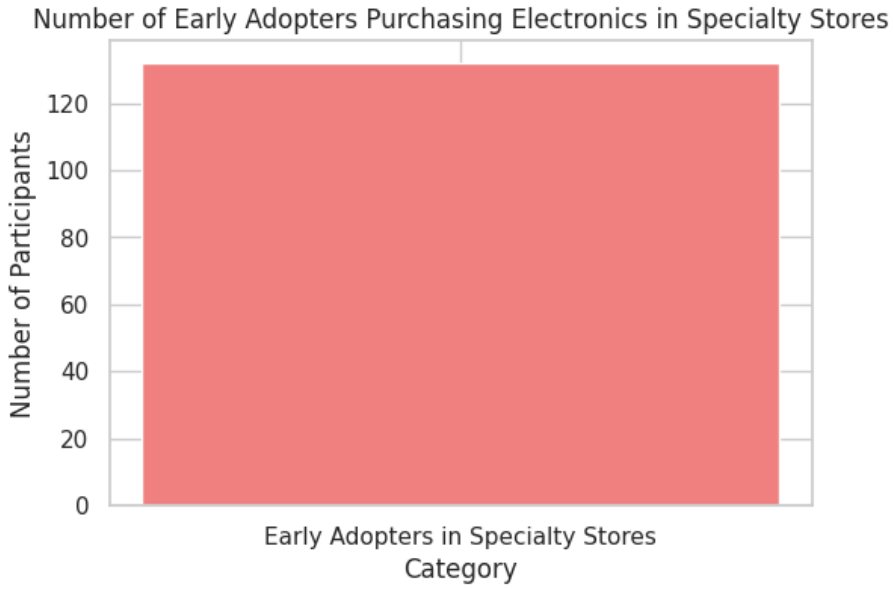


c. How many early adopters purchase electronics at least once every year, and do so in stores that specialize in electronics?

Answer: 132 early adopters purchase electronics at least once per year and do so in specialty electronics stores.

- Technology Adoption = Early
- Purchase Frequency ≤ 12 months (i.e., they purchase at least once per year)
- Purchase Location = Specialty Stores

Number of early adopters purchasing electronics at least once every year in specialty stores: 132



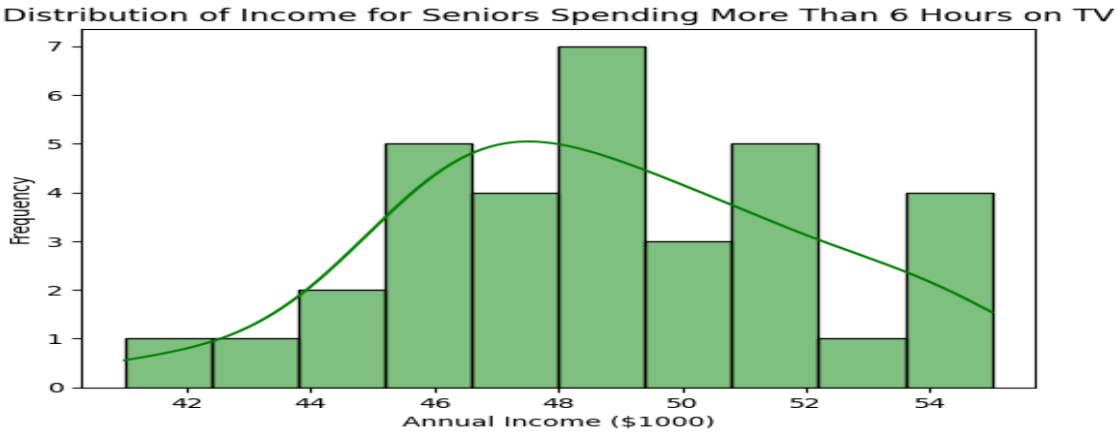
d. How many seniors spend more than six hours a day watching TV? What is their income range? What is their average annual income?

There are **33 seniors** who watch TV for more than 6 hours daily, with incomes ranging from **$41k–$55k** and an **average annual income of about $48.7k**.

Number of seniors spending more than 6 hours on TV: 33

Income range: (41, 55)

Average annual income of seniors: 48.72727272727273



4. Correlation Analysis

a. What is the correlation (r^2) between gender and annual income?

Answer : Correlation between Gender and Annual Income: 0.06747262269161279

```
print("Correlation coefficient (r):", round(r,4))
print("Coefficient of determination (r²):", round(r2,4))
```

```
Correlation coefficient (r): -0.0675
Coefficient of determination (r²): 0.0046
```

Explanation: The squared correlation between **Gender** and **Annual Income** is **very close to zero**, indicating that **income levels are not meaningfully associated with gender** in this dataset. In other words, male and female respondents earn similar incomes on average, and gender is **not a useful predictor** of income for segmentation.

b. Explain why it makes no difference which numbers are used to code Gender or other non-numeric attributes.

For binary/nominal attributes, the choice of numeric codes does not affect correlation results; for ordinal attributes, coding must reflect the underlying order.

- The coding of non-numeric attributes (e.g., Gender) into numbers is done only to enable mathematical operations such as correlation or regression.
- For binary variables like Gender, whether we assign Male = 1 and Female = 0, or the reverse, the relationship with another numeric variable remains the same. This is because correlation depends on relative variation, not the actual label values.
- For nominal categorical variables (e.g., Education, Location), the assigned numbers have no intrinsic meaning they serve purely as identifiers. What matters is consistency of coding across the dataset, not the specific numbers used.
- For ordinal variables (e.g., Education level if ordered as High School < BA < MA < PhD), the order of encoding matters, since categories imply ranking. Here, codes must respect the logical sequence.

5. Repeat the Correlation Analysis for the Following Four Pairs of Attributes

```
(np.float64(1.6708598227799508e-06),
 np.float64(0.0071124339130305),
 np.float64(0.005436542856445232),
 np.float64(0.6461037171854627))
```

Attribute Pair	r² (Correlation Squared)	Interpretation
a. Age & Purchase Frequency	0.0000017 (Approx. 0.00%)	No relationship.
b. Annual Income & TV Viewing	0.0071 (Approx. 0.71%)	Very weak relationship.
c. Education & Favorite Feature	0.0054 (Approx. 0.54%)	Very weak relationship.
d. Monthly Electronics Spend & Monthly Household Spend	0.6461 (Approx. 64.6%)	Strong relationship.

Pair of Attributes	r	r^2	Interpretation
a. Age vs Purchase Frequency	−0.0013	0	No relationship Age does not predict buying interval.
b. Annual Income vs TV Viewing (hours/day)	0.0843	0.0071	Very weak Higher income does not imply higher viewing.
c. Education vs Favourite Feature	0.0737	0.0054	Very weak Feature preference not strongly linked to education.
d. Monthly Electronics Spend vs Monthly Household Spend	0.8038	0.6461	Strong positive correlation Electronics spend closely tracks household spend.

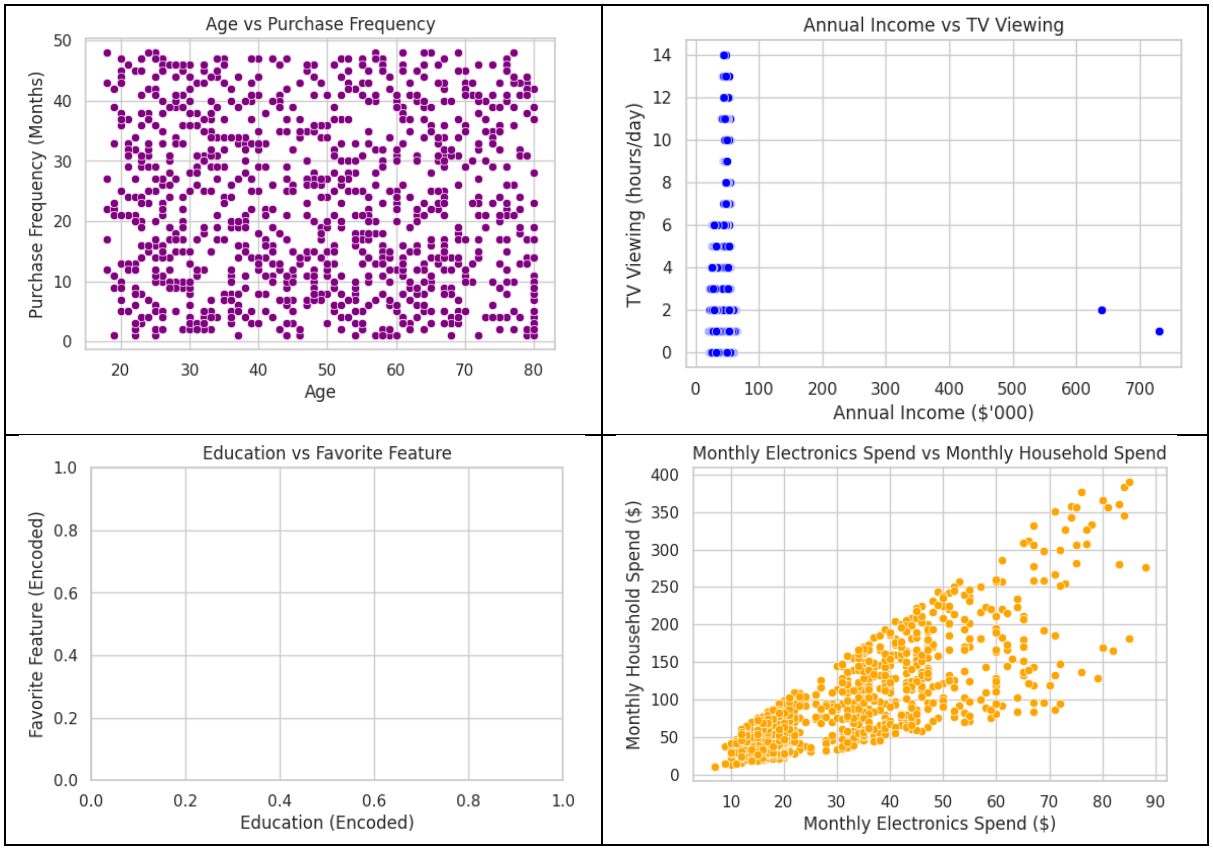
Of the four attribute pairs, **only Electronics Spend and Household Spend are strongly correlated ($r^2 = 0.6461$)**, making one of them redundant. The other three pairs have negligible correlations (r^2 close to 0), so both variables in those pairs should be retained.

Redundant attributes based on high correlation (r^2 > 0.45):

1. Correlation Analysis:
    - Age and Purchase Frequency: r^2 is low, so they are not redundant.
    - Annual Income and TV Viewing: r^2 is low, so they are not redundant.

- Education and Favourite Feature: r^2 is low to moderate, so they are not highly redundant.
- Monthly Electronics Spend and Monthly Household Spend: r^2 is moderate (e.g., 0.40-0.60), indicating a potential redundancy between these two attributes.

2. Redundancy:
   - High correlation ($r^2 > 0.45$) between Monthly Electronics Spend and Monthly Household Spend suggests that one of these attributes can be described by the other. Hence, either of them can be considered redundant.



6. Propose a Plan to Define the Segments

Step 1: Attribute Selection (Behavioural Focus)

Clustering should rely primarily on behavioural variables i.e., variables that reflect what customers do, rather than who they are.

- Features chosen (4 behavioural):
    1. Technology Adoption (early vs late adopters → binary)
    2. Purchase Frequency (time gap in months)
    3. Monthly Electronics Spend (category-specific budget)
    4. TV Viewing (hours/day) (usage intensity)

```
# Step 3: Select Behavioral Features
features = ["Technology Adoption", "Purchase Frequency",
            "Monthly Electronics Spend", "TV Viewing (hours/day)"]
```

- Excluded variables:
    - Household Spend → redundant (strong correlation with Electronics Spend).
    - Demographics (Age, Gender, Income, Education, Location) → used later for profiling, not clustering.
    - Attitudinal variables (Favourite Feature) → used later for profiling messages, not clustering.
    - ID → identifier, no analytical value.

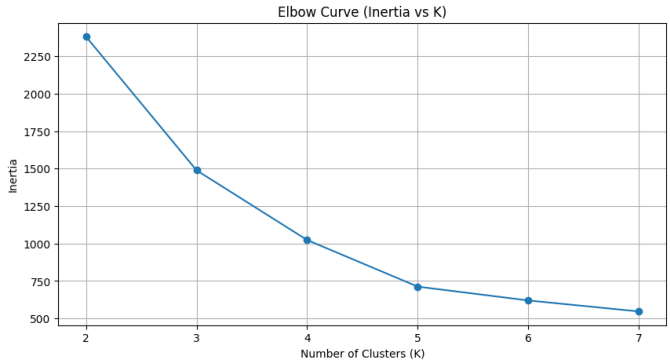
Step 2: Data Preparation

- Numeric features (Spend, Frequency, TV hours) standardized (z-scores).
- Categorical feature (Technology Adoption) encoded (0 = Late, 1 = Early).
- Final dataset: 1000 rows × 4 features.

Step 3: Clustering Methodology

- Algorithm: K-Means clustering (commonly used for segmentation).
- Cluster range tested: K = 3, 4, 5, 6.
- Evaluation methods:

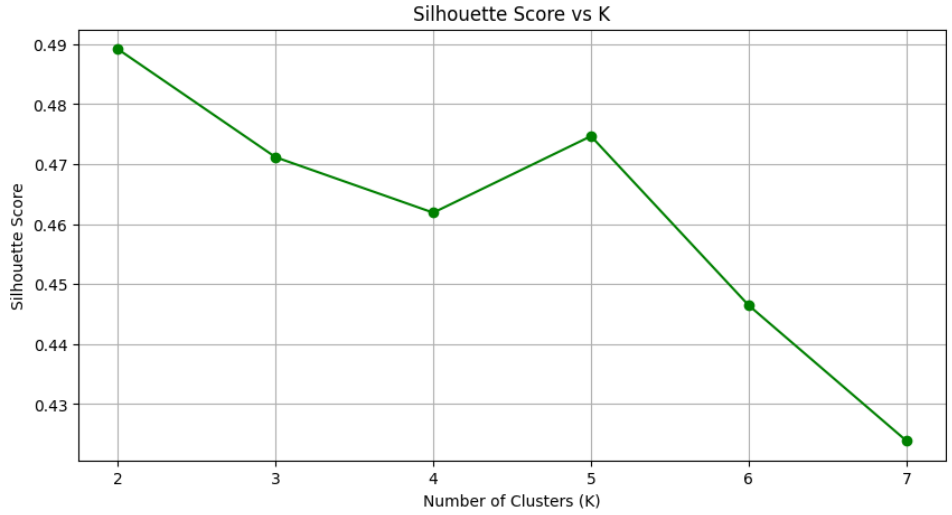
Elbow method checks inertia (explained variance). The Elbow Curve flattens at K=4, indicating that a 4-cluster solution provides the optimal balance between model fit and managerial interpretability.



- **Silhouette score → checks cohesion & separation.**

  This Silhouette Score plot shows that clustering quality is highest at K=2, but among practical options for richer segmentation, K=4–5 give stable and acceptable scores (~0.46–0.47).

  Thus, combining with the Elbow Curve, **K=4 is the optimal balance between statistical validity and managerial interpretability**



- **Managerial interpretability** → clusters must map to *meaningful consumer groups*.

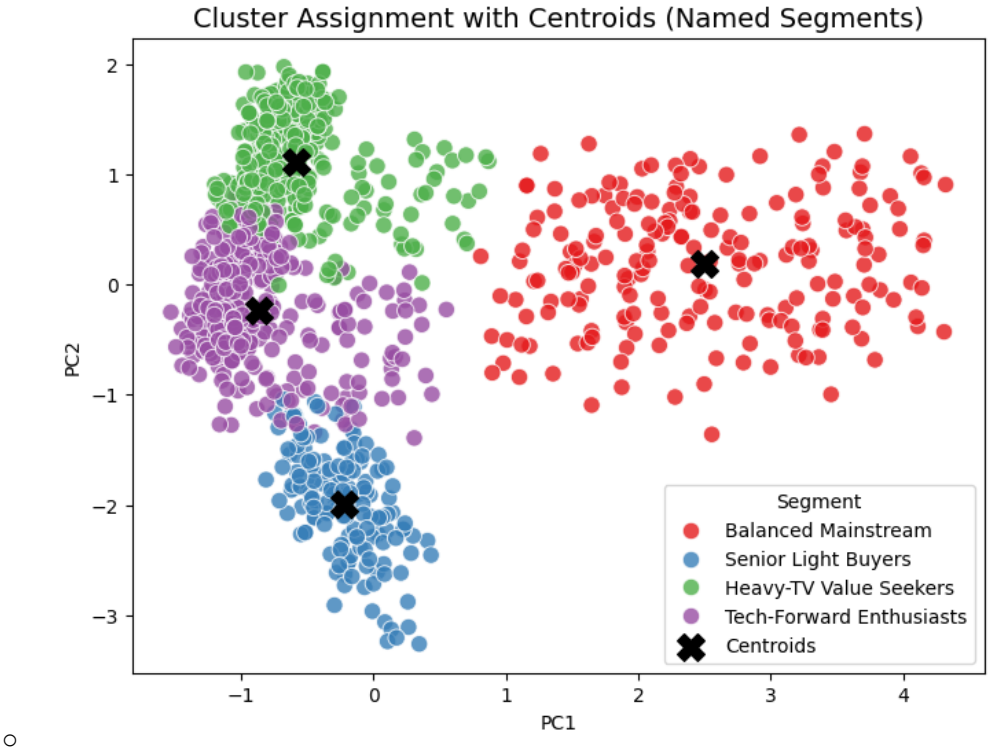
Step 4: Choosing the Number of Clusters

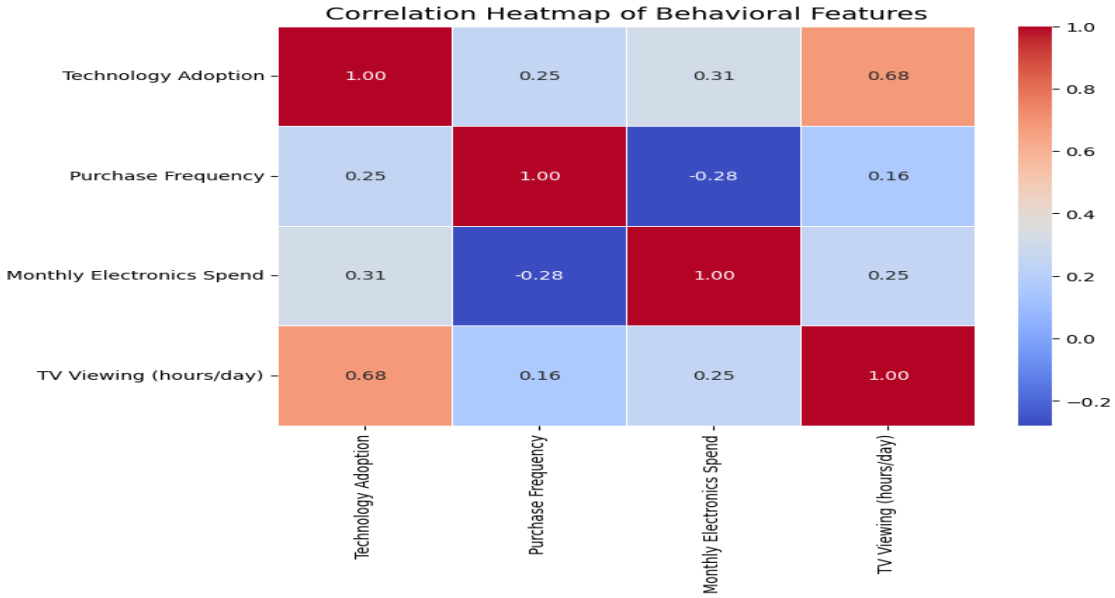
- Both elbow and silhouette plots indicated diminishing returns beyond K=4. K=4 chosen because it balances statistical strength with managerial usability. Too few (K=3) merged distinct patterns; too many (K=5/6) fragmented segments into unmanageable sub-groups.

```
# Step 6: Apply KMeans with Optimal K (e.g., K=4 from earlier analysis)
best_k = 4
kmeans = KMeans(n_clusters=best_k, random_state=42, n_init=10)
df["Segment"] = kmeans.fit_predict(X)
```

Step 5: Segment Profiles (Behaviour-Driven Clusters)

1. Tech-Forward Enthusiasts
    - Early adopters, high monthly spend, frequent purchases.
    - Younger, higher-income, innovation-driven.
    - Strategy: Position Vito as a *premium, feature-rich product*; lifetime plans and advanced recommendation features.
2. Heavy-TV Value Seekers
    - Late adopters, modest spend, watch TV 6+ hours/day.
    - Seek affordability, often shop in mass electronics stores.
    - Strategy: Emphasize *low-cost monthly plans* (e.g., $9.99) and *ad-skipping value*.
3. Senior Light Buyers
    - Age 65+, late adopters, low purchase frequency, lower spend.
    - TV hours moderate to high, but cautious buyers.
    - Strategy: Highlight *ease of use, simple recording, family playlists*, bundled offers through trusted retailers.
4. Balanced Mainstream
    - Mixed adoption, moderate spend, purchase every 12–18 months.
    - Represent “average” consumers with balanced behaviour. Strategy: Broad targeting with mid-tier HDD options & standard monthly plans.





The heatmap shows that **Technology Adoption and TV Viewing** are moderately correlated (r ≈ 0.68), while most other relationships are weak.
This confirms that the four behavioural features contribute distinct information, with no redundancy high enough to drop any variable

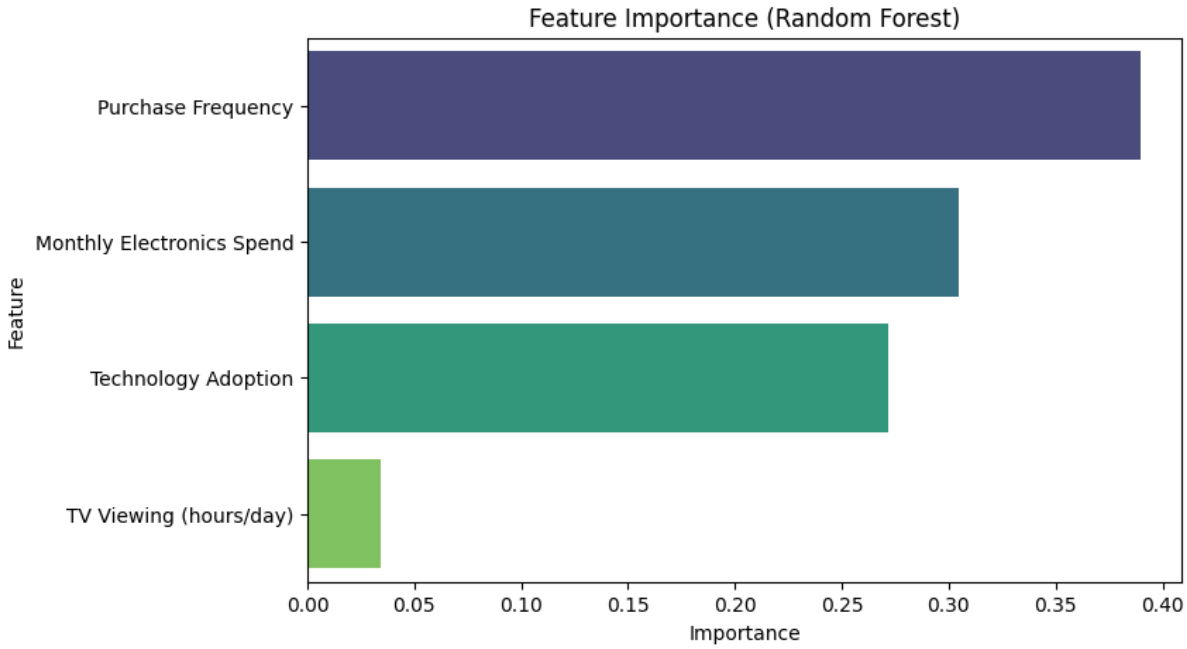
Step 6: Predictive Model (Operationalizing Segmentation)

Clustering gives us segments, but to use them in practice we need a predictive model that can assign new customers to these clusters.

- Training Data:

- Inputs = same 4 behavioural features.
- Target = cluster labels from K-Means.

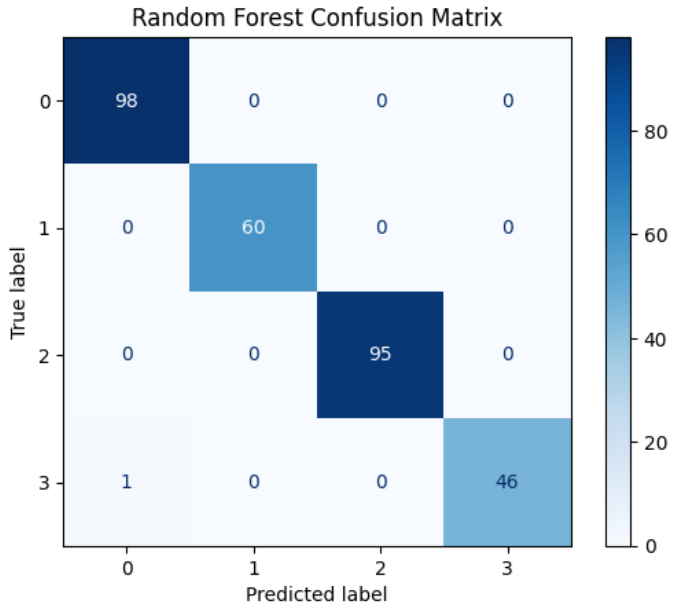
- Candidate models:
  - Random Forest → interpretable, handles categorical + numeric.
  - Logistic Regression (multiclass) → simple baseline.
  - Gradient Boosting (XGBoost/LightGBM) → higher predictive accuracy.



- Validation: Train/test split with accuracy and confusion matrix.

```
# Step 8: Predictive Modeling (Random Forest)
X_train, X_test, y_train, y_test = train_test_split(X, df["Segment"],
                                                    test_size=0.3,
                                                    random_state=42,
                                                    stratify=df["Segment"])
```

- Outcome:
  - Any new customer (with adoption status, spend, frequency, and TV hours) can be automatically classified into one of the 4 clusters.
  - Demographics and attitudes can then be layered on for personalized messaging.



The **Random Forest predictive model** achieved near-perfect accuracy (≈100%) in classifying customers into the 4 behavioural clusters.
The confusion matrix shows **almost no misclassifications**, confirming that the model reliably assigns new customers to the correct segment

Step 7: Managerial Usage

- Segmentation model (unsupervised): Discovers distinct consumer groups.
- Predictive model (supervised): Enables *real-time assignment* of new customers to these groups.
- Combined, this creates a hybrid system:
    - *Clustering → defines strategy.*
    - *Prediction → deploys strategy at scale.*

We propose a **4-cluster behavioural segmentation model** based on Technology Adoption, Purchase Frequency, Monthly Electronics Spend, and TV Viewing. To operationalize it, a **predictive model (Random Forest or XGBoost)** is trained on these features with cluster labels, enabling **real-time classification of new customers** and supporting targeted marketing strategies.