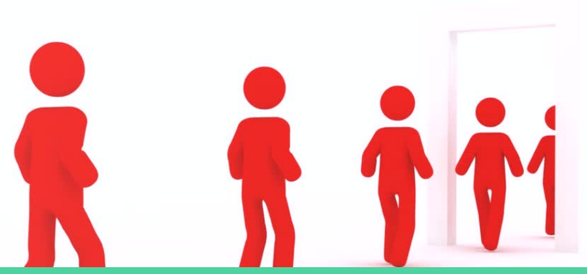
MARKET RESEARCH

UNSUPERVISED CLUSTERING OF A GROCERY FIRM'S CUSTOMER DATABASE TO FORM CUSTOMER SEGMENTATION.

Data Architecture



GROUP 6 Kohei | Leopold | Elisabeth | Rahul | Ekiomoado | Rauf

ABOUT THE DATASET

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases
5524	1957	Graduation	Single	58138	0	0	04-09-2012	58	635	88	546	172	88	88	3	8	10
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	11	1	6	2	1	6	2	1	1
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	426	49	127	111	21	42	1	8	2
6182	1984	Graduation	Together	26646	1	0	10-02-2014	26	11	4	20	10	3	5	2	2	0
5324	1981	PhD	Married	58293	1	0	19-01-2014	94	173	43	118	46	27	15	5	5	3
7446	1967	Master	Together	62513	0	1	09-09-2013	16	520	42	98	0	42	14	2	6	4
965	1971	Graduation	Divorced	55635	0	1	13-11-2012	34	235	65	164	50	49	27	4	7	3
6177	1985	PhD	Married	33454	1	0	08-05-2013	32	76	10	56	3	1	23	2	4	0
4855	1974	PhD	Together	30351	1	0	06-06-2013	19	14	0	24	3	3	2	1	3	0
5899	1950	PhD	Together	5648	1	1	13-03-2014	68	28	0	6	1	1	13	1	1	0
1994	1983	Graduation	Married	null	1	0	15-11-2013	11	5	5	6	0	2	1	1	1	0
387	1976	Basic	Married	7500	0	0	13-11-2012	59	6	16	11	11	1	16	1	2	0
2125	1959	Graduation	Divorced	63033	0	0	15-11-2013	82	194	61	480	225	112	30	1	3	4
8180	1952	Master	Divorced	59354	1	1	15-11-2013	53	233	2	53	3	5	14	3	6	1
2569	1987	Graduation	Married	17323	0	0	10-10-2012	38	3	14	17	6	1	5	1	1	0
2114	1946	PhD	Single	82800	0	0	24-11-2012	23	1006	22	115	59	68	45	1	7	6
9736	1980	Graduation	Married	41850	1	1	24-12-2012	51	53	5	19	2	13	4	3	3	0
4939	1946	Graduation	Together	37760	0	0	31-08-2012	20	84	5	38	150	12	28	2	4	1
6565	1949	Master	Married	76995	0	1	28-03-2013	91	1012	80	498	0	16	176	2	11	4
2278	1985	2n Cycle	Single	33812	1	0	03-11-2012	86	4	17	19	30	24	39	2	2	1

only showing top 20 rows

ABOUT THE DATASET

NUMBER OF DATAPOINTS: 2240

5 ROWS × 29 COLUMNS

```
root
 |-- ID: integer (nullable = true)
 |-- Year_Birth: integer (nullable = true)
 |-- Education: string (nullable = true)
 |-- Marital_Status: string (nullable = true)
 |-- Income: integer (nullable = true)
 |-- Kidhome: integer (nullable = true)
 |-- Teenhome: integer (nullable = true)
 |-- Dt_Customer: string (nullable = true)
 |-- Recency: integer (nullable = true)
 |-- MntWines: integer (nullable = true)
 |-- MntFruits: integer (nullable = true)
 |-- MntMeatProducts: integer (nullable = true)
 |-- MntFishProducts: integer (nullable = true)
 |-- MntSweetProducts: integer (nullable = true)
 |-- MntGoldProds: integer (nullable = true)
 |-- NumDealsPurchases: integer (nullable = true)
 |-- NumWebPurchases: integer (nullable = true)
 |-- NumCatalogPurchases: integer (nullable = true)
 |-- NumStorePurchases: integer (nullable = true)
 |-- NumWebVisitsMonth: integer (nullable = true)
 |-- AcceptedCmp3: integer (nullable = true)
 |-- AcceptedCmp4: integer (nullable = true)
 |-- AcceptedCmp5: integer (nullable = true)
 |-- AcceptedCmp1: integer (nullable = true)
...
 |-- Z_CostContact: integer (nullable = true)
 |-- Z_Revenue: integer (nullable = true)
 |-- Response: integer (nullable = true)
```

UNSUPERVISED CLUSTERING

About the dataset :

The dataset consists of 2240 datapoints and 29 attributes
It can be categorized into the following subsets

Customer's Information

- ID
- Year_Birth
- Education
- Marital_Status
- Income
- Kidhome
- Teenhome
- Dt_Customer
- Recency
- Complain

Products

Amount spent on diffrent probucts in last 2 years.

- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts
- MntSweetProducts
- MntGoldProds

Promotion

- NumDealsPurchases
- AcceptedCmp1
- AcceptedCmp2
- AcceptedCmp3
- AcceptedCmp4
- AcceptedCmp5
- Response

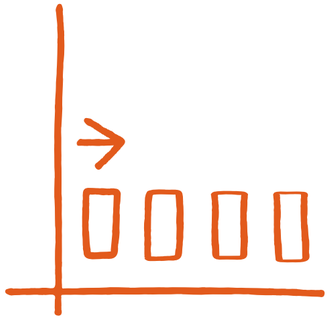
Place

- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases
- NumWebVisitsMonth

OBJECTIVE: UNSUPERVISED CLUSTERING OF CUSTOMER RECORDS FOR SEGMENTATION.

THE PIPELINE

- HDFS FOR STORAGE
- APACHE SPARK FOR ML
- PLOTLY AND MATPLOTLIB FOR ANALYSIS



DATA CLEANING

- There are missing values in income
- Dt_Customer that indicates the date a customer joined the database is not parsed as DateTime
- There are some categorical features in our data frame. So we will need to encode them into numeric forms later.

SOME INSIGHTS

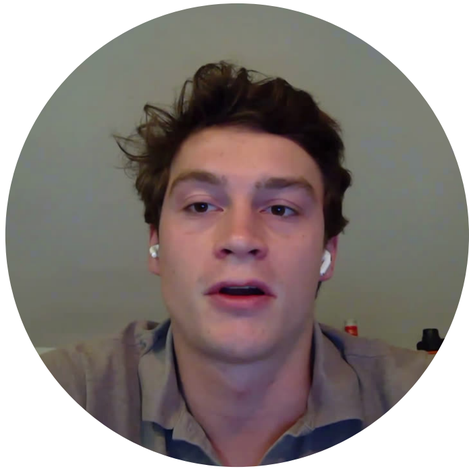
```
The newest customer's enrolment date in the records: 2014-06-29
The oldest customer's enrolment date in the records: 2012-07-30
```

Marital_Status	count
Married	857
Together	573
Single	471
Divorced	232
Widow	76
Alone	3
YOLO	2
Absurd	2

```
Total categories in the feature Education:
```

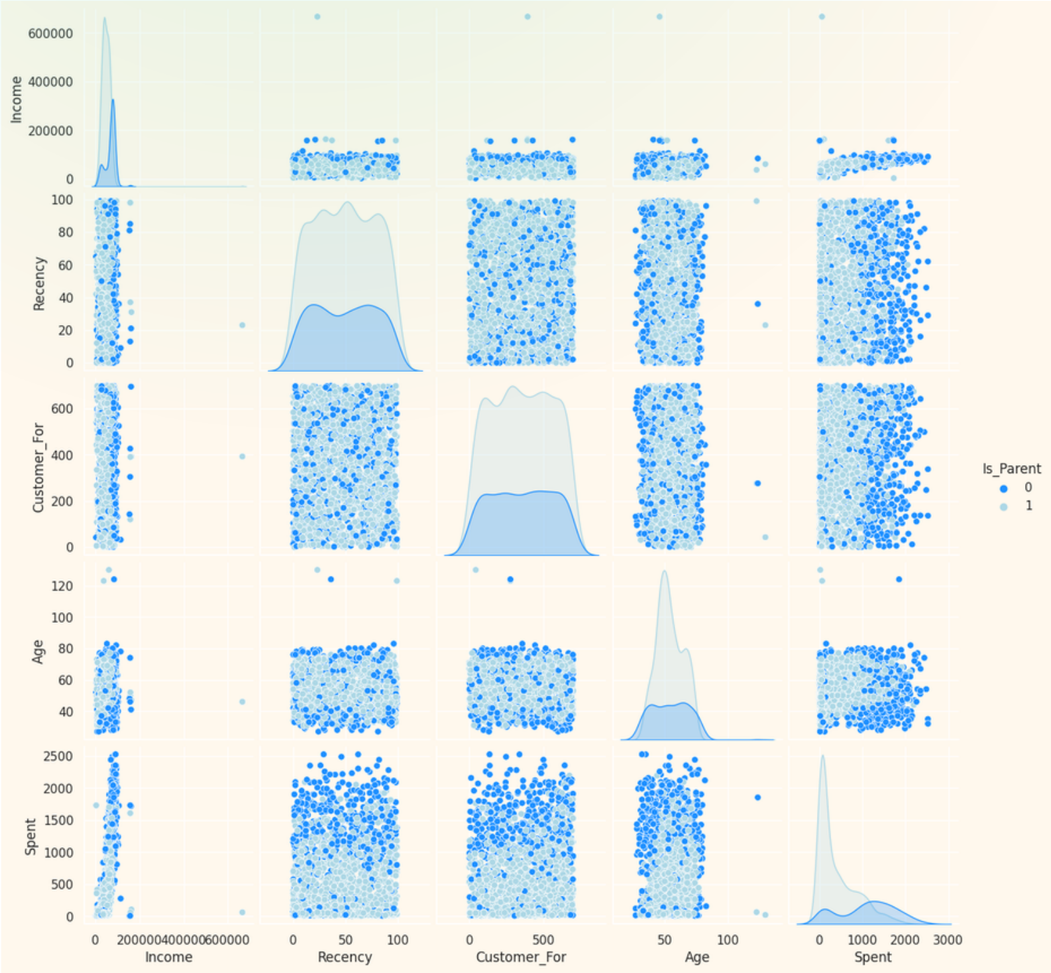
Education	count
Graduation	1116
PhD	481
Master	365
2n Cycle	200
Basic	54

FEATURE ENGINEERING



- Extract "Age" from "Year_Birth".
- Calculate total "Spent" over two years.
- Derive "Living_With" from "Marital_Status".
- Count "Children" for kids and teenagers.
- Define "Family_Size" for household clarity.
- Indicate "Is_Parent" status.
- Simplify "Education" into three categories.
- Drop redundant features.

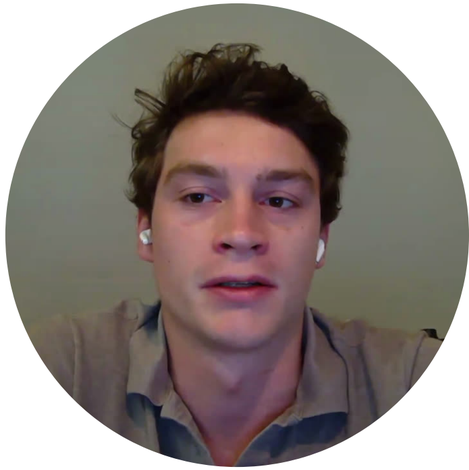
OUTLIERS



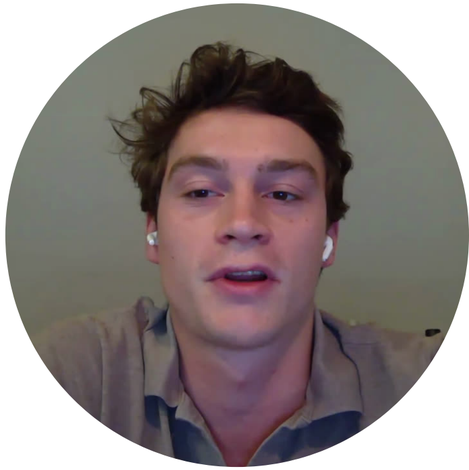
CLEARLY, THERE ARE A FEW OUTLIERS IN THE **INCOME** AND **AGE** FEATURES.

THE OUTLIERS IN THE DATA WILL BE DELETED.

THE TOTAL NUMBER OF DATA-POINTS AFTER REMOVING THE OUTLIERS ARE: 2212



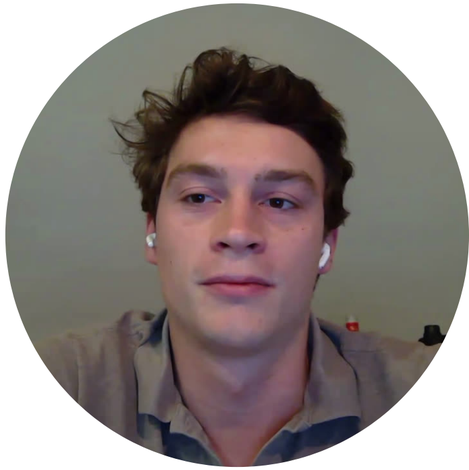
LABEL ENCODING



USING *STRINGINDEXER*

CATEGORICAL VARIABLES IN THE DATASET:
[**'EDUCATION'**, **'LIVING_WITH'**]

FINAL COLUMNS



DEMOGRAPHIC INFORMATION:

- 'INCOME'
- 'AGE'
- 'CUSTOMER_FOR'
- 'EDUCATION'

PURCHASE BEHAVIOR:

- 'RECENCY'
- 'NUMDEALSPURCHASES'
- 'NUMWEBPURCHASES'
- 'NUMCATALOGPURCHASES'
- 'NUMSTOREPURCHASES

ONLINE ACTIVITY:

- 'NUMWEBVISITSMONTH'

'FAMILY CHARACTERISTICS:

- 'KIDHOME'
- 'TEENHOME'
- 'CHILDREN'
- 'FAMILY_SIZE'
- 'IS_PARENT'
- 'LIVING_WITH'

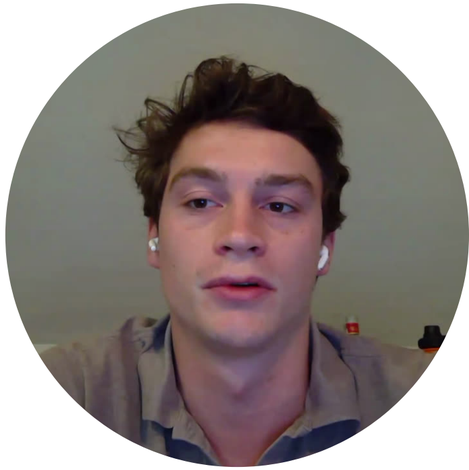
PRODUCT PREFERENCES:

- 'WINES'
- 'FRUITS'
- 'MEAT'
- 'FISH'
- 'SWEETS'
- 'GOLD'

FINANCIAL METRICS:

- 'SPENT'

VECTOR ASSEMBLER



ASSEMBLE ALL THE FEATURES USING VECTOR ASSEMBLER

```
#assemble all the features with VectorAssembler
assembler = VectorAssembler(inputCols=['Income','Kidhome', 'Teenhome', 'Recency', 'Wines', 'Fruits', 'Meat', 'Fish', 'Sweets', 'Gold', 'NumDealsPurchases',
                                        'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'Customer_For', 'Age', 'Spent',
                                        'Children', 'Family_Size', 'Is_Parent', 'Living_With', 'Education'], outputCol="features")
```

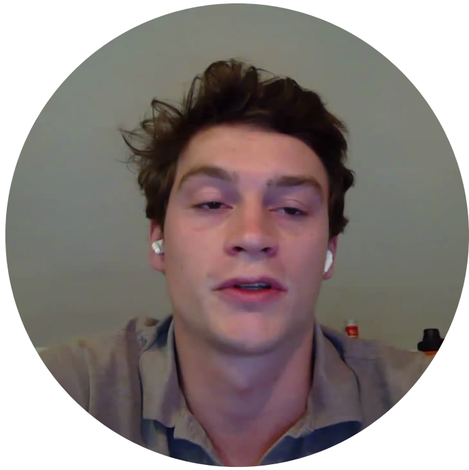
STANDARD SCALER

APPLYING STANDARD SCALER TO THE VECTOR FEATURES COLUMN

```
# Create a StandardScaler transformer object
scaler = StandardScaler(inputCol="features", outputCol="scaled_features")
df = scaler.fit(df).transform(df)
```

PRINCIPAL COMPONENT ANALYSIS (PCA)

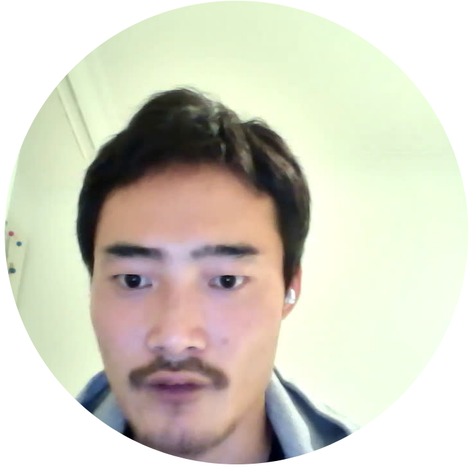
REDUCTION TO 3 DIMENSIONS



```
pca = PCA(k=3, inputCol="scaled_features", outputCol="pca_features")
df_3d = pca.fit(df).transform(df)
```

CLUSTERING

CLUSTERING THE DATA INTO 4 GROUPS

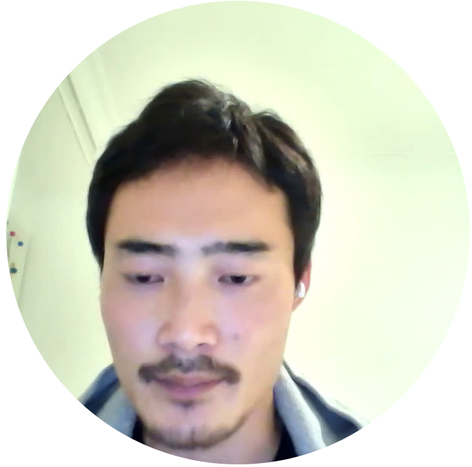


```
kmeans = KMeans(k=4, seed=221, maxIter=20, featuresCol="pca_features")

model = kmeans.fit(df_3d)
```

```
model.summary.clusterSizes
```

[609, 552, 512, 539]

CLUSTERING



CLUSTER CENTERS

```
[ 0.96648876 -4.14113021 -0.9796162 ]
[-1.91664218 -7.76965135 -0.80083877]
[-5.23649985 -4.93432097 -0.23720261]
[ 1.52107134 -6.5821081   0.36131407]
```

CLUSTERING

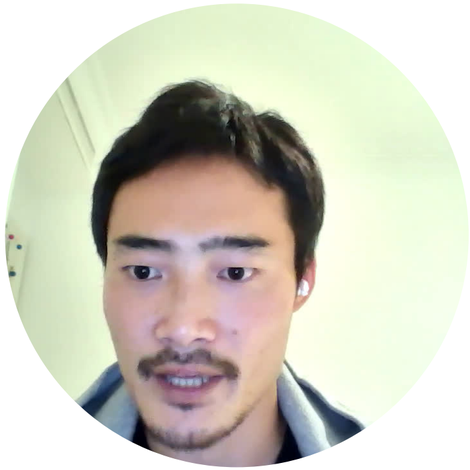
CLUSTER EVALUATIOR

```
evaluator = ClusteringEvaluator(featuresCol="scaled_features")

# Evaluate clustering by computing Silhouette score
silhouette = evaluator.evaluate(predictions)
print(f"Silhouette with squared euclidean distance = {silhouette}")

# Evaluate clustering by computing Within Set Sum of Squared Errors.
WSSSE = model.summary.trainingCost
print(f"Within Set Sum of Squared Errors = {WSSSE}")
```

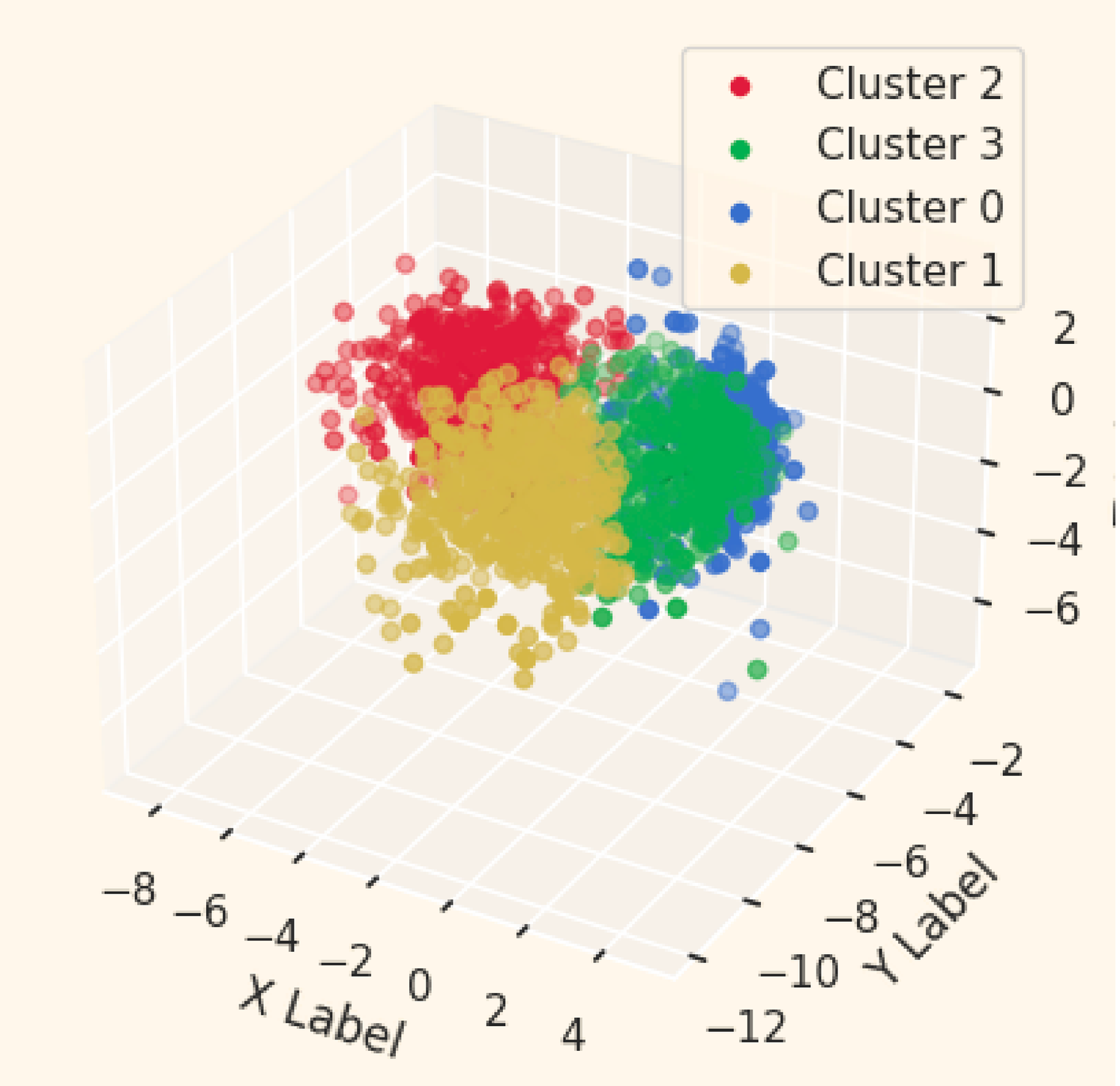
```
Silhouette with squared euclidean distance = 0.257886274446046
Within Set Sum of Squared Errors = 7550.4451503696755
```

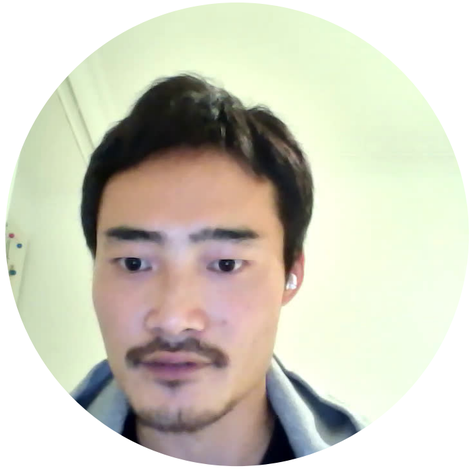
	x	y	z	label
0	-6.038836	-5.663533	-2.881266	2
1	1.819890	-5.847703	1.094035	3
2	-3.668800	-5.093755	-0.172052	2
3	1.600339	-4.369462	-0.038134	0
4	-0.397761	-6.002835	-0.295160	3
...	...	...	...	...
2207	-3.376048	-8.261926	-0.930023	1
2208	2.046670	-9.838168	0.979424	3
2209	-3.719522	-3.931759	-0.994706	2
2210	-2.540641	-7.476411	1.323551	1
2211	1.679223	-7.512706	-0.224476	3

2212 rows × 4 columns



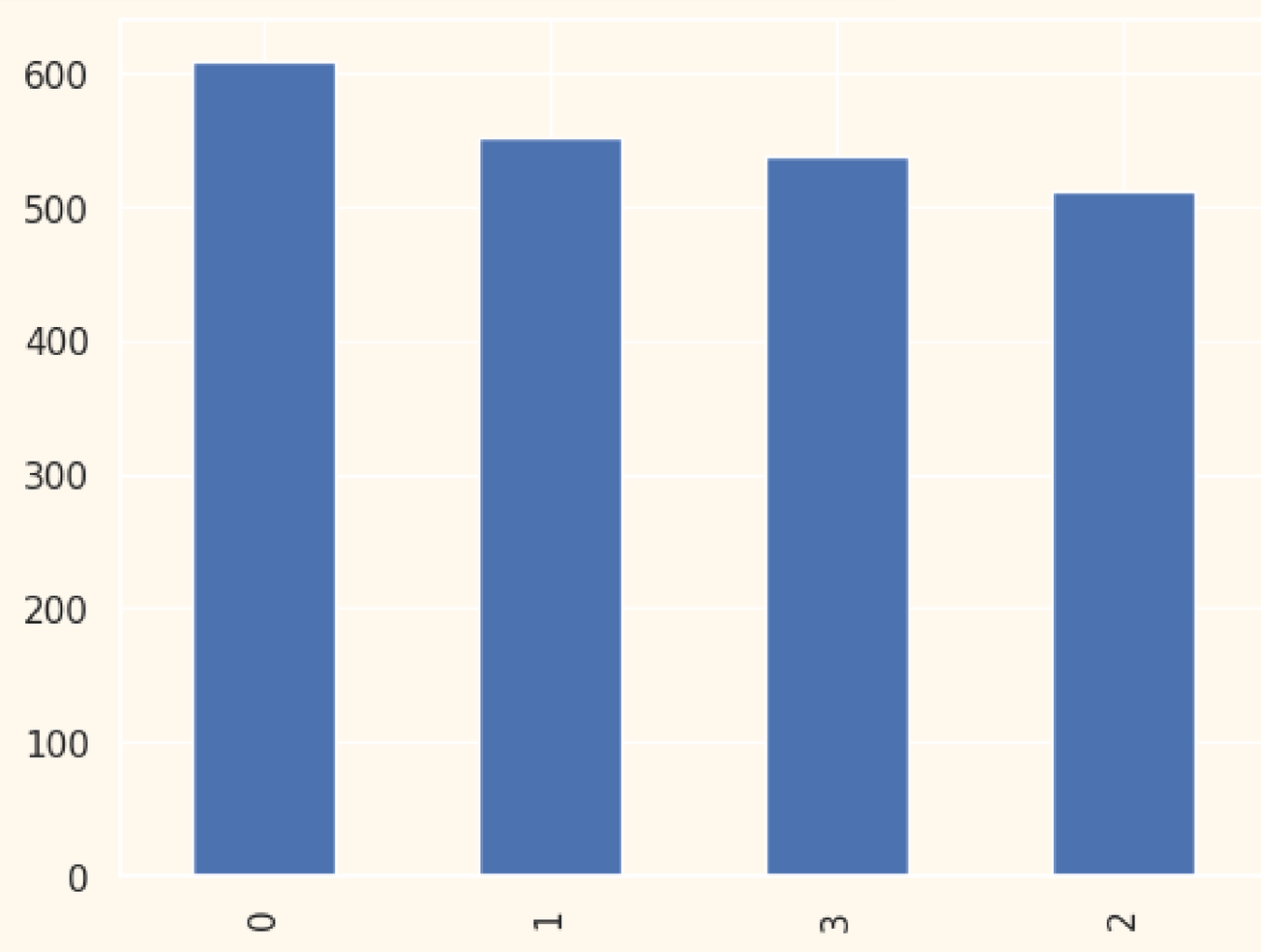
3D ANALYSIS OF CLUSTERS (PCA)

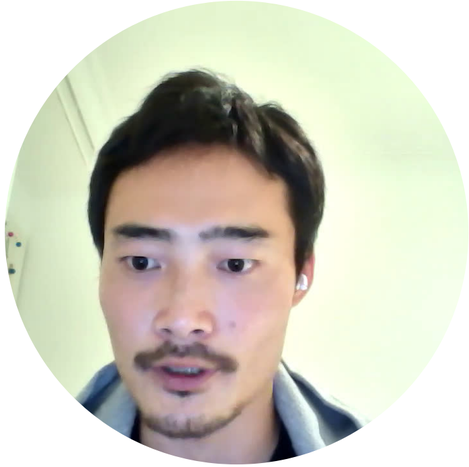




3D ANALYSIS OF CLUSTERS (PCA)

ANALYSIS OF CLUSTERS





ANALYSIS OF CLUSTERS

- Cluster 0 (Blue): Low income and low spending.
- Cluster 1 (Yellow): Moderate income, varied spending, some high spenders.
- Cluster 2 (Purple): High income, high spending.
- Cluster 3 (Red): High income, moderate spending.
- Outliers: Present mainly in the yellow cluster.

GRAPHS

GRAPH OVERVIEW:

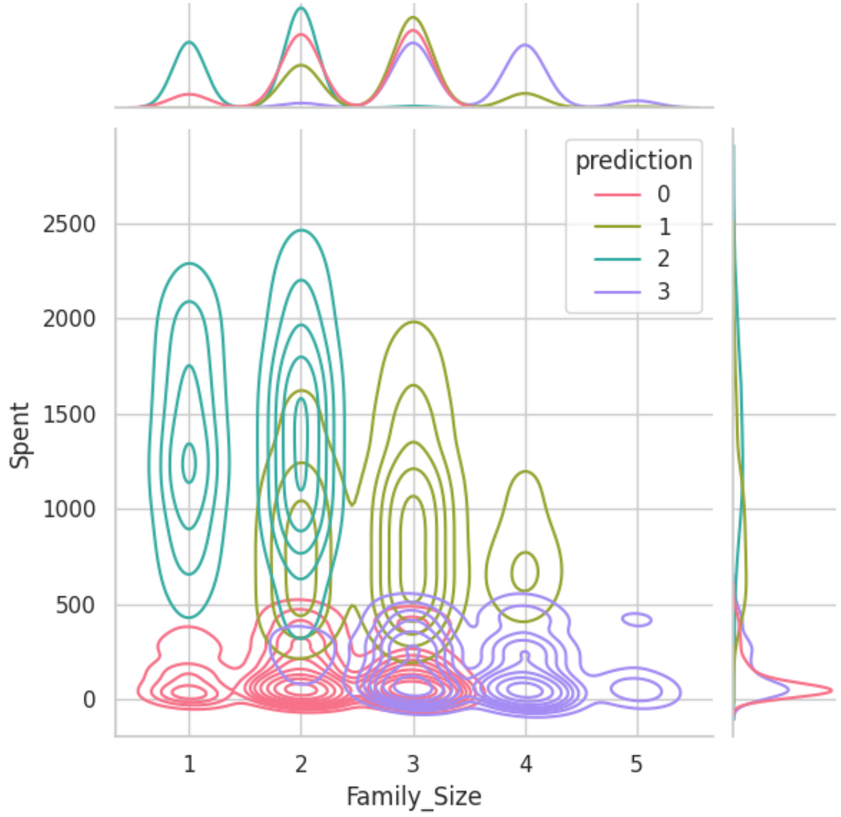
- TYPE: RIDGE AND SCATTER PLOT COMBINATION
- AXES: (X-AXIS) VS. (Y-AXIS)
- CLUSTERS: FOUR DISTINCT GROUPS (0 TO 3) DISTINGUISHED BY COLORS

KEY INSIGHTS:

- RIDGE PLOTS ILLUSTRATE THE DENSITY AND SPREAD OF SPENDING FOR VARYING LEVELS OF THE X-AXIS VARIABLE.
- SCATTER PLOT OVERLAY PROVIDES SPECIFIC DATA POINTS, SHOWING THE VARIATION OF Y AGAINST THE X-AXIS VARIABLE.
- DIFFERENT CLUSTERS INDICATE DISTINCT BEHAVIORAL SEGMENTS.

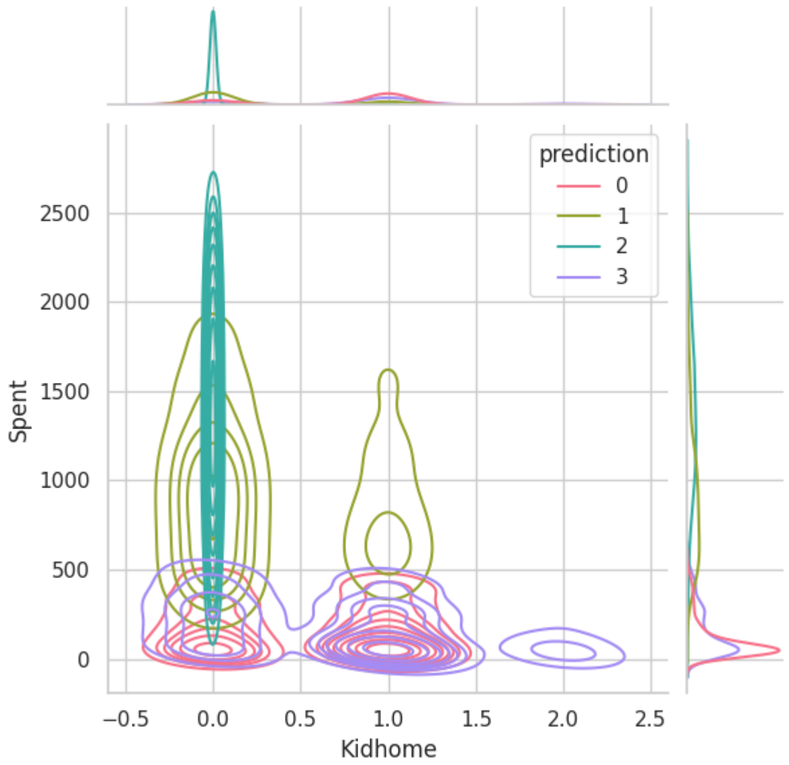
PURPOSE:

- TO UNDERSTAND HOW THE X-AXIS VARIABLE INFLUENCES SPENDING.
- TO PINPOINT KEY SEGMENTS FOR TARGETED ENGAGEMENT.
- TO INFORM STRATEGIC DECISIONS BASED ON DATA-DRIVEN INSIGHTS.



PROFILING CLUSTERS

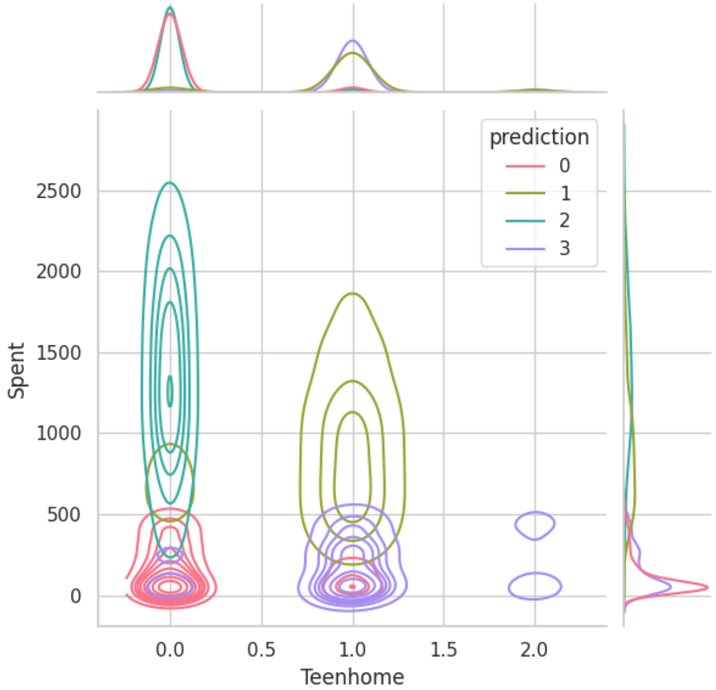
GRAPH WITH 'KIDHOME' ON X-AXIS:



CLUSTER 0: LOW 'KIDHOME', LOW SPENDING.
CLUSTER 1: MODERATE 'KIDHOME', VARIED SPENDING.
CLUSTER 2: HIGH 'KIDHOME', LOW-MODERATE SPENDING.
CLUSTER 3: LOW-MODERATE 'KIDHOME', VERY HIGH SPENDING.

PROFILING CLUSTERS

GRAPH WITH 'TEENHOME' ON X-AXIS:

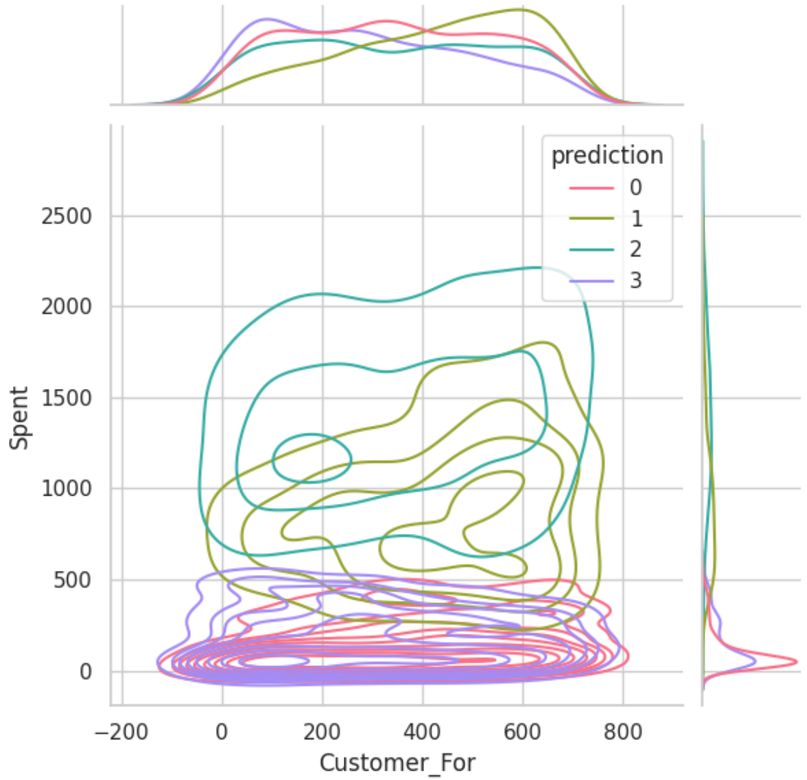


- CLUSTER 0: LOW 'TEENHOME', LOW SPENDING.
- CLUSTER 1: MODERATE 'TEENHOME', WIDE SPENDING RANGE.
- CLUSTER 2: HIGH 'TEENHOME', LOWER SPENDING.
- CLUSTER 3: MODERATE 'TEENHOME', VERY HIGH SPENDING.

PROFILING CLUSTERS

GRAPH WITH 'CUSTOMER_FOR' ON X-AXIS:

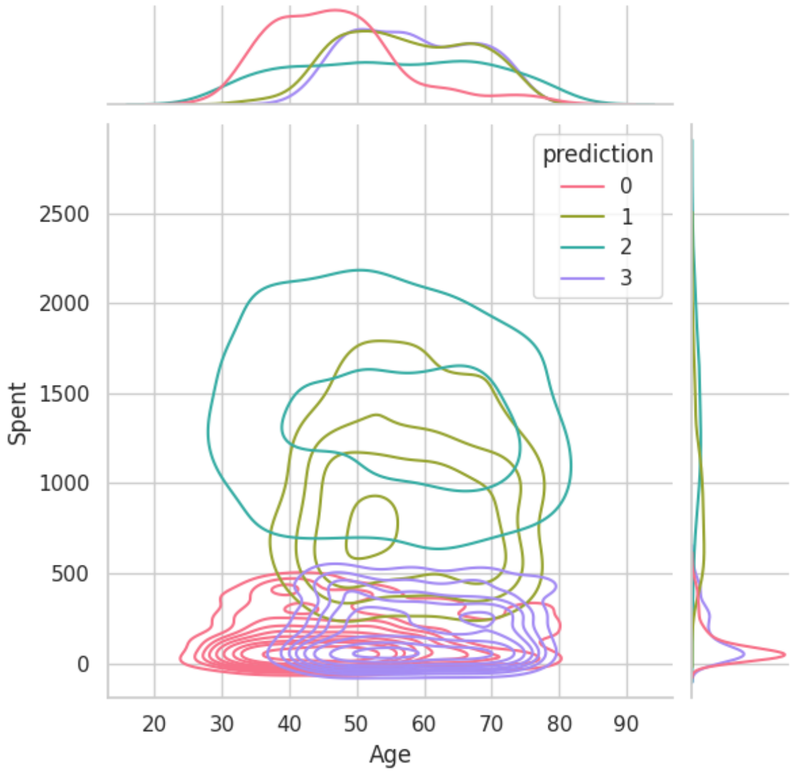
- CLUSTER 0: MIXED 'CUSTOMER_FOR', LOW SPENDING.
- CLUSTER 1: BROAD 'CUSTOMER_FOR' RANGE, MODERATE SPENDING.
- CLUSTER 2: NEWER CUSTOMERS, VARIED SPENDING.
- CLUSTER 3: MIXED 'CUSTOMER_FOR', VERY HIGH SPENDING.



PROFILING CLUSTERS

GRAPH WITH 'AGE' ON X-AXIS:

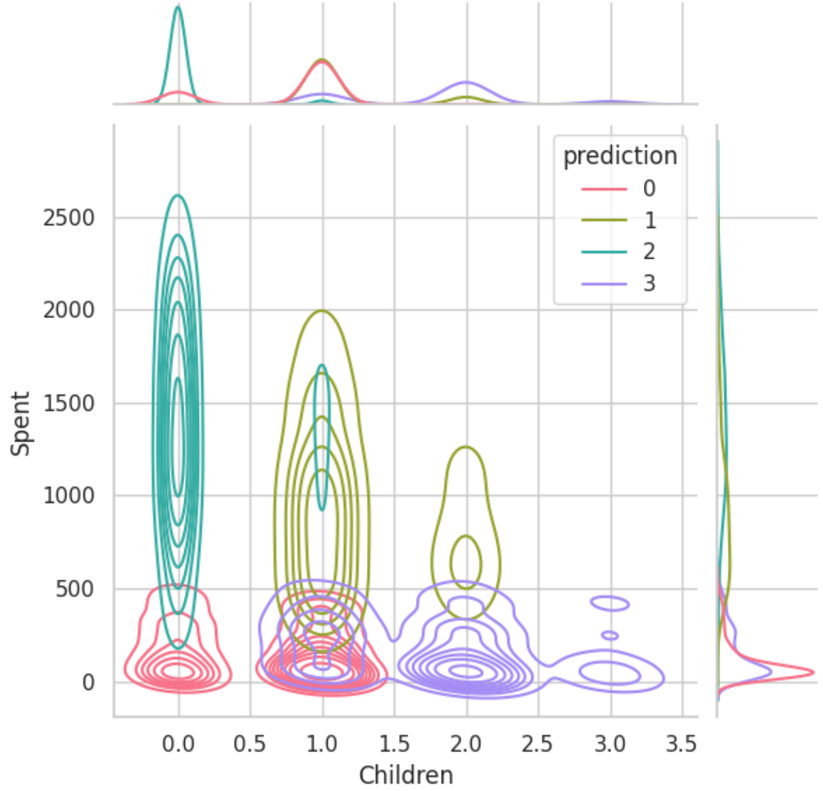
- CLUSTER 0: YOUNGER AGES, LOW SPENDING.
- CLUSTER 1: WIDE AGE RANGE, MODERATE-HIGH SPENDING.
- CLUSTER 2: MIDDLE-OLDER AGES, VARIED SPENDING.
- CLUSTER 3: BROAD AGE RANGE, VERY HIGH SPENDING.



PROFILING CLUSTERS

GRAPH WITH 'CHILDREN' ON X-AXIS:

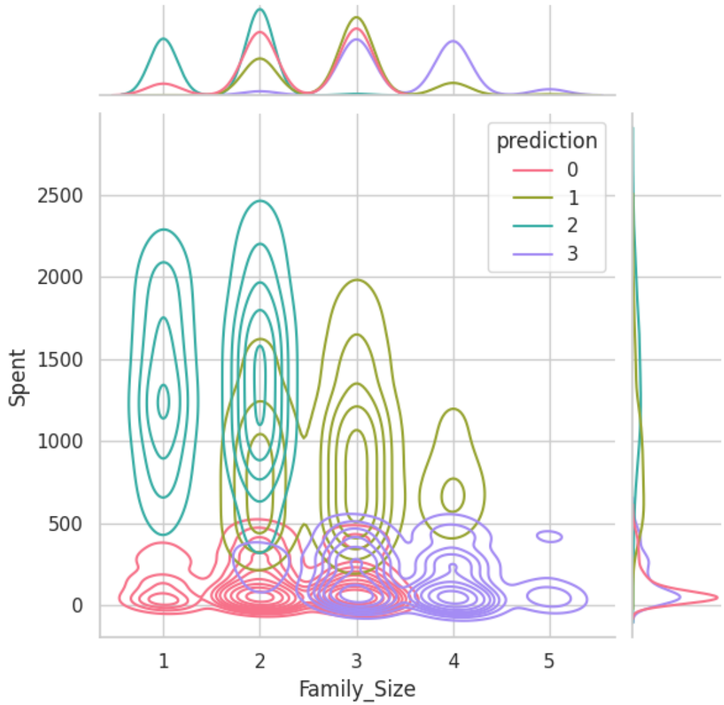
- CLUSTER 0: LOW 'CHILDREN', LOW SPENDING.
- CLUSTER 1: MODERATE 'CHILDREN', VARIED SPENDING.
- CLUSTER 2: HIGHER 'CHILDREN', MODERATE SPENDING.
- CLUSTER 3: LOW-MODERATE 'CHILDREN', VERY HIGH SPENDING.



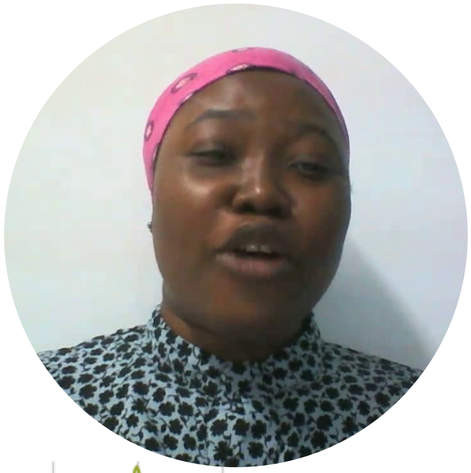
PROFILING CLUSTERS

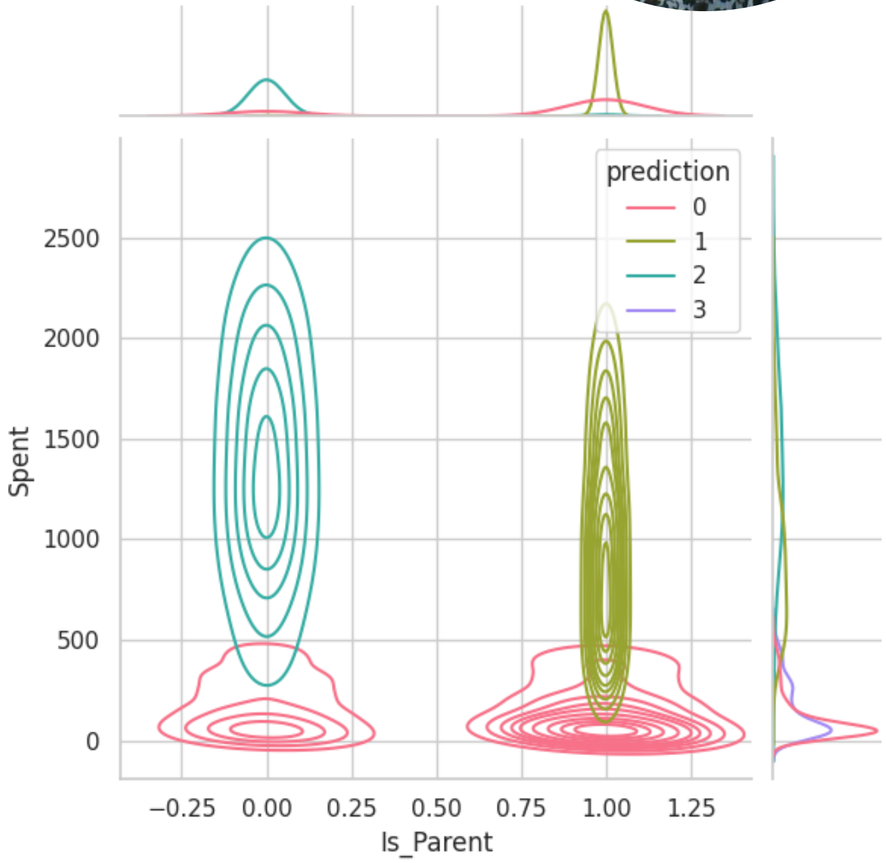
GRAPH WITH 'FAMILY_SIZE' ON X-AXIS:

- CLUSTER 0: SMALLER FAMILIES, LOWER SPENDING.
- CLUSTER 1: VARIED FAMILY SIZES, WIDE SPENDING RANGE.
- CLUSTER 2: LARGER FAMILIES, MODERATE SPENDING.
- CLUSTER 3: MID-RANGE FAMILY SIZES, VERY HIGH SPENDING.



PROFILING CLUSTERS





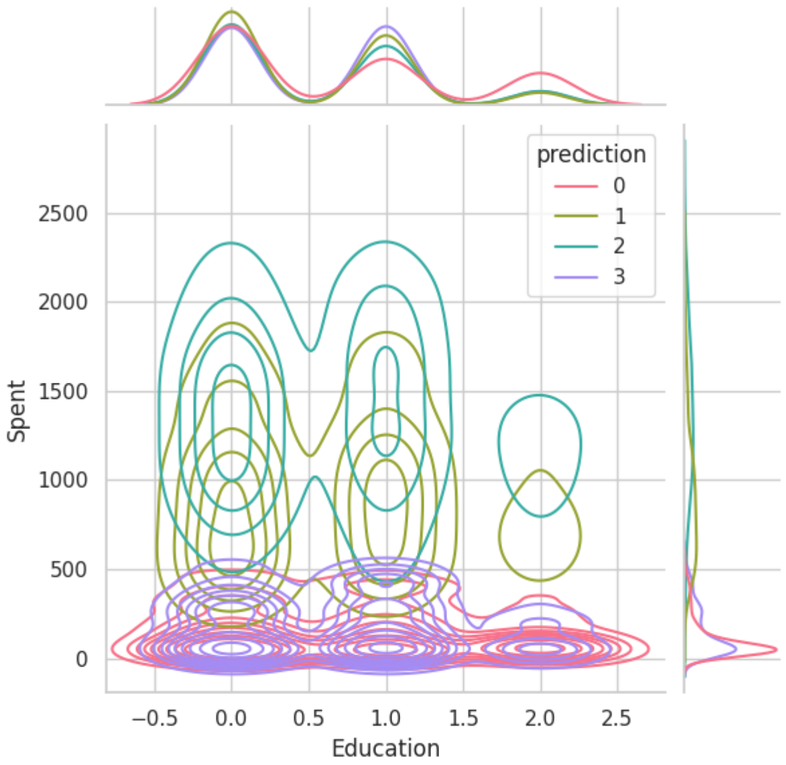
GRAPH WITH 'IS_PARENT' ON X-AXIS:

- CLUSTER 0: LIKELY NON-PARENTS, LOWER SPENDING.
- CLUSTER 1: MIX OF PARENTS AND NON-PARENTS, VARIED SPENDING.
- CLUSTER 2: MAINLY PARENTS, MODERATE SPENDING.
- CLUSTER 3: MAINLY PARENTS, VERY HIGH SPENDING.

PROFILING CLUSTERS

GRAPH WITH 'EDUCATION' ON X-AXIS:

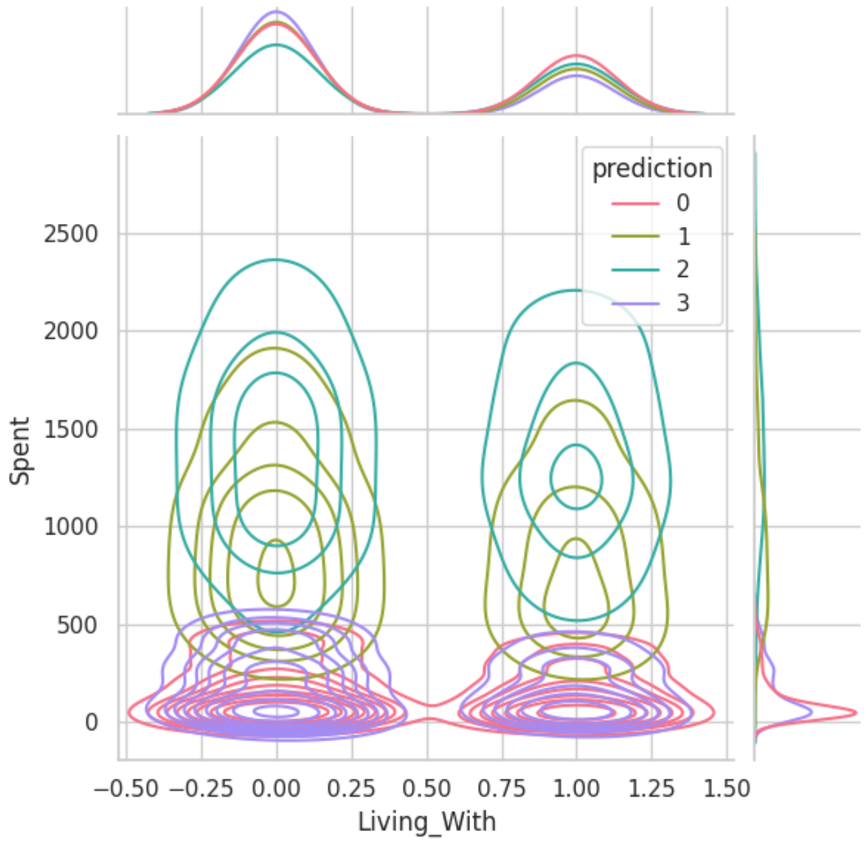
- CLUSTER 0: LOWER EDUCATION, LOWER SPENDING.
- CLUSTER 1: VARIED EDUCATION, MODERATE-HIGH SPENDING.
- CLUSTER 2: HIGHER EDUCATION, MODERATE SPENDING.
- CLUSTER 3: MIXED EDUCATION, VERY HIGH SPENDING.



PROFILING CLUSTERS

GRAPH WITH 'LIVING_WITH' ON X-AXIS:

- CLUSTER 0: LOWER 'LIVING_WITH', LOWER SPENDING.
- CLUSTER 1: VARIED 'LIVING_WITH', WIDE SPENDING RANGE.
- CLUSTER 2: HIGHER 'LIVING_WITH', MODERATE SPENDING.
- CLUSTER 3: MID-RANGE 'LIVING_WITH', VERY HIGH SPENDING.



CLUSTER 0

CHARACTERISTICS: THIS CLUSTER TYPICALLY REPRESENTS INDIVIDUALS OR SMALL FAMILIES WITH FEWER DEPENDENTS, SUCH AS CHILDREN OR TEENAGERS. THEY MAY INCLUDE YOUNGER CUSTOMERS, SINGLE ADULTS, OR COUPLES WITHOUT CHILDREN.

BEHAVIOR: THESE CUSTOMERS TEND TO SPEND LESS ON GROCERIES, WHICH COULD BE DUE TO LOWER DISPOSABLE INCOME OR FEWER HOUSEHOLD MEMBERS TO PROVIDE FOR. THEY MAY PRIORITIZE ESSENTIAL PURCHASES OR BUDGET ITEMS OVER PREMIUM PRODUCTS.

LOW SPEND, LOW DEPENDENTS/COMMITMENTS

CLUSTER 1

Characteristics: Cluster 1 is a diverse segment that includes customers with varying family sizes and living situations. It may comprise middle-aged adults, mixed households, and families with children or teenagers.

Behavior: Their spending patterns are varied, with some making moderate purchases and others spending more liberally. This suggests that this group may include both budget-conscious customers and those willing to spend more on groceries when needed.

MODERATE TO HIGH SPEND, MIXED DEPENDENTS/COMMITMENTS

CLUSTER 2

Characteristics: This cluster appears to consist of larger families or individuals with higher educational levels or commitments. There may be a higher number of dependents or cohabitants, such as children, teenagers, or extended family members.

Behavior: Despite the potential for higher spending due to more significant household needs, this group shows moderate expenditure, which could indicate price sensitivity or a focus on value-for-money purchases.

MODERATE SPEND, LARGER FAMILIES OR HIGHER COMMITMENTS

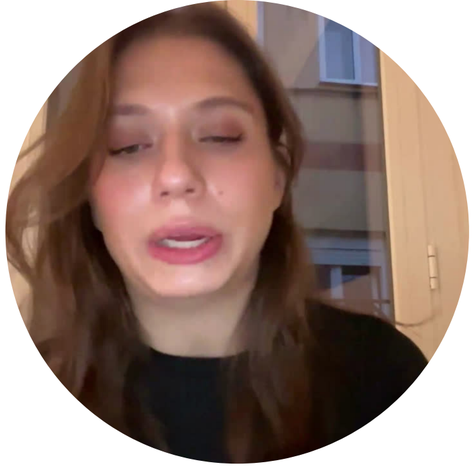
CLUSTER 3

Characteristics: Customers in this cluster may not be defined by a specific demographic such as age or family size, but rather by their high spending behavior.

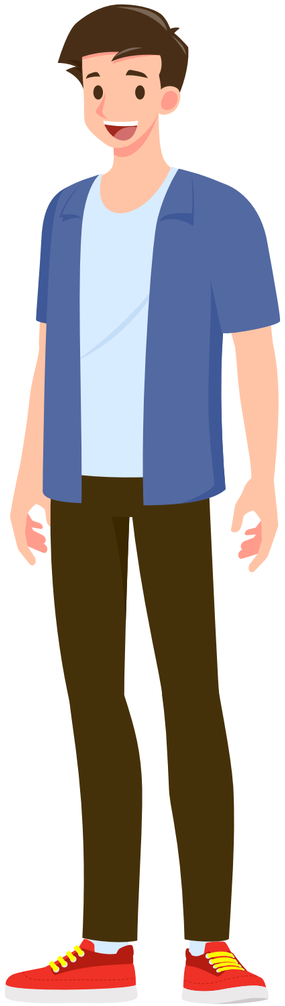
Behavior: This suggests a group with higher disposable income or a preference for premium products. They may be less sensitive to price, opting for convenience, quality, or luxury items. These could be long-term, loyal customers or those who make bulk purchases.

HIGH SPEND, AFFLUENT OR LESS PRICE-SENSITIVE

CLUSTER 0 : "BUDGET-CONSCIOUS INDIVIDUALS"



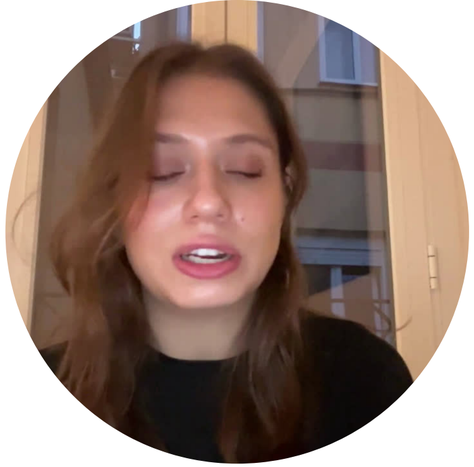
MEET ALEX, A YOUNG PROFESSIONAL LIVING INDEPENDENTLY. WITH A FOCUS ON ESSENTIALS, ALEX IS BUDGET-CONSCIOUS AND APPRECIATES PROMOTIONS AND DISCOUNTS. TARGETING STRATEGIES SHOULD EMPHASIZE VALUE FOR MONEY AND OCCASIONAL DEALS TO CAPTURE ALEX'S ATTENTION.



CLUSTER 1 : "DIVERSE HOUSEHOLD DYNAMICS"

INTRODUCING THE THOMPSONS, A DIVERSE FAMILY WITH DIFFERENT NEEDS. SOME FAMILY MEMBERS ARE BUDGET-CONSCIOUS, WHILE OTHERS ARE WILLING TO SPLURGE ON QUALITY. A SEGMENTED MARKETING APPROACH IS KEY FOR THE THOMPSONS, ENSURING TAILORED MESSAGES TO ADDRESS THE VARIED PREFERENCES WITHIN THE FAMILY.

CLUSTER 2 : "EDUCATED FAMILIES, VALUE SEEKERS"



MEET THE JOHNSONS, A FAMILY WITH CHILDREN AND A FOCUS ON EDUCATION. DESPITE HAVING SIGNIFICANT HOUSEHOLD NEEDS, THEY REMAIN VALUE-ORIENTED. TARGETING THE JOHNSONS SHOULD INVOLVE FAMILY-SIZED PACKAGES, BULK DEALS, AND PROMOTIONS RELATED TO EDUCATION AND BACK-TO-SCHOOL ESSENTIALS.

CLUSTER 3 : "AFFLUENT AFICIONADOS"

INTRODUCING THE PARKERS, A GROUP WITH A PENCHANT FOR PREMIUM PRODUCTS AND A TASTE FOR THE FINER THINGS IN LIFE. LESS PRICE-SENSITIVE, THE PARKERS APPRECIATE EXCLUSIVITY AND LOYALTY REWARDS. MARKETING TO THIS CLUSTER SHOULD EMPHASIZE LUXURY, CONVENIENCE, AND EXCLUSIVE OFFERINGS TO FOSTER LONG-TERM LOYALTY.