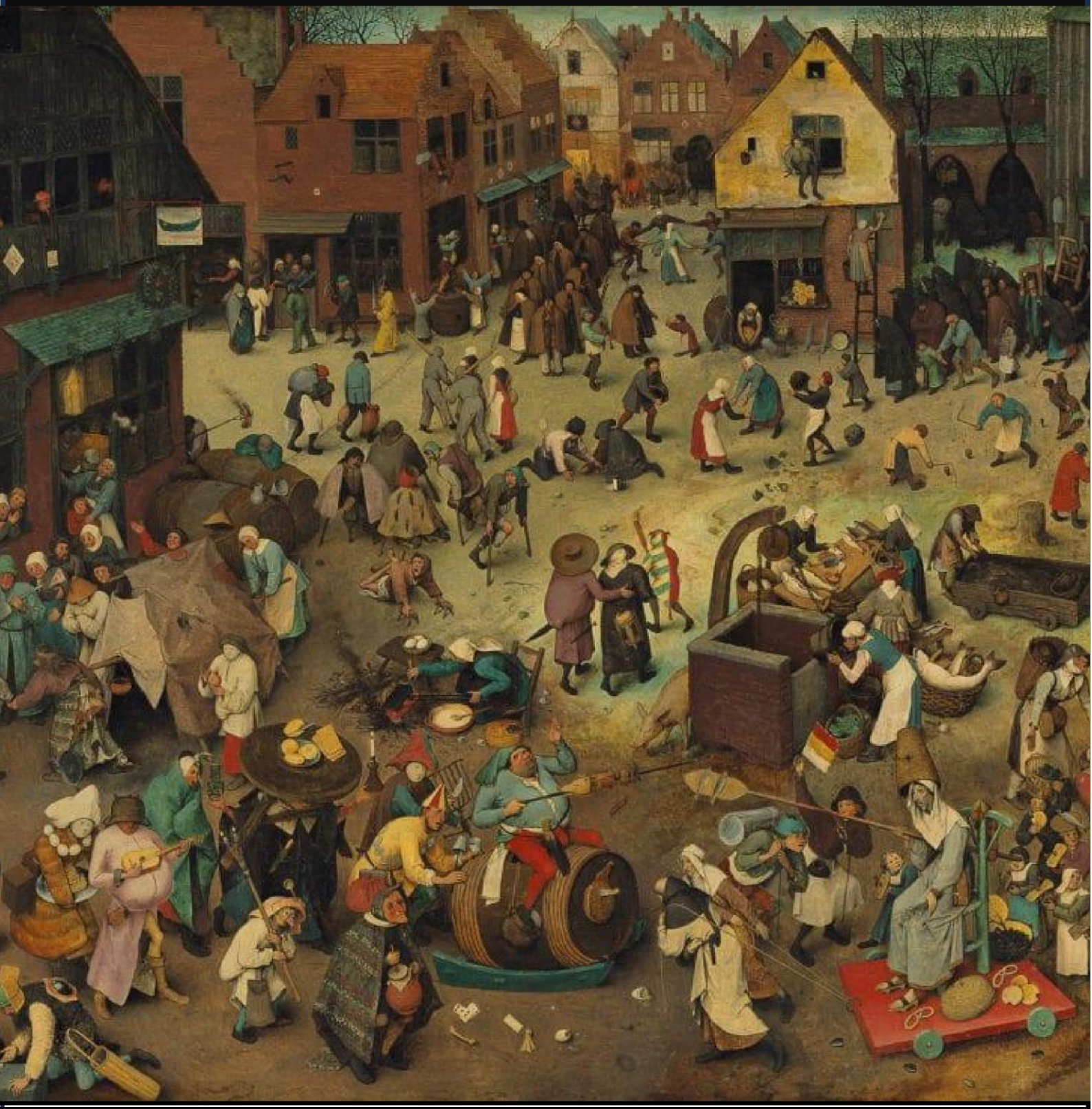
Coherent Consumer Behavioral Analysis



Group: 4

Project Summary

Batch Details	PGPDSE (Post Graduate Program in Data Science and Engineering) Gurgaon, July 2023 Batch
Team Members	Harsh Bhalla, Krishan Kumar, Divyansh Ganotra, Yash Sharma, Shubhrant Shah, Animesh Srivastava, Akash Roy, Sahib Nijhavan
Domain of the Project	Customer Analytics
Proposed Project Title	'Coherent Consumer Behavioral Analysis'
Group Number	4
Team Leader	Harsh Bhalla
Mentor Name	Mr. Sourab Reddy

Date: 15th March, 2024

Contents

Introduction

A good understanding of customers is extremely important for running a successful business. KYC or know your customer is what, makes all the difference for many companies. KYC helps them do their best in creating, communicating, and delivering their offerings by tailoring them to their customer's needs. And that makes customer analytics the most important part of both marketing analytics and the marketing function of a company in general. But understanding and meeting customers' needs is easier said than done. In fact, customer analytics is a very broad area. It may include a wide range of characteristics of customers and their behavior and numerous different outcomes and performance indicators that the business might be interested in. In this project we've decided to focus on one of the most fundamental marketing frameworks that of segmentation, targeting and positioning known as the STP framework. The STP framework is the most logical choice as it applies to all areas of business and marketing activities.

Problem Statement

In the realm of fast-moving consumer goods (FMCG), the pressing challenge is harnessing the power of customer analytics to extract meaningful insights from vast datasets. Recognizing the importance of customer analytics in marketing, which includes guiding decisions, enhancing satisfaction, and refining strategies, effectively implementing a range of techniques such as cluster analysis, dimensionality reduction, descriptive statistics, elasticity modeling remains a complex task. This project aims to navigate this challenge by utilizing FMCG customer data, striving for a seamless integration of these analytical methods. The objective is to achieve a holistic understanding of customer segmentation, behavior interpretation, and precise predictions for informed and adaptive marketing strategies, fostering sustainable business growth.

Abstract

The problem at hand involves leveraging Customer Analytics through a multi-step approach. First, the challenge lies in effectively segmenting customers using cluster analysis and dimensionality reduction, employing Python and packages like NumPy, SciPy, and scikit-learn. The second phase demands applying descriptive statistics to interpret segmented customers' behavior, forming hypotheses for subsequent modeling. The third stage involves elasticity modeling for purchase probability exploring various elasticities through advanced regression techniques. Model deployment through the pickle package adds an additional layer of complexity to address.

Data Dictionary

Variable	Data Type	Range	Description
ID	Numerical	Integer	Shows a unique identifier of a customer.
Day	Numerical	Integer	Day when the customer has visited the store
Incidence	Categorical	{0, 1}	Purchase Incidence 0: The customer has not purchased an item from the category of interest 1: The customer has purchased an item from the category of interest
Brand	Categorical	{0, 1, 2, 3, 4, 5}	Shows which brand the customer has purchased 0: No brand was purchased 1, 2, 3, 4, 5: Brand ID
Quantity	Numerical	integer	Number of items bought by the customer from the product category of interest
Last_Inc_Brand	Categorical	{0, 1, 2, 3, 4, 5}	Shows which brand the customer has purchased on their previous store visit 0: No brand was purchased 1, 2, 3, 4, 5: Brand ID
Last_Inc_Quantity	Numerical	Integer	Number of items bought by the customer from the product category of interest during their previous store visit

Price_1	Numerical	Real	Price of an item from Brand 1 on a particular day
Price_2	Numerical	Real	Price of an item from Brand 2 on a particular day
Price_3	Numerical	Real	Price of an item from Brand 3 on a particular day
Price_4	Numerical	Real	Price of an item from Brand 4 on a particular day
Price_5	Numerical	Real	Price of an item from Brand 5 on a particular day
Promotion_1	Categorical	{0, 1}	Indicator whether Brand 1 was on promotion or not on a particular day 0: There is no promotion 1: There is promotion
Promotion_2	Categorical	{0, 1}	Indicator of whether Brand 2 was on promotion or not on a particular day 0: There is no promotion 1: There is promotion
Promotion_3	Categorical	{0, 1}	Indicator of whether Brand 3 was on promotion or not on a particular day 0: There is no promotion 1: There is promotion
Promotion_4	Categorical	{0, 1}	Indicator of whether Brand 4 was on promotion or not on a particular day 0: There is no promotion 1: There is promotion
Promotion_5	Categorical	{0, 1}	Indicator of whether Brand 5 was on promotion or not on a particular day 0: There is no promotion 1: There is promotion
Sex	Categorical	{0, 1}	Biological sex (gender) of a customer. 0: male 1: female
Marital status	Categorical	{0, 1}	Marital status of a customer. 0: single 1: non-single (divorced / separated / married / widowed)
Age	Numerical	Integer	The age of the customer in years
Education	Categorical	{0, 1, 2, 3}	Level of education of the customer 0: other / unknown 1: high school 2: university 3: graduate school
Income	Numerical	Real	Self-reported annual income in US dollars of the customer.
Occupation	Categorical	{0, 1, 2}	Category of occupation of the customer 0: unemployed / unskilled 1: skilled employee / official 2: management / self-employed / highly qualified employee

Dataset Description

a. Segmentation dataset: The dataset consists of information about the purchasing behavior of 2,000 individuals from a given area when entering a physical 'FMCG' store. All data has been collected through the loyalty cards they use at checkout. The data has been preprocessed and there are no missing values. In addition, the volume of the dataset has been restricted and anonymized to protect the privacy of the customers.

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
```

b. Purchase dataset: The dataset consists of information about the purchases of chocolate candy bars of 500 individuals from a given area when entering a physical 'FMCG' store in the period of 2 years. All data has been collected through the loyalty cards they use at checkout. The data has been preprocessed and there are

no missing values. In addition, the volume of the dataset has been restricted and anonymized to protect the privacy of the customers.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 58693 entries, 0 to 58692
Data columns (total 24 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 58693 non-null  int64
 1   Day                58693 non-null  int64
 2   Incidence          58693 non-null  int64
 3   Brand              58693 non-null  int64
 4   Quantity           58693 non-null  int64
 5   Last_Inc_Brand     58693 non-null  int64
 6   Last_Inc_Quantity  58693 non-null  int64
 7   Price_1            58693 non-null  float64
 8   Price_2            58693 non-null  float64
 9   Price_3            58693 non-null  float64
 10  Price_4            58693 non-null  float64
 11  Price_5            58693 non-null  float64
 12  Promotion_1        58693 non-null  int64
 13  Promotion_2        58693 non-null  int64
 14  Promotion_3        58693 non-null  int64
 15  Promotion_4        58693 non-null  int64
 16  Promotion_5        58693 non-null  int64
 17  Sex                58693 non-null  int64
 18  Marital status     58693 non-null  int64
 19  Age                58693 non-null  int64
 20  Education          58693 non-null  int64
 21  Income             58693 non-null  int64
 22  Occupation         58693 non-null  int64
 23  Settlement size    58693 non-null  int64
dtypes: float64(5), int64(19)
memory usage: 10.7 MB
```

Methodology

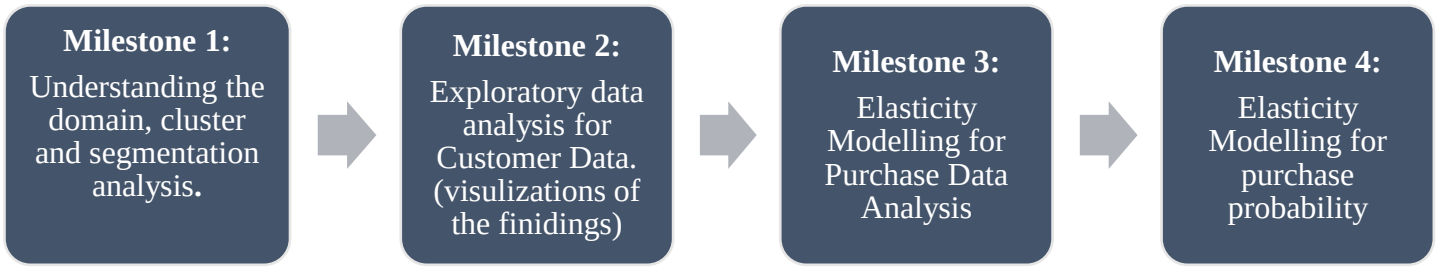
Begin by thoroughly studying and comprehending relevant theoretical concepts encompassing marketing fundamentals and the rationale behind choosing specific models for customer analytics. Establish a solid understanding of the chosen methodologies and their applicability throughout the project.

a. Cluster Analysis and Dimensionality Reduction: Implement cluster analysis and dimensionality reduction techniques using popular Python packages - NumPy, SciPy, and scikit-learn. Explore both hierarchical and flat clustering methods, emphasizing the K-means algorithm. Visualize data effectively to enhance comprehension. Utilize Principal Components Analysis (PCA) through the scikit-learn package for dimensionality reduction. Integrate the outcomes of both techniques to gain a comprehensive insight into customer segmentation. Execute model deployment using the pickle package for practical applicability.

b. Descriptive Statistics and Behavioral Interpretation: Proceed to the exploratory phase of analysis by applying descriptive statistics to the segmented customer data. Interpret customer behavior by obtaining intuitive insights through descriptive statistics, categorized by brand and segment. Visualization of findings will play a crucial role in forming hypotheses about customer segments, setting the groundwork for subsequent modelling activities.

c. Elasticity Modelling for In-depth Insights: Participate in the analysis of elasticity, emphasizing purchase likelihood, brand selection, and purchase volume. Move beyond fixed metrics and engage in intricate computations of purchase probability elasticity, brand preference's own price elasticity, brand choice cross-

price elasticity, and purchase quantity elasticity. Utilize logistic regression techniques through well-established libraries such as scikit-learn. Utilize this stage to investigate models, extracting varied insights to improve decision-making and formulate effective strategies.



Fundamental Marketing Framework

a. STP Framework: STP is a fundamental marketing framework. STP framework lays out the process of exploring potential customers and understanding them. It can be applied to all areas of business and marketing activities.

b. Segmentation: The process of dividing a population of customers into groups that share similar characteristics.

Observations within the same group would have comparable purchasing behavior.

Observations within the same group would respond similarly to different marketing activities.

c. Targeting: The process of evaluating potential profits from each segment and deciding which segments to focus on. Targeting is generally considered in 'advertising' territory.

Selecting ways to promote your products. You can target one segment on TV and another online.

Examining customers' perception (involves psychology and usually budget constraints).

d. Positioning**:** What product characteristics do the customers from a certain segment need?

Shows how a product should be **presented** to the customers and through **what channels**.

e. Customers**:** The clients of our business are individuals rather than organizations (B2C). The data we used came from Fast-moving consumer goods company (FMCG). The advantage of B2C model in terms of data science is that we have much more data points.

f. Marketing Mix**:** Develop the best product or service (purchase probability) and offer it at the right price (brand choice probability) through the right channels (purchase quantity).

We try to answer these fundamental questions about positioning and marketing mix:

1. Will the customer buy a product from a particular product category when they entry the shop?
    - To buy or not to buy – Purchase Probability.
2. Which brand is the customer going to choose?
    - Which brand to choose – Brand choice Probability
3. How many units is the customer going to purchase?
    - How many quantities – Purchase quantity.

4P's of Marketing:

1. Product characteristics: Product features; Branding; Packaging.
2. Price of offering: Product cost; Long term price changes; Discounts.
3. Promotion: Price reduction; Display; Feature
4. Place or channel of offering: – Intensive Distribution, Selective Distribution, Exclusive Distribution

Price Elasticity: It measures how a variable of interest (purchase behavior) changes when the price changes.

Price Elasticity

$$E = \frac{\frac{\Delta Y}{Y}}{\frac{\Delta P}{P}}$$

here, Y: economic variable of interest (number of units sold), P: price

Types of Price Elasticity:

1. Own Price Elasticity – Price elasticity with respect to the same product.
2. Cross Price Elasticity – Price elasticity with respect to another product.

Supply and Demand: The cheaper the product the higher the demand.

$$Revenue_i = P_i * Q_i$$

here, Q: quantity. P: price

We can use Price Elasticity concept to find the point at which price times quantity is optimal.

K-means Clustering:

1. Choose the number of clusters
2. Specify cluster seeds
3. Calculate the centroid (geometrical center)
4. Repeat until the centroids stop changing

Principal Components Analysis – PCA:

The goal of PCA is to find the best possible subspace which explain most of the variance. Most commonly it is used to reduce the dimensionality (number of features) of a problem.

K-means with PCA:

1. Reduce Dimensionality with PCA
2. Perform K-means with PCA scores as features
3. Visualize and interpret clusters

Price Elasticity of Purchase Probability:

It quantifies the change in probability of purchase of a product with a given change in its price.

Own-price elasticity of purchase probability

$$E = beta * price * (1 - Pr(purchase))$$

Customer Analytics

The dataset comprises information on the purchasing patterns of 2000 individuals who visited a physical FMCG store. The data was gathered by analyzing the usage of loyalty cards during the checkout process.

We performed customer segmentation by applying hierarchical and flat clustering techniques for dividing customers into groups. We also use Principal Component Analysis (PCA) to reduce the dimensionality of the problem, as well as combining PCA an K-means for an even more professional customer segmentation.

EXPLORE DATASET

We imported the 'segmentation data' dataset and did descriptive analysis to explore the dataset. We look at the data to gain some insight.

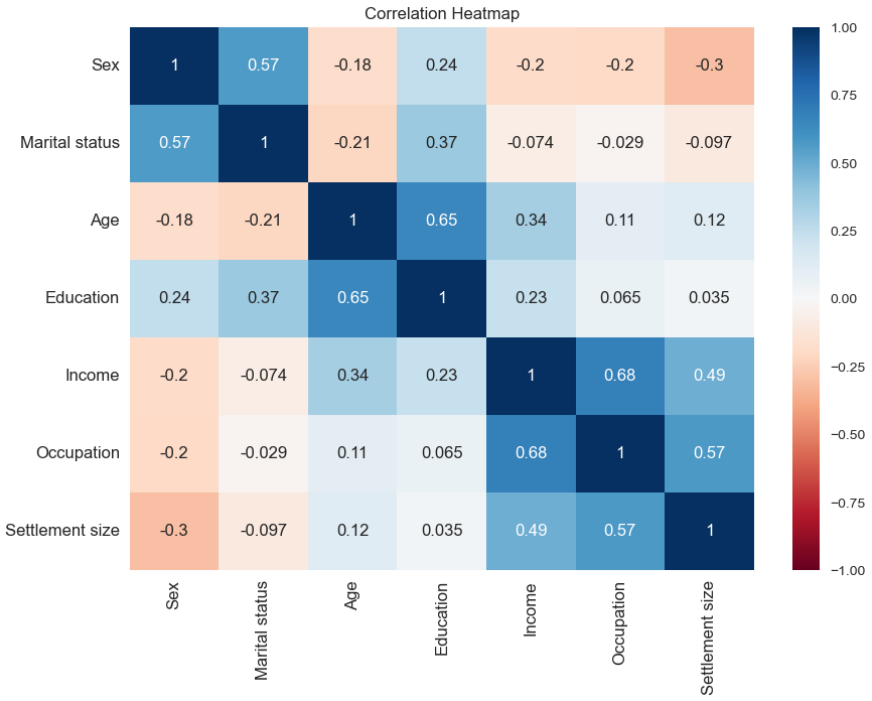
	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
count	2000.000000	2000.000000	2000.000000	2000.00000	2000.000000	2000.000000	2000.000000
mean	0.457000	0.496500	35.909000	1.03800	120954.419000	0.810500	0.739000
std	0.498272	0.500113	11.719402	0.59978	38108.824679	0.638587	0.812533
min	0.000000	0.000000	18.000000	0.00000	35832.000000	0.000000	0.000000
25%	0.000000	0.000000	27.000000	1.00000	97663.250000	0.000000	0.000000
50%	0.000000	0.000000	33.000000	1.00000	115548.500000	1.000000	1.000000
75%	1.000000	1.000000	42.000000	1.00000	138072.250000	1.000000	1.000000
max	1.000000	1.000000	76.000000	3.00000	309364.000000	2.000000	2.000000

CORRELATION ESTIMATION

Then we compute Pearson correlation coefficient for the features in our data set.

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
Sex	1.000000	0.566511	-0.182885	0.244838	-0.195146	-0.202491	-0.300803
Marital status	0.566511	1.000000	-0.213178	0.374017	-0.073528	-0.029490	-0.097041
Age	-0.182885	-0.213178	1.000000	0.654605	0.340610	0.108388	0.119751
Education	0.244838	0.374017	0.654605	1.000000	0.233459	0.064524	0.034732
Income	-0.195146	-0.073528	0.340610	0.233459	1.000000	0.680357	0.490881
Occupation	-0.202491	-0.029490	0.108388	0.064524	0.680357	1.000000	0.571795
Settlement size	-0.300803	-0.097041	0.119751	0.034732	0.490881	0.571795	1.000000

Now we plot the correlations using a Heat Map. Heat Maps are a great way to visualize correlations using color coding. We use RdBu as a color scheme and set the range from -1 to 1, as it is the range of the Pearson Correlation.



In this case we find the range from -0.25 to 0.68, they are the minimum and maximum correlation indices between our features.

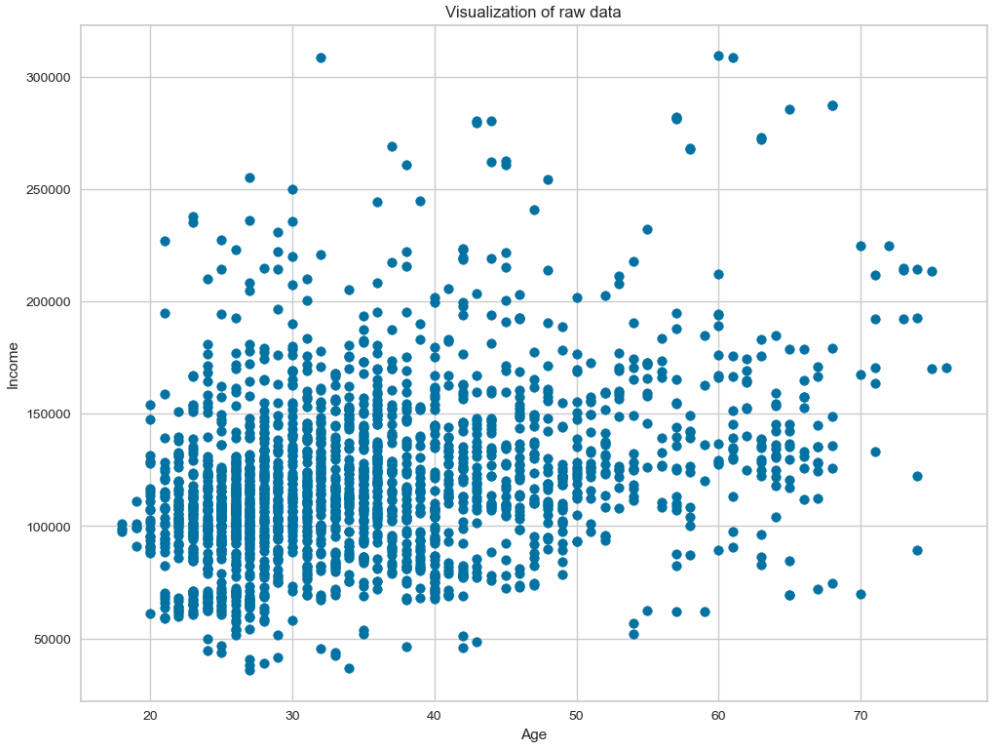
Inference:

1. Age and Education exhibit a strong correlation with a coefficient of 0.65.
2. Income and Occupation exhibit a strong correlation with a coefficient of 0.68.

Exploring the correlation between the features of the consumers is the first step to identify similar consumers and putting them together in groups, which is the essence of segmentation.

VISUALIZE DATA

We will create a scatterplot using 2000 data points, visualizing the relationship between Age and Income.



STANDARDIZATION

Our segmentation models will be based on similarities and differences between individual consumers on the features that characterize them. We'll quantify these similarities and differences. We want to treat all the features equally and we can achieve that by transforming the features in such a way that their values fall within the same numerical range. Otherwise, in our case Income would be considered much more important than Education for instance. This is what is also referred to as bias.

We apply Standardizing technique on data so that all features have equal weight. This is important for modelling.

```
array([[-0.91739884, -0.99302433,  2.65361447, ...,  0.09752361,
         0.29682303,  1.552326  ],
       [ 1.09003844,  1.00702467, -1.18713209, ...,  0.78265438,
         0.29682303,  1.552326  ],
       [-0.91739884, -0.99302433,  1.11731585, ..., -0.83320224,
        -1.26952539, -0.90972951],
       ...,
       [-0.91739884, -0.99302433, -0.41898277, ..., -0.90695688,
        -1.26952539, -0.90972951],
       [ 1.09003844,  1.00702467, -1.01643224, ..., -0.60332923,
        -1.26952539, -0.90972951],
       [-0.91739884, -0.99302433, -0.93108232, ..., -1.3789866 ,
        -1.26952539, -0.90972951]])
```

CLUSTERING

The main goal of clustering is to group individual observations so that the observation from one group are very similar to each other. In addition, we'd like them to be very different from the observations in other groups.

Clustering is of two types:

1. Hierarchical Clustering
    a. Divisive clustering
    b. Agglomerative clustering
2. Flat Clustering

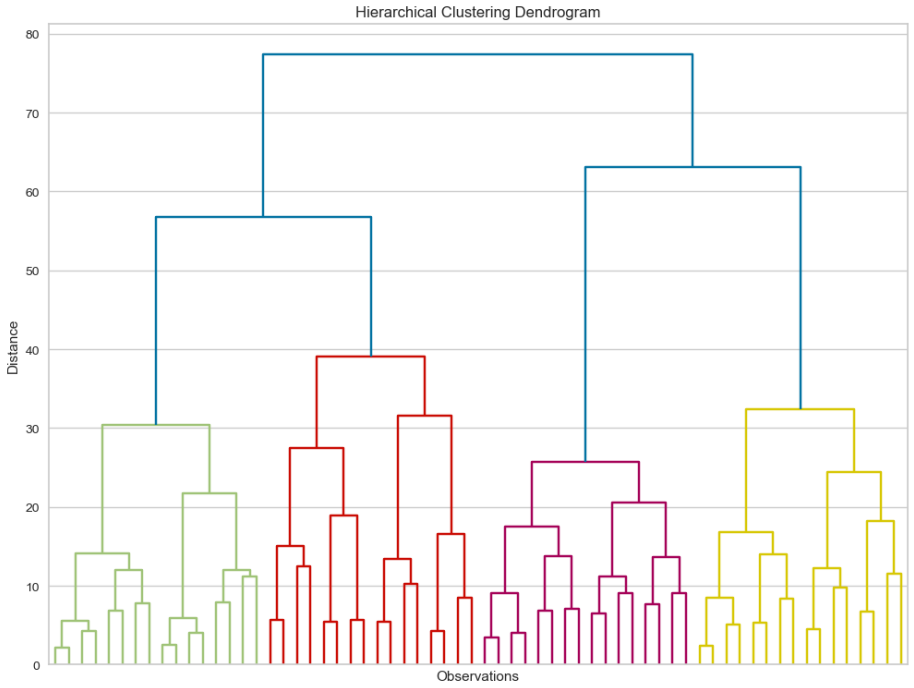
Measure of distances between observations?

1. Euclidean distance
2. Manhattan distance
3. Maximum distance
4. Ward Method

a. HIERARCHICAL CLUSTERING

Performed Hierarchical Clustering. The results are returned as a linkage matrix. We plot the results from the Hierarchical Clustering using a Dendrogram.

Dendrogram is tree like hierarchical representation of points. We truncate the dendrogram for better readability. And set level p to 5 for showing only the last 5 merged clusters. We also omit showing the labels for each point. To find the clusters we find a horizontal line on the dendrogram on which to cut. We did this by finding the longest vertical line un-intercepted by a horizontal line from the dendrogram.



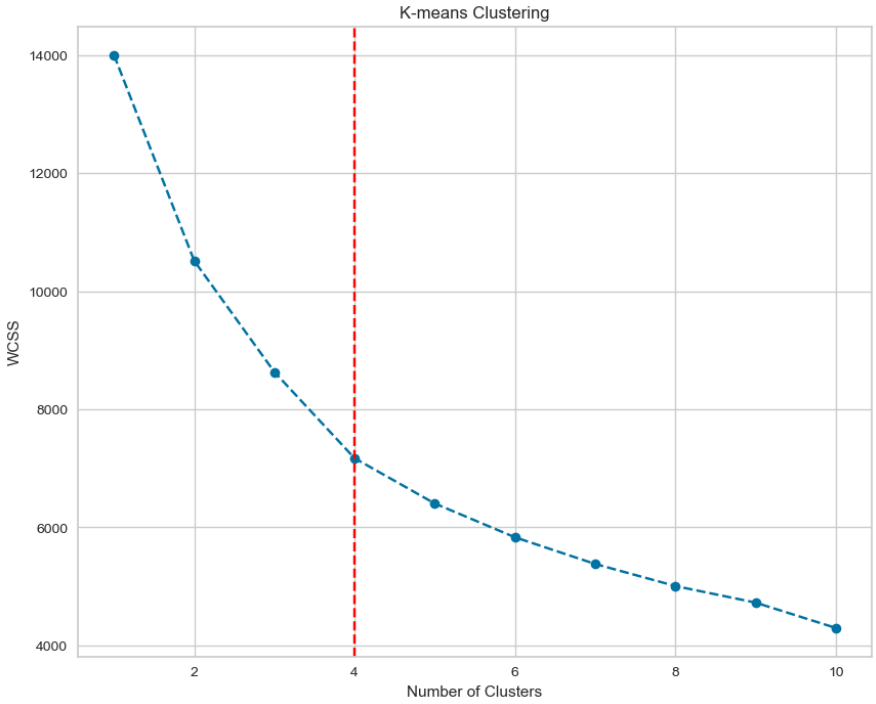
b. K-MEANS CLUSTERING

Performed K-means clustering. We consider 1 to 10 clusters, so for loop runs 10 iterations. In addition, we run the algorithm at many different starting points using k-means ++. K-means ++ is an initialization algorithm that finds the best starting points for centroids.

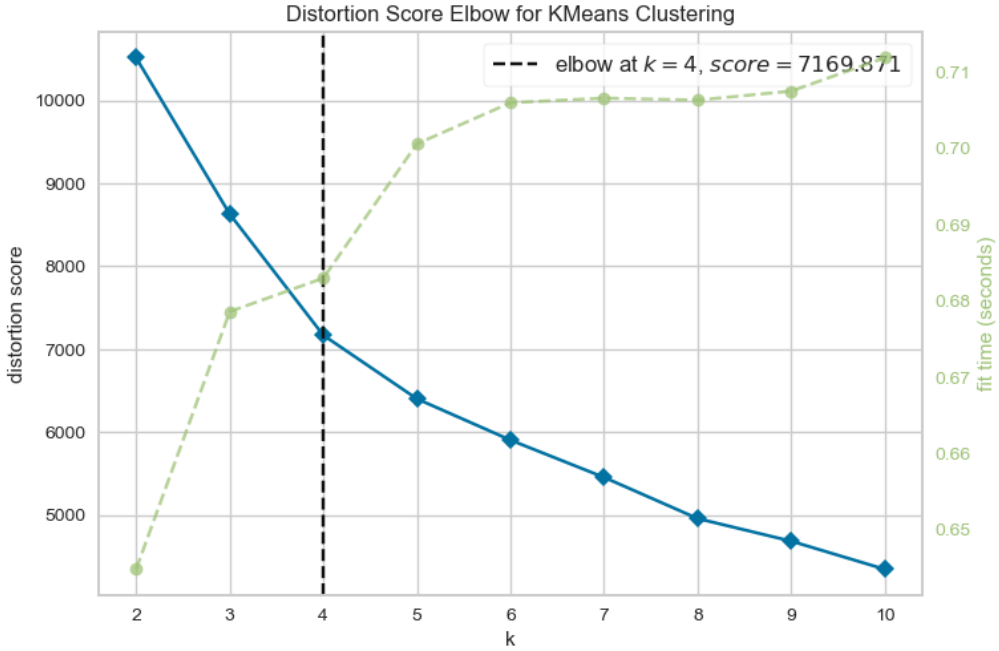
Then we plot the Within Cluster Sum of Squares (WCSS) for the different number of clusters. From this plot we choose the number of clusters. Depending on the shape of this graph we determine the number of clusters, using Elbow method.

$$\mathrm{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \text{cluster } j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2,$$

where $\bar{\mathbf{x}}_j$ is the sample mean in cluster j



ELBOW PLOT



We conducted a K-means analysis and choose cluster count of 4, partitioning our data into these four clusters.

RESULT

We create a new data frame with the original features and add a new column with the assigned clusters for each point. And calculated the mean values for the clusters

Segment K-means	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	N Obs	Prop Obs
0	0.501901	0.692015	55.703422	2.129278	158338.422053	1.129278	1.110266	263	0.1315
1	0.352814	0.019481	35.577922	0.746753	97859.852814	0.329004	0.043290	462	0.2310
2	0.853901	0.997163	28.963121	1.068085	105759.119149	0.634043	0.422695	705	0.3525
3	0.029825	0.173684	35.635088	0.733333	141218.249123	1.271930	1.522807	570	0.2850

For segment 0, sex ratio 50% for males and 50% for females. 69% are married. Their avg age is 56. Education level is highest compared to other segments. Annual income is also highest compared to other segments. We labelled this segment as Prosperous.

For segment 1, 35% are females and almost all are single. Their avg age is 36. Their education level is low on average as compared to other segments. Annual income is lowest compared to other segments. And they live in small cities. We labelled this segment them Marginalized.

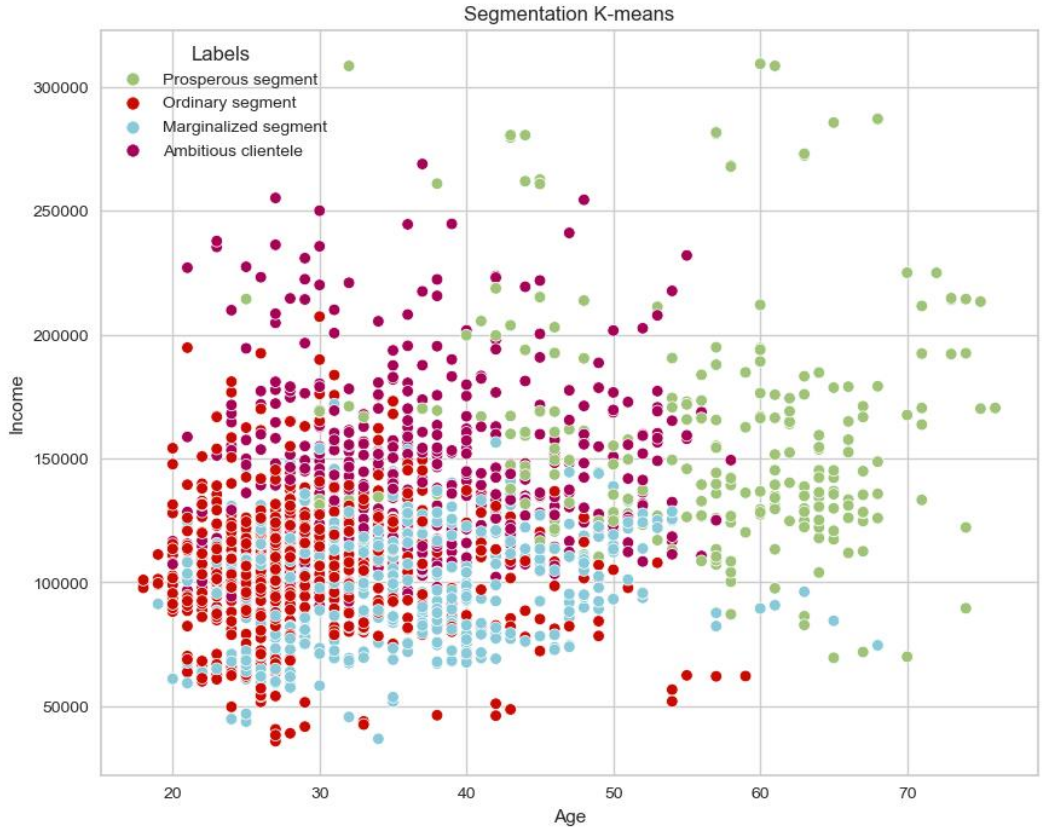
For segment 2, 85% are females and all are in relationships. Their avg age in 29. They have medium level of education, average income and middle management jobs. They seem equally distributed between small, mid-sized and big cities. They are avg in most parameters so we labelled this segment as Ordinary.

For segment 3, it comprises almost entirely of men, less than 20% of whom are in relationships. Avg age is 36. They relatively have low Education paired with high level of annual income. Most of this segment lives in big or middle-sized cities. We labelled this segment as Ambitious clientele.

Now we compute the size and proportions of the four clusters compared to entire dataset. And also, how much whole population each cluster represents.

Segment K-means	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	N Obs	Prop Obs
Prosperous segment	0.501901	0.692015	55.703422	2.129278	158338.422053	1.129278	1.110266	263	0.1315
Marginalized segment	0.352814	0.019481	35.577922	0.746753	97859.852814	0.329004	0.043290	462	0.2310
Ordinary segment	0.853901	0.997163	28.963121	1.068085	105759.119149	0.634043	0.422695	705	0.3525
Ambitious clientele	0.029825	0.173684	35.635088	0.733333	141218.249123	1.271930	1.522807	570	0.2850

We plot the results from the K-means algorithm. Each point in our dataset is plotted with the color of the clusters it has been assigned to.



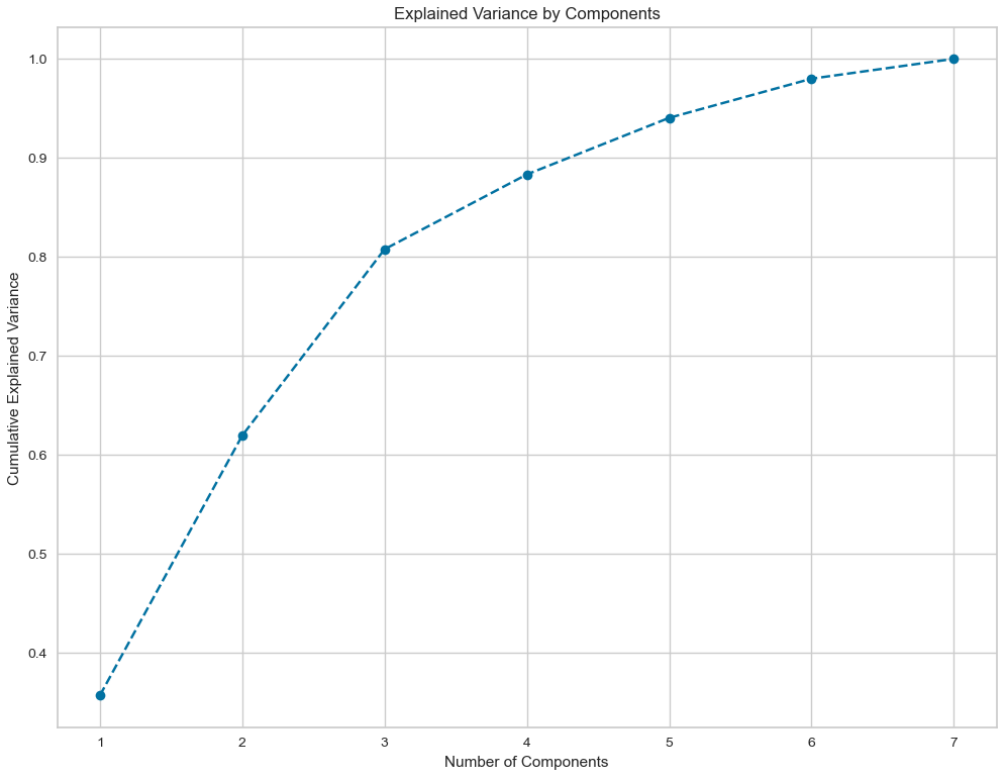
PRINCIPAL COMPONENT ANALYSIS

PCA is utilized to identify a subset of components, specifically seven in our scenario, that capture the variance within each of the seven individual components. These components are organized in descending order of significance, reflecting the importance of each in explaining the dataset's variance. PCA applies a linear transformation on the dataset which created seven new variables. Together these seven components explain 100% of the variable of the data.

```
array([0.35696328, 0.26250923, 0.18821114, 0.0755775 , 0.05716512,
       0.03954794, 0.02002579])
```

- First component explains, 36% of variance.
- Second component explains, 26% of variance.
- Third component explains, 19% of variance.
- Fourth component explains, 7% of variance.
- Fifth component explains, 6% of variance.
- Sixth component explains, 4% of variance.
- Seventh component explains, 2% of variance.

Now we plot the cumulative variance explained by total number of components. On this graph we choose the subset of components we want to keep. We will keep around 80 % of the explained variance.



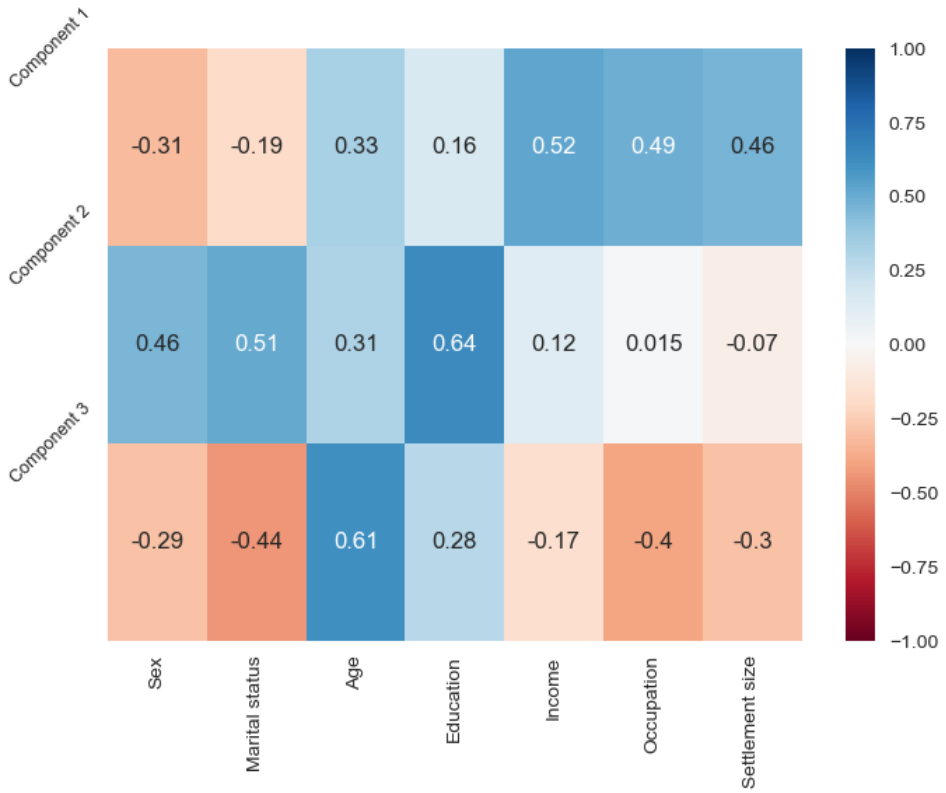
We choose three components and fit the model on our data with the selected components.

PCA RESULTS

The components attribute shows the loadings of each component on each of the seven original features. The loadings are the correlations between the components and the original features. All values are in the range -1 to 1, as they are essentially correlations.

	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
Component 1	-0.314695	-0.191704	0.326100	0.156841	0.524525	0.492059	0.464789
Component 2	0.458006	0.512635	0.312208	0.639807	0.124683	0.014658	-0.069632
Component 3	-0.293013	-0.441977	0.609544	0.275605	-0.165662	-0.395505	-0.295685

Plotting heat map for Principal Components against original features. We use the RdBu color scheme and set borders to -1 and 1.



Here is what our new components represents:

For Component 1 we see positive correlation between Age, Occupation, Settlement size. These are related to career of a person, so this component shows the Career focus of the individual.

For Component 2 we see positive correlation between Sex, Marital status and Education. These are related to education and lifestyle of the individual.

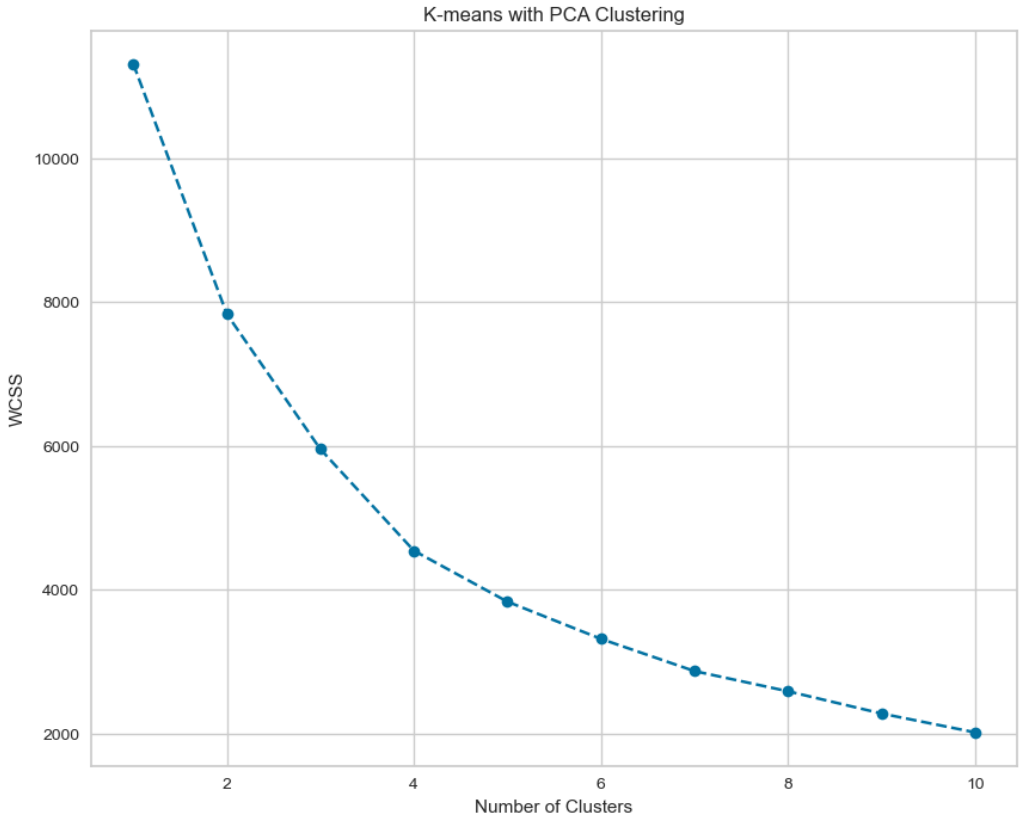
For Component 3 we see positive correlation for Age and negative correlation for Marital status and Occupation. This indicates the experience a person has, no matter if work experience or life experience.

Our original data refers to the original seven features i.e., 7 dimensional. We transform it into three dimensional using PCA transform method. The values we obtained are PCA scores.

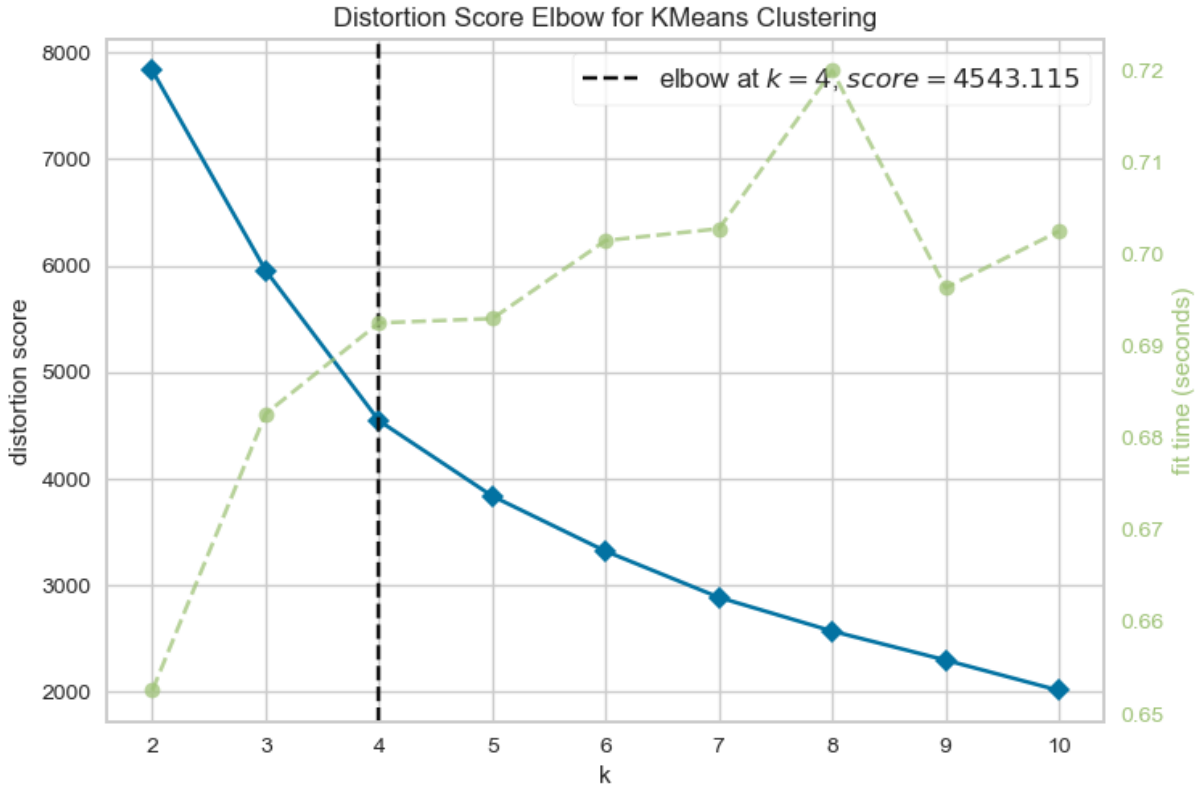
```
array([[ 2.51474593,  0.83412239,  2.1748059 ],
       [ 0.34493528,  0.59814564, -2.21160279],
       [-0.65106267, -0.68009318,  2.2804186 ],
       ...,
       [-1.45229829, -2.23593665,  0.89657125],
       [-2.24145254,  0.62710847, -0.53045631],
       [-1.86688505, -2.45467234,  0.66262172]])
```

K MEANS CLUSTERING WITH PCA

We fit K-means using the transformed data from the PCA. Plot the Within Cluster Sum of Squares for the K-means PCA model.



ELBOW PLOT



We have chosen four clusters, so we run K-means with number of clusters equals four.

K MEANS CLUSTERING WITH PCA RESULTS

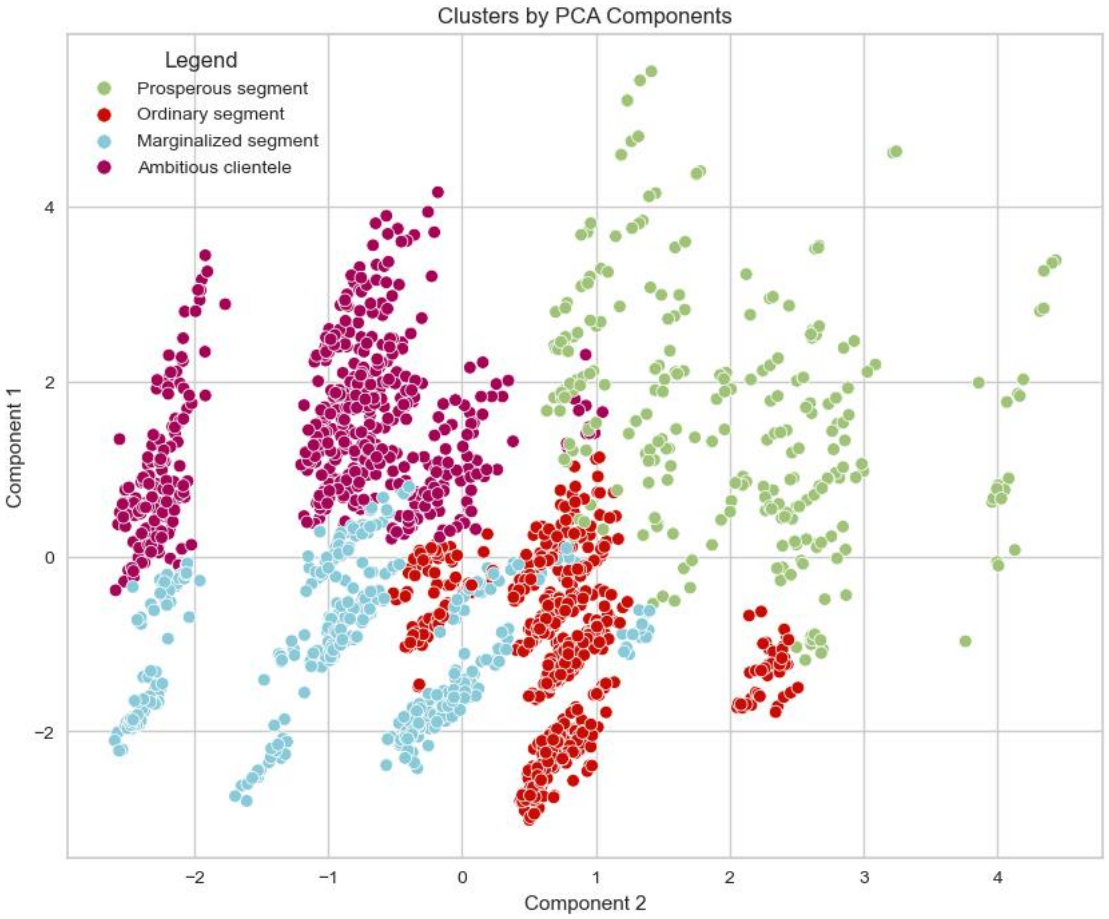
We create a new data frame with the original features and add the PCA scores and assigned clusters. The last column we add contains the PCA K-means clustering labels. We also calculate the means by segments.

- Component 1 is Career
- Component 2 is Education Lifestyle
- Component 3 is Experience

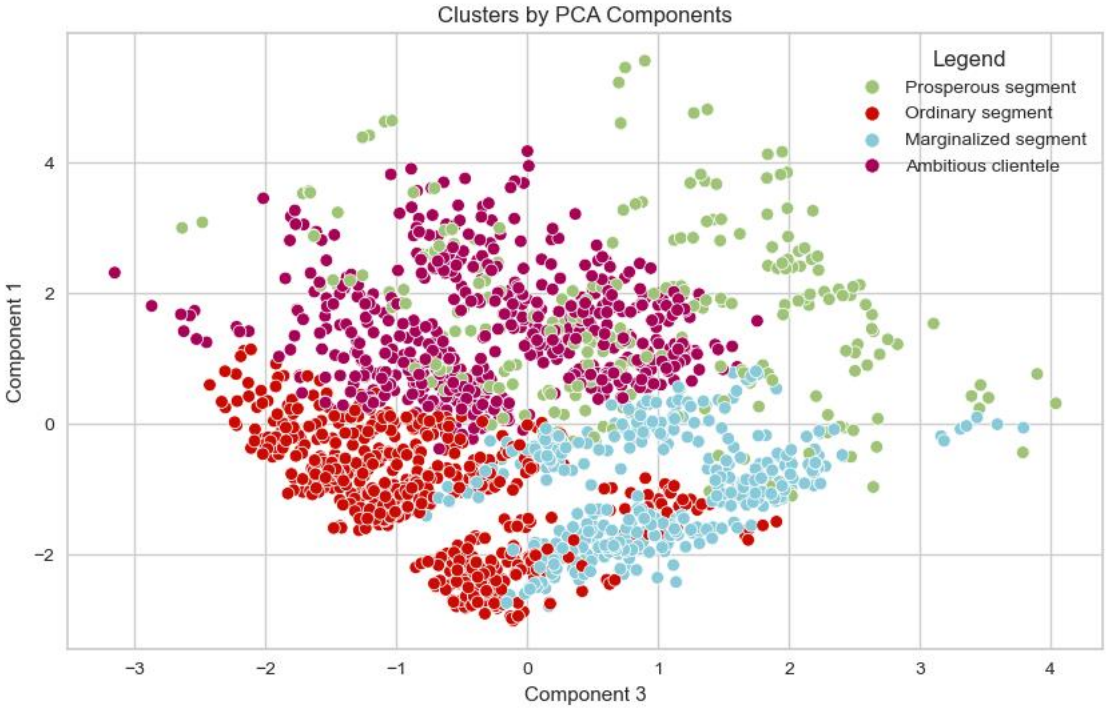
Segment K-means PCA	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Component 1	Component 2	Component 3
0	0.900289	0.965318	28.878613	1.060694	107551.500000	0.677746	0.440751	-1.107019	0.703776	-0.781410
1	0.027444	0.168096	35.737564	0.734134	141525.826758	1.267581	1.480274	1.372663	-1.046172	-0.248046
2	0.306522	0.095652	35.313043	0.760870	93692.567391	0.252174	0.039130	-1.046406	-0.902963	1.003644
3	0.505660	0.690566	55.679245	2.128302	158019.101887	1.120755	1.101887	1.687328	2.031200	0.844039

Here, Segment 0 is Ordinary segment, Segment 1 is Ambitious clientele, Segment 2 is Marginalized segment and Segment 3 is Prosperous segment.

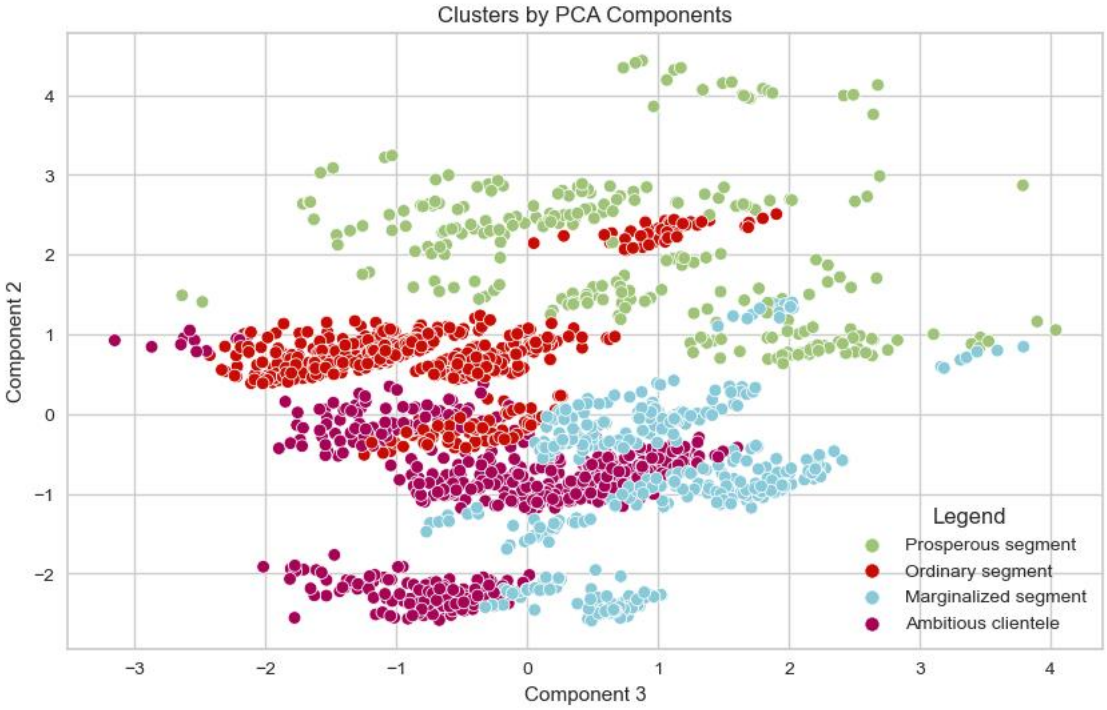
Now we plot data by PCA components. The Y axis is the first component, X axis is the second component.



We plot data by PCA components. The Y axis is the first component, X axis is the third component.



We plot data by PCA components. The Y axis is the second component, X axis is the third component.



DATA EXPORT

Subsequently, essential objects for Purchase Analytics are preserved. They are exported as pickle files, encompassing the Scaler, PCA, and Kmeans PCA entities. These objects are vital for preprocessing and segmenting the purchase dataset.

Exploratory Data Analysis

DATA PREVIEW

	ID	Day	Incidence	Brand	Quantity	Last_Inc_Brand	Last_Inc_Quantity	Price_1	Price_2	Price_3	Price_4	Price_5	Promotion_1	Promotion_2	Prc
0	200000001	1	0	0	0	0	0	1.59	1.87	2.01	2.09	2.66	0	1	
1	200000001	11	0	0	0	0	0	1.51	1.89	1.99	2.09	2.66	0	0	
2	200000001	12	0	0	0	0	0	1.51	1.89	1.99	2.09	2.66	0	0	
3	200000001	16	0	0	0	0	0	1.52	1.89	1.98	2.09	2.66	0	0	
4	200000001	18	0	0	0	0	0	1.52	1.89	1.99	2.09	2.66	0	0	

NUMBER OF ROWS AND COLUMNS

```
Rows: 59113
Columns: 24
```

DATA INFORMATION

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59113 entries, 0 to 59112
Data columns (total 24 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID                 59113 non-null  int64
 1   Day                59113 non-null  int64
 2   Incidence          59113 non-null  int64
 3   Brand              59113 non-null  int64
 4   Quantity           59113 non-null  int64
 5   Last_Inc_Brand     59113 non-null  int64
 6   Last_Inc_Quantity  59113 non-null  int64
 7   Price_1            58955 non-null  object
 8   Price_2            58936 non-null  float64
 9   Price_3            58913 non-null  float64
 10  Price_4            58928 non-null  float64
 11  Price_5            58974 non-null  float64
 12  Promotion_1        59113 non-null  int64
 13  Promotion_2        59113 non-null  int64
 14  Promotion_3        59113 non-null  int64
 15  Promotion_4        59113 non-null  int64
 16  Promotion_5        59113 non-null  int64
 17  Sex                59113 non-null  int64
 18  Marital status     59113 non-null  int64
 19  Age                58923 non-null  float64
 20  Education          59113 non-null  int64
 21  Income             58804 non-null  float64
 22  Occupation         59113 non-null  int64
 23  Settlement size    59113 non-null  int64
dtypes: float64(6), int64(17), object(1)
memory usage: 10.8+ MB
```

COLUMNS IN THE DATASET

```
Index(['ID', 'Day', 'Incidence', 'Brand', 'Quantity', 'Last_Inc_Brand',
       'Last_Inc_Quantity', 'Price_1', 'Price_2', 'Price_3', 'Price_4',
       'Price_5', 'Promotion_1', 'Promotion_2', 'Promotion_3', 'Promotion_4',
       'Promotion_5', 'Sex', 'Marital status', 'Age', 'Education', 'Income',
       'Occupation', 'Settlement size'],
      dtype='object')
```

DATA DESCRIPTION

	count	mean	std	min	25%	50%	75%	max
ID	59113.0	2.000003e+08	144.076884	2.000000e+08	2.000001e+08	2.000003e+08	2.000004e+08	2.000005e+08
Day	59113.0	3.494792e+02	212.039054	1.000000e+00	1.610000e+02	3.430000e+02	5.300000e+02	7.300000e+02
Incidence	59113.0	2.492514e-01	0.432583	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Brand	59113.0	8.443659e-01	1.633605	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00
Quantity	59113.0	6.910832e-01	1.497595	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.500000e+01
Last_Inc_Brand	59113.0	8.404919e-01	1.631744	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00
Last_Inc_Quantity	59113.0	2.478643e-01	0.431776	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Price_2	58936.0	1.780978e+00	0.170862	1.260000e+00	1.580000e+00	1.880000e+00	1.890000e+00	1.900000e+00
Price_3	58913.0	2.006792e+00	0.046864	1.870000e+00	1.970000e+00	2.010000e+00	2.060000e+00	2.140000e+00
Price_4	58928.0	2.159913e+00	0.089842	1.760000e+00	2.120000e+00	2.170000e+00	2.240000e+00	2.260000e+00
Price_5	58974.0	2.654814e+00	0.098283	2.110000e+00	2.630000e+00	2.670000e+00	2.700000e+00	2.800000e+00
Promotion_1	59113.0	3.437315e-01	0.474957	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Promotion_2	59113.0	3.156328e-01	0.464771	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Promotion_3	59113.0	4.288397e-02	0.202597	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Promotion_4	59113.0	1.180620e-01	0.322684	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Promotion_5	59113.0	3.581277e-02	0.185825	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Sex	59113.0	3.854990e-01	0.486717	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Marital status	59113.0	3.931115e-01	0.488445	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Age	58923.0	3.879197e+01	12.044119	1.800000e+01	3.000000e+01	3.600000e+01	4.600000e+01	7.500000e+01
Education	59113.0	1.101281e+00	0.652362	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	3.000000e+00
Income	58804.0	1.218840e+05	40667.903573	3.824700e+04	9.554100e+04	1.179710e+05	1.385250e+05	3.093640e+05
Occupation	59113.0	7.747196e-01	0.663397	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00
Settlement size	59113.0	6.570636e-01	0.794678	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	2.000000e+00

INFERENCES:

1. Day: The average purchase day is around 349 out of 730 days in the dataset, with a standard deviation of approximately 212.
2. Incidence: The mean incidence is about 25%, indicating that on average, 25% of customers make a purchase.
3. Brand: We have total 5 Brands.
4. Quantity: The average quantity purchased is 0.69, with a standard deviation of 1.50.
5. Last_Inc_Brand: Similar to "Brand," but related to the last purchase.
6. Last_Inc_Quantity: Similar to "Quantity," but related to the last purchase.
7. Price_1, Price_2, Price_3, Price_4, Price_5: Average prices for different products.
8. Promotion_1, Promotion_2, Promotion_3, Promotion_4, Promotion_5: Incidence of different promotions.
9. Sex, Marital Status: It is a number category column {0,1}
10. Age: The average age of customers is approximately 38.79 years, with a standard deviation of about 12.05.
11. Education: It is a number category column which divided int 3 part
12. Income: The average annual income is approximately 121,841.25, with a standard deviation of about 40,643.12.
13. Occupation, Settlement Size: It is a number category column {0,1,2}

NULL VALUES

```
ID                     0
Day                    0
Incidence              0
Brand                  0
Quantity               0
Last_Inc_Brand         0
Last_Inc_Quantity      0
Price_1              158
Price_2              177
Price_3              200
Price_4              185
Price_5              139
Promotion_1            0
Promotion_2            0
Promotion_3            0
Promotion_4            0
Promotion_5            0
Sex                    0
Marital status         0
Age                  190
Education              0
Income               309
Occupation             0
Settlement size        0
dtype: int64
```

INFERENCES:

1. Price_1 column has 158 missing values.
2. Price_2 column has 177 missing values.
3. Price_3 column has 200 missing values.
4. Price_4 column has 185 missing values.
5. Price_5 column has 139 missing values.
6. The Age column has 190 missing value.
7. The Income column has the highest number of missing values, with 309 entries.

NULL VALUE TREATMENT

```
def fill_missing_values_by_id(df):
    # Fill missing Age with mode for each group (ID)
    df['Age'] = df.groupby('ID')['Age'].transform(lambda x: x.fillna(x.mode().iloc[0] \
                                                                     if not x.mode().empty else None))

    return df

# Example usage
data = fill_missing_values_by_id(df_purchase)
```

INFERENCE:

We treated 190 missing values for Age colum. In this code we efficiently handles missing values in the 'Age' column by replacing them with the mode within each 'ID' group, ensuring that missing values are imputed based on relevant subgroups within the dataset.

```
def fill_missing_income_by_id(df):
    # Group by ID and fill missing Income with mode for each group
    df['Income'] = df.groupby('ID')['Income'].transform(lambda x: x.fillna(x.mode().iloc[0] \
                                                                    if not x.mode().empty else None))

    return df

# Example usage
data = fill_missing_income_by_id(df_purchase)
```

INFERENCES:

We treated 309 missing values for Income column.In this code we efficiently handles missing values in the 'Income' column by imputing them with the mode of each respective 'ID' group, thereby enhancing the completeness and utility of the dataset for subsequent analyses.

DROPING ALL NAN VALUES WITHIN THE DATAFRAME

```
ID                  0
Day                 0
Incidence           0
Brand               0
Quantity            0
Last_Inc_Brand      0
Last_Inc_Quantity   0
Price_1             0
Price_2             0
Price_3             0
Price_4             0
Price_5             0
Promotion_1         0
Promotion_2         0
Promotion_3         0
Promotion_4         0
Promotion_5         0
Sex                 0
Marital status      0
Age                 0
Education           0
Income              0
Occupation          0
Settlement size     0
dtype: int64
```

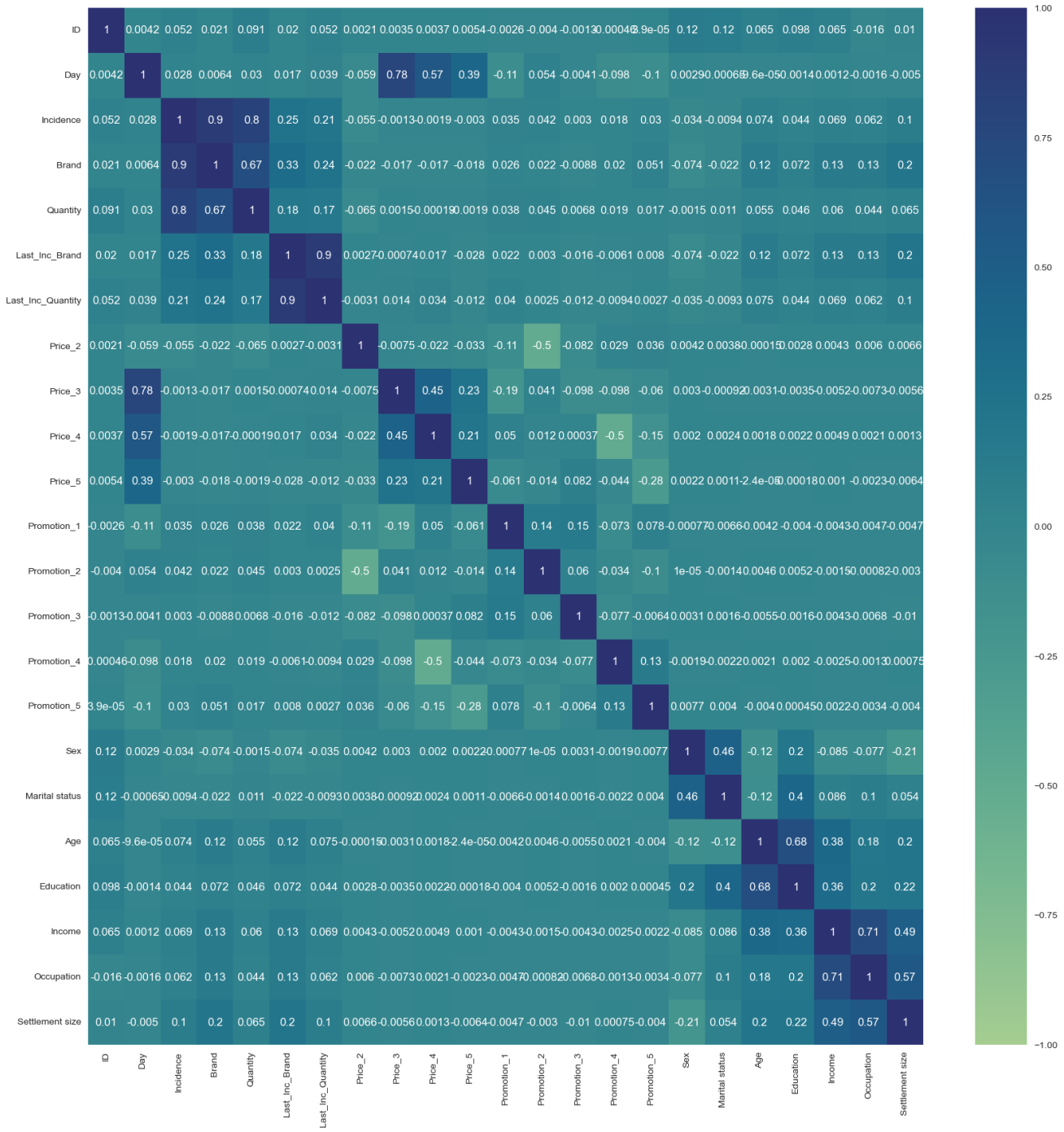
INFERENCE:

We performed missing value treatment for all columns. Now our data frame is ready for Analysis.

DATA DESCRIPTION AFTER NULL VALUE TREATMENT

	count	mean	std	min	25%	50%	75%	max
ID	58694.0	2.000003e+08	144.315842	2.000000e+08	2.000001e+08	2.000003e+08	2.000004e+08	2.000005e+08
Day	58694.0	3.494306e+02	212.045789	1.000000e+00	1.610000e+02	3.430000e+02	5.300000e+02	7.300000e+02
Incidence	58694.0	2.493952e-01	0.432667	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Brand	58694.0	8.442941e-01	1.633073	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00
Quantity	58694.0	6.919617e-01	1.498724	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.500000e+01
Last_Inc_Brand	58694.0	8.408526e-01	1.631666	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00
Last_Inc_Quantity	58694.0	2.480833e-01	0.431904	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Price_2	58694.0	1.781002e+00	0.170865	1.260000e+00	1.580000e+00	1.880000e+00	1.890000e+00	1.900000e+00
Price_3	58694.0	2.006789e+00	0.046868	1.870000e+00	1.970000e+00	2.010000e+00	2.060000e+00	2.140000e+00
Price_4	58694.0	2.159950e+00	0.089821	1.760000e+00	2.120000e+00	2.170000e+00	2.240000e+00	2.260000e+00
Price_5	58694.0	2.654795e+00	0.098275	2.110000e+00	2.630000e+00	2.670000e+00	2.700000e+00	2.800000e+00
Promotion_1	58694.0	3.438341e-01	0.474991	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Promotion_2	58694.0	3.156370e-01	0.464773	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Promotion_3	58694.0	4.279824e-02	0.202404	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Promotion_4	58694.0	1.178655e-01	0.322452	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Promotion_5	58694.0	3.586397e-02	0.185953	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
Sex	58694.0	3.858827e-01	0.486807	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Marital status	58694.0	3.931066e-01	0.488444	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
Age	58694.0	3.879408e+01	12.052403	1.800000e+01	3.000000e+01	3.600000e+01	4.600000e+01	7.500000e+01
Education	58694.0	1.101578e+00	0.652493	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	3.000000e+00
Income	58694.0	1.218412e+05	40643.123123	3.824700e+04	9.554100e+04	1.179710e+05	1.385250e+05	3.093640e+05
Occupation	58694.0	7.742018e-01	0.663242	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00
Settlement size	58694.0	6.559274e-01	0.794175	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	2.000000e+00

CORRELATION BETWEEN FEATURES USING HEATMAP



INFERENCE:

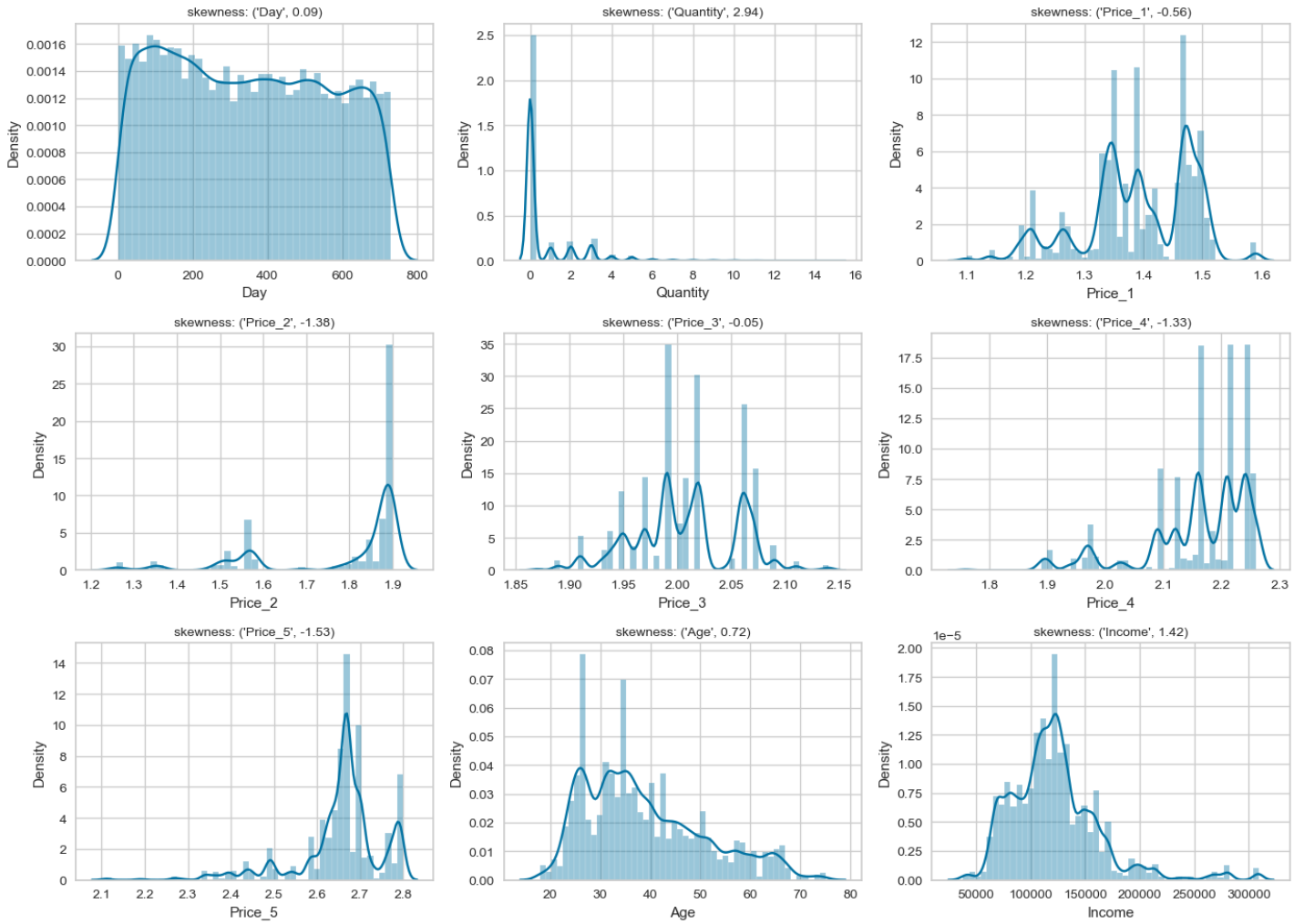
Perfect Positive Correlation (1)
Strong Positive Correlation (0.7 to 0.9)
No Correlation (0)
Strong Negative Correlation (-0.7 to -0.9)

Heatmap shows following correlations:

1. Day and Price_3 has strong positive correlation.
2. Incidence and Brand has strong positive correlation.
3. Incidence and Quality has strong positive correlation.
4. Brand and Quantity has strong positive correlation.
5. Last_Inc_Brand and Last_Inc_Quantity has strong positive correlation.
6. Age and Education has strong positive correlation.
7. Income and Occupation has strong positive correlation.

UNIVARIATE ANALYSIS

Plot for Numerical Columns

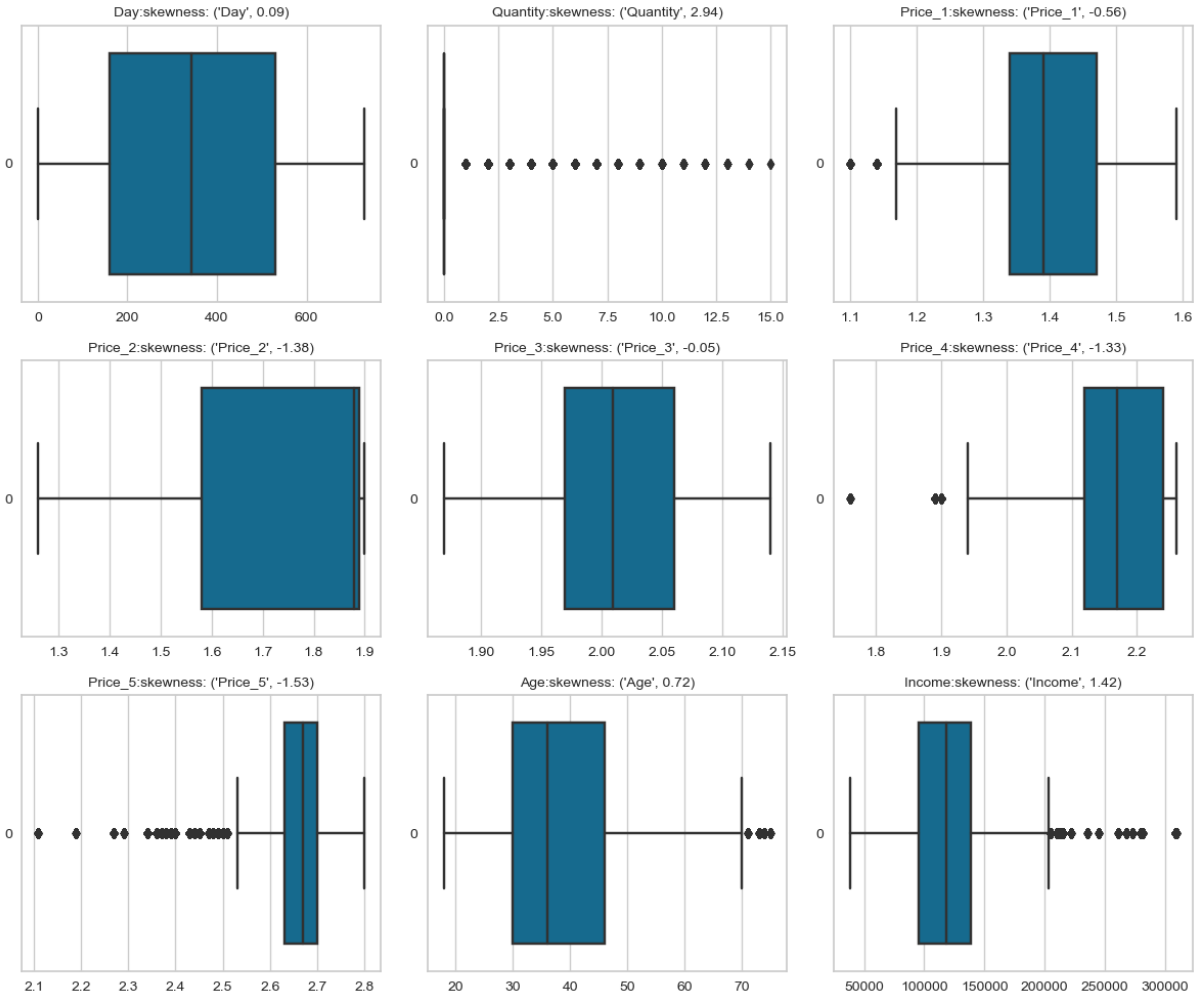


INFERENCE:

Perfectly Symmetric (Skewness = 0)
Moderately Skewed (0.5 to 1 or -0.5 to -1)
Highly Skewed (1 to 2 or -1 to -2)
Very Highly Skewed (Greater than 2 or less than -2)

1. Day column skewness value is 0.09 hence it is Perfectly Symmetric.
2. Quality column skewness value is 2.94 hence it is Very Highly Right Skewed.
3. Price_1 column skewness value is -0.56 hence it is Moderately Left Skewed.
4. Price_2 column skewness value is -1.38 hence it is Highly Skewed.
5. Price_3 column skewness value is -0.05 hence it is Perfectly Symmetric.
6. Price_4 column skewness value is -1.33 hence it is Highly Skewed.
7. Price_5 column skewness value is -1.53 hence it is Highly Skewed.
8. Age column skewness value is 0.72 hence it is Moderately Skewed.
9. Income column skewness value is 1.42 hence it is Highly Skewed.

BOXPLOT

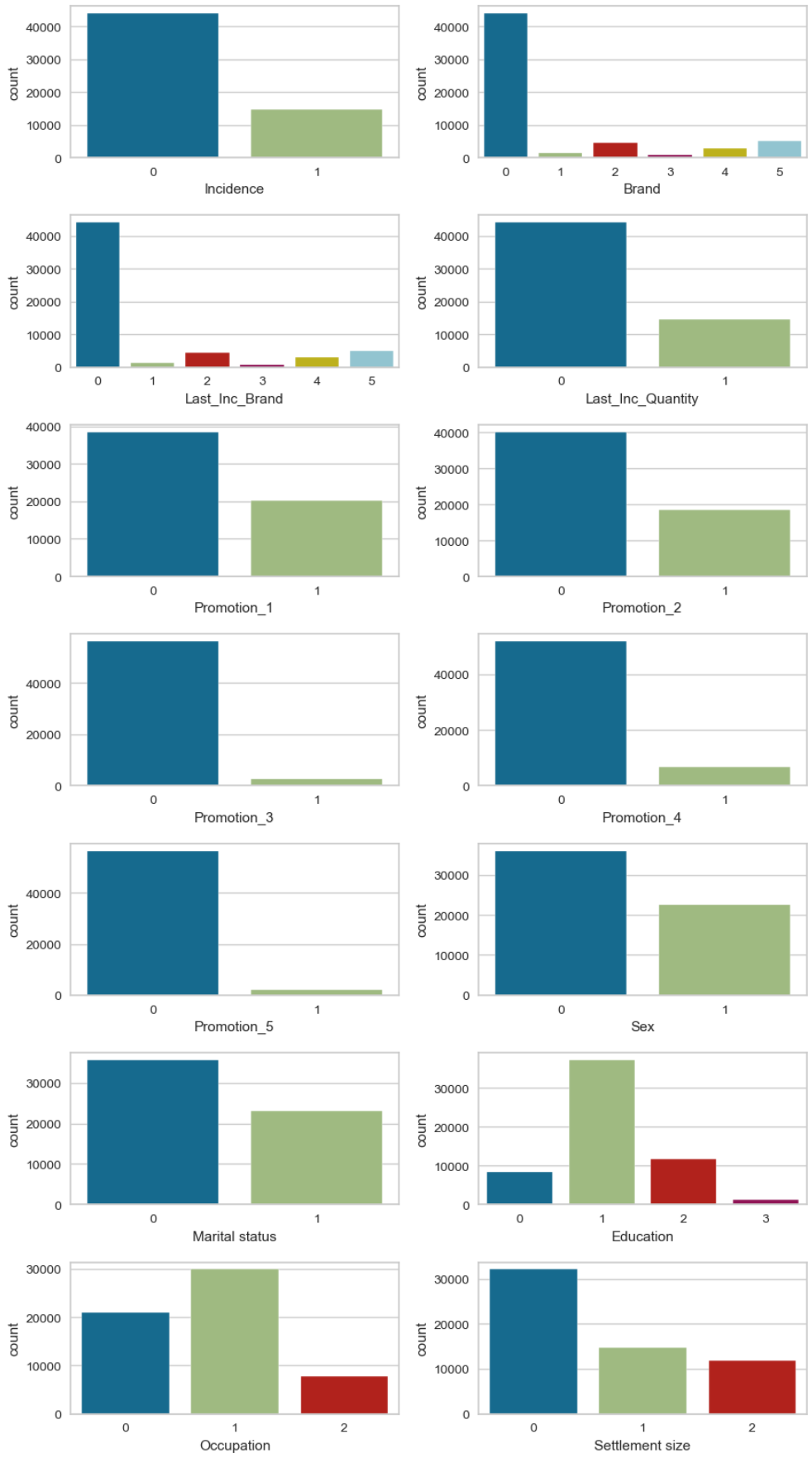


INFERENCE:

Perfectly Symmetric (Skewness = 0)
Moderately Skewed (0.5 to 1 or -0.5 to -1)
Highly Skewed (1 to 2 or -1 to -2)
Very Highly Skewed (Greater than 2 or less than -2)

1. Day column skewness value is 0.09 hence it is perfectly Symmetric.
2. Quality column skewness value is 2.94 hence it is Very Highly Right Skewed.
3. Price_1 column skewness value is -0.56 hence it is Moderately Left Skewed.
4. Price_2 column skewness value is -1.38 hence it is Highly Skewed.
5. Price_3 column skewness value is -0.05 hence it is Perfectly Symmetric.
6. Price_4 column skewness value is -1.33 hence it is Highly Skewed.
7. Price_5 column skewness value is -1.53 hence it is Highly Skewed.
8. Age column skewness value is 0.72 hence it is Moderately Skewed.
9. Income column skewness value is 1.42 hence it is Highly Skewed.

PLOT FOR CATEGORICAL COLUMNS



INFERENCE:

1. Incidence column has two categories 0 and 1. And count of category 0 is highest.
2. Brand column has six categories 0, 1, 2, 3, 4 and 5. And count of category 0 is highest.
3. Last_Inc_Brand column has six categories 0, 1, 2, 3, 4 and 5. And count of category 0 is highest.
4. Last_Inc_Quantity column has two categories 0 and 1. And count of category 0 is highest.
5. Promotion_1 column has two categories 0 and 1. And count of category 0 is highest.
6. Promotion_2 column has two categories 0 and 1. And count of category 0 is highest.
7. Promotion_3 column has two categories 0 and 1. And count of category 0 is highest.
8. Promotion_4 column has two categories 0 and 1. And count of category 0 is highest.
9. Promotion_5 column has two categories 0 and 1. And count of category 0 is highest.
10. Sex column has two categories 0 and 1. And count of category 0 is highest.
11. Marital status column has two categories 0 and 1. And count of category 0 is highest.
12. Education column has four categories 0, 1, 2 and 3. And count of category 1 is highest.
13. Occupation column has three categories 0, 1 and 2. And count of category 1 is highest.
14. Settlement size column has three categories 0, 1 and 2. And count of category 0 is highest.

Purchase Analysis

Now we explore both the descriptive and predictive analysis of the purchase behavior of customer, including models for purchase incidence, brand choice, and purchase quantity.

While analyzing the dataset we find that we don't have an equal number of records per customer or an equal number of records per day. So descriptive statistics would neither be useful nor appropriate.

CHECKING FOR MISSING VALUES

```
ID                  0
Day                 0
Incidence           0
Brand               0
Quantity            0
Last_Inc_Brand      0
Last_Inc_Quantity   0
Price_1             0
Price_2             0
Price_3             0
Price_4             0
Price_5             0
Promotion_1         0
Promotion_2         0
Promotion_3         0
Promotion_4         0
Promotion_5         0
Sex                 0
Marital status      0
Age                 0
Education           0
Income              0
Occupation          0
Settlement size     0
dtype: int64
```

There are no missing values.

DATA SEGMENTATION
We start by loading our pickled objects in order to segment the purchase data set.

STANDARDIZATION
We standardize the purchase data in the same way we did the segmentation data, using the standard scaler.

PCA
We apply PCA on the purchase data and obtain 3 principal components for each row in the table.

K-means PCA
Based on the principal components, we use the predict method from PCA to segment the purchase data into the four segments.

We created a duplicate of the current data frame and designate it as the "purchase predictors" data frame. This copy will serve as the original reference for any modifications to the predictors data frame. Then we introduced a new column in the predictors data frame to include segment information.

Descriptive Analysis by Segmentation

DATA ANALYSIS BY CUSTOMER

	ID	Day	Incidence	Brand	Quantity	Last_Inc_Brand	Last_Inc_Quantity	Price_1	Price_2	Price_3	Price_4	Price_5	Promotion_1	Promotion_2	Prc
0	200000001	1	0	0	0	0	0	1.59	1.87	2.01	2.09	2.66	0	1	
1	200000001	11	0	0	0	0	0	1.51	1.89	1.99	2.09	2.66	0	0	
2	200000001	12	0	0	0	0	0	1.51	1.89	1.99	2.09	2.66	0	0	
3	200000001	16	0	0	0	0	0	1.52	1.89	1.98	2.09	2.66	0	0	
4	200000001	18	0	0	0	0	0	1.52	1.89	1.99	2.09	2.66	0	0	

Now we are interested in having one record per individual to analyze the data on an individual level. So, we create a new data frame for creating a summary of each customer's purchasing behavior.

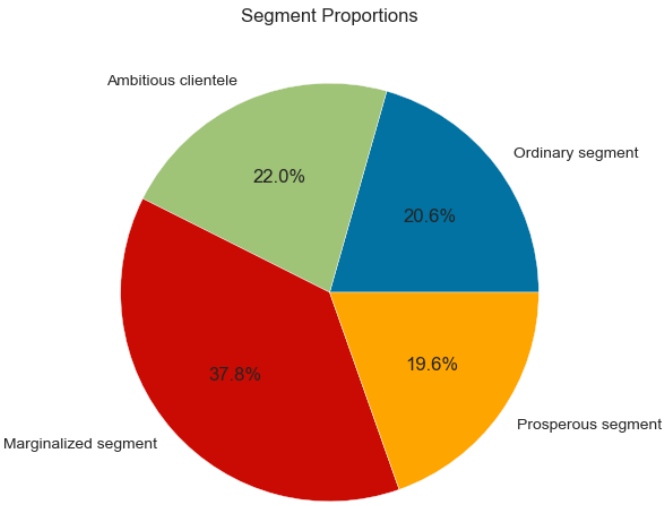
We add columns with Purchase occasions by customer ID, number of purchases per customer ID, average number of purchases by customer ID and the segment each customer.

Here are first 5 customers in the data set.

ID	N_Visits	N_Purchases	Average_N_Purchases	Segment
200000001	101	9	0.089109	2.0
200000002	87	11	0.126437	3.0
200000003	97	10	0.103093	2.0
200000004	85	11	0.129412	2.0
200000005	111	13	0.117117	1.0

SEGMENT PROPORTIONS

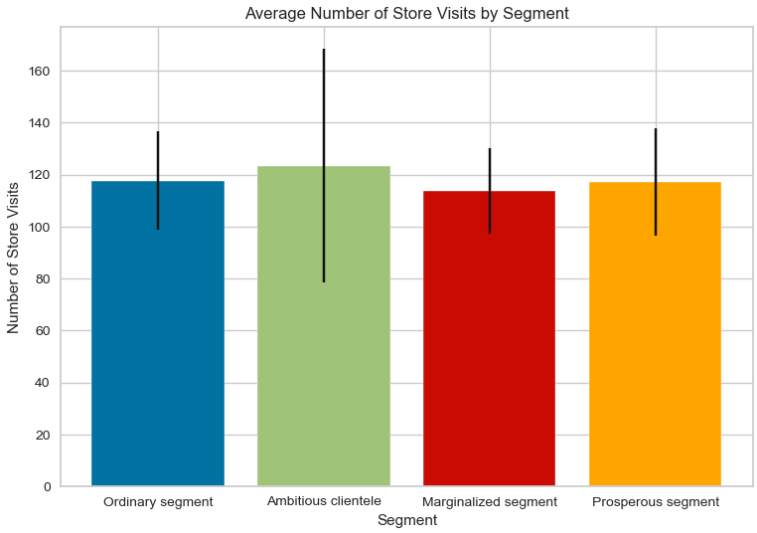
We calculate the proportions of each segment and set the appropriate column name. Then plot the segment proportions as a pie chart. Now we can easily see how the store visitors are distributed across segments, which is the largest segment and the relative sizes of each segment.



Marginalized segments the largest, with 37.8%. Ambitious clientele segment comes second, with 22%. Ordinary segment and Prosperous segment are almost equally distributed, with 20% each.

PURCHASE OCCASION AND PURCHASE INCIDENCE

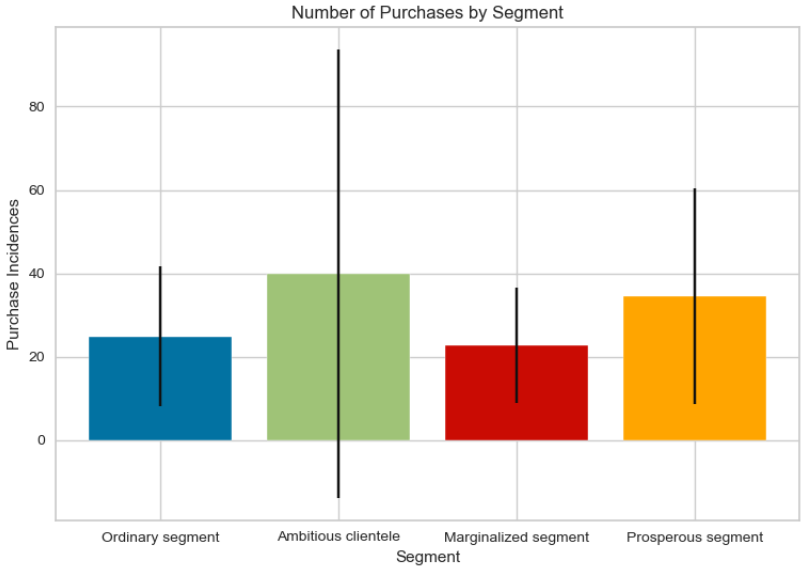
We calculated the mean by the four segments. It will help us determine the average customer behavior in each segment. Then we calculate the standard deviation by segments. It will help us determine how homogeneous each of the segments.
We plot the average number of store visits for each of the four segments using a bar chart. We display the standard deviation as a straight line. The bigger the length, the higher the standard deviation is.



The height of each bar represents the mean store visits. The vertical line indicates the dispersion of the data points or how big the standard deviation is.

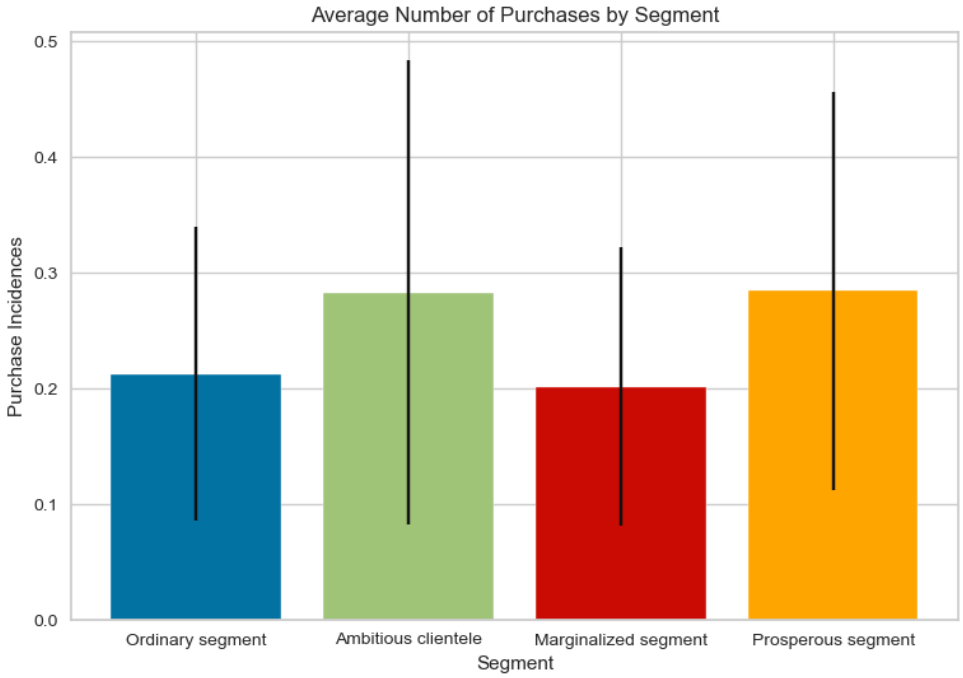
Bar chart indicates that Marginalized segment visits the store least often while Ambitious clientele visits it most. However, the standard deviation amongst customers from the second segment is quite high. This implies that the customers in this segment are at least homogeneous, i.e. least alike when it comes to how often they visit the grocery store.

Now we display the average number of purchases by segments. They will help us understand how often each group buys chocolate candy bars.



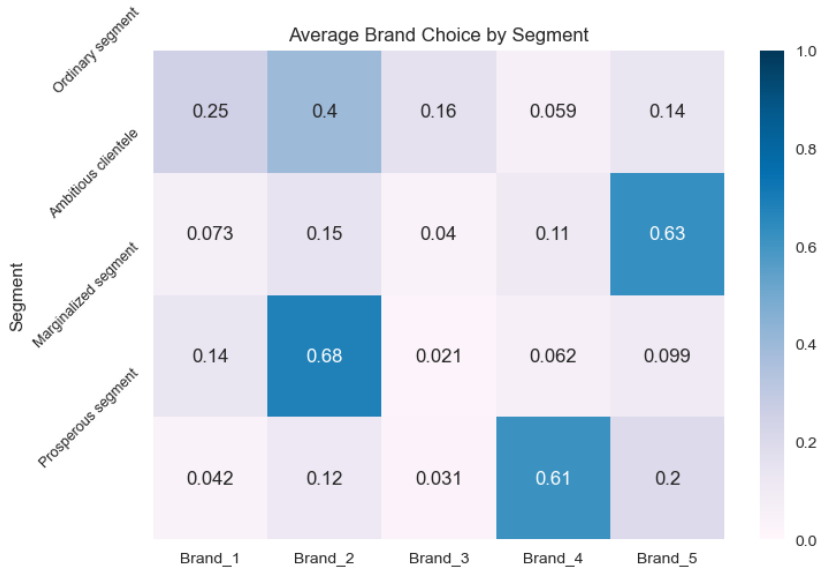
We observe that the Ambitious clientele segment buys product more often. However, once again we see that its standard deviation is the highest it might be that a part of the segment buys products very frequently and another part less so. Although consumers in this segment have a somewhat similar income, the way they might want to spend their money might differ.
The most homogeneous segment appears to be that of the Marginalized segment. This is signified by the segment having the lowest standard deviation or shortest vertical line.
The Ordinary segment seems consistent with about 25 average purchases and a standard deviation of 30.

Now we display the average number of purchases.



BRAND CHOICE

We select only rows where incidence is one i.e., we are only interested in the times a purchase was made. Here we make dummies for each of the five brands, so to find the average brand choice by segment.



The five brands are arranged in ascending order of price. Brand 1 is the cheapest brand, while brand 5 is the most expensive one.

- Ordinary segment is most heterogeneous, they have preference for Brand 2 but weak preference for Brand 1, Brand 3, Brand 4 and Brand 5.
- Ambitious clientele segment shows a strong preference for Brand 5, the most expensive brand.
- Marginalized segment shows an extremely strong preference for Brand 2 to almost 70%.
- Prosperous segment shows a strong preference for Brand 4.

REVENUE

We compute the revenue for brand 1. For each entry where Brand 1 was purchased, we multiply the price of the brand for that particular day by the quantity of the product purchased.

Similarly, we compute revenue for brand 2, brand 3, brand 4 and brand 5.

Now we compute the total revenue for each of the segments. We simply sum the revenue for each of the five brands. And we further modify our table to include the segment proportions. We also add the labels for the segments.

Segment	Revenue Brand 1	Revenue Brand 2	Revenue Brand 3	Revenue Brand 4	Revenue Brand 5	Total Revenue	Segment Proportions
Ordinary segment	2611.19	4768.52	3909.17	861.38	2439.75	14590.01	0.206
Ambitious clientele	736.09	1746.42	664.75	2363.84	19440.92	24952.02	0.220
Marginalized segment	2258.90	13955.14	716.25	1629.31	2230.50	20790.10	0.378
Prosperous segment	699.47	1298.23	731.35	14185.57	5509.69	22424.31	0.196

It is interesting to see the size of the segment compared to the revenue they bring.

Findings indicate that the Ordinary segment contributes the smallest total revenue of 14,590, representing 20% of the total. On the other hand, the Ambitious clientele segment boasts the highest total revenue of 24,952, constituting 22% of the overall revenue. The Marginalized segment contributes a total revenue of 20,790, making up 38% of the total. Lastly, the Prosperous segment generates a total revenue of 22,424, accounting for 19.6% of the overall revenue.

Purchase Probability Model

We use a statistical model that estimates purchase probability for each customer at each shopping trip. Then we will calculate price elasticity of purchase probability under different conditions.
When a customer visits the store, we call that a purchase occasion. The customer may or may not buy a product from the product category we're interested in. If there's an observation in our dataset, we know the customer visited the shop, then the incidence variable indicates whether a purchase was actually made, 1 stands for a purchase and 0 for no purchase.

This output could be used in two different ways:
First case as probability estimate, here the output is between 0 and 1. If we get an output of 0.85, we would consider that there is 85% chance of purchase. Alternatively, if we get 0.15, we'd have a 15% chance of purchase.

Second case as classifier, so if the output number is below 0.5, we usually classify it as 0, that would translate to no purchase. Alternatively, if we get number above 0.5, we classify it as 1, that would translate to purchase.

We use logistic regression to determine the probability of purchase. The main reason is that it is straight forward approach which can be easily interpreted and is widely understood. A logistic regression is a classification method which outputs a probability between 0 and 1.

In order to predict we need input and output variables. Purchase probability is influenced by Price. Our Y is Incidence, as we want to predict the purchase probability for our customers. Our dependent variable is based on the average price of chocolate candy bars. Therefore, we create a new data frame, containing the mean across the five brand prices.

We create a Logistic Regression model using sklearn. Then we fit the model with our X or price and our Y or incidence. The model quantifies the exact relationship between price and purchase probability.

```
model_purchase.coef_
```

```
array([[-2.34854135]])
```

The coefficients for price are negative, signaling that with an increase in price, the purchase probability decreases.

PRICE ELASTICITY OF PURCHASE PROBABILITY

The price elasticity of purchase probability is % change in purchase probability in response to 1% change in price.

Here we see the prices for the five different brands, which is an important factor in determining purchase probability. It informs the price range, for which we will be exploring purchase probability.

	Price_1	Price_2	Price_3	Price_4	Price_5
count	58693.000000	58693.000000	58693.000000	58693.000000	58693.000000
mean	1.392074	1.780999	2.006789	2.159945	2.654798
std	0.091139	0.170868	0.046867	0.089825	0.098272
min	1.100000	1.260000	1.870000	1.760000	2.110000
25%	1.340000	1.580000	1.970000	2.120000	2.630000
50%	1.390000	1.880000	2.010000	2.170000	2.670000
75%	1.470000	1.890000	2.060000	2.240000	2.700000
max	1.590000	1.900000	2.140000	2.260000	2.800000

The prices vary across the brands, with mean prices ranging from approximately 1.39 to 2.65 units.
The standard deviations are relatively small compared to the mean prices, indicating a relatively consistent pricing strategy within each brand. The minimum and maximum prices for each brand indicate the range of prices observed in the dataset.

We introduce the price range for which we'll examine the purchase probability. We choose a price range between 0.5 and 3.49, which somewhat expands the actual observed price range, which is from 1.1 to 2.8.

We then predict the purchase probability for our newly defined price range. The result is a 2 x 300 array. Create price elasticities master data frame. It will contain all the elasticities we calculate during the purchase analytics.

If the percentage change is greater than 100%, we say that the output or purchase probability is called elastic. On the other hand, for changes less than 100% it is inelastic. So, if the elasticity has a value smaller than one, in absolute terms, we say it is inelastic. If it is greater than one, we say it is elastic.
Furthermore, we can spot where the customer becomes inelastic from our data frame. We observe this happens at the 1.25 mark.

```
pd.options.display.max_rows = None
df_price_elasticities
```

70	1.20	-0.892452
71	1.21	-0.914393
72	1.22	-0.936696
73	1.23	-0.959362
74	1.24	-0.982392
75	1.25	-1.005785
76	1.26	-1.029541
77	1.27	-1.053660
78	1.28	-1.078142
79	1.29	-1.102985
80	1.30	-1.128189
81	1.31	-1.153752

Now we plot the price elasticity of purchase probability of the average customer. We observe that the price elasticities are all negative, i.e. the price elasticity decreases as price increases. The decrease in price is slow in the range between 0.5 and 1.1, and then it became steeper after the 1.1 mark. Thus, indicating the inverse proportionality between price and purchase probability.

Conclusion, with prices lower than 1.25, we can increase our product price without losing too much in terms of purchase probability. For prices higher than 1.25, we have more to gain by reducing our prices.

PURCHASE PROBABILITY BY SEGMENTS

Segment 0 – Ordinary segment

```
model_incidence_segment0.coef_
```

```
array([[-1.50834589]])
```

Segment 1 – Ambitious clientele

```
model_incidence_segment_1.coef_
```

```
array([[-1.71386567]])
```

Segment 2 – Marginalized segment

```
model_incidence_segment2.coef_
```

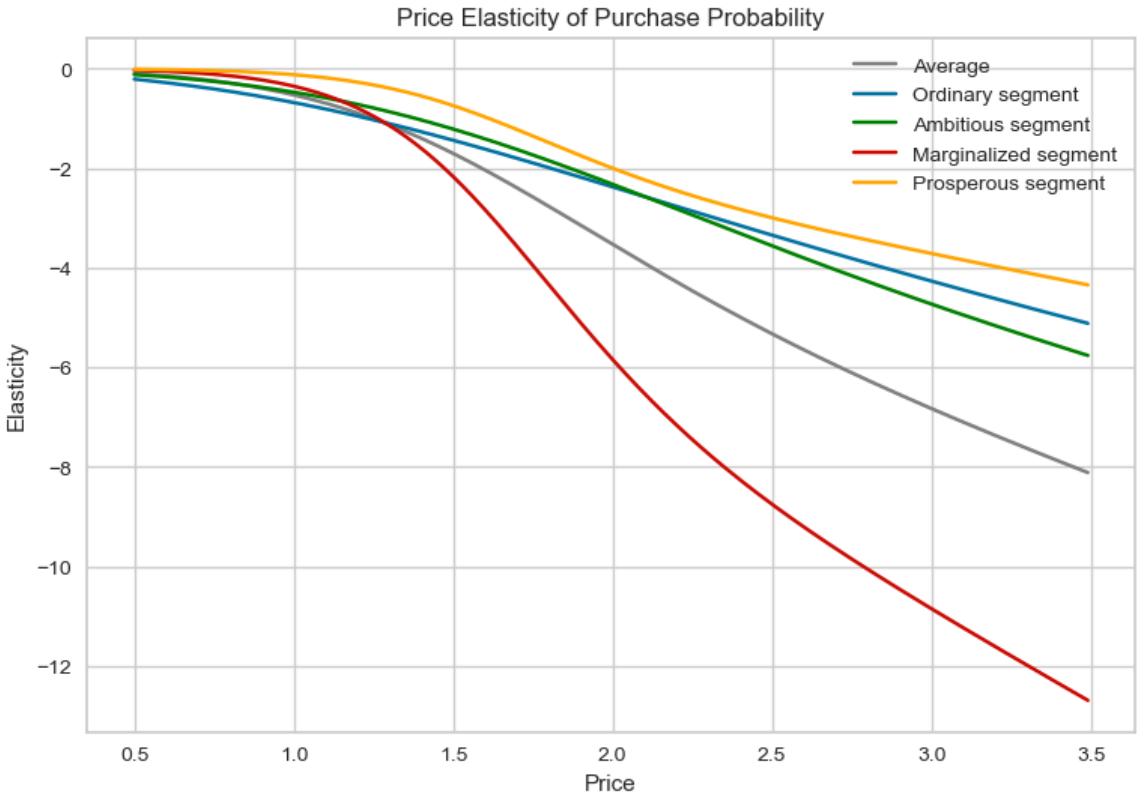
```
array([[-3.6405426]])
```

Segment 3 – Prosperous segment

```
model_incidence_segment3.coef_
```

```
array([[-1.24558371]])
```

Now we display all elasticities of purchase probability on the same plot.



Ordinary segment: We'd consider raising prices within 0.5 to 1.24 range, especially as customers in the ordinary segment are highly price-sensitive. So, until prices point 1.24, purchase probability is inelastic.

Marginalized segment: We'd increase prices if we were in the 0.5 to 1.39 range, so, until prices point 1.39, purchase probability is inelastic.

Ambitious segment: We'd increase prices if we were in the 0.5 to 1.27 range, and think about decreasing them afterwards. so, until prices point 1.27, purchase probability is inelastic.

Prosperous segment: We'd increase prices if we were in the 0.5 to 1.62 range and think about decreasing them afterwards. Customers in the prosperous segment are the least price-sensitive, so, until prices point 1.62, purchase probability is inelastic.

PURCHASE PROBABILITY WITH PROMOTION FEATURE

The product's price could be temporarily reduced or other types of promotions such as display or feature may come into play. So, product promotion may affect purchase probability. Therefore, for this model we incorporate a promotion feature to see its effect on elasticity. We calculate the mean promotion across all brands.

For model estimation we use Logistic Regression. The model quantifies the exact relationship between price, promotion, and probability of purchase. The resulting coefficients are -1.49 for price and 0.56 for promotion.

```
model_incidence_promotion.coef_
```

```
array([[-1.49402073,  0.56146774]])
```

The coefficient for promotion is positive. Therefore, there is a positive relationship between promotion and purchase probability, meaning that with increase in promotion, the purchase probability also increases.

PRICE ELASTICITY WITH PROMOTION

Now we calculate elasticity of purchase probability with promotion.

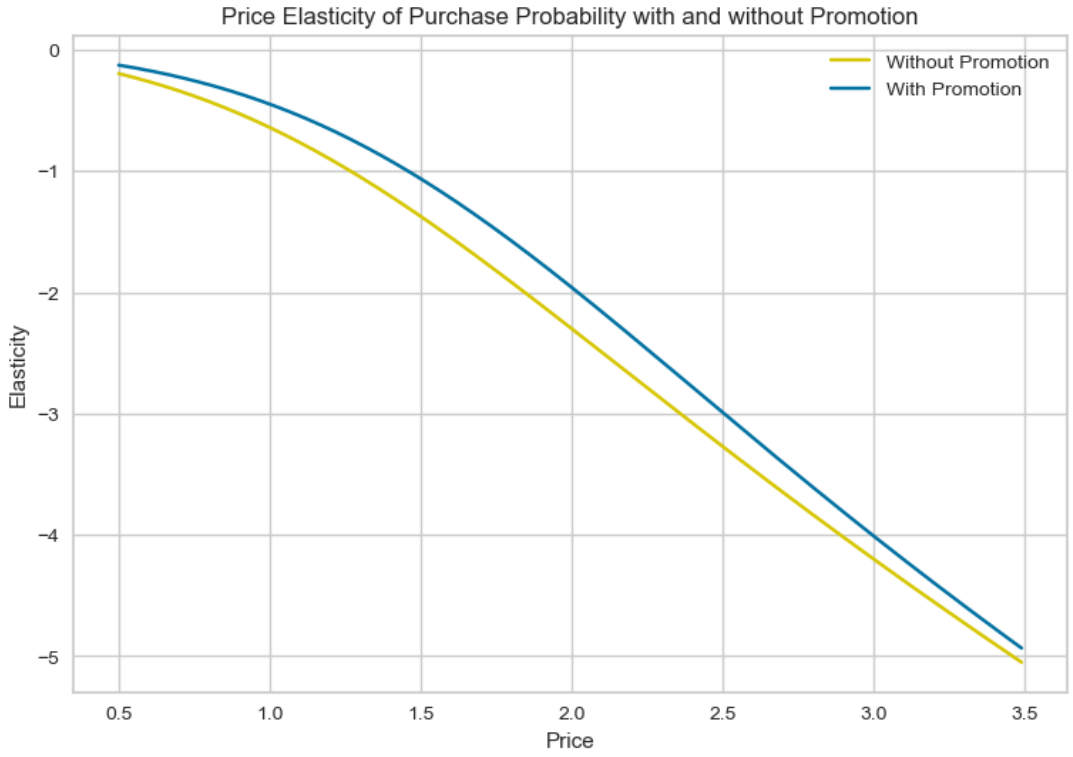
	Price_Point	Mean_PE	PE_Segment_0	PE_Segment_1	PE_Segment_2	PE_Segment_3	Elasticity_Promotion_1
0	0.50	-0.096453	-0.211181	-0.119076	-0.030920	-0.010587	-0.125745
1	0.51	-0.100525	-0.217751	-0.123258	-0.032687	-0.011192	-0.129861
2	0.52	-0.104724	-0.224429	-0.127534	-0.034540	-0.011827	-0.134057
3	0.53	-0.109053	-0.231216	-0.131905	-0.036485	-0.012493	-0.138332
4	0.54	-0.113515	-0.238112	-0.136372	-0.038524	-0.013191	-0.142688
5	0.55	-0.118115	-0.245118	-0.140936	-0.040662	-0.013923	-0.147127
6	0.56	-0.122854	-0.252235	-0.145600	-0.042904	-0.014691	-0.151648
7	0.57	-0.127737	-0.259463	-0.150364	-0.045254	-0.015495	-0.156253
8	0.58	-0.132767	-0.266804	-0.155231	-0.047716	-0.016339	-0.160943
9	0.59	-0.137947	-0.274258	-0.160201	-0.050295	-0.017222	-0.165719
10	0.60	-0.143281	-0.281825	-0.165277	-0.052998	-0.018147	-0.170582

PRICE ELASTICITY WITHOUT PROMOTION

Now we calculate elasticity of purchase probability without promotion.

	Price_Point	Mean_PE	PE_Segment_0	PE_Segment_1	PE_Segment_2	PE_Segment_3	Elasticity_Promotion_1	Elasticity_Promotion_0
0	0.50	-0.096453	-0.211181	-0.119076	-0.030920	-0.010587	-0.125745	-0.195653
1	0.51	-0.100525	-0.217751	-0.123258	-0.032687	-0.011192	-0.129861	-0.201775
2	0.52	-0.104724	-0.224429	-0.127534	-0.034540	-0.011827	-0.134057	-0.207999
3	0.53	-0.109053	-0.231216	-0.131905	-0.036485	-0.012493	-0.138332	-0.214326
4	0.54	-0.113515	-0.238112	-0.136372	-0.038524	-0.013191	-0.142688	-0.220758
5	0.55	-0.118115	-0.245118	-0.140936	-0.040662	-0.013923	-0.147127	-0.227294
6	0.56	-0.122854	-0.252235	-0.145600	-0.042904	-0.014691	-0.151648	-0.233936
7	0.57	-0.127737	-0.259463	-0.150364	-0.045254	-0.015495	-0.156253	-0.240685
8	0.58	-0.132767	-0.266804	-0.155231	-0.047716	-0.016339	-0.160943	-0.247541
9	0.59	-0.137947	-0.274258	-0.160201	-0.050295	-0.017222	-0.165719	-0.254505
10	0.60	-0.143281	-0.281825	-0.165277	-0.052998	-0.018147	-0.170582	-0.261577

Now comparing Price Elasticity of Purchase probability with and without promotion with a plot. The two lines represents the elasticity of purchase probability given maximum and minimum promotional activities.



This graph tells us that elasticity curve with promotion sits above its respective no promotion counterpart for the entire price range. Customers are less price sensitive to price changes when there are promotion activities. According to this model, it would be more beneficial to have a higher original price and constant promotion rather than a lower original price.

References

1. https://data.world/sheakher/customer-activity-analysis
2. https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/
3. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#:~:text=The%20k%2Dmeans%20algorithm%20uses,the%20different%20properties%20of%20clusters.
4. https://www.analyticsvidhya.com/blog/2022/07/principal-component-analysis-beginner-friendly/
5. https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/
6. Elasticity vs. Inelasticity of Demand: What's the Difference? (investopedia.com)
7. https://en.wikipedia.org/wiki/Price_elasticity_of_demand
8. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
9. https://www.techtarget.com/searchcustomerexperience/definition/customer-segmentation
10. https://www.khanacademy.org/economics-finance-domain/microeconomics/elasticity-tutorial/price-elasticity-tutorial/a/price-elasticity-of-demand-and-price-elasticity-of-supply-cnx
11. https://corporatefinanceinstitute.com/resources/economics/cross-price-elasticity/