Mid-Semester – MBA Project Report



**Work Integrated Learning Programmes Division (WILPD),**

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI,**

**VIDYA VIHAR, PILANI, RAJASTHAN - 333031.**

**(March, 2024)**

ANALYTICAL HORIZONS: UNRAVELING CUSTOMER INSIGHTS THROUGH ADVANCED ANALYTICS, MACHINE LEARNING, AND DEEP LEARNING IN MARKETING STRATEGIES

**Course No:** S2-23\_MBAZG622T

**Mid Semester Project Report**

**Student Name: Koustav Dutta**

**BITS ID: 2022mb21009**

**Program: Business Analytics**

**Project Area: Florida, USA**

**Dissertation / Project Work carried out at:**

**Deloitte, India**



**Work Integrated Learning Programmes Division (WILPD),**

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI,**

**VIDYA VIHAR, PILANI, RAJASTHAN - 333031.**

**(March, 2024)**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Section No.** | **Section Name** | **Page No.** |
| 1 | Requirement Statement / Problem Statement | 4 |
| 2 | Project Objectives | 4-5 |
| 3 | Project Scope | 5 |
| 4 | Methodology | 6 |
| 5 | Plan vs Progress | 7 |
| 6 | Detailed description of project completed till Mid-Semester Report | 7-28 |
| 7 | Resource Requirements and their availability | 29 |
| 8 | Risks and Mitigations | 29-30 |
| 9 | Issues and Resolutions | 30 |
| 10 | Conclusions and Recommendations | 30-31 |
| 11 | References | 31-32 |
| 12 | Glossary | 32-33 |
| 13 | Summary of how the feedback for Project Outline have been addressed | 33 |

# Requirement Statement / Problem Statement

Company X, being a global retail giant, faces the challenge of adapting to the dynamic preferences and behaviors of its diverse customer base. In the pursuit of sustaining and growing market share, Company X acknowledges the need to leverage cutting-edge analytics methodologies to gain deeper insights into customer behavior, preferences, and purchasing patterns.

**Challenges:**

1. **Customer Segmentation:** Company X aims to improve its understanding of customer segments by addressing the challenge of effectively categorizing its diverse customer base. The existing methods may not be providing granular insights required for targeted marketing strategies.
2. **Data Overload:** With vast amounts of data generated from various touchpoints, Company X encounters the challenge of extracting meaningful insights. The sheer volume and complexity of data hinder the identification of critical patterns and trends.
3. **Predictive Modeling:** Anticipating future customer behavior is essential for proactive decision-making. However, Company X faces challenges in developing robust predictive models to forecast purchase probability, brand choices, and purchase quantities accurately.
4. **Elasticity Modeling:** Understanding the elasticity of customer choices concerning price changes, brand variations, and product availability is vital for strategic pricing and inventory decisions. Company X needs to enhance its modeling capabilities to capture these nuances effectively.
5. **Integration of Deep Learning:** While Company X recognizes the potential of deep learning in predicting future customer behavior, the challenge lies in seamlessly integrating these advanced techniques into existing analytics frameworks and business processes.

# Project Objectives

The project objectives for "Advanced Customer Analytics using Marketing Strategies, Machine Learning & Deep Learning" for Company X are as follows:

1. **Customer Segmentation Enhancement:**
   * Utilizing advanced analytics techniques such as Cluster Analysis and Dimensionality Reduction to redefine and enhance customer segmentation, allowing for more precise targeting and tailored marketing strategies.
2. **Data Optimization and Descriptive Statistics:**
   * Applying Descriptive Statistics to optimize the use of vast datasets, providing a clearer understanding of customer behavior and preferences, and improving the overall quality of data-driven insights.
3. **Elasticity Modeling:**
   * Developing elasticity models for purchase probability, brand choice, and purchase quantity. This aims to provide Company X with a nuanced understanding of how customers respond to changes in factors such as pricing, product availability, and brand variations.
4. **Deep Learning Implementation:**
   * Integrating Deep Learning techniques to predict future customer behavior accurately. Leveraging neural networks and deep learning algorithms will allow Company X to forecast customer actions and preferences in a more sophisticated manner.
5. **Operational Integration and Change Management:**
   * Supporting Company X in implementing changes based on the analytics findings. Ensure seamless integration of insights into existing business processes and provide guidance on change management for a smooth transition.
6. **Continuous Improvement Framework:**
   * Establishing a framework for continuous monitoring and improvement of customer analytics initiatives. This includes setting up feedback loops, refining models based on performance, and adapting strategies as market dynamics evolve.

By achieving these objectives, the project aims to empower Company X with advanced analytical capabilities, enabling the company to make data-driven decisions, enhance customer experiences, and optimize its overall business strategies for sustained success in the competitive retail sector.

# Project Scope

The scope of work to be done in this project includes as mentioned below:

1. **Project Goals:**
   * Understand and enhance customer segmentation through advanced analytics techniques.
   * Develop elasticity models for purchase probability, brand choice, and purchase quantity.
   * Implement deep learning to predict future customer behavior.
   * Optimize marketing strategies based on customer insights.
   * Facilitate seamless integration of analytics findings into existing business processes.
   * Establish a continuous improvement framework for ongoing enhancements.
2. **Timeline:**
   * The project will commence immediately and is scheduled to conclude by the end of 2nd week of May. Specific milestones and deadlines for key project phases will be outlined in the detailed project plan.
3. **Expected Results:**
   * Improved and refined customer segmentation strategies.
   * Accurate elasticity models providing insights into purchase probability, brand choice, and purchase quantity.
   * Successful implementation of deep learning models for predicting future customer behavior.
   * Enhanced marketing strategies tailored to specific customer segments.
   * Seamless integration of analytics insights into operational processes.

# Methodology

1. **Project Initiation and Planning:**
   * Develop a detailed project plan outlining key milestones, tasks, and timelines.
   * Establish communication mechanisms with my supervisor.
2. **Data Collection and Preparation:**
   * Gather relevant data from company sources, including customer transactions, demographics, and historical interactions.
   * Conduct data cleaning and preprocessing to ensure data quality and consistency.
   * Explore and understand the data through exploratory data analysis (EDA).
3. **Customer Segmentation:**
   * Apply Cluster Analysis and Dimensionality Reduction techniques to identify meaningful customer segments.
   * Utilize statistical methods to validate the robustness of the segmentation.
4. **Descriptive Statistics:**
   * Apply descriptive statistics to provide a comprehensive overview of customer behavior, purchasing patterns, and demographic characteristics.
   * Create visualizations to communicate key descriptive findings.
5. **Elasticity Modeling:**
   * Develop models for purchase probability, brand choice, and purchase quantity using appropriate statistical and machine learning techniques.
   * Validate and fine-tune models to ensure accuracy and reliability.
6. **Deep Learning Implementation:**
   * Implement deep learning algorithms, such as neural networks, to predict future customer behavior.
   * Train and validate models using historical data, optimizing hyperparameters for better performance.
7. **Continuous Improvement Framework:**
   * Establish a framework for continuous monitoring, feedback collection, and model refinement.
8. **Documentation and Knowledge Transfer:**
   * Document the entire methodology, including data sources, preprocessing steps, modeling techniques, and key findings.
9. **Project Evaluation and Closure:**
   * Evaluate project outcomes against predefined objectives and success criteria with the help of supervisor.
   * Generate a comprehensive project closure report, highlighting achievements, lessons learned, and recommendations for future initiatives.

# Plan vs Progress

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Start Date | End Date | Progress |
| Data Collection & Gathering | 2024-02-05 | 2024-02-07 | Done |
| Data Understanding & Analysis | 2024-02-08 | 2024-02-12 | Done |
| Research & Understanding of Marketing Techniques | 2024-02-14 | 2024-02-20 | Done |
| Identifying the Problems and digging deep | 2024-02-21 | 2024-02-25 | Done |
| Working with Dummy Data & EDA Techniques | 2024-02-26 | 2024-02-28 | Done |
| Deeper Literature Survey | 2024-02-30 | 2024-03-04 | Done |
| EDA and Clustering Algos on original data | 2024-03-05 | 2024-03-08 | Done |
| Identifying problems & shortcomings | 2024-03-10 | 2024-03-12 | Done |
| Hybrid Approach of algorithm on data | 2024-03-13 | 2024-03-16 | Done |
| Drawing Conclusions and insights | 2024-03-17 | 2024-03-19 | Done |
| Thinking and connecting the next steps | 2024-03-20 | 2024-03-21 | In Progress |
| Report Creation | 2024-03-22 | 2024-03-28 | Done |

# Detailed description of project completed till Mid-Semester Report

# 6.1. Functional Information & Requirements

The project objectives revolve around the realm of Customer Analytics, delving into the core principles of marketing modeling theory such as segmentation, targeting, positioning, and price elasticity. Here's an alternative formulation of the objectives:

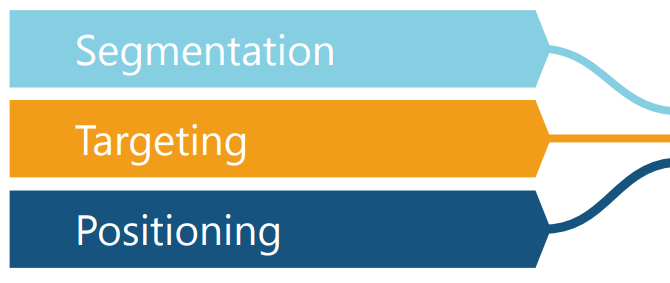
**a. Customer Analytics:** The initial phase of the project entails an in-depth exploration of customer segmentation techniques. This involves the application of both hierarchical and flat clustering methodologies to categorize customers into distinct groups based on shared characteristics. Additionally, Principal Components Analysis (PCA) is employed to streamline the data and facilitate a more nuanced understanding of customer segments. Furthermore, a hybrid approach combining PCA with K-means clustering is adopted to achieve a more granular segmentation, providing deeper insights into customer behavior.

**b. Purchase Analytics:** The subsequent phase of the project delves into the descriptive and predictive analysis of customer purchase behavior. This encompasses the development of models to analyze purchase incidence, brand choice, and purchase quantity, offering valuable insights into consumer preferences and buying patterns. Moreover, advanced deep learning techniques are leveraged to enhance predictive capabilities, enabling accurate forecasts of future purchasing trends.

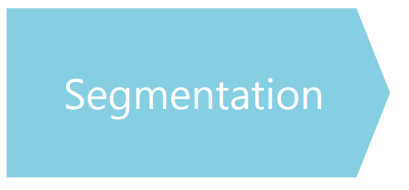
* The **STP framework,** an integral component of marketing strategy, plays a pivotal role in guiding the project's approach:

1. **Segmentation:** The segmentation aspect involves partitioning the heterogeneous market into homogeneous groups of consumers sharing similar characteristics or behaviors. By segmenting the market, marketers can tailor their strategies to meet the diverse needs and preferences of different consumer segments. Segmentation criteria may encompass demographic, psychographic, geographic, or behavioral variables, allowing for a more targeted and effective marketing approach [1].
2. Top of Form
3. **Targeting:** Following segmentation, targeting involves the strategic selection of specific customer segments to focus marketing efforts on. This process entails assessing various segments based on criteria such as size, growth potential, profitability, and alignment with the company's offerings. By identifying the most lucrative and compatible segments, marketers can allocate resources effectively and tailor their marketing strategies to resonate with the chosen target audience.
4. **Positioning:** Positioning is the strategic process of establishing a distinctive and favorable perception of the company's products or services in the minds of consumers relative to competitors. It involves crafting a unique value proposition and effectively communicating it through branding, messaging, and marketing channels. Successful positioning sets the company apart from competitors and reinforces its value proposition, ultimately shaping consumer perceptions and driving competitive advantage.

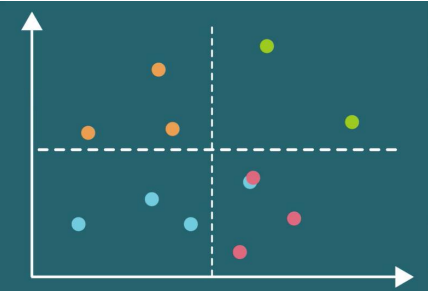
The STP framework serves as a strategic guidepost for marketers, providing a systematic approach to identifying, prioritizing, and engaging with target market segments. By leveraging segmentation, targeting, and positioning strategies, businesses can optimize their marketing efforts to better meet the needs and preferences of their target audience while effectively differentiating their offerings in the marketplace. This holistic approach enables companies to drive sustainable growth and profitability by aligning their marketing strategies with consumer demand and competitive dynamics.



STP is a fundamental marketing framework. It can be applied to all areas of business and marketing activities.

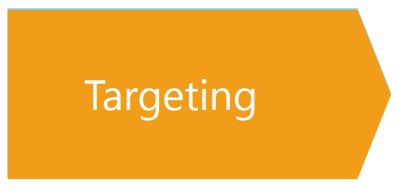


The process of dividing a population of customers into groups that share similar characteristics.

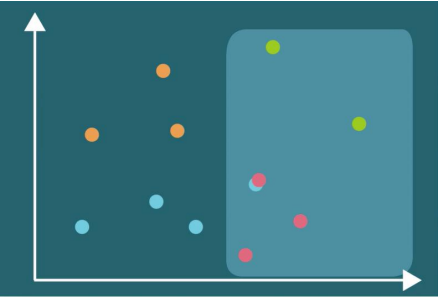


Observations within the same group would have comparable purchasing behavior.

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



The process of evaluating potential profits from each segment and deciding which segments to focus on.



Examining customers’ perception. (Involves psychology and usually budget constraints

Selecting ways to promote our products. We can **target** one segment on TV and another online



What product characteristics do the customers from a certain segment need?

**Marketing Mix Model**

* The **Marketing Mix Model (MMM)** stands as a strategic framework utilized by marketers to scrutinize and refine various facets of their marketing endeavors, with the overarching aim of maximizing return on investment (ROI) and accomplishing business objectives. Rooted in the foundational concept of the marketing mix, the model encompasses four key elements, often referred to as the four Ps: Product, Price, Place, and Promotion. Here's a concise overview of each component within the Marketing Mix Model:

1. **Product:**
   * The product dimension of the marketing mix encapsulates the tangible or intangible offerings presented by a company to its target market. This encompasses not only the physical product itself but also factors such as product attributes, quality, packaging, branding, and customer service. Within the Marketing Mix Model framework, marketers delve into how different product attributes and strategies influence consumer demand, satisfaction levels, and overall business performance.
2. **Price:**
   * Price delineates the monetary value customers are willing to exchange for the product or service offered by the company. Pricing strategies encompass a spectrum of approaches, spanning premium pricing, penetration pricing, discount pricing, and value-based pricing, among others. Within the Marketing Mix Model, marketers evaluate the ramifications of pricing decisions on sales volume, revenue generation, profitability margins, and market share acquisition.
3. **Place:**
   * Place, often termed distribution, pertains to the channels and methodologies employed by companies to make their products or services accessible to consumers. This encompasses decisions relating to distribution channels, intermediaries, logistics, inventory management, and the establishment of retail or online presences. Within the Marketing Mix Model framework, marketers scrutinize the efficacy of distribution strategies in reaching target demographics, maximizing market coverage, and ensuring product availability and convenience for consumers.
4. **Promotion:**
   * Promotion encompasses the gamut of activities and communication endeavors employed by companies to foster awareness and drive engagement with their products or services among target audiences. This encompasses advertising, public relations initiatives, sales promotions, direct marketing campaigns, social media endeavors, and other promotional tactics. Within the Marketing Mix Model framework, marketers analyze the impact of promotional activities on brand visibility, customer interaction levels, sales conversion rates, and overall marketing efficacy.

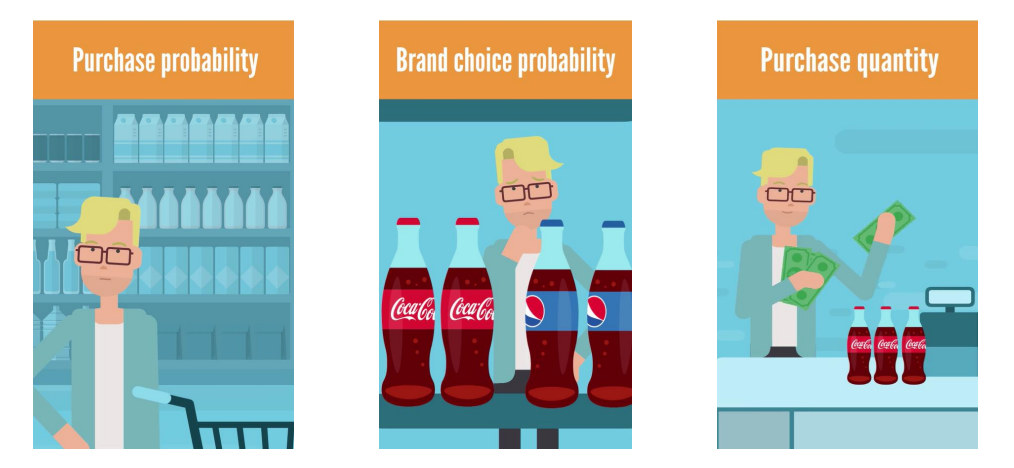
The Marketing Mix Model amalgamates insights from each of these four components, facilitating data-driven decision-making and the optimization of marketing strategies. By scrutinizing the performance of various marketing mix elements and their interplay over time, companies can deploy resources more judiciously, pinpoint avenues for enhancement, and elevate their overall marketing efficacy and ROI.

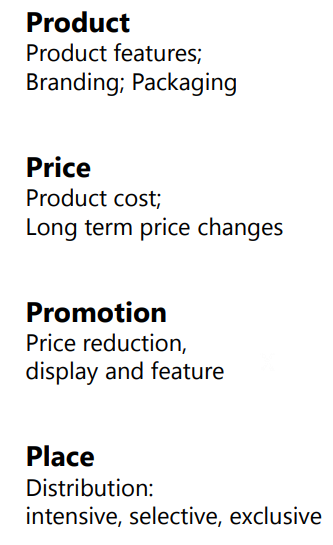
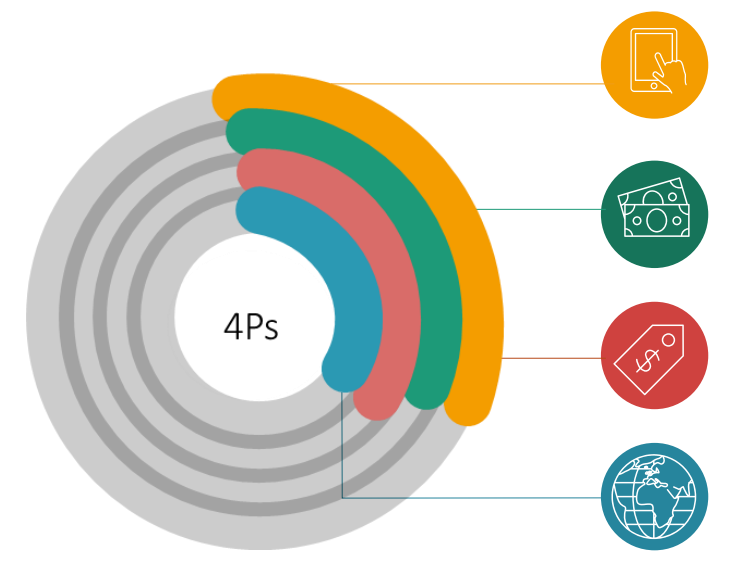


In fact, this process is so important, that it has a framework of its own called: **Marketing Mix**.

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

The aim is to develop the best product or service and offer it at the *right price* through the *right channels*.





**Physical and Online Retailers**



|  |  |  |
| --- | --- | --- |
| Location | Many locations, returning customers. | One location, many customers. |
| Data | Fewer customers at a particular store, due to the physical restriction. | More data points and more diverse customer information. |
| Returns | In physical stores, customers can see the product itself. Returns are less likely. | Products are returned more often, as customers cannot see and test an item. |
| Purchase History | Gather information through loyalty cards. | Database with all past purchases of customers. |
| Brand Choice | Unavailable in physical stores. We assume the customer has considered all competitor brands. | We may have data for all products that the customer has looked at and which competing products a customer has considered. |
| Ratings and Reviews | Unavailable in physical stores. | Different items could be reviewed and rated (significant features for predictive modeling). |

**Price Elasticity**

Price elasticity is a critical concept in economics and marketing, representing the responsiveness of demand for a product or service to changes in its price. It measures the percentage change in quantity demanded relative to a percentage change in price. The formula for **price elasticity of demand (PED)** is as follows:



The price elasticity coefficient can be positive, negative, or zero:

• If PED>1, demand is elastic, meaning that a small change in price results in a relatively larger change in quantity demanded. Products with elastic demand are sensitive to price changes, and reducing prices may lead to increased revenue.

• If PED<1, demand is inelastic, indicating that changes in price have a proportionally smaller impact on quantity demanded. Products with inelastic demand are less sensitive to price changes, and altering prices may have a limited effect on revenue.

• If PED=1, demand is unit elastic, meaning that changes in price result in proportionate changes in quantity demanded. Unit elastic demand indicates a balanced response to price changes, with revenue remaining constant.

Understanding price elasticity is crucial for businesses when making pricing decisions. By analyzing price elasticity, companies can determine the optimal pricing strategy to maximize revenue and profitability. Additionally, price elasticity insights can inform product positioning, market segmentation, and promotional efforts [3].

In addition to price elasticity of demand (PED), there are two other types of elasticity that are commonly used in economics and marketing: own price elasticity and cross-price elasticity.

1. **Own Price Elasticity:**
   * Own price elasticity measures the responsiveness of quantity demanded for a particular product to changes in its own price, holding all other factors constant. It helps businesses understand how sensitive consumers are to changes in the price of a specific product.



* + Own price elasticity can be used by businesses to determine the optimal pricing strategy for a product. For example, if own price elasticity is elastic (greater than 1), reducing the price may lead to a larger increase in quantity demanded, potentially resulting in higher total revenue.

Conversely, if own price elasticity is inelastic (less than 1), reducing the price may lead to a smaller increase in quantity demanded, and businesses may need to consider other strategies to boost sales.

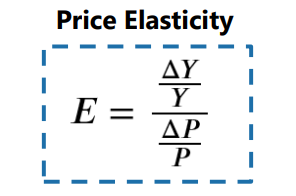
1. **Cross-Price Elasticity:**
   * Cross-price elasticity measures the responsiveness of quantity demanded for one product to changes in the price of another related product, while holding all other factors constant. It helps businesses understand how the demand for one product is affected by changes in the price of another product.
   * The formula for cross-price elasticity of demand (XED) is:



* + Cross- Cross-price elasticity can provide valuable insights into the relationship between substitute or complementary products. For example, if the cross-price elasticity between two products is positive, it suggests that they are substitutes, meaning that an increase in the price of one product leads to an increase in demand for the other product. Conversely, if the cross-price elasticity is negative, it indicates that the products are complements, meaning that an increase in the price of one product leads to a decrease in demand for the other product [4].

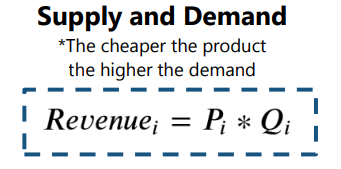
By understanding both own price elasticity and cross-price elasticity, businesses can make informed pricing decisions, anticipate changes in consumer behavior, and develop effective marketing strategies to maximize revenue and profitability.

If we want to generalize the concept of Price Elasticity, Price elasticity measures how a variable of interest changes when the price changes.



Y: Economic variable of Interest

P: Price



Q: Quantity

P: Price

**Own Price Elasticity**



Price elasticity with respect to the same product

**Cross Price Elasticity**



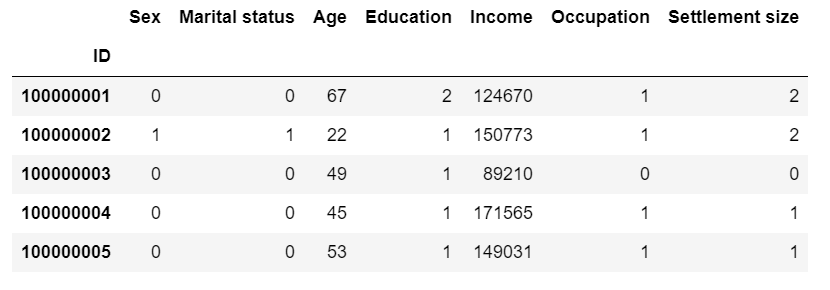
Price elasticity with respect to another product

**6.2. Customer Analytics Segmentation Dataset Collection & Analysis**

The first part of the project focuses on **Customer Analytics** by performing customer segmentation. It involves the application of hierarchical and flat clustering techniques for dividing customers into groups. It also features applying the Principal Components Analysis (PCA) to reduce the dimensionality of the problem, as well as combining PCA and K-means for an even deeper insightful customer segmentation. The Dataset used for this purpose is from our client – Company X (due to privacy reasons, name of Client cannot be revealed) which is a FMCG company. The dataset was originally kept in AWS S3 Buckets of the client. With special permission from client, I have gathered sample of the original data which was already preprocessed by the Data Engineering Team of the client. The sample dump of the original dataset is given in the form of .CSV file. Given below is the information about the dataset fields, their meanings, significance. The Coding of the entire thesis has been done using **Python programming language** on **Anaconda Platform Virtual Environment** in **Jupyter IDE**.



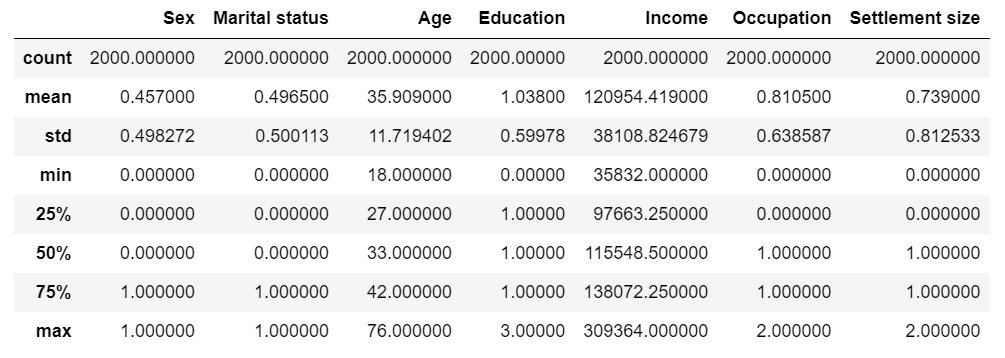
**a. Sample Data**



We have to refer to the “Segmentation Data Legend” in order to understand the meaning of the above data.

**b. Statistical Analysis**

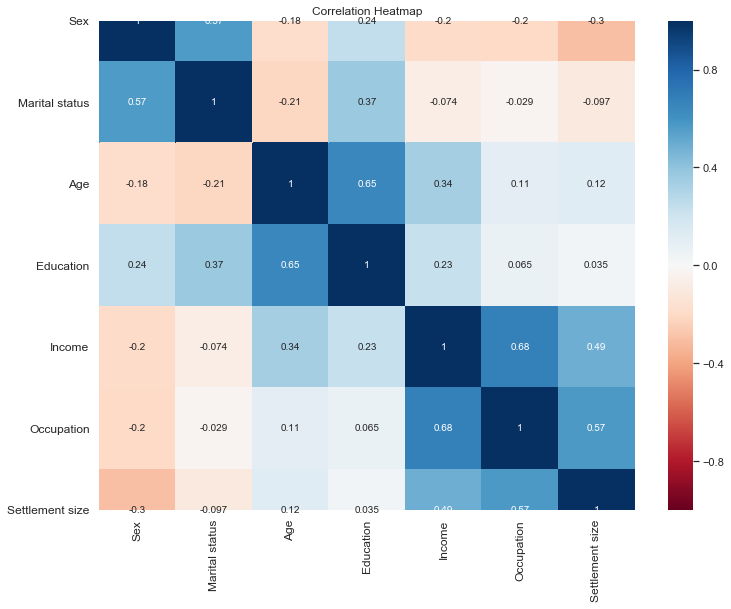
After this, we will analyse the Statistical Description of the dataset. This is done in order to gain a deeper insight about the mean, standard deviation, spread of the data and various other factors. Given below is the statistical description of the dataset [9]:



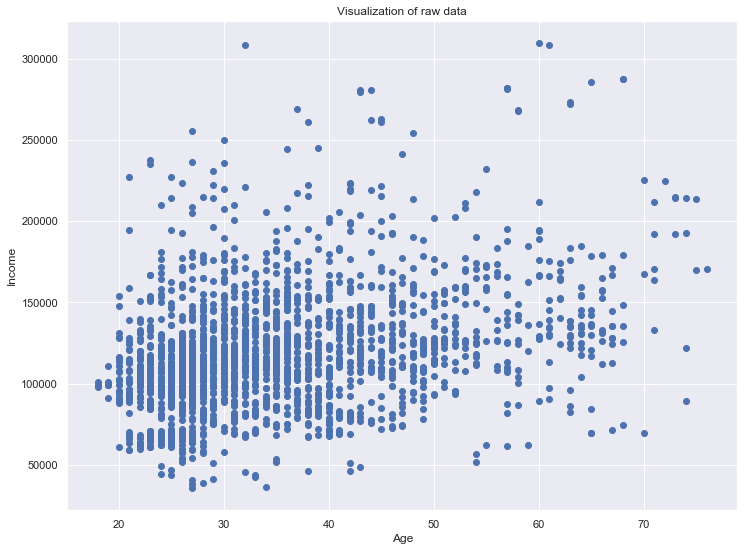
**c. Correlation Map Analysis**

Next, the task in line is to find the Correlation between the various Features of the dataset. Correlation chart visualizes the strength and direction of relationships between the variables involved in our dataset. By plotting correlations between the variables, we can identify which variables are most strongly associated with each other. This helps in understanding which customer attributes are most relevant for segmentation.

Correlation maps aid in feature selection by highlighting variables that have a significant impact on customer segmentation. Variables with high correlations with the dependent variable are likely to be important for distinguishing between segments. We can prioritize these variables for further analysis and segmentation modeling [10].



1. **Data Visualization (Scatter Plot)**

****

**6.3. Clustering & Segmentation of Customers**

The process we follow is to make use of “Hierarchical Clustering” technique in order to find the “Number of Clusters” from the dataset. Next, we move on to “Flat Clustering” or “K-Means Clustering” technique in order to perform the Segmentation of the customers in the dataset [5].

1. **Hierarchical Clustering**

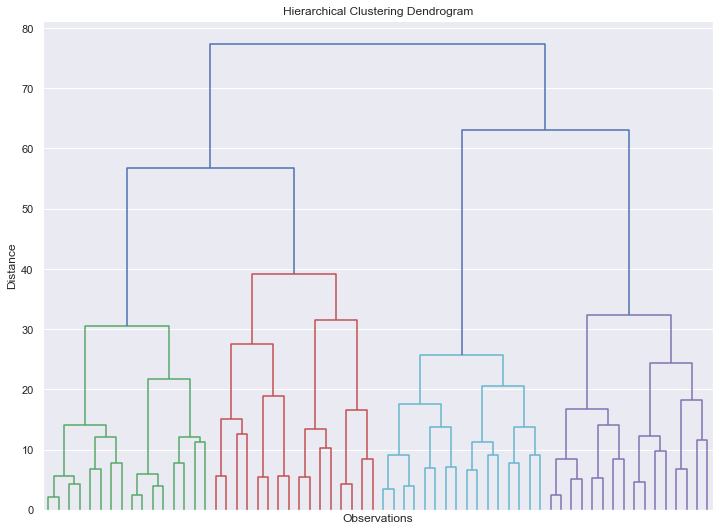
Hierarchical clustering is a popular technique used in data mining and exploratory data analysis to group similar data points into clusters based on their characteristics. Unlike K-means clustering, which requires the number of clusters to be specified in advance, hierarchical clustering creates a hierarchical tree-like structure, known as a dendrogram, to represent the relationships between data points. Hierarchical clustering begins by treating each data point as a separate cluster. Then, it iteratively merges the closest pairs of clusters based on a similarity or distance measure until all data points belong to a single cluster or until a predetermined number of clusters is reached [11].

A distance metric, such as Euclidean distance, Manhattan distance, or correlation distance, is used to measure the dissimilarity or similarity between data points. This metric determines how clusters are merged or split based on their proximity in the feature space. We are using **“Euclidean Distance”** in this project.

* The hierarchical clustering process results in a dendrogram, which is a tree-like diagram that illustrates the hierarchical relationships between clusters and data points.
* At the bottom of the dendrogram, each data point is represented as a single cluster. As the algorithm progresses, clusters are merged, and branches in the dendrogram represent the fusion of clusters.
* The length of each branch in the dendrogram corresponds to the distance between the clusters being merged. Longer branches indicate greater dissimilarity between clusters, while shorter branches represent closer similarity.

The dendrogram allows us to visualize the hierarchical structure of the data and identify clusters at different levels of granularity. By cutting the dendrogram at a certain height, we can determine the number of clusters or hierarchical levels that best represent the data [12].

The choice of where to cut the dendrogram depends on the specific application and the desired level of clustering granularity. Different cut heights yield different numbers of clusters, enabling us to explore the data at various levels of detail.



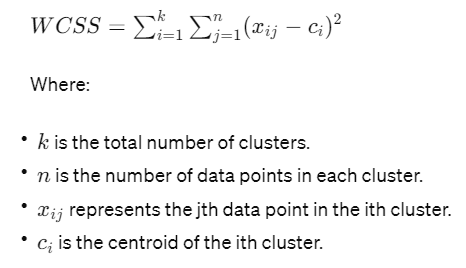
1. **K-Means Clustering (with K-Means ++)**

K-means clustering stands out as a widely utilized unsupervised machine learning method employed for clustering and segmentation endeavors. Its primary objective revolves around dividing a dataset into a predefined number of clusters, with each cluster being epitomized by its centroid. An enhancement over the standard K-means algorithm, K-means++, enriches the initialization phase to yield more precise clustering outcomes [13].

* The K-means clustering process commences with the random initialization of cluster centroids. Subsequently, it iteratively assigns each data point to the nearest centroid and updates the centroids based on the mean of the data points affiliated with each cluster.
* This iterative process persists until convergence is achieved, characterized by minimal alterations in centroids or reaching a stipulated number of iterations.

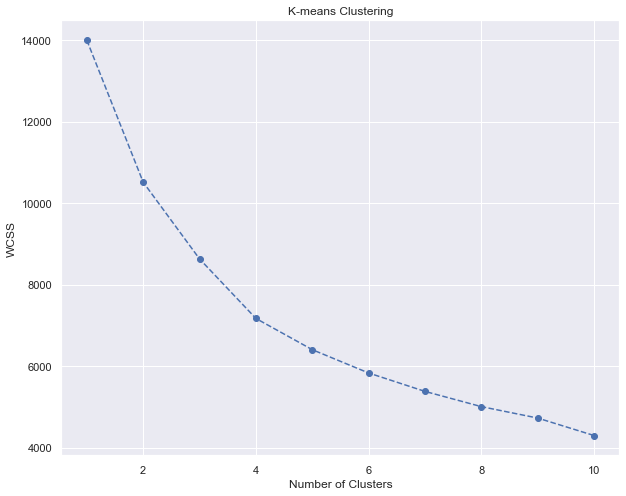
Within-Cluster Sum of Squares (WCSS) serves as a metric gauging the compactness of clusters. It quantifies the summation of squared distances between each data point and its designated cluster centroid. Throughout the K-means algorithm iterations, the goal is to minimize WCSS as centroids undergo adjustments to refine cluster assignments.

* In the realm of customer segmentation, K-means clustering proves instrumental in identifying discrete customer groups predicated on shared attributes such as demographics, purchasing behaviors, or preferences.
* WCSS assumes significance in assessing clustering quality by discerning how closely grouped data points are around their respective centroids. Reduced WCSS values connote more tightly knit and distinct clusters.



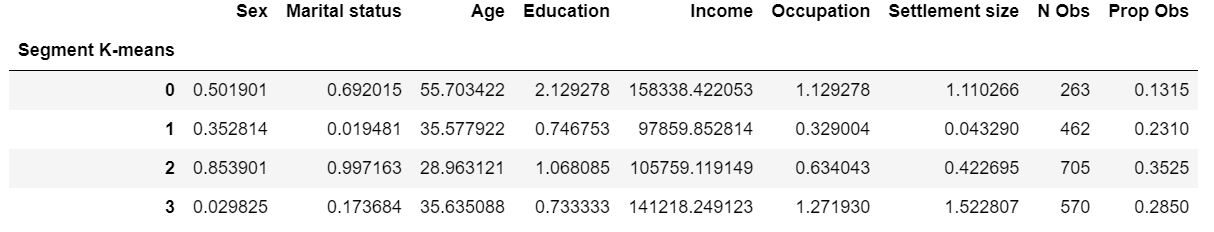
K-means++ improves upon the random initialization step of K-means by selecting initial centroids that are well spread out and representative of the data distribution.

Instead of selecting centroids randomly, K-means++ selects the first centroid randomly and subsequent centroids with a probability proportional to the square of the distance from the nearest centroid.

****

Depending on the shape of the above graph, we make a decision about the number of clusters. We make use of the **“Elbow Method”** for this purpose. Thus, the number of clusters comes out to be 4, since there is change in slope at that point.

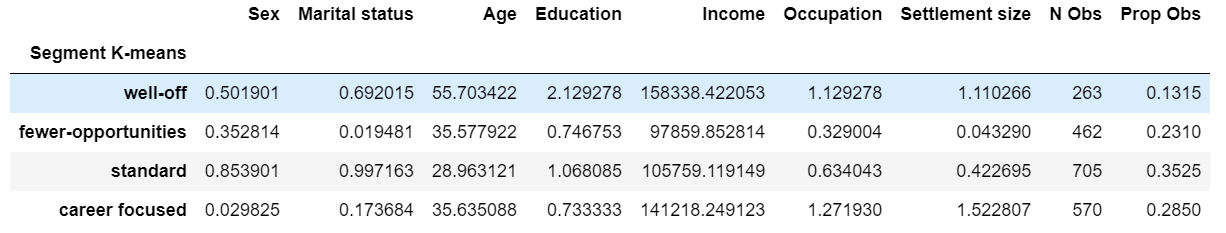
Next, we calculate the mean values for the fields and also add a new column with the assigned clusters for each point corresponding to the fields.



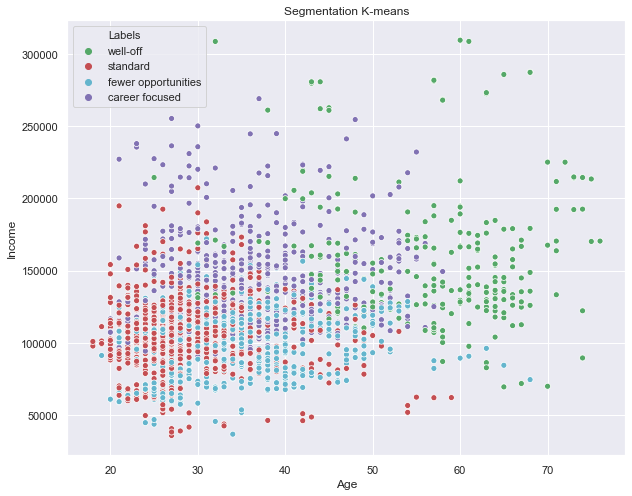
After carefully analyzing the above results, we dig deeper insights, observe the values of the various fields (Age, Education, Income, Occupation, Settlement Size etc) [7].

* We see that, for the **Segment 0**, there is maximum income, high value of occupation, settlement size, education etc and name the segment as **“Well-Off”**.
* Similarly, we observe the values of the fields and see, that for **Segment 1**, the value of Occupation is least and the other values of income, education, settlement size etc are also quite less. We name it as **“Fewer Opportunities”.**
* For **Segment 2**, we notice that the values for all the fields are in the middle range and have a balance, hence, we name the segment as **“Standard”.**
* Finally, on observing the values of various fields corresponding to **Segment 3**, we see that, the values of Occupation & Settlement Size are maximum among all the other segments. The age group is also in the middle range and Income is also quite high. Hence, we name this segment as **“Career Focused”.**

Further, we also find out the proportion of customers under each of the segments and observe that, Segment 2 or “Standard” Segment is having the highest proportion of customers followed by Segment 3, Segment 1 and Segment 0. Thus, the percentage of customers from Well-Off Segment is quite less. This analysis can provide a deeper insight and help the client in taking decisions and strategizing its way forward in the market [8].



Next, we take a step forward and plot the Customer Data mapping them with their corresponding segments.



The main problem with the above plotted graph after performing K-Means Clustering based Segmentation is that the graph does not effectively segregate the data points and do not segment them properly into clear cut clusters [14]. Apart from these, we can list out a number of problems that can occur if we want to make our method of research flexible for any kind of dataset going forward. Since, we are working on a sample dataset for our project, the problems may not be clearly visible but, as this process is going to be extended for use in real-life dataset by the client – Company X, so, here are the problems that can be faced by K-means Clustering method for Segmentation of the customers:

1. **Sensitive to Initial Centroid Selection:**
   * K-means clustering is sensitive to the initial selection of centroids. Random initialization can lead to different clustering results in each run, impacting the stability and reliability of the segmentation.
2. **Assumption of Circular Clusters:**
   * K-means assumes that clusters are spherical or isotropic and have equal variance. This assumption may not hold true for complex or irregularly shaped clusters in real-world data, leading to suboptimal segmentation.
3. **Impact of Outliers:**
   * Outliers or noise in the data can significantly affect the clustering results in K-means. Since K-means aims to minimize the within-cluster sum of squares, outliers may distort cluster boundaries and influence centroid positions.
4. **Difficulty with High-Dimensional Data:**
   * K-means struggles with high-dimensional data, where the curse of dimensionality can lead to increased computational complexity and reduced clustering performance. High-dimensional data may also exhibit sparsity, making it challenging to define meaningful distance metrics.

To address these challenges and improve the effectiveness of customer segmentation, we have devised an innovative hybrid approach of using **Principal Component Analysis (PCA) with K-means clustering**. As per the work results and dealing with the dataset and performing a wide variety of tests, following are the challenges and problems which can be handled efficiently by this technique:

1. **Dimensionality Reduction:**
   * PCA reduces the dimensionality of the data by transforming the original features into a lower-dimensional space of principal components. By capturing the most significant variance in the data, PCA helps mitigate the curse of dimensionality and enhances the performance of K-means clustering.
2. **Feature Extraction:**
   * PCA extracts underlying patterns and structures from high-dimensional data, allowing K-means to operate on a more compact and informative feature space. This reduces the impact of noise and irrelevant features on the clustering process, resulting in more robust segmentation.
3. **Improved Cluster Interpretation:**
   * PCA simplifies the data representation while preserving as much variance as possible. This makes it easier to interpret and visualize the resulting clusters, as the principal components represent meaningful combinations of the original features.
4. **Enhanced Stability and Consistency:**
   * PCA + K-means offers more stable and consistent clustering results compared to standalone K-means. By reducing the sensitivity to initial centroid selection and the impact of outliers, PCA pre-processing improves the convergence and reliability of the clustering algorithm.

Overall, the combination of PCA and K-means clustering addresses the limitations of K-means for customer segmentation by reducing dimensionality, extracting relevant features, and enhancing the stability and interpretability of the segmentation results. Thus, the sophisticated approach is as follows.

1. **Principal Component Analysis (PCA)**

**Principal Component Analysis (PCA)** is a fundamental technique used for dimensionality reduction and feature extraction. Its role in customer segmentation, especially when adopting a hybrid approach with K-means clustering, is instrumental in enhancing the effectiveness and efficiency of the segmentation process [15].

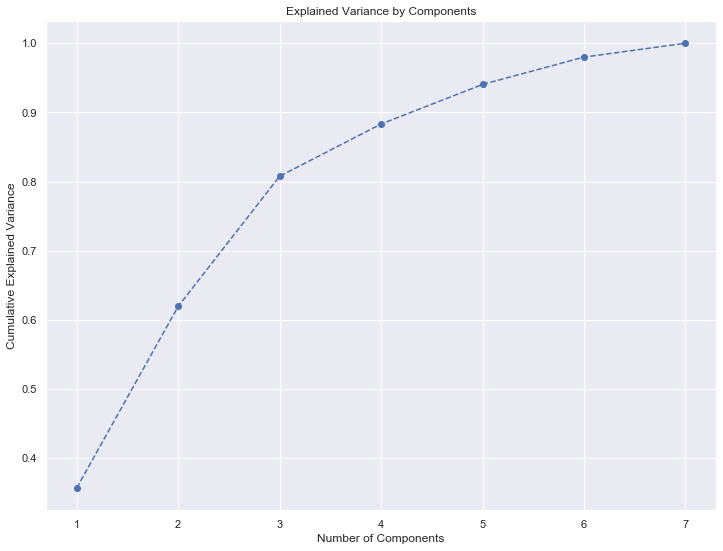
PCA is a mathematical method that transforms high-dimensional data into a lower-dimensional representation while preserving the maximum variance in the data. It achieves this by identifying the principal components, which are linear combinations of the original features that capture the most significant sources of variation. By reducing the dimensionality of the data, PCA simplifies complex datasets and facilitates visualization, interpretation, and analysis.

In our Customer Segmentation process, PCA plays a crucial role in preprocessing and preparing the data before integrating the data into the pipeline of K-means ++ Clustering Algorithm.

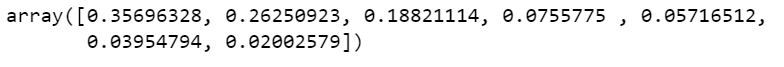
1. **Dimensionality Reduction:** The Customer dataset of Company X contains numerous variables, such as Sex, Marital Status, Age, Education, Income, Occupation and Settlement size. PCA reduces the dimensionality of the dataset by transforming them into a lower-dimensional space of principal components. This reduces the computational complexity of subsequent clustering algorithms and mitigates the curse of dimensionality.
2. **Feature Extraction:** PCA extracts underlying patterns and structures from our high-dimensional customer data, thus helping in customer segmentation. By capturing the most significant sources of variance, PCA helps focus the clustering algorithm (K-Means ++) on the most informative aspects of the data, leading to more meaningful and interpretable segmentation results.
3. **Noise Reduction:** PCA can help mitigate the impact of noise and irrelevant features in our customer data by emphasizing the principal components that explain the most variance. This enhances the robustness of the segmentation process and improves the quality of the resulting customer segments.

Given below is the plot showing the Cumulative Explained Variance vs Number of Components from the Customer Dataset.

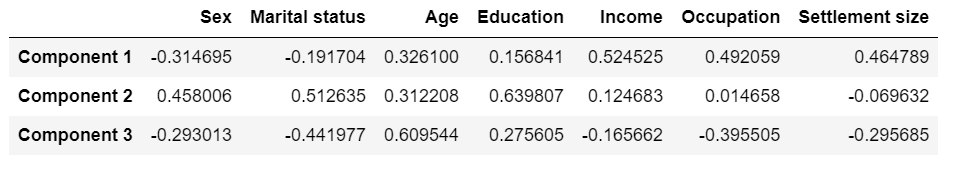
Top of Form

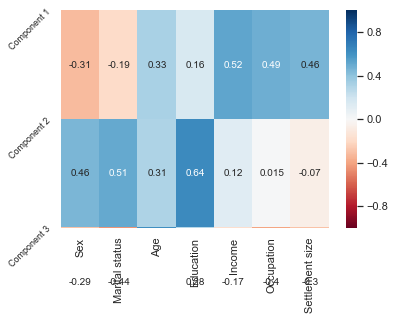


The above plot is based upon the individual variances explained by each of the components from the dataset as shown below:



The components above are arranged in descending order of importance. Components’ variance always sums up to 1. Thus, we find a subset of components, while preserving variance. The Rule of Thumb is to keep 80% of the variance. Thus, we choose **the Number of Components to be kept as 3**. Next, we observe the Component Matrix having the factor loadings, i.e., the **Correlation between the Original variable and the Components**. The values of the Component Matrix is in the range [-1,1] as it shows the Pearson Correlation values between the PCA Components and the Features of the Dataset. We can also view the Heat Map for Principal Components against original features.

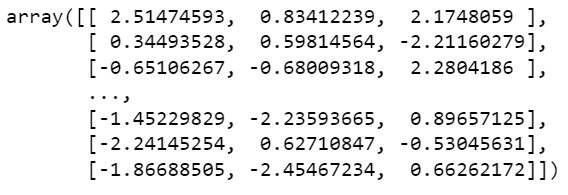




From the above Correlation Heatmap between the PCA Components and the various original features of the Customer Dataset, we observe that:

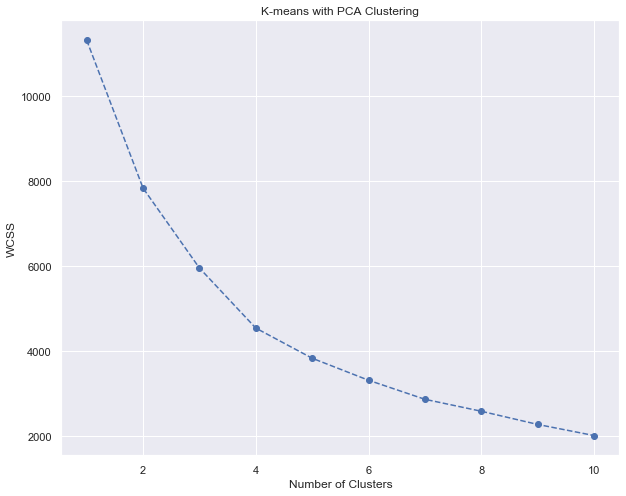
* There is high correlation between the Component 1 and the features like Age, Education, Income, Occupation and Settlement Size indicating “**Career Focused”** quality of individual.
* For the Component 2, we see that, the correlation is prominent for the features like Sex, Marital Status, Education, and correlation is not there much for the Career Focused quality. This is more focused on **“Education & Lifestyle”** of the individuals.
* The Correlation values between the Component 3 and the fields like Marital Status, Age, Occupation focuses on the **“Experience”** of individuals.

After this, the 7-Dimensional Customer Data needs to be transformed into 3-Dimensional Customer Data where each row represents each of the datapoints and the three columns represent the 3 PCA Components representing the entire data.



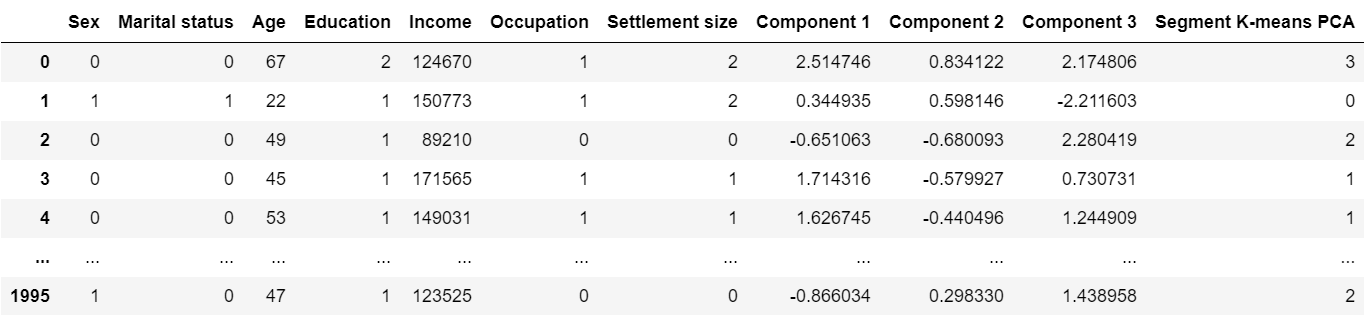
1. **K-Means Clustering with PCA**

Now, we have done the Dimensionality Reduction of the Customer Dataset and extracted 3 Components which explains the maximum variance of the entire data thus, representing the whole of the data and its variation. So, we will plot the **“Within Cluster Sum of Squares (WCSS)”** for the **K-means PCA model**. Here we make a decision about the number of clusters using this innovative hybrid algorithm [16].

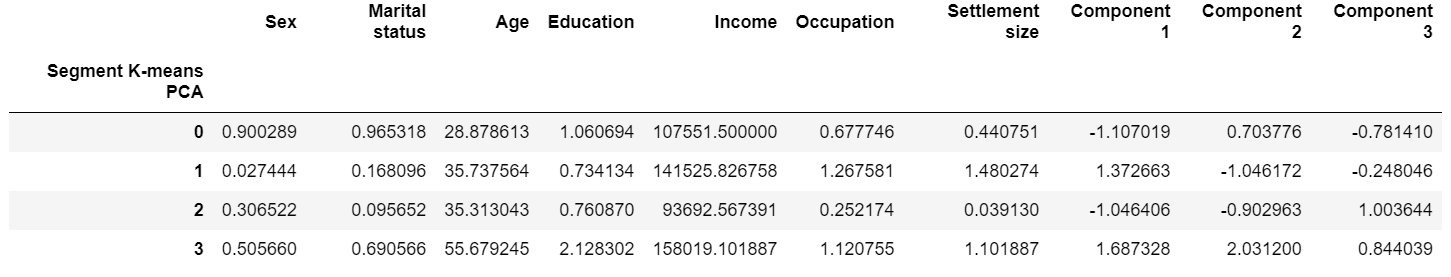


Thus, from the above graph, we observe that the significant change in slope occurs at the **“Elbow Point”** so, we have chosen **four clusters**, and therefore, we run K-means with number of clusters equals four.

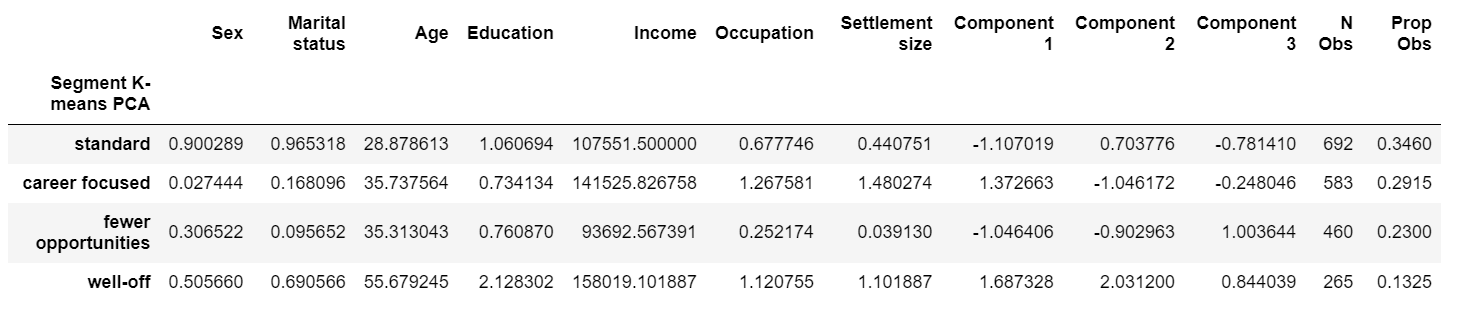
After fitting our Customer Data with the (K-Means + PCA) Model, we Map the 3 Components corresponding to the original dataset and also map the Segments of the data corresponding to this data.



Next, we calculate the means of the field values by segments.



Next, we do a careful analysis of the 3 components columns corresponding the mean values of the original columns and try to extract the meanings represented by combination of the components [17]. Give below is the resultant table. This also shows the Number of Observations for each of the Segments and the Proportion of Observations for the segments:



Career Education Experience

Taking reference from the previously shown Correlation Heatmap between the PCA Components and the original features of the Customer Dataset, we do a demographic analysis taking into consideration the “Segment K-Means PCA” column of the above table which corresponds to the three component values against the mean values if the original columns. Thus, we observe that:

* All the three values of the three components corresponding to the **Segment 3** are high positive factor loadings, thus, it represents the **“Well Off”** Segment. This signifies that this cluster or segment is having good position wrt Career, Education & Lifestyle and Experience in all the three categories.
* The values corresponding to the **Segment 2** in the 3 component columns have high positive value for only the Component 3 which represents the Experience. The values for the other two components which represents Career and Education & Lifestyle are quite less in negative. So, this Segment 2 represents the **“Fewer Opportunities”** section.
* Similarly, as see from the table above, the value of Component 1 corresponding to **Segment 1** is a high positive one. But, for the other two Components, the values are negative. This represents the Segment 1 as representing the **“Career Focused”**  Segment.
* For the **Segment 0**, all the values of the three components are more or less same. Thus, this segment represents the **“Standard”** Segment.

Finally, we plot the data by PCA components. The Y axis of the graph represents the first component, and the X axis represents the second component.



With the successful implementation of Customer Segmentation using the innovative PCA + K-Means Hybrid Algorithm, our focus regarding the future of the thesis now shifts towards delving deeper into Purchase Analytics, an essential aspect of understanding customer behavior. We will have to perform exploration and analysis, with the application of Descriptive Statistics as the foundation of our investigation. We will have to interpret customers' behavior and offer invaluable insights. Slowly, we will make use of elasticity modeling, where traditional metrics are redefined to encompass purchase probability, brand choice, and purchase quantity. Finally, we can also take into consideration the use of Deep Learning to predict future behaviors of customers, unlocking a realm of possibilities and empowering Company X to anticipate and adapt to evolving customer dynamics with confidence and precision.

# Resource Requirements and their availability

1. **Project Lead (You):**
   * Responsibilities: Overall project management, data analysis, modeling, and implementation of advanced analytics techniques and machine learning.
   * Availability: Full-time commitment throughout the project duration.
2. **Supervisor:**
   * Responsibilities: Guidance, mentorship, and periodic review of project progress.
   * Availability: Regular check-ins as per the agreed-upon schedule, with additional support as needed.
3. **Company Database Admin:**
   * Responsibilities: Access to relevant databases, assistance with data extraction and preprocessing.
   * Availability: Coordination with the database admin as needed during the initial stages of the project.
4. **Hardware (Laptop):**
   * Requirements: High-performance laptop with specifications suitable for data analysis and machine learning tasks.
   * Availability: Full-time access to the required hardware throughout the project duration.
5. **Software:**
   * Requirements: Specialized software for data analysis (e.g., Python with libraries like Pandas, NumPy, scikit-learn), machine learning (TensorFlow, PyTorch) (Visualization (Tableau, Power BI)).
   * Availability: Installation and ongoing access to the necessary software for the entire project duration.

# Risks and Mitigations

**Risks and Mitigation Plan for Advanced Customer Analytics Project at Company X:**

1. **Lack of Stakeholder Engagement:**
   * **Risk:** Limited stakeholder engagement may lead to misunderstandings regarding project goals and outcomes.
   * **Mitigation Plan:** Regular communication channels can be established with my supervisor, including project updates, progress reports, and clarification sessions to address any queries or concerns.
2. **Biased Data Interpretation:**
   * **Risk:** Unintentional bias in interpreting analytics results may lead to inaccurate conclusions.
   * **Mitigation Plan:** Incorporating a peer review process, involving discussions with a supervisor or a cross-functional team, to ensure objectivity and minimize bias in the interpretation of analytics findings.
3. **Data Quality and Security Concerns:**
   * **Risk**: Issues related to data quality, integrity, or security may compromise the accuracy of analytics models.
   * **Mitigation Plan**: Implementing robust data quality checks, work closely with Company’s Data teams to ensure data security protocols, and adhering to industry best practices for data integrity throughout the project lifecycle.
4. **Insufficient Resources:**
   * **Risk:** Inadequate resources, such as time, may hinder the project's effectiveness.
   * **Mitigation Plan:** Conducting a thorough resource assessment at the project outset and implementing time management strategies to ensure optimal utilization throughout the project.
5. **Unforeseen Technological Challenges:**
   * **Risk:** Technological issues or compatibility challenges may arise during the implementation of machine learning and deep learning models.
   * **Mitigation Plan:** Conducting thorough technology assessments, collaborating with Supervisor, and establishing contingency plans to address any unforeseen technological challenges promptly.

# Issues and Resolutions

One of the primary challenges encountered was the availability of incomplete and masked preprocessed data, leading to difficulties in the analysis due to problems faced in interpretation and understanding. The interpretation of results from sophisticated analytics models such as Unsupervised Machine Learning have been challenging, particularly in understanding the underlying factors driving predictions or segmentations. Ensuring the explainability and transparency of results was crucial for actionable insights. Also, the application of advanced analytics techniques such as machine learning posed challenges in terms of implementation complexity and computational resources required. Handling and training complex models have resulted in prolonged processing times and resource constraints.

By implementing techniques and steps such as **Data Quality Assessment and Enhancement**, **Interpretability Techniques for Unsupervised Learning Models, Stakeholder Collaboration and Communication, Incremental Model Training and Resource Optimization** and **Simplification of Analytics Models,** the challenges associated with incomplete data, interpretation difficulties, and complexities in implementing advanced analytics techniques were mitigated. This has facilitated in the generation of actionable insights and informed decision-making to achieve the project objectives effectively.

Top of Form

# Conclusions and Recommendations

Customer segmentation into four distinct segments - **"Well off"**, **"Career Focused"**, **"Standard"**, and **"Fewer Opportunities"** provides valuable insights into the diverse needs, preferences, and behaviors of Company X’s customer base. By understanding the characteristics and motivations of each segment, Company X can tailor its marketing, pricing, and product strategies to better serve the unique needs of each group.

1. **Targeted Marketing:** Segment-specific marketing campaigns can be developed to effectively target each customer segment with personalized messaging and offers. This targeted approach enhances customer engagement and loyalty by resonating with the specific interests and preferences of each segment.
2. **Product Customization:** By understanding the distinct needs and preferences of each segment, Company X can customize its product offerings to better meet the demands of different customer groups. This customization can include the introduction of new products, variations in packaging sizes, or the development of exclusive lines tailored to specific segments.
3. **Optimized Pricing Strategies:** Pricing strategies can be optimized based on the willingness to pay and price sensitivity of each segment. For example, premium pricing may be suitable for the "Well off" segment, while discount pricing or value bundles may appeal more to the "Fewer Opportunities" segment. Dynamic pricing algorithms can be implemented to adjust prices in real-time based on demand and segment characteristics.
4. **Improved Customer Experience:** By catering to the unique needs and preferences of each segment, Company X can enhance the overall customer experience and satisfaction. This can lead to increased customer retention, positive word-of-mouth, and higher customer lifetime value.

By implementing these strategies, Company X can leverage customer segmentation to drive business growth, enhance customer satisfaction, and maintain a competitive edge in the retail market.

# References

1. Kotler, P., & Keller, K. L. (2016). Marketing Management (15th ed.). Pearson. [Kotler & Keller, 2016]
2. Wedel, M., & Kamakura, W. A. (2012). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Springer. [Wedel & Kamakura, 2012]
3. Nagle, T. T., Hogan, J. E., & Zale, J. (2016). The Strategy and Tactics of Pricing: A Guide to Growing More Profitably (6th ed.). Routledge. [Nagle et al., 2016]
4. Monroe, K. B., & Cox, J. L. (2005). Pricing: Making Profitable Decisions (3rd ed.). McGraw-Hill/Irwin. [Monroe & Cox, 2005]
5. Dolnicar, S., & Grün, B. (2008). A review of unquestioned standards in using cluster analysis for data-driven market segmentation. The data-driven market segmentation literature, 35(1), 70-78. [Dolnicar & Grün, 2008]
6. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666. [Jain, 2010]
7. Fader, P. S. (2012). Customer Centricity: Focus on the Right Customers for Strategic Advantage. Wharton Digital Press. [Fader, 2012]
8. Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. Journal of Marketing, 80(6), 97-121. [Wedel & Kannan, 2016]
9. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning. [Hair et al., 2019]
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. [Hastie et al., 2009]
11. Raschka, S., & Mirjalili, V. (2019). Python Machine Learning (3rd ed.). Packt Publishing. [Raschka & Mirjalili, 2019]
12. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer. [James et al., 2013]
13. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. ACM Computing Surveys (CSUR), 31(3), 264-323. [Jain et al., 1999]
14. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. [Bishop, 2006]
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. [Pedregosa et al., 2011]
16. McKinney, W. (2018). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media. [McKinney, 2018]
17. Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press. [Kelleher et al., 2015]

# Glossary

Here's a glossary of technical terms used in the project report:

1. **Customer Analytics:** The process of collecting, analyzing, and interpreting customer data to understand customer behavior, preferences, and trends.
2. **Segmentation:** Dividing customers into distinct groups based on shared characteristics, such as demographics, behavior, or purchasing patterns.
3. **Cluster Analysis:** A data analysis technique used to group similar data points into clusters or segments based on their attributes or characteristics.
4. **Dimensionality Reduction:** A process of reducing the number of features or variables in a dataset while preserving its important information and structure.
5. **Descriptive Statistics:** Statistical techniques used to summarize and describe the main features of a dataset, such as mean, median, mode, and standard deviation.
6. **Elasticity Modeling:** Analyzing the sensitivity of customer behavior, such as purchase probability or brand choice, to changes in factors like price or marketing efforts.
7. **Machine Learning:** A branch of artificial intelligence that enables computers to learn from data and make predictions or decisions without being explicitly programmed.
8. **Deep Learning:** A subset of machine learning that involves training artificial neural networks with multiple layers to learn hierarchical representations of data.
9. **Predictive Modeling:** Building statistical models or machine learning algorithms to predict future outcomes or behaviors based on historical data.
10. **K-Means Clustering:** A popular unsupervised machine learning algorithm used for partitioning a dataset into a predetermined number of clusters.
11. **Principal Component Analysis (PCA):** A technique used for dimensionality reduction by transforming the original variables into a new set of orthogonal variables called principal components.
12. **WCSS (Within-Cluster Sum of Squares):** A metric used to evaluate the homogeneity or compactness of clusters in clustering algorithms such as K-means.
13. **Data Preprocessing:** The process of cleaning, transforming, and preparing raw data for analysis by addressing issues such as missing values, outliers, and normalization.
14. **Feature Extraction:** The process of selecting or creating new features from raw data that are most relevant or informative for a particular analysis or prediction task.
15. **Interpretability:** The degree to which the results of a model or analysis can be easily understood and explained by humans.
16. **Model Evaluation:** Assessing the performance and accuracy of predictive models using various metrics and techniques such as cross-validation or confusion matrices.
17. **Data Visualization:** Representing data graphically to visually explore patterns, trends, and relationships within the data.
18. **Stakeholder:** Individuals or groups with an interest or stake in the outcomes of the project, such as executives, customers, or employees.

# Summary of how the feedback for Project Outline have been addressed

The feedback given by the Faculty Examiner for the Project Outline – “GOOD”. This showed that the methods followed, and plan drawn out by me to execute the project was correct. And so, likewise, I have followed the steps and developed the thesis of the project and executed it properly too.

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**Work Integrated Learning Programmes Division (WILPD)**

# Particulars of Student

|  |  |
| --- | --- |
| Student Id | 2022mb21009 |
| Student Name | Koustav Dutta |
| Student E-Mail Address | [2022mb21009@wilp.bits-pilani.ac.in](mailto:2022mb21009@wilp.bits-pilani.ac.in) / [koustavdutta.dgp@gmail.com](mailto:koustavdutta.dgp@gmail.com) |
| Employing Organization and Location | Deloitte Touche Tohmatsu India LLP |
| Programme Name | Mid Semester Project Report |
| Semester | Final Semester |
| Project Title | ANALYTICAL HORIZONS: UNRAVELING CUSTOMER INSIGHTS THROUGH ADVANCED ANALYTICS, MACHINE LEARNING, AND DEEP LEARNING IN MARKETING STRATEGIES |

# Particulars of the Supervisor and Additional Examiner

|  |  |  |
| --- | --- | --- |
|  | **Supervisor** | **Additional Examiner** |
| Name | Mr. Aritra Bhaumik | Mrs. Juhi Khandelwal |
| Qualification | Master’s Degree (Economics), University of Calcutta | B.Tech (IT) ; MBA (Finance), IMT Ghaziabad |
| Designation | Senior Consultant (SA&MA: A&C, Deloitte India) | Manager (SA&MA: A&C, Deloitte India) |
| Employing Organization and Location | Deloitte Touche Tohmatsu India LLP, Bengaluru | Ministry of home affairs, Govt of India |
| Phone No (with STD code) | +91 9739018000 | +91 8073181664 |
| Email Address | [abhaumik@deloitte.com](mailto:abhaumik@deloitte.com) | [jkhandelwal@deloitte.com](mailto:jkhandelwal@deloitte.com) |

**Remarks of the Supervisor on Mid-Semester Project Report**

I have thoroughly reviewed the Mid Semester Project Report and am pleased with the comprehensive coverage of Customer Analytics domain. The well-structured representation and thoughtful consideration, as well as research of potential algorithms and processes, indicate a robust approach. It's evident that significant effort has been invested in understanding the complexities of customer segmentation and the application of advanced analytics techniques.

I particularly appreciate the clarity and coherence in presenting the methodology and findings. Your ability to effectively communicate complex concepts and analytical methodologies is impressive and contributes to the overall professionalism of the report.

If you have any further questions or need clarification, feel free to reach out.

Thank you in advance for your diligence and commitment to the project. Regards.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Signature of Student** | **Signature of Supervisor** | **Signature of Additional Examiner** |
| **Name: Mr. Koustav Dutta** | **Name: Mr. Aritra Bhaumik** | **Name: Mrs. Juhi Khandelwal** |