**Architecture Design**

**Clustering for Customer Segmentation & Understanding.**

|  |  |
| --- | --- |
| Written by | Parekh Kathan Chiragkumar |
| Version | 1.0 |
| Date | 05-06-2023 |

**Document Control**

### **Change Record**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | 07-11-2021 | Keshava Prasaath.K |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

### **Approval Status:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

**Index**

|  |  |
| --- | --- |
| **Content** | **Page No** |
| Abstract | 4 |
| 1. Introduction | 4 |
| 1.1 What is Architecture Design? | 4 |
| 1.2 Scope | 4 |
| 1.3 Constraints | 4 |
| 2. Technical Specification | 5 |
| 2.1 Dataset | 5 |
| 2.2 Logging | 7 |
| 2.3 New Feature Generation | 7 |
| 2.4 Deployment | 7 |
| 3. Technology Stack | 8 |
| 4. Proposed Solution | 8 |
| 5. Architecture Description | 8 |
| 6. User Input/Output Workflow | 10 |

**Abstract**

This study proposes a machine learning-based approach for segmenting mall customers. The objective is to group customers into distinct segments based on demographic information, purchasing patterns, and visitation behavior. A comprehensive dataset comprising customer attributes and transactional data is collected. Various machine learning algorithms, including clustering and classification techniques, are employed to analyze the dataset and identify meaningful customer segments. The results reveal the presence of distinct segments with unique traits, such as spending patterns and visitation frequency. The findings provide valuable insights for mall owners and retailers to tailor marketing campaigns and enhance the overall customer experience. Additionally, the study explores the potential of incorporating additional data sources, such as social media and mobile app usage, to improve segmentation accuracy. The application of machine learning in mall customer segmentation enables data-driven decision-making and personalized experiences in the retail industry.

**1. Introduction**

**1.1 What is Architecture Design?**

The goal of Architecture Design (AD) or a low-level design document is to give the internal design of the actual program code for the `Bike Share Prediction System`. AD describes the class diagrams with the methods and relation between classes and program specification. It describes the modules so that the programmer can directly code the program from the document.

**1.2 Scope**

Architecture Design(AD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. And the complete workflow.

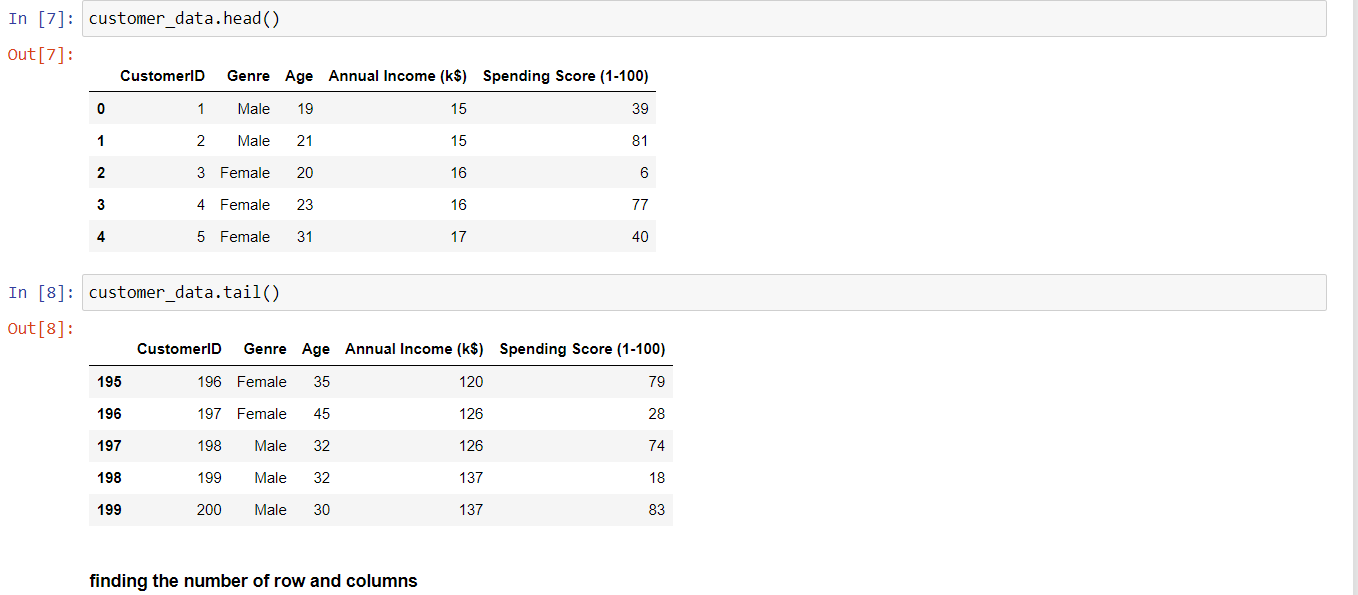
**1.3 Constraints**

We only predict the expected casual and registered customers based on the weather condition and date information.

**2. Technical Specification**

**2.1 Dataset**

This data set is created only for the learning purpose of the customer segmentation concepts , also known as market basket analysis . I will demonstrate this by using unsupervised ML technique (KMeans Clustering Algorithm) in the simplest form.



**2.2 Logging**

We should be able to log every activity done by the user

* The system identifies at which step logging require.
* The system should be able to log each and every system flow.
* Developers can choose logging methods. Also can choose database logging.
* The system should be not be hung even after using so much logging. Logging just because we can easily debug issuing so logging is mandatory to do.

**2.3 DataBase**

The system needs to store every request into the database and we need to store it in such a way that it is easy to retain and look into the records.

The system should capture every data that any user gave and the prediction that has been made by that input.

**3. Technology Stack**

|  |  |
| --- | --- |
| Backend | Python,Sklearn,Seaborn,Matplotlib |
| Deployment | **joblib** |

**4. Proposed Solution**

We will use performed EDA to find the important relation between different attributes and will use a machine-learning algorithm to predict the future sales demand. The client will be filled the required feature as input and will get results through the web application. The system will get features and it will be passed into the backend where the features will be validated and preprocessed and then it will be passed to a hyperparameter tuned machine learning model to predict the final outcome.

**5** **Architecture detail**



**5.1Data Gathering**

Data source: <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>

Train and Test data are stored in .csv format.

**5.2 Raw Data Validation**

Raw data validation is a crucial step in customer segmentation using machine learning. It involves assessing the quality, accuracy, completeness, and consistency of the dataset. This validation ensures that the data is reliable, representative, and free from errors or inconsistencies. Key considerations include data quality, accuracy, completeness, consistency, formatting, privacy, security, and sample size. By conducting thorough validation, potential issues can be identified and addressed, leading to improved reliability and accuracy in the customer segmentation analysis.

**5.3 Model Building**

**Model Building of K-means Clustering:**

**K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct clusters. The process of building a K-means clustering model involves the following steps:**

1. **Data Preprocessing: Prepare the dataset by performing necessary preprocessing steps, such as removing outliers, scaling variables, and handling missing values. Ensure that the data is in a suitable format for clustering analysis.**
2. **Determining the Number of Clusters (K): Select an appropriate value for the number of clusters, K, based on domain knowledge or using techniques such as the elbow method or silhouette analysis. This determines the desired level of granularity in the clustering results.**
3. **Initialization: Initialize K cluster centroids randomly or using a specific method**

**5.9 Model Saving**

Using Joblib

**5.10 GitHub**

The whole project directory will be pushed into the GitHub repository.

**6. User Input / Output Workflow.**

****