High Level Design Document

Introduction

This High Level Design (HLD) document outlines the architecture and core components for **ClusterVision - Customer Segmentation Visualizer**. The project is a Streamlit-based tool enabling users to perform K-Means clustering on retail customer data, visualize clusters, and evaluate clustering quality using Elbow and Silhouette methods. The tool provides interactive visualizations and exportable results.

1. System Architecture Overview

Architecture Summary:
ClusterVision is a modular, client-server web application built with Streamlit. It processes user-uploaded data, applies unsupervised clustering, visualizes results, and allows data export.

Module	Role
User Interface (UI)	Collects user input, displays visualizations, manages interactions
Data Processing	Loads, validates, and preprocesses customer data
Clustering Engine	Performs K-Means clustering, computes Elbow & Silhouette metrics
Visualization	Generates interactive cluster plots and evaluation charts
Export Module	Enables export of clustered data and visualizations

2. Component Interactions

Sequence Step	Interaction Description
1. User uploads data	UI → Data Processing: Data is uploaded and validated
2. Data preprocessing	Data Processing → Clustering Engine: Cleaned data is passed for clustering
3. Clustering & evaluation	Clustering Engine: Runs K-Means, computes metrics
4. Visualization	Clustering Engine → Visualization: Results sent for plotting
5. User interaction	UI ↔ Visualization: User explores clusters, adjusts parameters
6. Export results	UI → Export Module: User exports clustered data/plots

3. Data Flow Overview

Data Source/Target	Data Type	Flow Description
User → UI	CSV/Excel data	User uploads customer dataset

UI → Data Processing	Raw data	Data is validated and preprocessed
Data Processing → Clustering Engine	Cleaned data	Data used for clustering and evaluation
Clustering Engine → Visualization	Cluster labels, metrics	Data for plots and evaluation charts
Visualization → UI	Plots, tables	Interactive display to user
UI → Export Module	Clustered data/plots	User exports results

4. Technology Stack

Layer/Function	**Technology/Framework**
Web UI	Streamlit
Data Processing	Pandas, NumPy
Machine Learning	scikit-learn (KMeans, metrics)
Visualization	Streamlit, Matplotlib/Plotly
Export	Pandas (to CSV/Excel), Streamlit download
Language	Python 3.x

5. Scalability & Reliability

- **Scalability:** Designed for moderate datasets; can be containerized for deployment. For large-scale or concurrent users, deploy behind a WSGI server or scale horizontally.
- **Reliability:** Input validation and error handling ensure robust operation. Stateless design allows easy recovery and redeployment.
- **Security:** User data is processed in-memory and not persisted; recommend secure deployment practices for sensitive data.

End of Document