Text

Description automatically generated

**Activity based**

**Project Report on**

**DWDM**

**Submitted to Vishwakarma University, Pune**

**Under the Initiative of**

**By**

**Arun Adhikari**

**SRN No : 202100406**

**Roll No : 38**

**Div : G**

**Third Year Engineering**

**Department of Computer Engineering**

**Faculty of Science and Technology**

**Academic Year**

**2023-2024**

**Problem Statement**

**Implement a k-means Clustering Algorithm.**

**Business Requirements**

**Stakeholders: Executive leadership, customers,**

**Data Sources: Customer database, content database**

import pandas as pd

from sklearn.preprocessing import StandardScaler

from scipy.cluster.hierarchy import linkage, dendrogram

import matplotlib.pyplot as plt

# Load and preprocess data (replace 'file\_path' with the actual file path)

file\_path = 'Online\_Retail.xlsx'

data = pd.read\_excel(file\_path, usecols=["InvoiceNo", "CustomerID", "Quantity", "UnitPrice", "InvoiceDate"])

data = data.dropna(subset=['CustomerID'])

# Calculate 'TotalSpent' for each transaction

data['TotalSpent'] = data['Quantity'] \* data['UnitPrice']

# Aggregate per CustomerID to create features for clustering

customer\_data = data.groupby('CustomerID').agg({

'TotalSpent': 'sum',

'InvoiceNo': 'nunique'

}).rename(columns={'InvoiceNo': 'PurchaseFrequency'})

# Standardize the features

scaler = StandardScaler()

customer\_data\_scaled = scaler.fit\_transform(customer\_data)

# Apply Hierarchical Clustering

linked = linkage(customer\_data\_scaled, method='ward') # 'ward' minimizes variance within clusters

# Plot the dendrogram

plt.figure(figsize=(15, 10))

dendrogram(linked, orientation='top', distance\_sort='ascending', show\_leaf\_counts=False)

plt.title('Dendrogram for Hierarchical Clustering')

plt.xlabel('Customer Index')

plt.ylabel('Distance')

plt.show()

