Title: Exploratory Data Analysis and Clustering Report

Student Name: Ellanki Sneha

Student Number: 23029784

GitHub Repository: [https://github.com/Ellankisneha/Clustering-And-Fitting.git]

**Introduction**

This report presents an exploratory data analysis (EDA) and clustering analysis performed on the Walmart sales dataset of 45 stores. The dataset contains information about weekly sales, holiday events, and other relevant factors across multiple stores. The primary objectives of this analysis are:

**Explore the distribution and relationships between variables in the dataset.**

Perform clustering analysis to identify any inherent groupings within the data.

**Dataset Description**

The dataset used in this analysis is sourced from Kaggle and contains weekly sales data for 45 Walmart stores. It includes information such as store number, date, weekly sales, holiday events, and more. A brief overview of the features is as follows:

Store: Store number

Date: Date of sales

Weekly\_Sales: Weekly sales in the store

Holiday\_Flag: Indicates if the week is a holiday week (1) or not (0)

Temperature: Temperature on the given date

Fuel\_Price: Fuel price on the given date

CPI: Consumer Price Index on the given date

Unemployment: Unemployment rate on the given date

**Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, aimed at understanding the underlying patterns and characteristics of a dataset. As part of this analysis, a histogram of the weekly sales distribution was created for the given dataset. The dataset contains information related to weekly sales, and the histogram provides insights into the distribution of sales amounts across the dataset. The histogram visually represents the frequency of different weekly sales amounts, allowing us to observe the central tendency, dispersion, and skewness of the sales data. By dividing the sales amounts into intervals (or bins) and plotting the frequency of observations within each interval, the histogram illustrates the overall pattern of sales distribution. Additionally, it helps identify any outliers or unusual patterns in the sales data.

**Histogram of Weekly Sales Distribution**

**A screenshot of a computer

Description automatically generatedA graph of a number of weeks with a number of holidays

Description automatically generatedA graph of a sales distribution

Description automatically generated with medium confidenceTo understand the distribution of weekly sales, a histogram is plotted:**

**A screenshot of a computer

Description automatically generatedScatter Plot of Temperature vs. Weekly Sales**

**A screenshot of a computer

Description automatically generatedA graph showing the temperature and the temperature

Description automatically generated with medium confidenceTo visualize the relationship between temperature and weekly sales, a scatter plot is generated:**

**Confusion Matrix:** It seems that the error is caused because the dataset contains non-numeric data, specifically in the 'Date' column. Correlation analysis requires numerical data, so we need to exclude non-numeric columns or convert them to numeric format. We'll exclude the 'Date' column from the correlation analysis for now. Here's the modified code

**A screenshot of a graph

Description automatically generatedA screenshot of a computer

Description automatically generated**

**A blue and black diagram

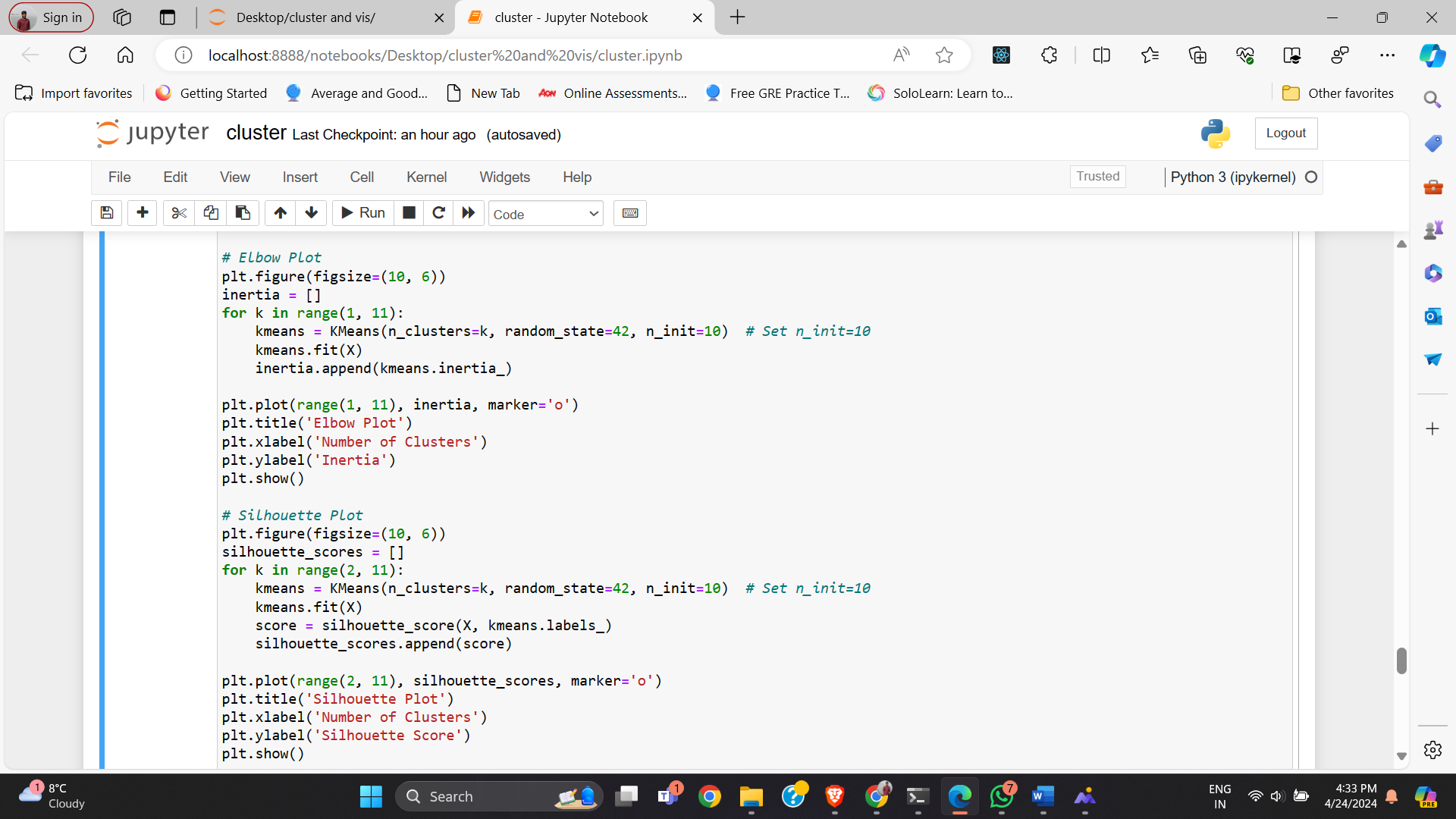
Description automatically generated with medium confidenceA blue and red squares with white text

Description automatically generated**

Clustering Analysis

A graph with a line

Description automatically generatedA graph with a line

Description automatically generatedElbow Clustering and Silhouette Plot