



NED UNIVERSITY OF ENGINEERING & TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & IT

CT-361: Artificial Intelligence & Expert System

CCP Report

Instructor: Dr. Waseemullah Nazir

Made By:

Laiba Shahid

CT-22061

Tooba Kashaf

CT-22068

Haiqa Siddiqua

CT-22071

Muhammad Zohaib Khan

CT-22072

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
BACHELORS OF SCIENCE IN COMPUTER SCIENCE

Complex Computing Problem Assessment Rubrics

Course Code: CT-361		Course Title: Artificial Intelligence & Expert System	
Criteria and Scales			
Excellent (3)	Good (2)	Average (1)	Poor (0)
Criterion 1: Understanding the Problem: How well the problem statement is understood by the student			
Understand the problem clearly and identify the underlying issues and functionalities.	Adequately understands the problem and identifies the underlying issues and functionalities.	Inadequately defines the problem and identifies the underlying issues and functionalities.	Fails to define the problem adequately and does not identify the underlying issues and functionalities.
Criterion 2: Research: The amount of research that is used in solving the problem			
Contains all the information needed for solving the problem	Good research leads to a successful solution	Mediocre research which may or may not lead to an adequate solution	No apparent research
Criterion 3: Code: How complete the code is along with the assumptions and selected functionalities			
Complete the code according to the selected functionalities of the given case with clear assumptions	Incomplete code according to the selected functionalities of the given case with clear assumptions	Incomplete code according to the selected functionalities of the given case with unclear assumptions	Wrong code and naming conventions
Criterion 4: Report: How thorough and well-organized is the solution			
All the necessary information is organized for easy use insolving the problem	Good information organized well could lead to a good solution	Mediocre information which may or may not lead to a solution	No report provided

Total Marks: _____

Teacher's Signature: _____

AI-Powered Data Exploration Platform

AI Assistant for Data Science

Overview

This **AI Assistant for Data Science** is a **Streamlit-powered** tool that helps users analyze, clean, and visualize datasets interactively. It includes voice-enabled queries alongside text input. It leverages **LangChain with Groq's LLM** (Gemma 2-9B) to provide AI-driven insights, automatic chart generation, and data cleaning capabilities and a chat interface.

Features & How to Use Them

1 Data Upload & Initial Setup

- **Supported Formats:** CSV, Excel (XLSX)
- **How to Use:**
 - Click "**Let's get started**"
 - Upload your dataset using the file uploader
 - The AI automatically analyzes the data and provides a summary

◆ **What It Does:**

- Displays the first few rows of data
- Explains column meanings
- Lists column data types
- Checks for missing values & duplicates
- Provides statistical summaries (mean, median, etc.)
- Identifies correlations, outliers, anomalies, and patterns

Exploratory Data Analysis

General Information about the Dataset

Data Overview

The first rows of your dataset look like this:

	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-11-07	45.1	50.09	44	44.9	44.9	117701670
1	2013-11-08	45.93	46.94	40.685	41.65	41.65	27925307
2	2013-11-11	40.5	43	39.4	42.9	42.9	16113941
3	2013-11-12	43.66	43.78	41.83	41.9	41.9	6316755
4	2013-11-13	41.03	42.87	40.76	42.6	42.6	8688325

The dataframe has the following columns:

- **Date:** The date of the stock's trading activity.
- **Open:** The opening price of the stock on that day.
- **High:** The highest price the stock reached during the trading day.
- **Low:** The lowest price the stock reached during the trading day.
- **Close:** The closing price of the stock on that day.
- **Adj Close:** The adjusted closing price, which takes into account stock splits, dividends, and other

- **Adj Close:** The adjusted closing price, which takes into account stock splits, dividends, and other corporate actions.
- **Volume:** The number of shares traded on that day.

Column Types

Column Type 0 Date object 1 Open float64 2 High float64 3 Low float64 4 Close float64 5 Adj Close float64 6 Volume float64

There are 30 missing values in this dataframe.

No, there are no duplicate values in the dataframe.

Download as CSV

Data Summarisation

	Open	High	Low	Close	Adj Close	Volume
count	2259	2259	2259	2259	2259	2259
mean	36.0203	36.6999	35.3395	36.0036	36.0036	21751860.6516
std	14.1185	14.3721	13.8287	14.09	14.09	19099883.6264
min	13.95	14.22	13.725	14.01	14.01	0
25%	25.55	26.215	24.9125	25.41	25.41	12335301.5
50%	35.42	36.1	34.82	35.49	35.49	16913051
75%	44.205	45.015	43.3275	44.135	44.135	24280821.5
max	78.36	80.75	76.05	77.63	77.63	269213085

The correlation matrix shows strong positive relationships between all numerical variables except for 'Volume', which has weak negative correlations with 'Open', 'High', 'Low', 'Close', and 'Adj Close'.

The data points from 2021-02-12 to 2021-07-29 in the 'Close' column appear to be outliers.

Quick Fixes for you!

The 'Date' column should be converted to datetime objects using `pd.to_datetime(df['Date'])`.

Anomalies Found!

The dataset has potential anomalies in the 'Volume' column, with several data points significantly exceeding the typical volume.

Patterns Found!

The dataset shows a general upward trend in stock prices over time. There are some fluctuations and periods of consolidation, but the overall direction is positive.

Clean Data

2 Data Cleaning

◆ **What It Does:**

- Removes duplicate rows
- Fills missing values using **forward-fill (ffill)** and **backward-fill (bfill)**
- Updates the dataset and **recomputes all analyses**

How to Use:

- Click the **"Clean Data"** button
- The AI reprocesses the data and shows updated statistics

Clean Data

Data cleaned successfully!

Cleaned Dataset Summary

Data Overview

The first rows of your dataset look like this:

index--p5bJXXpQgvPz6yvQMFiY	Date	Open	High	Low	Close	Adj Close	Volume
0	2013-11-07 00:00:00	45.1	50.09	44	44.9	44.9	117701670
1	2013-11-08 00:00:00	45.93	46.94	40.685	41.65	41.65	27925307
2	2013-11-11 00:00:00	40.5	43	39.4	42.9	42.9	16113941
3	2013-11-12 00:00:00	43.66	43.78	41.83	41.9	41.9	6316755
4	2013-11-13 00:00:00	41.03	42.87	40.76	42.6	42.6	8688325

The dataframe has the following columns:

- **Date:** The date of the stock data.
- **Open:** The opening price of the stock on that date.
- **High:** The highest price the stock reached on that date.
- **Low:** The lowest price the stock reached on that date.

- **Close:** The closing price of the stock on that date.
- **Adj Close:** The adjusted closing price, which takes into account stock splits and dividends.
- **Volume:** The number of shares traded on that date.

Column Types

Column	Type
Date	datetime64[ns]
Open	float64
High	float64
Low	float64
Close	float64
Adj Close	float64
Volume	float64

There are 0 missing values.

There are no duplicate values in the dataframe.

Data Summarisation

	Date	Open	High	Low	Close	Adj Close	Volume
count	2264	2264	2264	2264	2264	2264	2264
mean	2018-05-08 11:16:44.946996480	36.0598	36.7381	35.38	36.0427	36.0427	22004937.6555
min	2013-11-07 00:00:00	13.95	14.22	13.725	14.01	14.01	0
25%	2016-02-08 18:00:00	25.575	26.245	24.9263	25.4925	25.4925	12354670.75
50%	2018-05-08 12:00:00	35.45	36.1338	34.85	35.53	35.53	16972044.5
75%	2020-08-06 06:00:00	44.26	45.095	43.4224	44.1625	44.1625	24301233
max	2022-11-03 00:00:00	78.36	80.75	76.05	77.63	77.63	269213085
std	None	14.1279	14.3791	13.8403	14.0989	14.0989	19822946.1053

The numerical variables in the dataframe show strong positive correlations between 'Open', 'High', 'Low', 'Close', and 'Adj Close'. 'Date' has a moderate positive correlation, while 'Volume' has very low correlations with the other variables.

There are likely outliers in the 'Close' column of the dataframe.

Quick Fixes for you!

The dataframe appears to be in good shape with appropriate data types.

Anomalies Found!

The standard deviation being NaN is likely due to an unusual distribution of the data, causing a potential issue with the standard deviation calculation.

Patterns Found!

The dataset shows a general upward trend in stock prices, with fluctuations within a relatively consistent range.

3 Smart Column Analysis

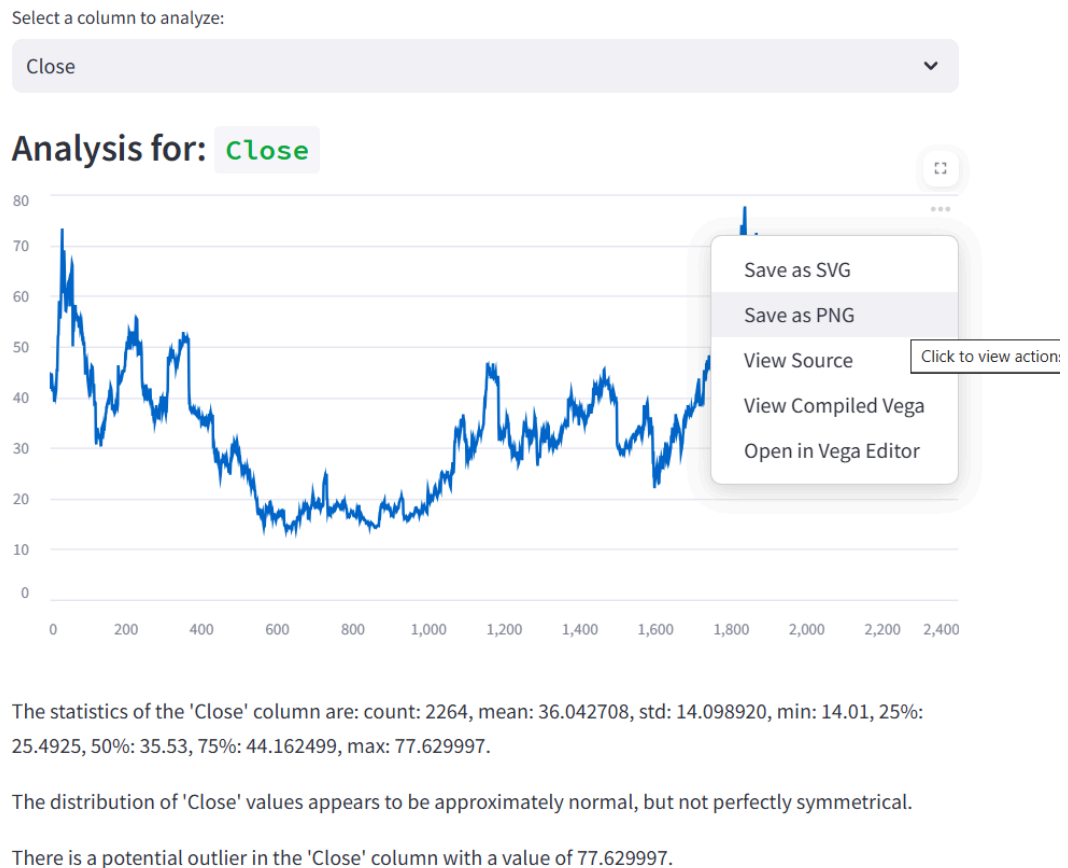
◆ What It Does:

- **Automatically suggests the best chart type** (bar, line, pie, histogram)
- Provides:
 - Statistical summary
 - Distribution analysis
 - Outlier detection
 - Trend/seasonality insights
 - Missing value assessment
 - Chart can be downloaded as options shown in the image.

- When the cursor is on the chart, more details are visible.

How to Use:

- Select a column from the dropdown
- The AI generates visualizations and insights



4 Variable Comparison Tool

◆ What It Does:

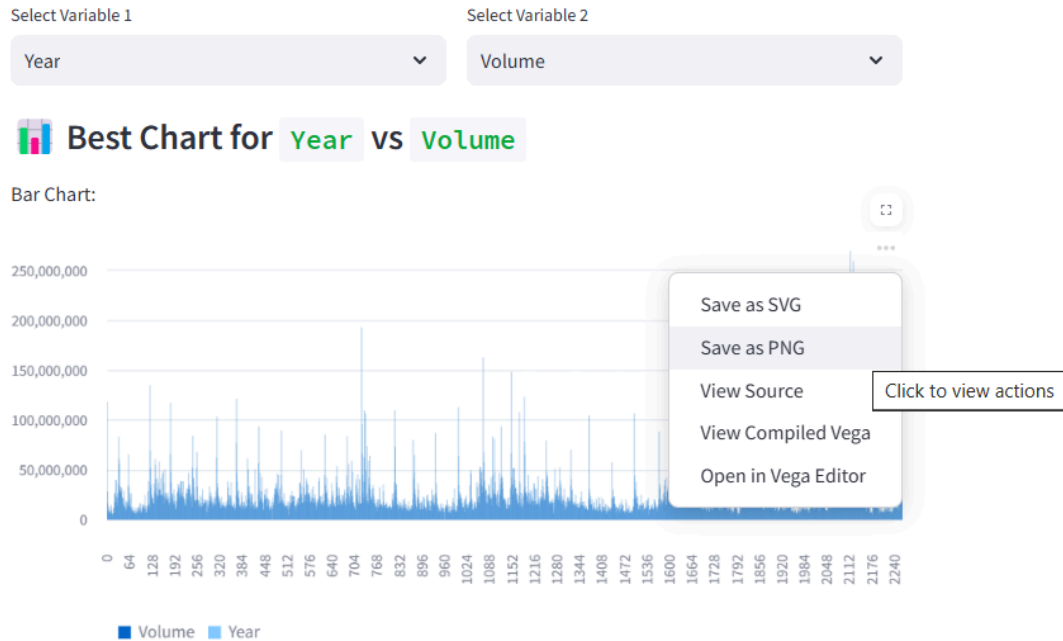
- **Automatically picks the best chart** (scatter, line, bar, histogram) for comparing two variables
- Provides **individual insights** for each variable
- Chart can be downloaded as options shown in the image.
- When the cursor is on the chart, more details are visible.

How to Use:

- Select **Variable 1** and **Variable 2** from dropdowns

- The AI generates side-by-side comparisons

Engages the user: Would you like to compare Variables Side-by-Side?



LLM Insights

Insights on Year

The dataframe contains data from multiple years, with 2020, 2014, 2015, 2016, 2019, and 2021 being the most represented years. 2013 has considerably less data.

Insights on Volume

The volume of trading is quite high, with an average of around 22 million shares traded per day. The minimum volume is 0, and the maximum volume is 269 million shares. The median volume is 16.97 million shares, indicating that half the days have a

5 AI Chat Interface (With Voice Input)



◆ Input Text Feature

How to Use Text Input:

1. Continue typing questions in the text box

Ask Anything About Your Data

 Welcome to the AI Assistant ChatBox!

Got burning questions about your data? You're in the right place! Just type in your queries, and let's unravel the mysteries of your data together!  

You: which year has the highest volume of sales?

AI: 2020

You: Share insights on high and low columns of the dataset?


AI: The dataset shows a wide price range, with a maximum high of 80.75 and a minimum low of 13.725. This suggests substantial price volatility during the observed period.

Ask a question about the data:


what period are you talking about?

AI: The period is from November 7th, 2013 to November 13th, 2013.

◆ Voice Features:


-  **Speak your questions** instead of typing
- **Dual input mode:** Switch between voice and text seamlessly
- **Automatic transcription:** Converts speech to text using Google's API

How to Use Voice Input:

1. Click the  **microphone button**
2. Allow browser microphone access when prompted
3. Speak clearly (5-10 second limit)
4. View automatically transcribed text
5. Get AI response as usual

Ask Anything About Your Data

 Welcome to the AI Assistant ChatBox!

Got burning questions about your data? You're in the right place! Speak or type your queries below! 



Speak your question:



 You said: what insights can you give about the high and low columns of the data set

AI: The maximum price reached was 80.75, and the minimum price was 13.725. The difference between the 'High' and 'Low' columns for each day shows the price range, with larger differences indicating greater price volatility.

Or type your question:

Ask a question about the data:

Technical Details

- **Backend:**
 - **LangChain Agents** (Pandas DataFrame Agent)
 - **Groq API** (Gemma 2-9B model)
- **Caching:**
 - **Session-state caching** for repeated analyses
 - **Forces recomputation after cleaning**
- **Error Handling:**
 - Gracefully handles missing data, unsupported operations

Ideal Use Cases

- ✓ **Exploratory Data Analysis (EDA)**
- ✓ **Quick data cleaning & preprocessing**
- ✓ **Automated visualization generation**

- ✓ Natural language queries on datasets
 - ✓ Educational tool for learning data science
-



Notes

- **Large datasets may take longer** for AI processing.
- **For best results**, ensure column names are descriptive.
- **The AI may occasionally misinterpret queries**—try rephrasing if needed.