Lesson 02 Demo 15

Data Integrity Using GenAI

Objective: To ensure the quality of the data using GenAI

Tools required: Julius AI

Prerequisites: None

Steps to be followed:

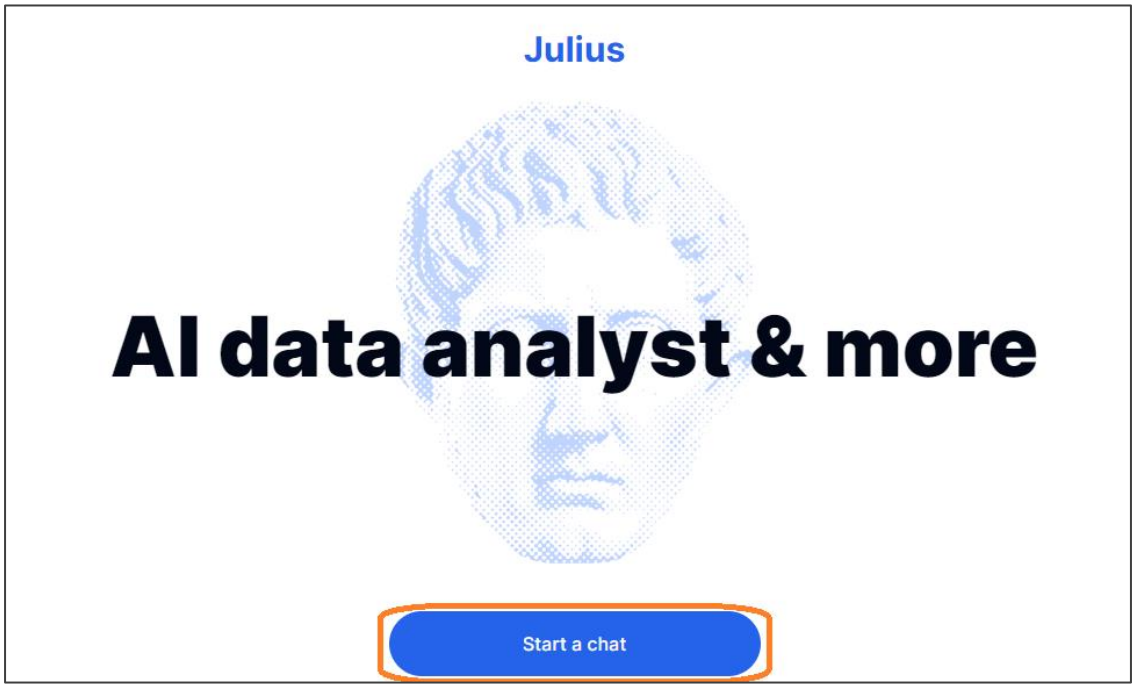
1. Download the dataset
2. Upload the dataset to Julius AI
3. Check the data integrity using a prompt

Step 1: Download the Dataset

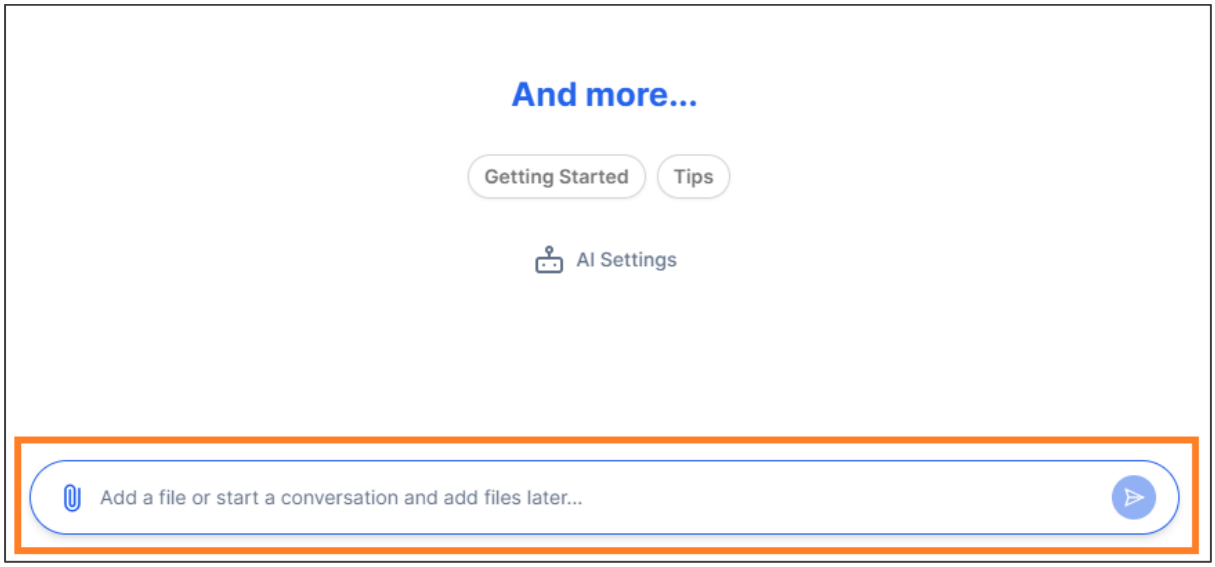
1.1 Download the Customer.csv from the LMS under the **Reference Materials** section

Step 2: Upload the dataset to Julius AI

2.1 Navigate to Julius AI
2.2 Click on Start a chat on the homepage

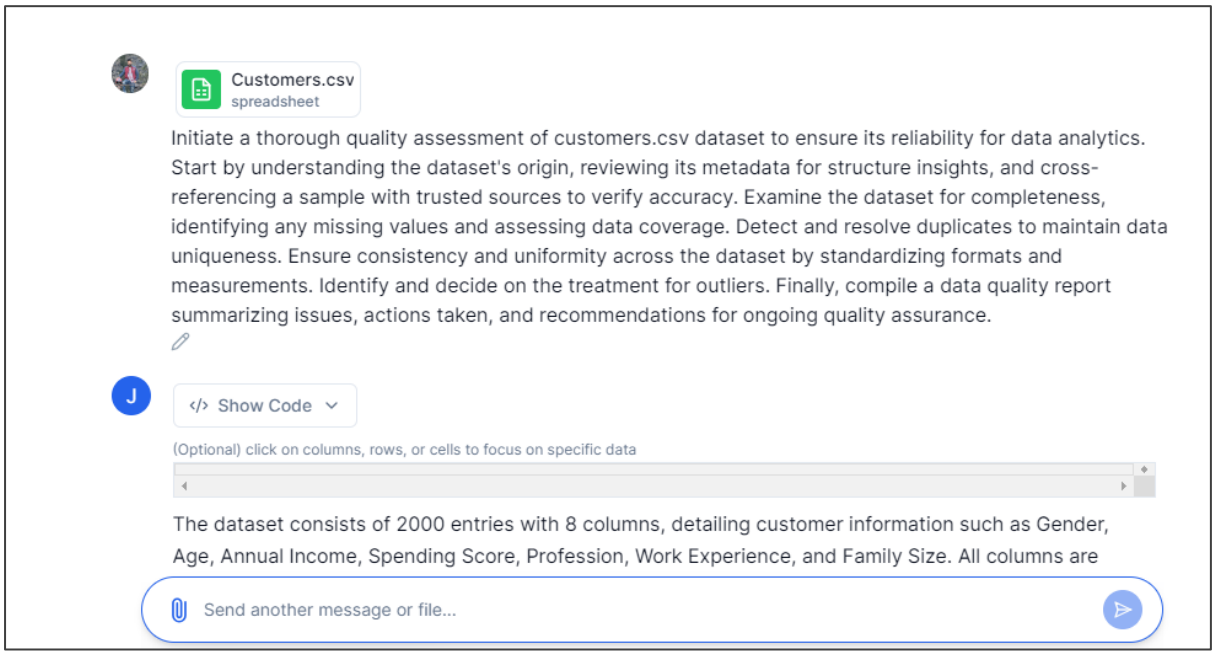


2.3 Click on **Add a file**... to upload the dataset



Step 3: Check the data integrity using a prompt

3.1 Create a prompt to ensure data integrity on the sample dataset uploaded to Julius AI



3.2 Check for data consistency, duplicates, and anomalies and identify missing values

(Optional) click on columns, rows, or cells to focus on specific data

The dataset consists of 2000 entries with 8 columns, detailing customer information such as Gender, Age, Annual Income, Spending Score, Profession, Work Experience, and Family Size. All columns are complete with no missing values, except for the Profession column, which has 35 missing entries. The data types are appropriate for the information represented, with integers for numerical values and objects (strings) for categorical data.

Next steps in the quality assessment include:

- Verifying the accuracy of a sample of data against trusted sources.
- Identifying and resolving any duplicate entries.
- Standardizing formats and measurements for consistency.
- Assessing and deciding on the treatment for outliers.
- Compiling a data quality report with findings and recommendations.

3.3 Check the accuracy of the data

Let's proceed with verifying the accuracy of a sample of data.

</> Show Code

0

	0
CustomerID	2000
Gender	2
Age	100
Annual Income ($)	1786
Spending Score (1-100)	101

The dataset does not contain any duplicate entries, ensuring uniqueness across all records. The CustomerID column has 2000 unique values, matching the total number of entries, which confirms that each customer has a unique identifier. The dataset includes a diverse range of ages, with 100 unique values, and a wide range of annual incomes, with 1786 unique values. The spending score, which ranges from 1 to 100, has 101 unique values, indicating a possible data entry error since the score should only range from 1 to 100.

3.4 Standardize formats and measurements for consistency and assess the dataset for outliers

No outliers were detected in the 'Age', 'Annual Income ($)', or 'Spending Score (1-100)' columns, indicating that the data points in these columns are within a reasonable range.

Next, we can proceed with compiling a data quality report summarizing the findings and recommendations based on the analysis conducted.

Show Code

{'Total Entries': 2000,
'Unique Customer IDs': 2000,
'Unique Genders': 2,
'Unique Ages': 100,
'Unique Annual Incomes': 1786,
'Unique Spending Scores': 101,
'Missing Professions': 35,
'Outliers in Age': 0,
'Outliers in Annual Income ($)': 0,
'Outliers in Spending Score (1-100)': 0}

3.5 Create a quality report summarizing the findings and recommendations based on the analysis conducted

Based on the analysis conducted, the data quality report highlights the following key points:

- The dataset contains a total of 2000 entries, each with a unique Customer ID.
- There are 2 unique genders represented in the dataset.
- The age of customers spans 100 unique values, indicating a wide range of customer ages.
- The annual income of customers is represented by 1786 unique values, showcasing a diverse economic background among the customers.
- The spending score, which should ideally range from 1 to 100, has 101 unique values, suggesting a potential data entry error.
- There are 35 missing entries in the 'Profession' column, which need to be addressed.
- No outliers were detected in the 'Age', 'Annual Income ($)', or 'Spending Score (1-100)' columns, indicating that the data within these columns falls within a reasonable range.

Recommendations for further action include investigating the cause of the extra unique value in the 'Spending Score (1-100)' column and addressing the missing values in the 'Profession' column.