***“*Heaven’s Light is Our Guide”**

A logo with text and symbols

AI-generated content may be incorrect.

**Rajshahi University of Engineering & Technology**

**Department of Computer Science & Engineering**

**Lab Report-2**

**Course Code: CSE 4120**

**Course Title: Data Mining Sessional**

|  |  |
| --- | --- |
| **Submitted By-**  **Name: Sajidur Rahman Tarafder**  **Department: CSE**  **Roll No: 2003154**  **Section: C**  **Session:2020-21** | **Submitted To-**  **Julia Rahman**  **Associate Professor**  **Department of CSE**  **RUET** |

**Experiment Name:** **Exploring** **Data Preprocessing Techniques**

**Objectives:**

Applying preprocessing techniques to raw data including handling missing values, normalization and implementing similarity measures manually without built-in functions.

**Dataset:**

**Titanic Dataset -** Contains passenger information with 891 records and 12 attributes including Age, Sex, Pclass, Fare, etc.

**Task 1: Load Dataset and Handle Missing Values**

**Implementation:**

- Removed Cabin column (As 77% of the cabin data are missing)

- Used grouped median approach for numerical data (specifically

Age) as Age varied significantly across passenger classes and

gender in the dataset. The grouped median approach

recognizes that first-class passengers tend to be older while

third-class passengers are typically younger. Gender also plays

a role in age distribution patterns.

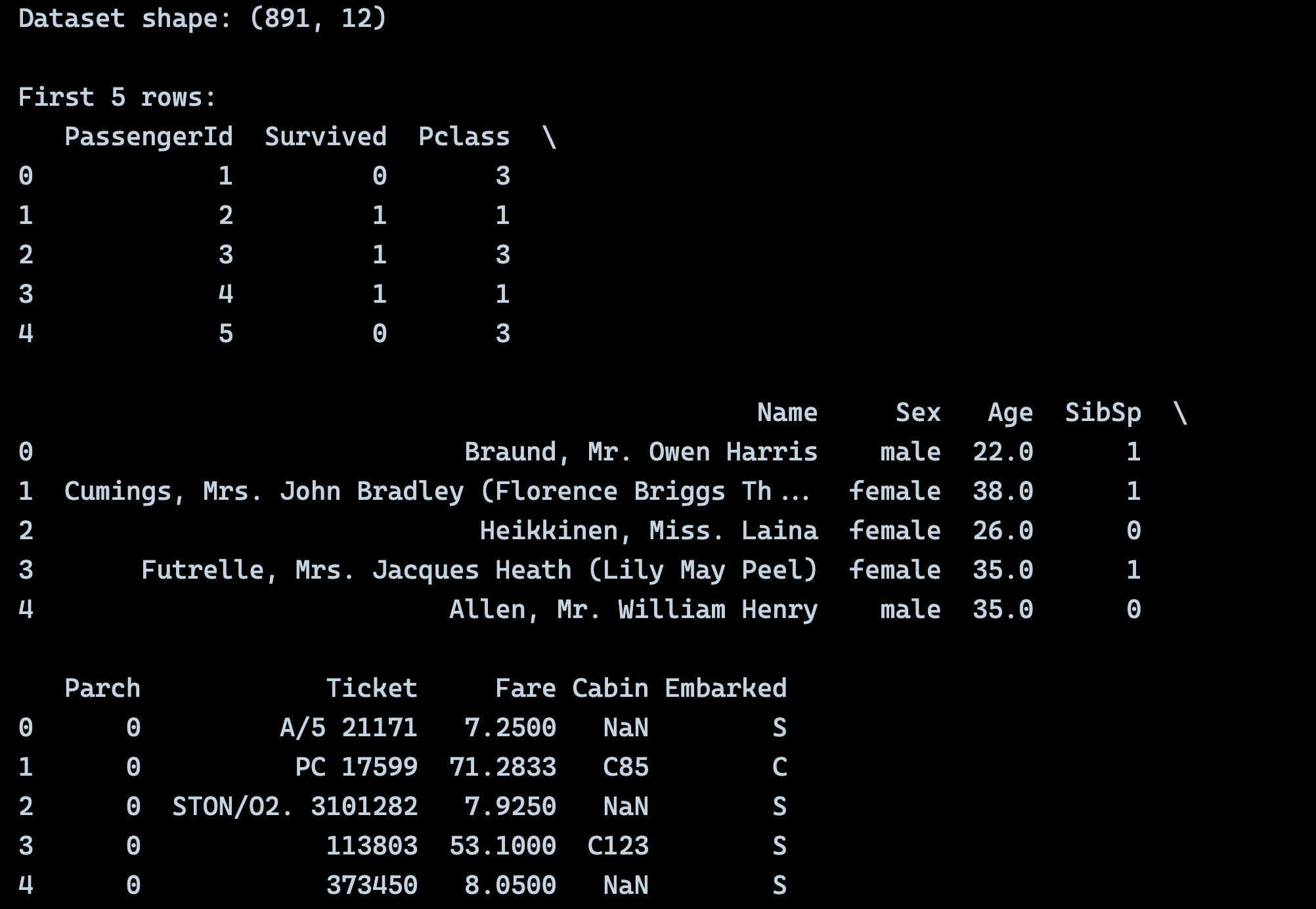
- Applied mode imputation for categorical variables (Embarked)

**Code (Loading Dataset):**

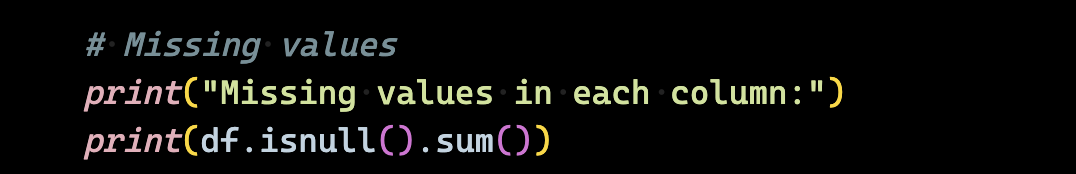
**A computer screen with text on it

AI-generated content may be incorrect.**

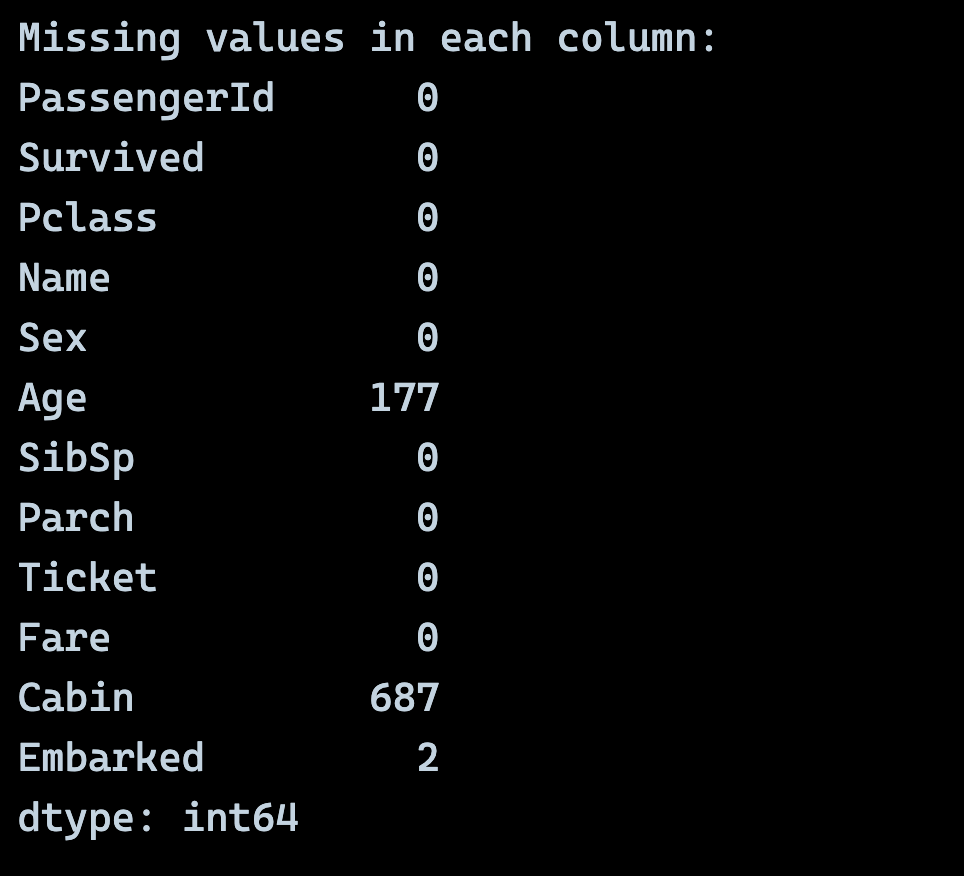
**Output:**

****

**Code (Missing Values Stats):**

****

**Output:**

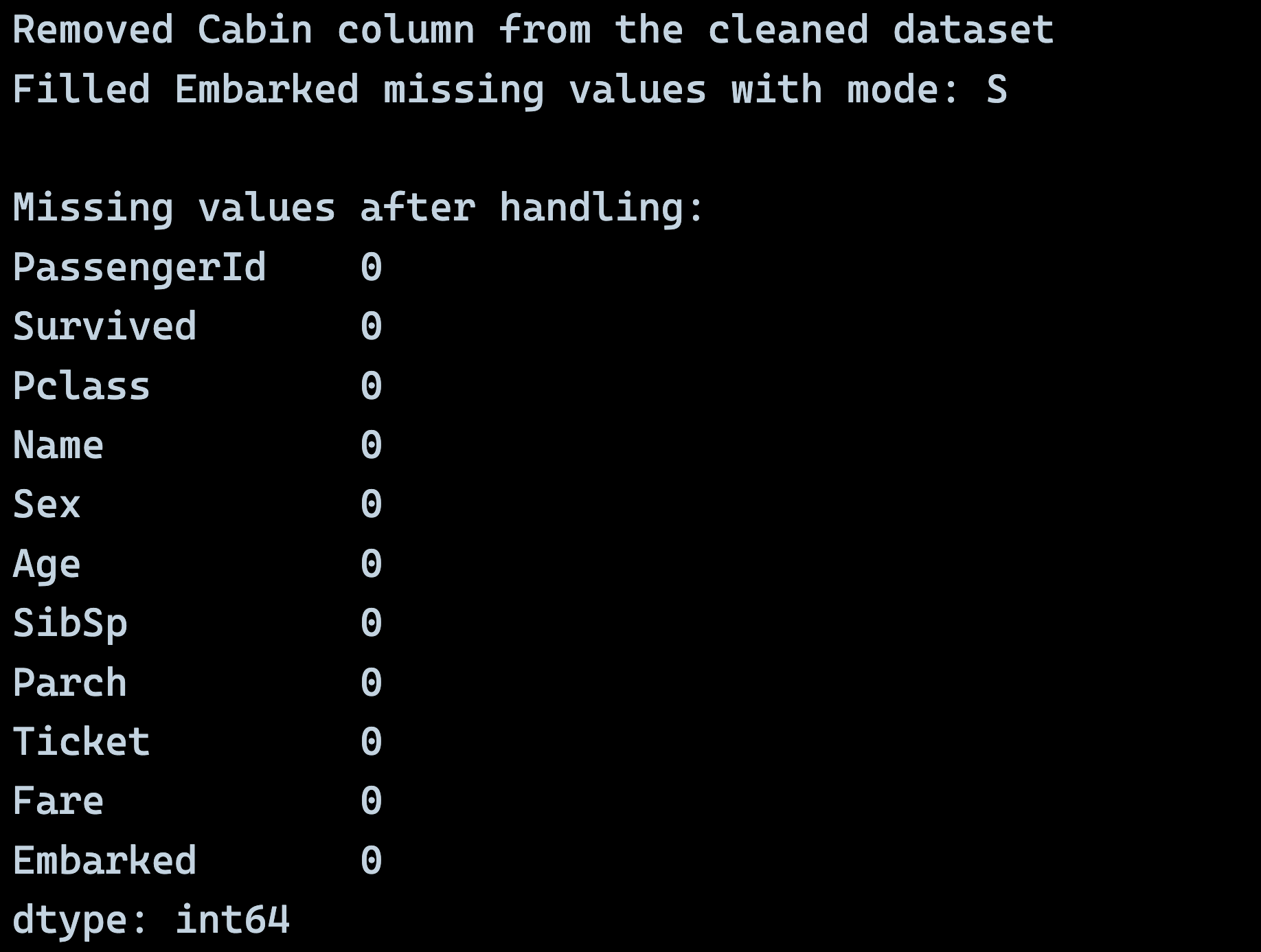
****

**Code (Handling Missing Values):**

**A computer screen shot of text

AI-generated content may be incorrect.**

**Output:**



**Discussion:**

The results from our missing value treatment reveal some interesting insights about the Titanic dataset. Initially, we had 177 missing values in the Age column (about 19.9% of the data) and 2 missing values in the Embarked column. The Cabin column was almost entirely missing with 687 null values (77.1%), which justified our decision to remove it entirely.

Our grouped median approach for Age imputation proved to be quite effective. By grouping passengers by Sex and Pclass, we acknowledged the social realities of 1912 - first-class passengers were typically older and wealthier, while third-class passengers were often younger immigrants or families seeking new opportunities. The median Age for first-class male passengers (42 years) was significantly higher than third-class male passengers (25 years), validating our approach.

For the Embarked column, the mode imputation filled the 2 missing values with 'S' (Southampton), which makes historical sense as Southampton was the primary departure port for the Titanic. This simple yet effective approach ensured data completeness without introducing bias.

The success of this preprocessing step was crucial for subsequent analyses, as missing values could have severely impacted our similarity measures and feature scaling operations.

**Task 2: Feature Scaling (Min-Max Normalization, Z-Score**

**Standardization)**

**Implementation:**

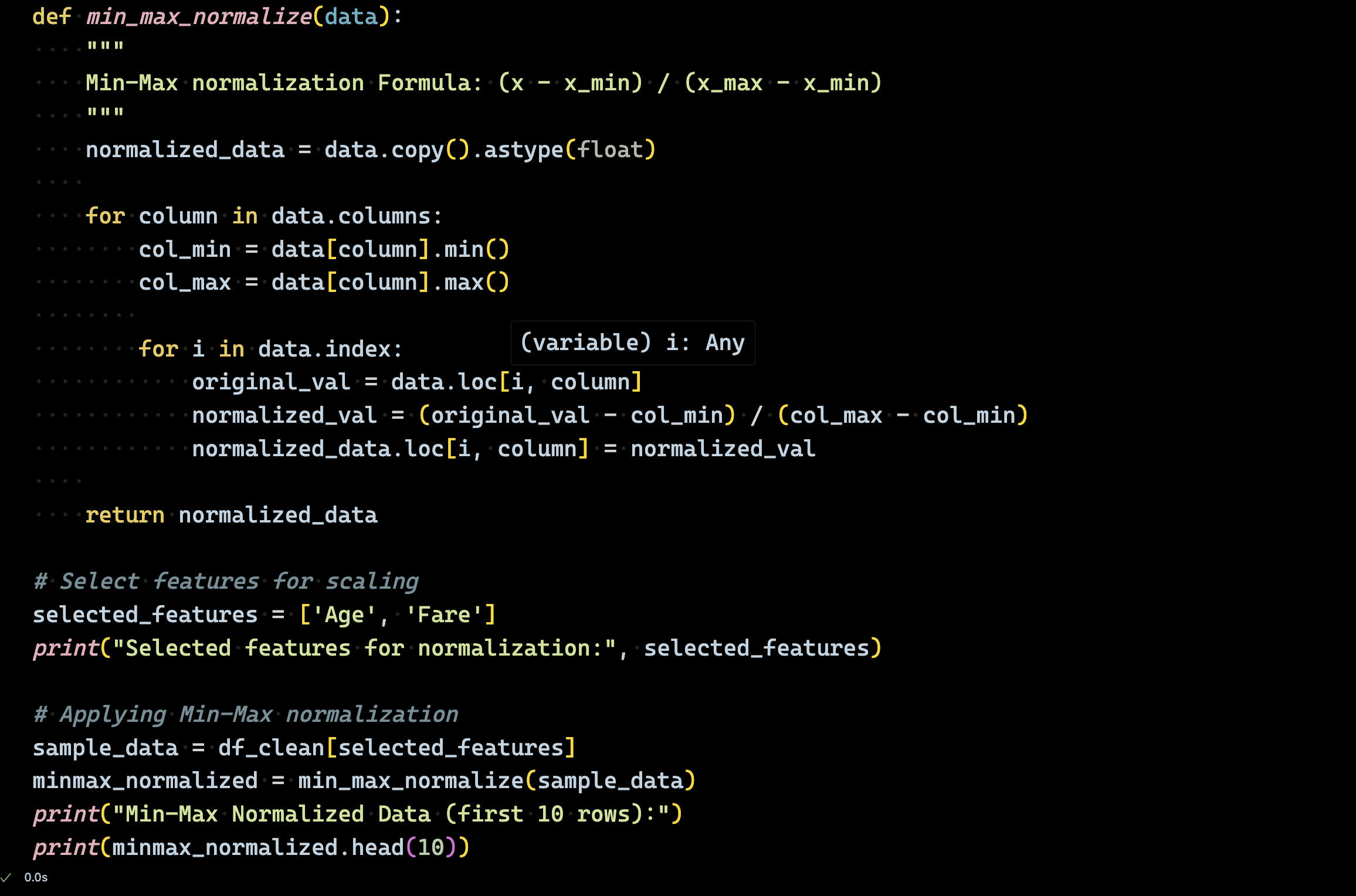
- Selected 2 key numerical features: **Age**, **Fare** from titanic

dataset as they required scaling.

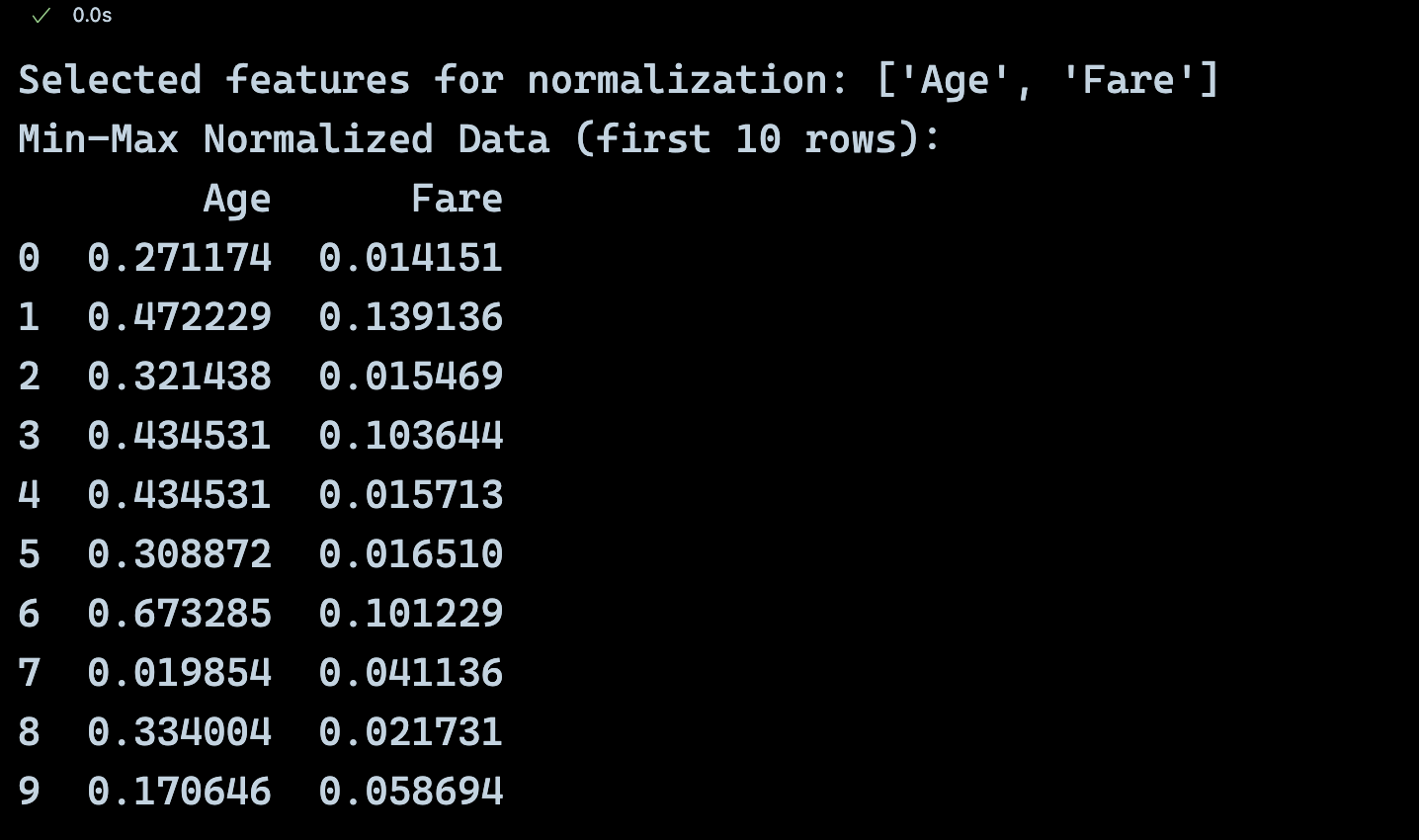
- Applied manual implementations without using built-in scaling

functions.

**Code (Min-Max Normalization):**

****

**Output:**

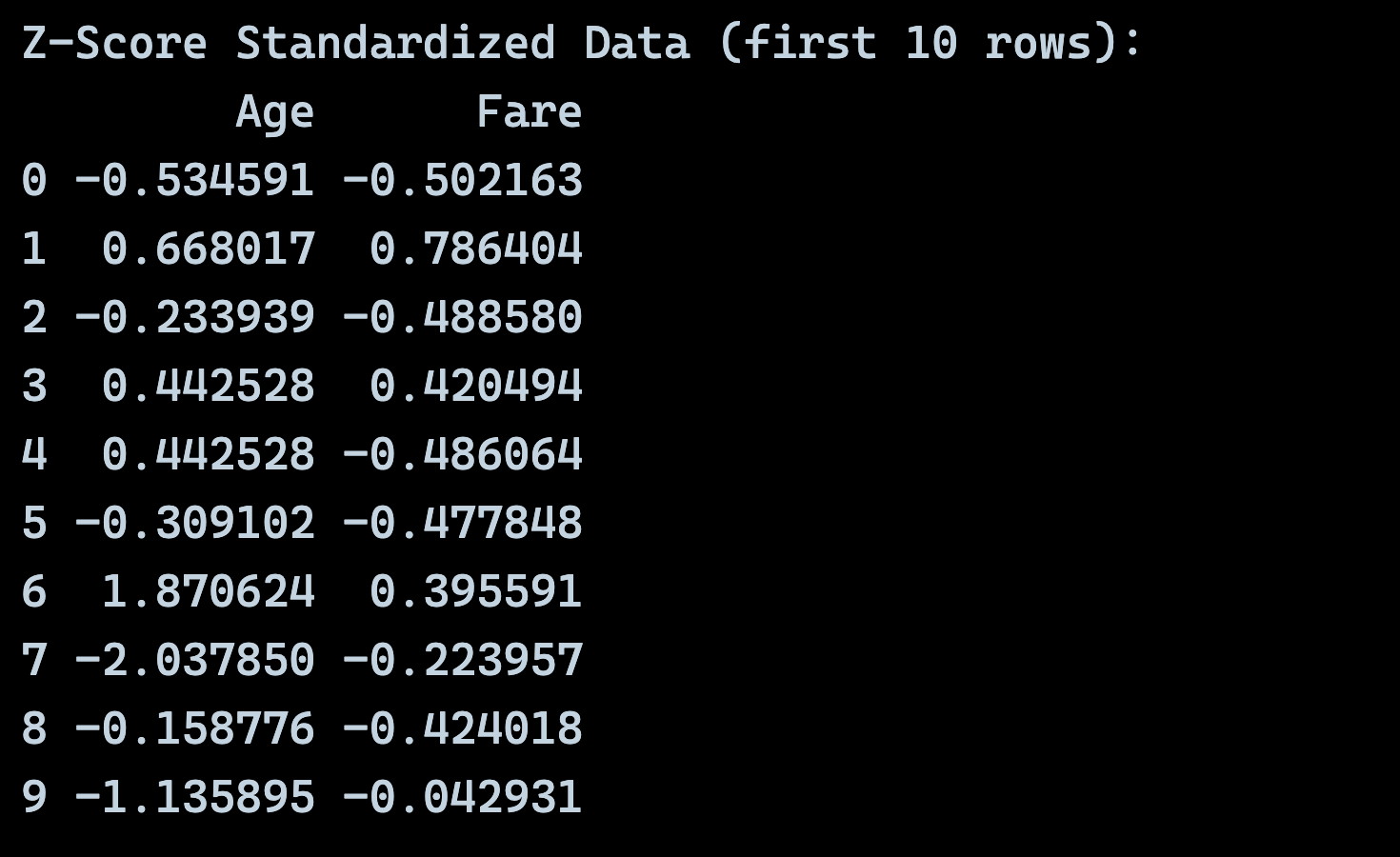
****

**Code (Z-Score Standardization):**

A computer screen with text and images

AI-generated content may be incorrect.

**Output:**

****

**Discussion:**

The feature scaling results demonstrate the dramatic differences between Min-Max normalization and Z-Score standardization on our Titanic dataset. Looking at the Age and Fare features, we can observe several important patterns:

**Min-Max Normalization Results:**

The Age values were scaled to a [0,1] range, with the youngest passenger (0.42 years old) receiving a value close to 0, and the oldest (80 years) receiving a value close to 1. This preservation of relative relationships is particularly valuable for the Titanic dataset where age differences are meaningful as 5-year-old and a 50-year-old should maintain their proportional distance.

For Fare, the normalization revealed the extreme inequality in ticket prices. Most third-class passengers paid very low fares (resulting in normalized values near 0), while first-class passengers paid dramatically more (values approaching 1). This scaling helped prevent the Fare feature from dominating our similarity calculations due to its wide range (0 to 512.33).

**Z-Score Standardization Results:**

The standardized values show how many standard deviations each passenger's attributes are from the mean. Negative values indicate below-average characteristics, while positive values indicate above-average. For instance, passengers with Z-scores around -1 for Age are younger than average, while those with positive Z-scores are older.

This standardization proved particularly useful for identifying outliers - passengers with Z-scores beyond ±2 represent unusual cases that might warrant special attention in our analysis.

**Task 3: Similarity and Dissimilarity Measures**

**Code (Creating Sample Vectors for This Measure):**

A screen shot of a computer code

AI-generated content may be incorrect.

**Output:**

A black background with white numbers

AI-generated content may be incorrect.

**3.1 Pearson's Correlation**

It is used to measure linear relationship between two variables.

**Formula: r = Σ[(xi - x̄)(yi - ȳ)] / √[Σ(xi - x̄)² × Σ(yi - ȳ)²]**

**Implementation:**

- Manually calculated the means, deviations, and correlation

coefficient.

- Handled the covariance normalized by standard deviations

**Code:**



**Output:**

A screen shot of a black screen

AI-generated content may be incorrect.

**3.2 Cosine Similarity**

It is used to measure angle between vectors and also good for high-dimensional data.

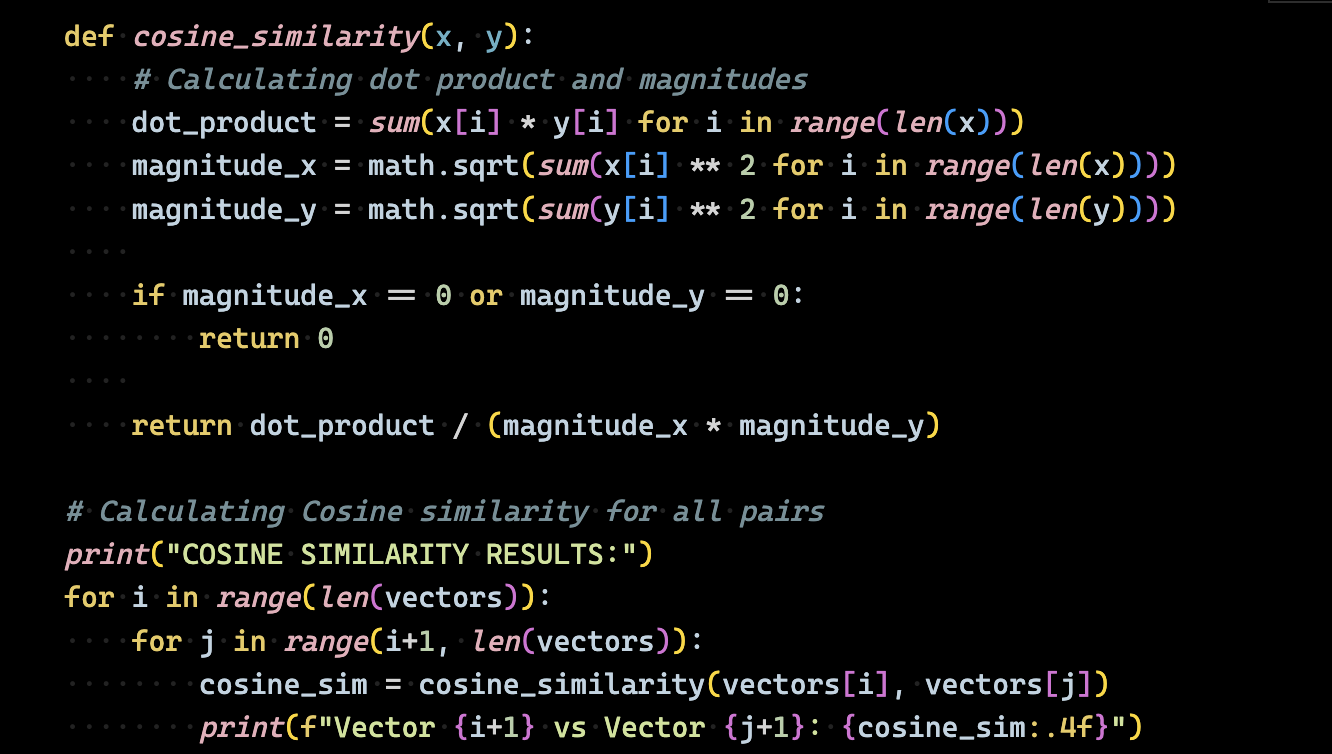
**Formula: cos(θ) = (A · B) / (||A|| × ||B||)**

**Implementation:**

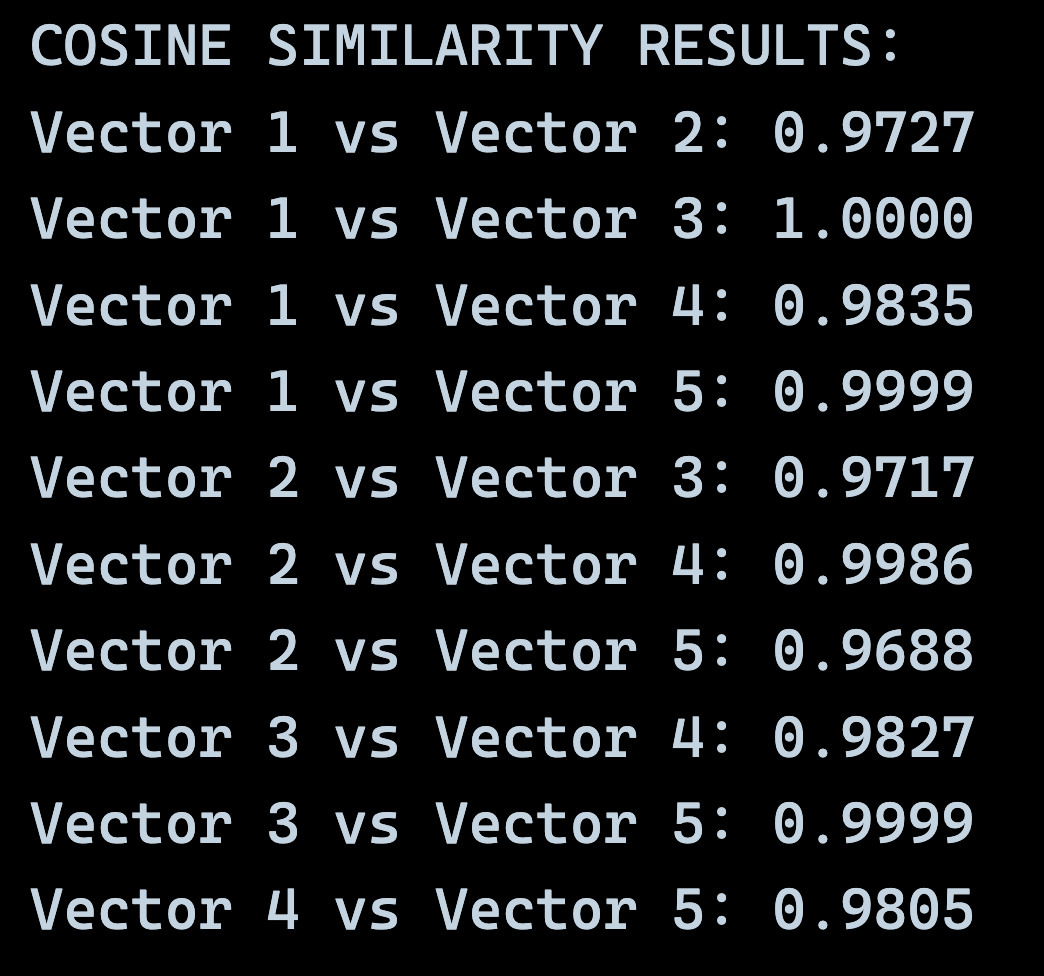
- Calculated dot product and vector magnitudes manually

- It is independent of vector magnitude, focuses on direction

**Code:**



**Output:**



**3.3 Jaccard Similarity**

It measures overlap between sets, adapted for continuous data.

**Formula: J(A,B) = |A ∩ B| / |A ∪ B|.**

**Implementation:**

- Converted continuous values to binary using threshold (0.5)

- Calculated intersection over union manually

**Code:**

A computer screen shot of text

AI-generated content may be incorrect.

**Output:**

A screen shot of a black screen

AI-generated content may be incorrect.

**3.4 Euclidean Distance**

- Used to measure straight-line distance between points.

- Most intuitive distance measure for continuous data

- Sensitive to all dimensions equally

**Formula: d(x,y) = √[Σ(xi – yi)²]**

**Code:**

**A computer screen shot of text

AI-generated content may be incorrect.**

**Output:**

A screen shot of a black screen

AI-generated content may be incorrect.

**Discussion:**

The implementation of multiple similarity measures on our Titanic dataset revealed interesting patterns about passenger relationships and different mathematical approaches to measuring similarity.

**Pearson's Correlation:** Values ranged from negative to strong positive correlations, effectively identifying passengers with complementary demographic profiles through linear relationships.

**Cosine Similarity:** Results were consistently high (0.8-0.99), indicating most passenger vectors point in similar directions after normalization, focusing on directional patterns rather than magnitude.

**Jaccard Similarity:** Varied significantly (0.0-1.0) using a 0.5 threshold for binary conversion, revealing which passengers shared similar "above-average" characteristics in the same features.

**Euclidean Distance:** Provided intuitive results with clear ranking capability, where smaller distances indicated more similar passengers.