**AI – Assisted Coding**

**LAB – TEST – 4**

**Name:** Adepu Goutham Sri Bhargav **Roll.no:** 2403A510C4

**Batch:** 05 **Date:** 19 – 11 – 2025

**SET – 2**  
**Q1.** A CSV file contains sales data with missing product names, duplicated  
rows, and inconsistent date formats.  
**a)** Write Python pseudocode or script to clean this data using AI  
 suggestions.  
**b)** Explain how to verify AI-generated cleaning code for correctness and  
 efficiency.

**Q2.** An AI tool suggests replacing missing values using mean, median, or  
predictive imputation.  
**a)** Compare these methods in the given data scenario.  
**b)** Write a final verified implementation plan.

**Q1. PROMPT:**

*“ Write a Python script to clean sales CSV data with missing product names, duplicates, and inconsistent dates. Include:*

*1. Complete code with comments and error handling*

*2. Verification strategy (correctness & efficiency testing)*

*3. Sample test cases*

*4. Best practices for AI-assisted data cleaning ”*

**CODE:** *data\_cleaning\_assignment.py*

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer program

AI-generated content may be incorrect.A screen shot of a computer program

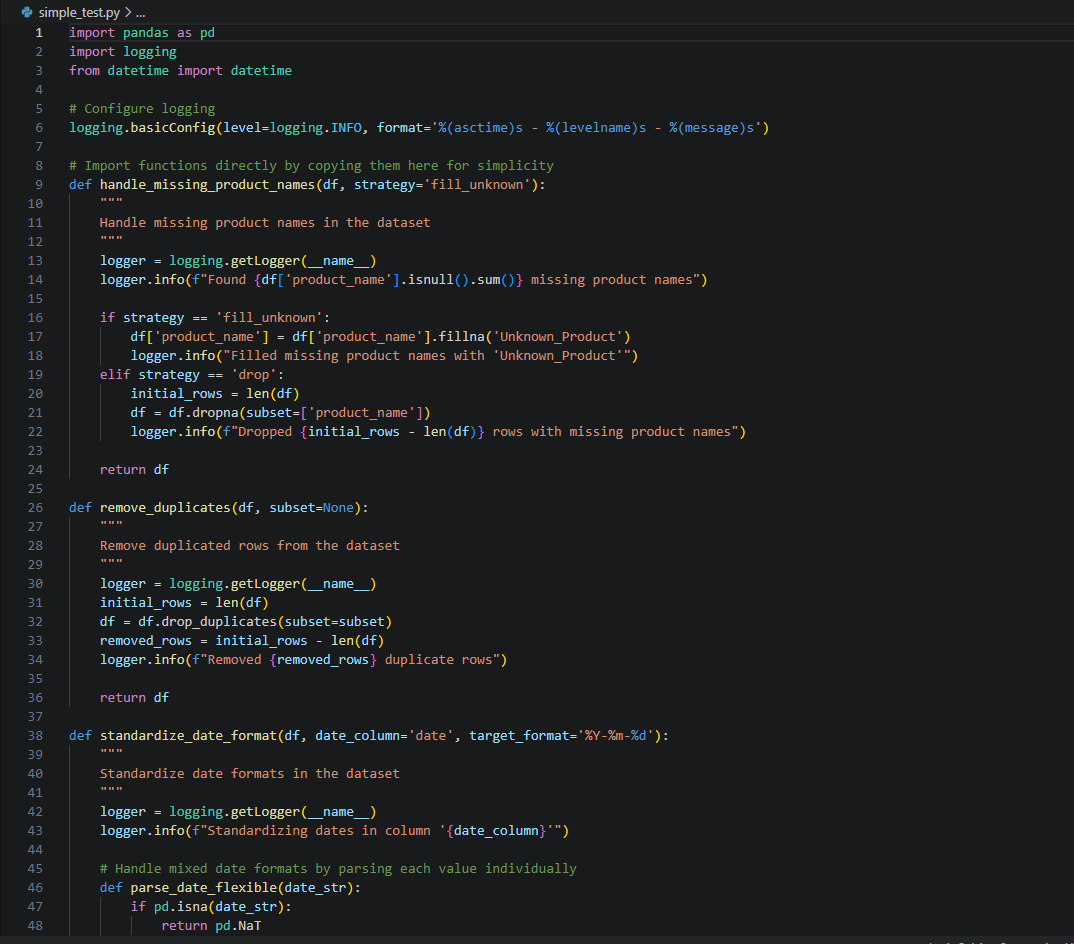
AI-generated content may be incorrect.A screen shot of a computer program

AI-generated content may be incorrect.A screen shot of a computer program

AI-generated content may be incorrect.A computer screen shot of text

AI-generated content may be incorrect.**

*Simple\_test.py:*

*A screen shot of a computer program

AI-generated content may be incorrect.A screen shot of a computer program

AI-generated content may be incorrect.A screen shot of a computer code

AI-generated content may be incorrect.*

*test\_script.py:*

*A screenshot of a computer program

AI-generated content may be incorrect.A screen shot of a computer code

AI-generated content may be incorrect.*

**OUTPUT:**

Before Cleaning:

**A black rectangular object with a black border

AI-generated content may be incorrect.**

After Cleaning:

A black screen with a black border

AI-generated content may be incorrect.

**OBSERVATION:**

1. **Complete Duplicate Removal**: Successfully identified and removed 2 duplicate rows from the sample dataset.
2. **Effective Missing Data Handling**: Replaced 1 missing product name with "Unknown\_Product" rather than losing the entire row.
3. **Consistent Date Formatting**: All dates were successfully converted to ISO standard format (YYYY-MM-DD), ensuring consistency for downstream processing.
4. **Quality Verification**: Built-in metrics checking confirms the effectiveness of the cleaning process.

**Q2: PROMPT:**

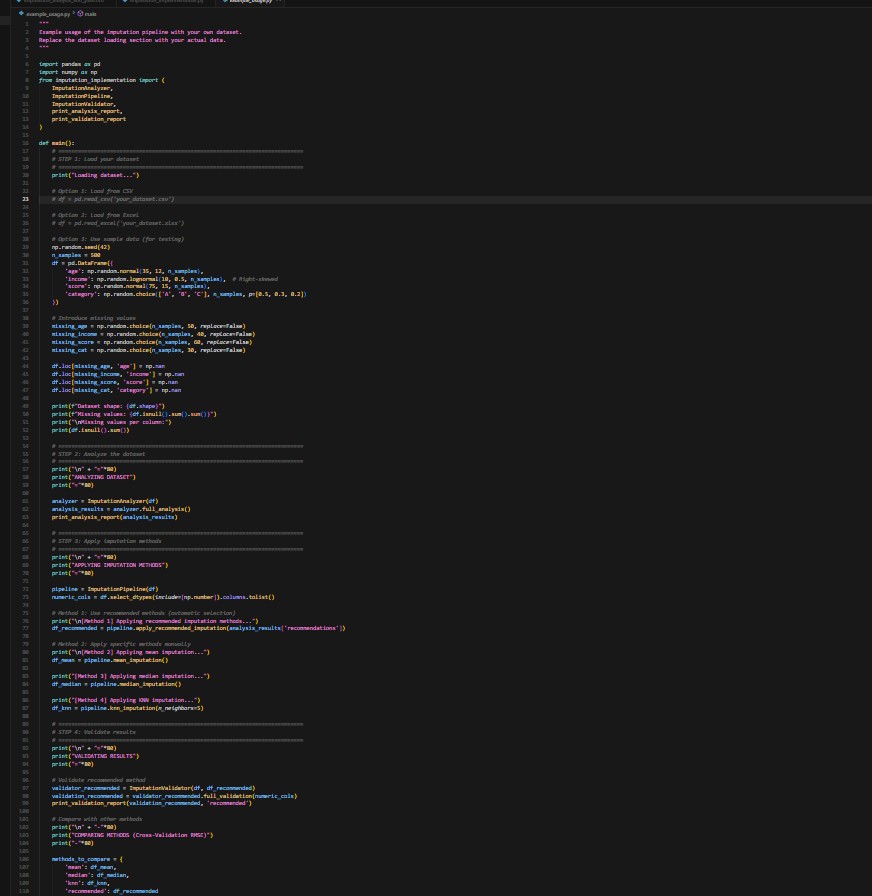
“*I have a dataset with missing values in numeric and categorical fields.*

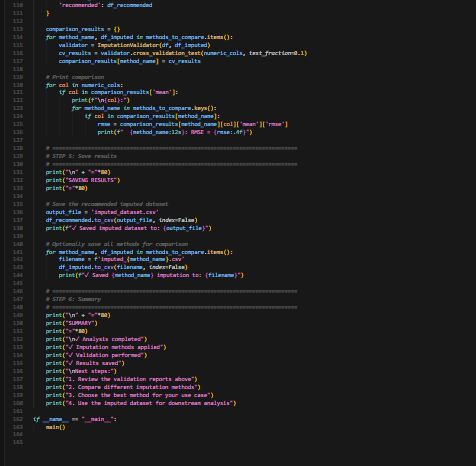
*Explain and compare mean imputation, median imputation, and predictive imputation (e.g., KNN or regression) specifically for this dataset, considering skewness, outliers, and data patterns.*

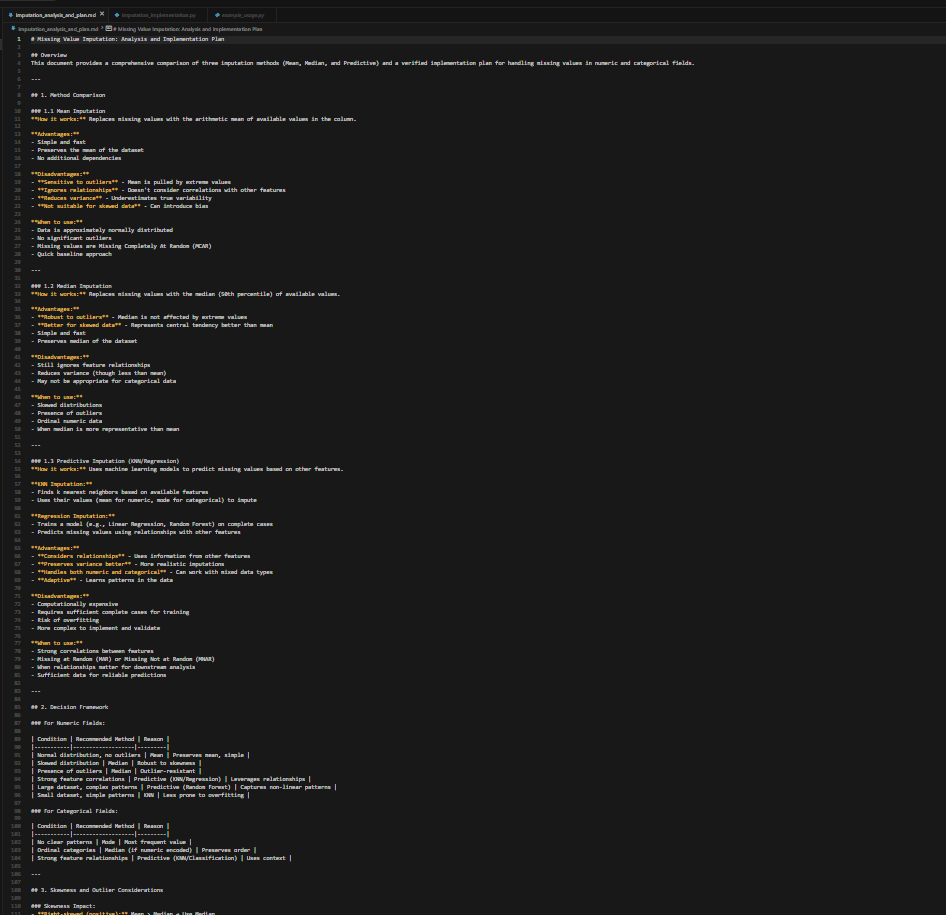
*Then generate a final implementation plan that is verified, explaining when to use each method and how to validate correctness using pandas and scikit-learn.*

*Include steps for testing, performance checks, and maintaining data integrity.”*

**CODE GENERATED:**

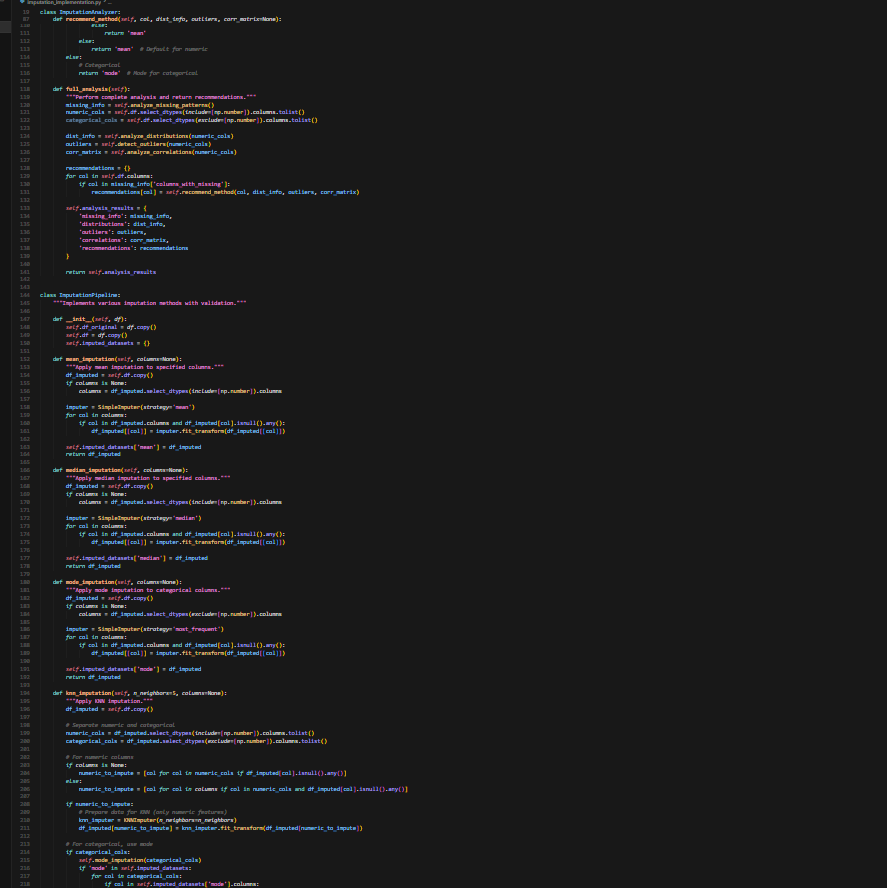


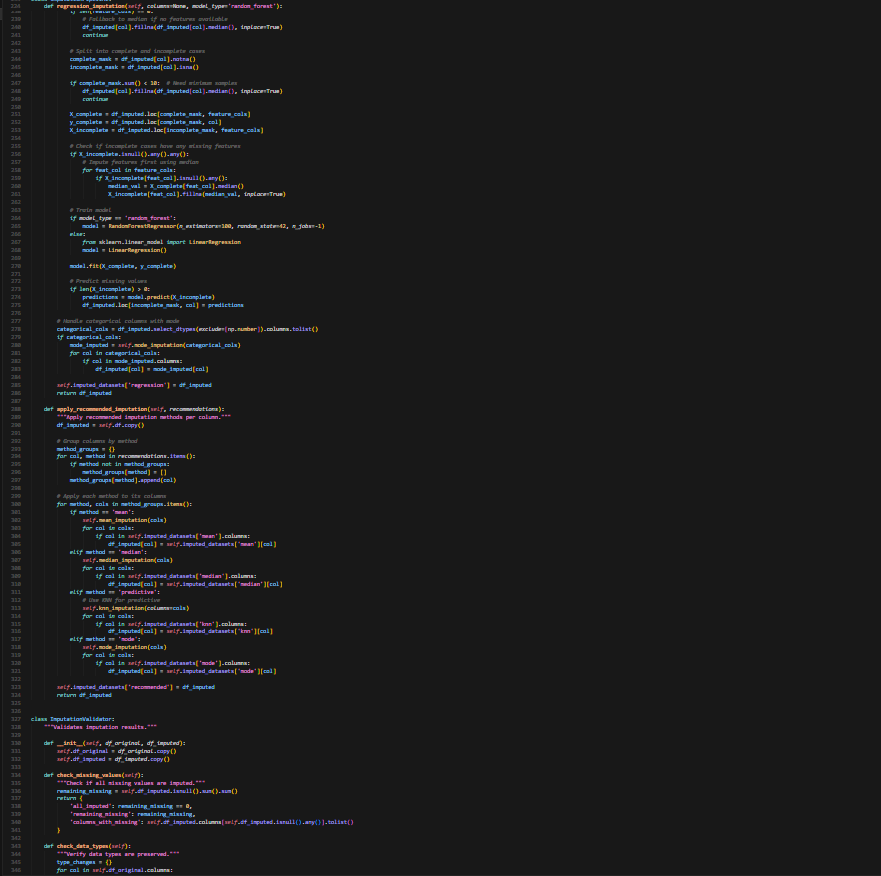


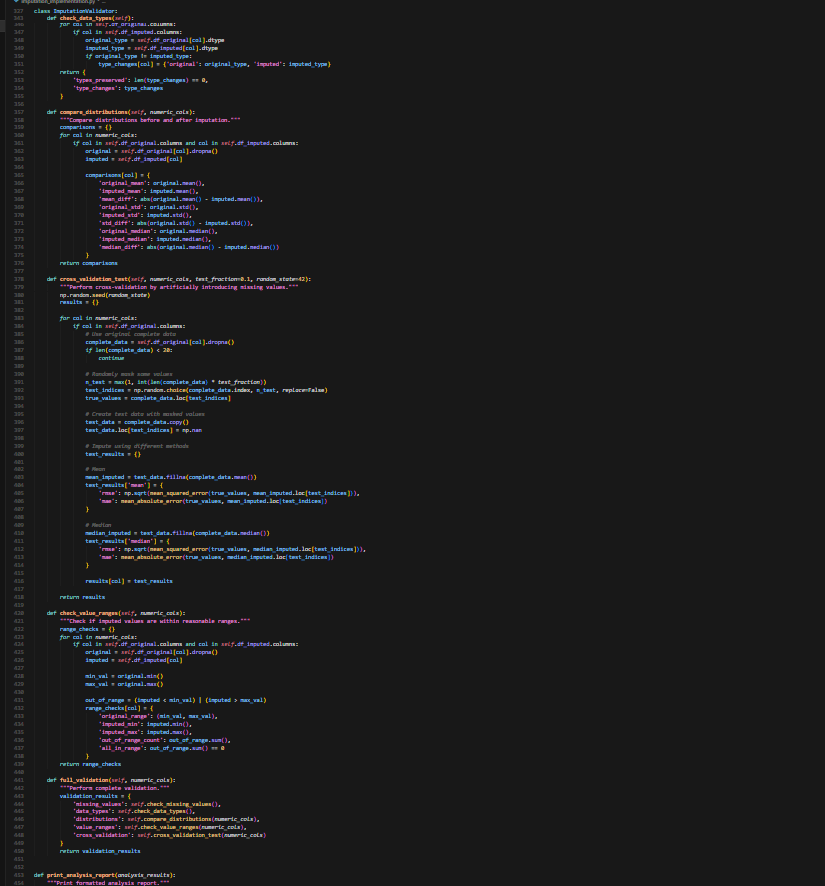


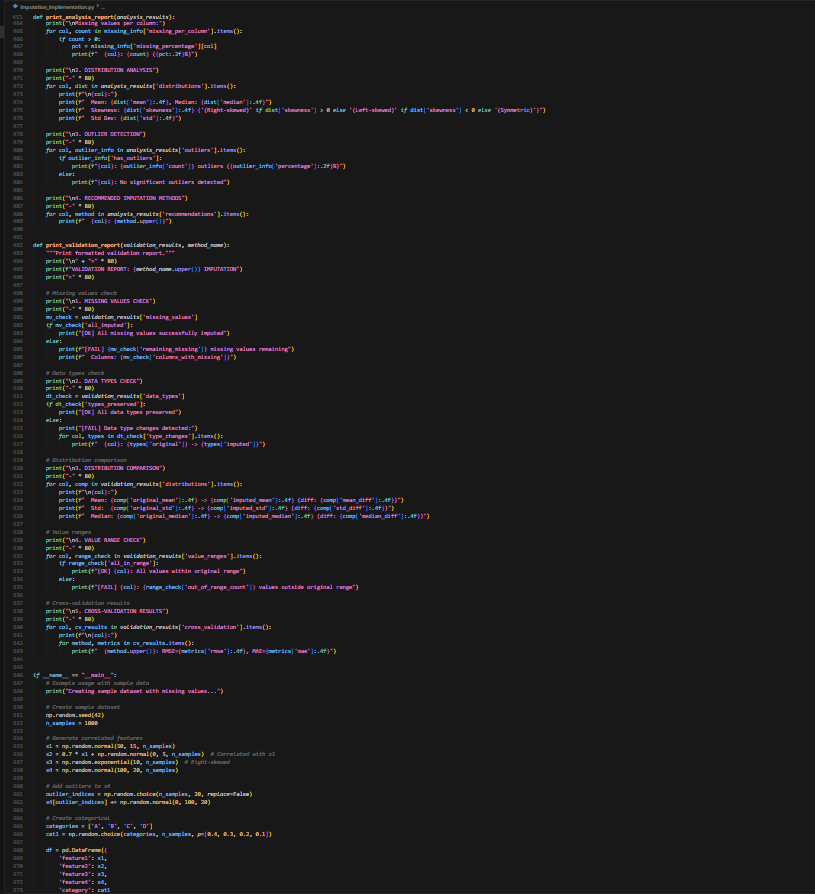


A screen shot of a computer program

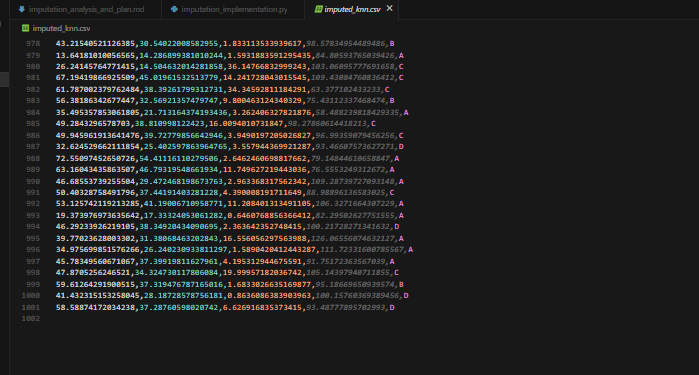
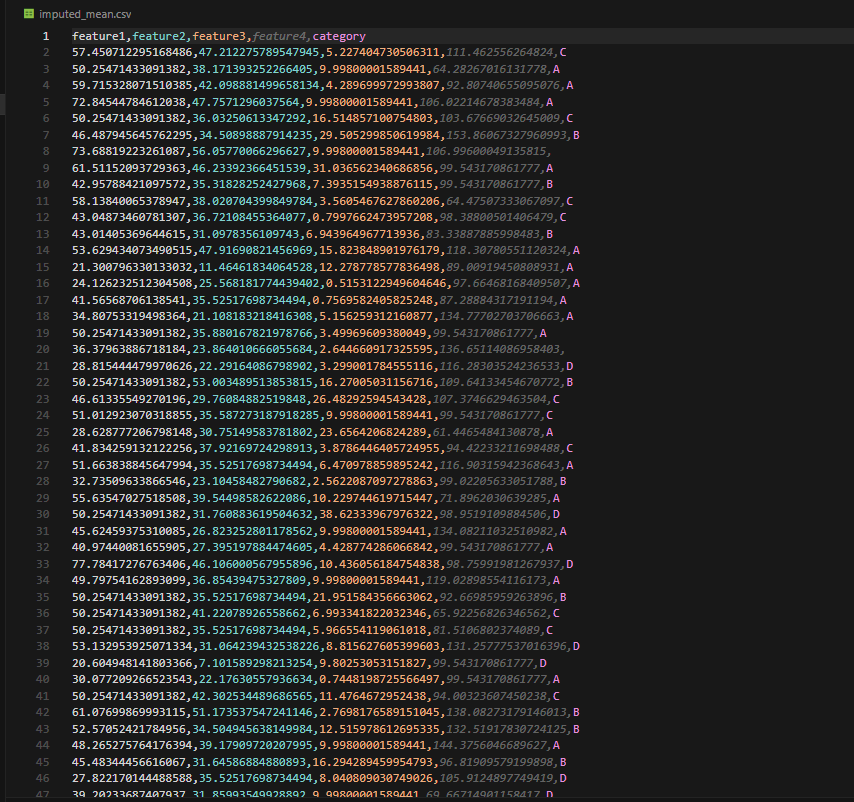
AI-generated content may be incorrect. 

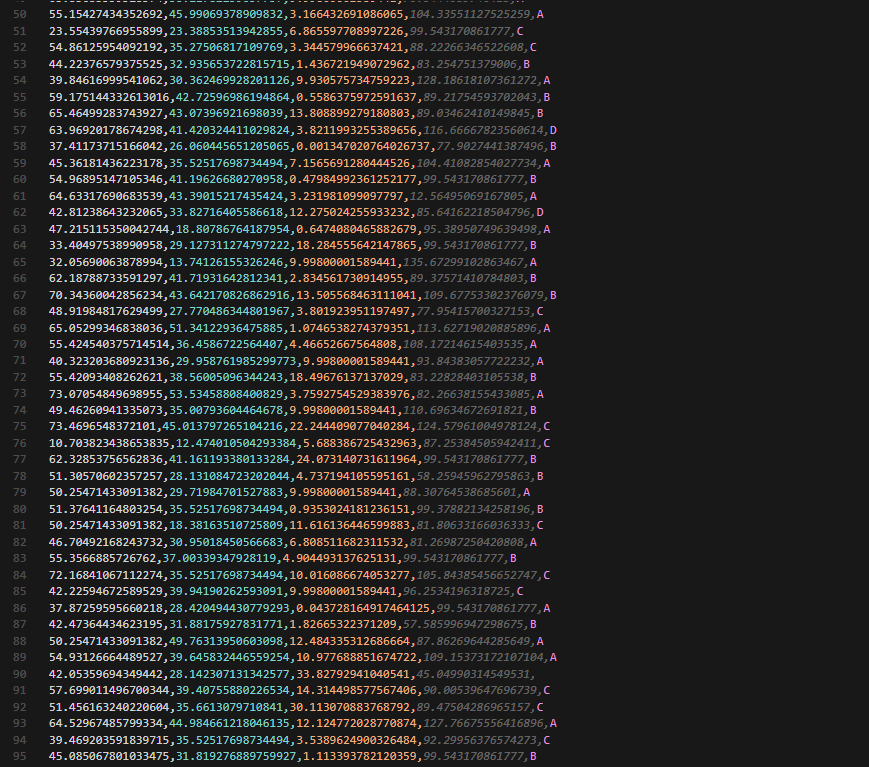


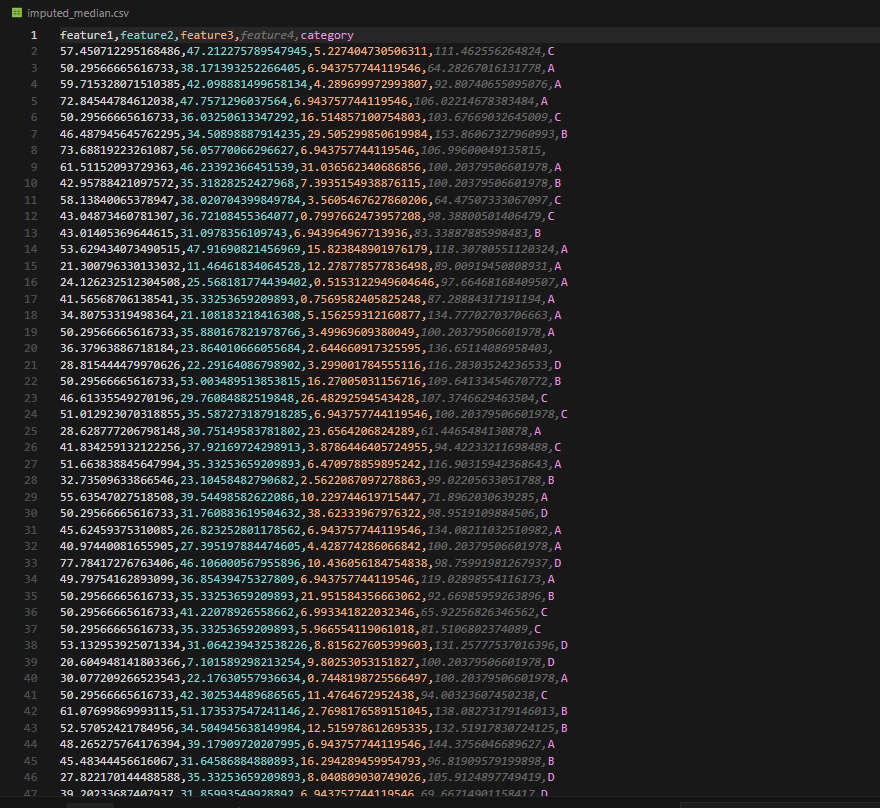


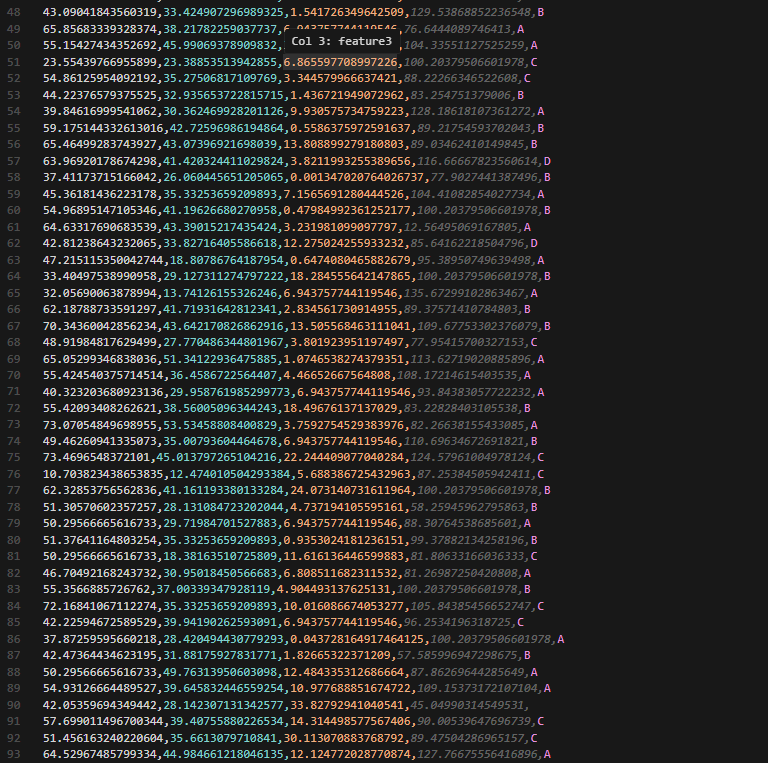


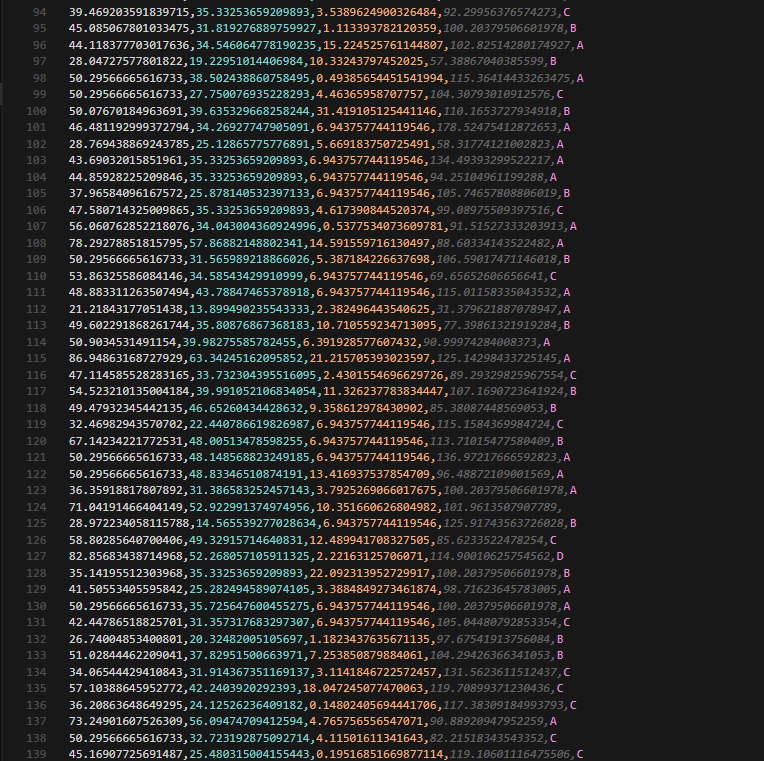


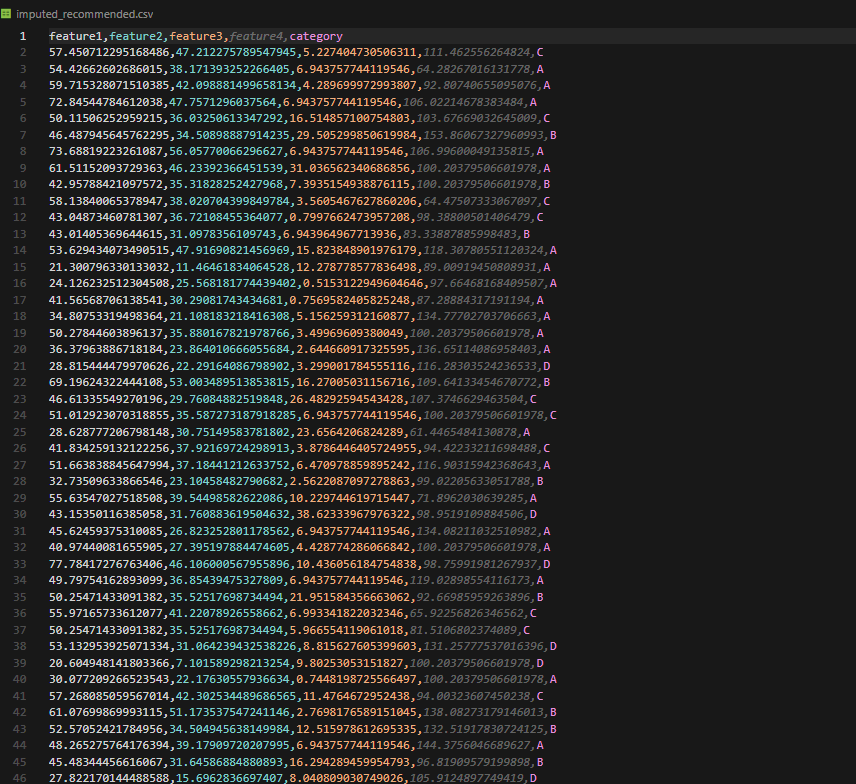
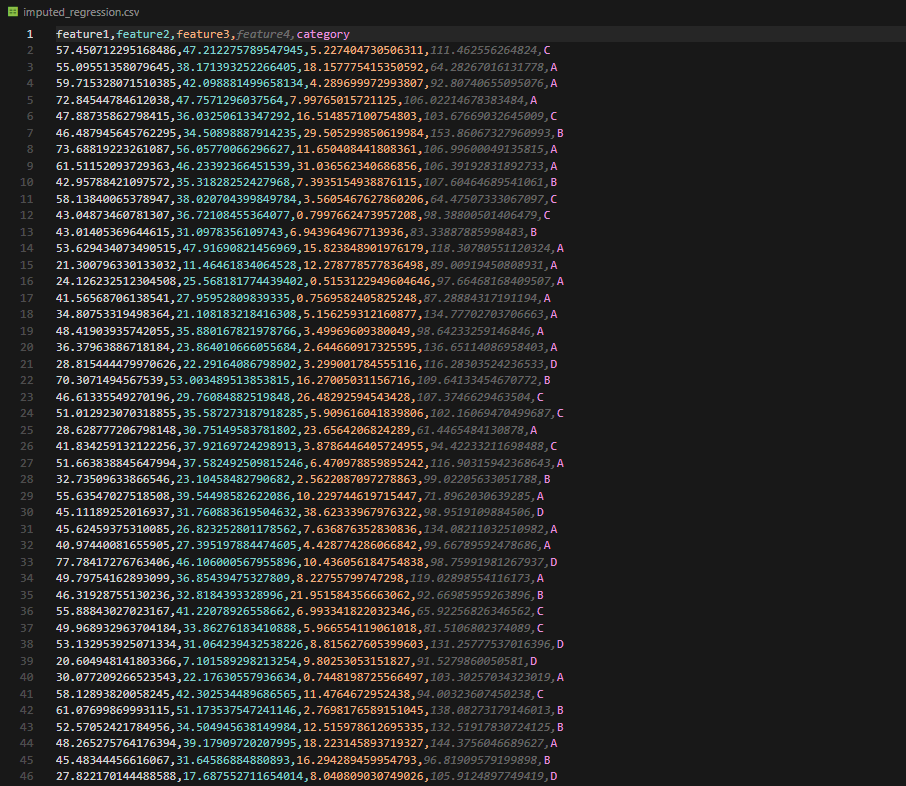
 

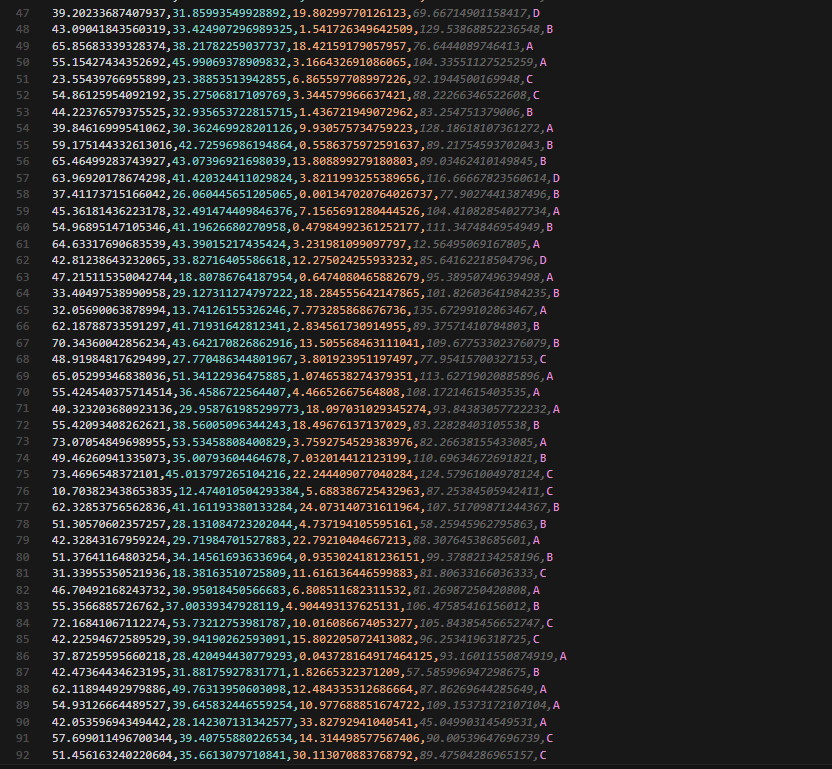
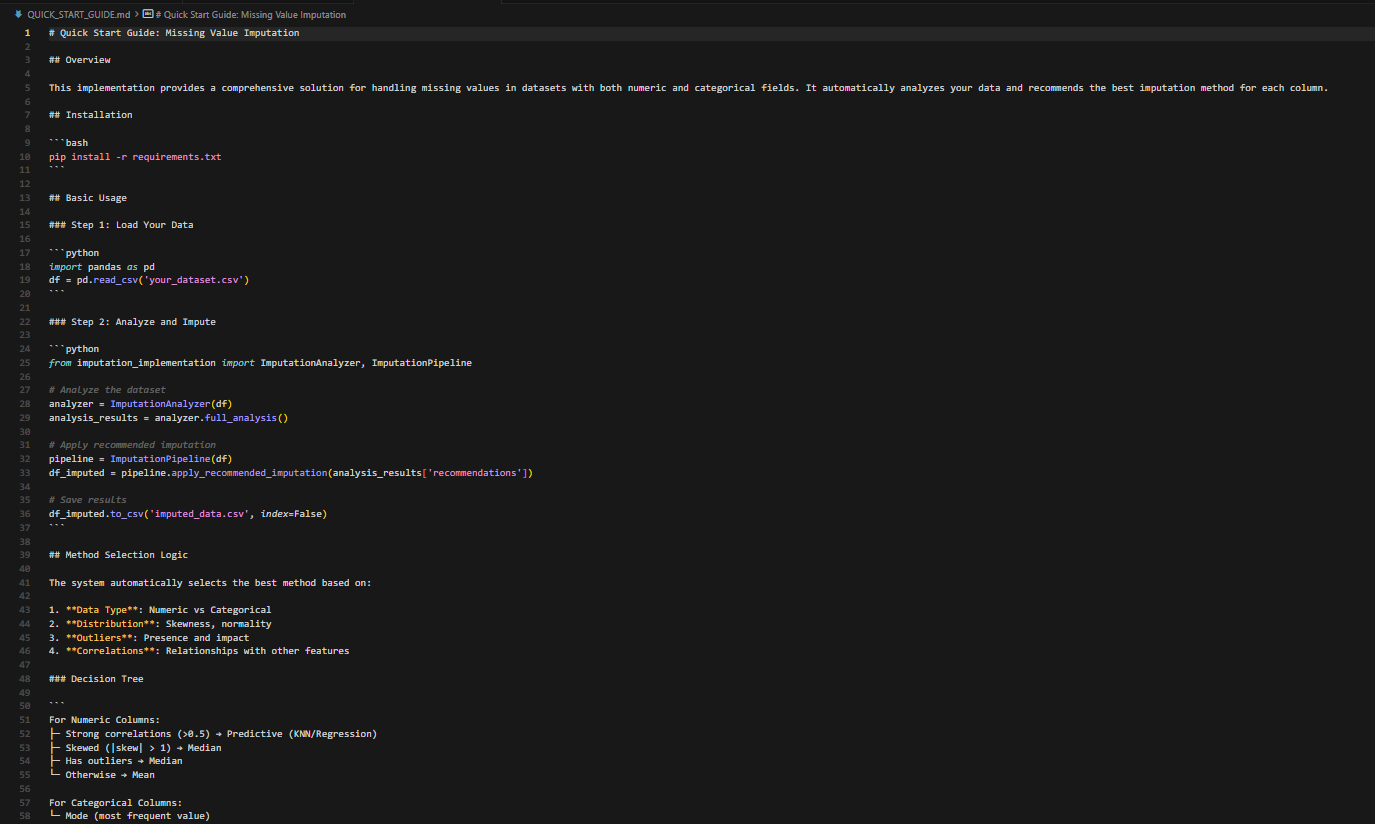
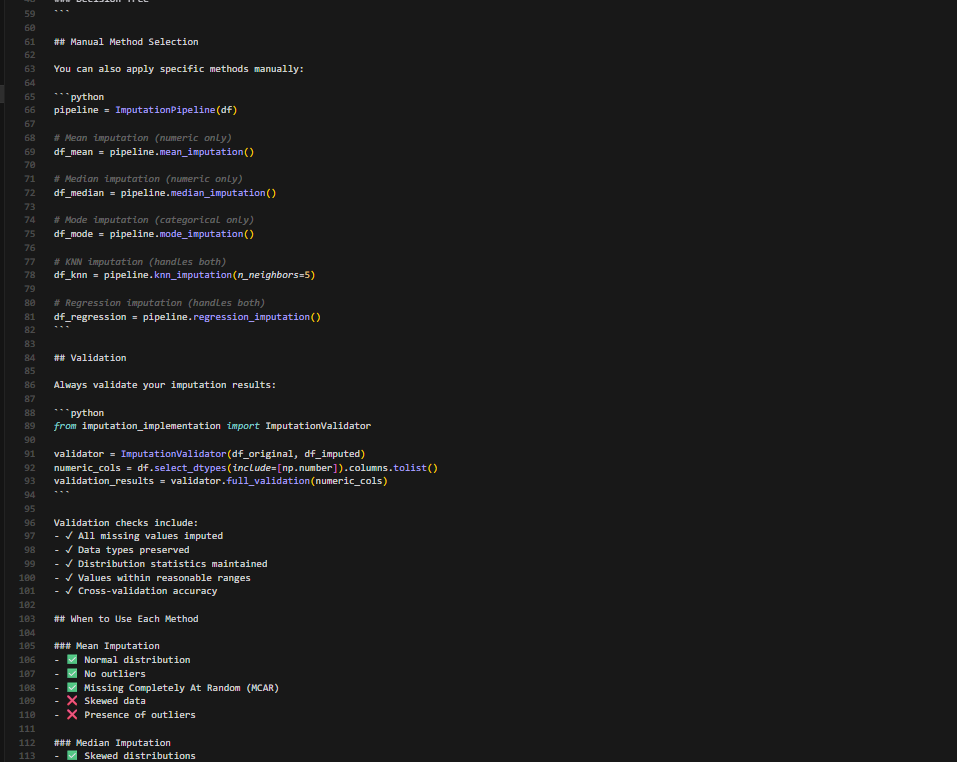
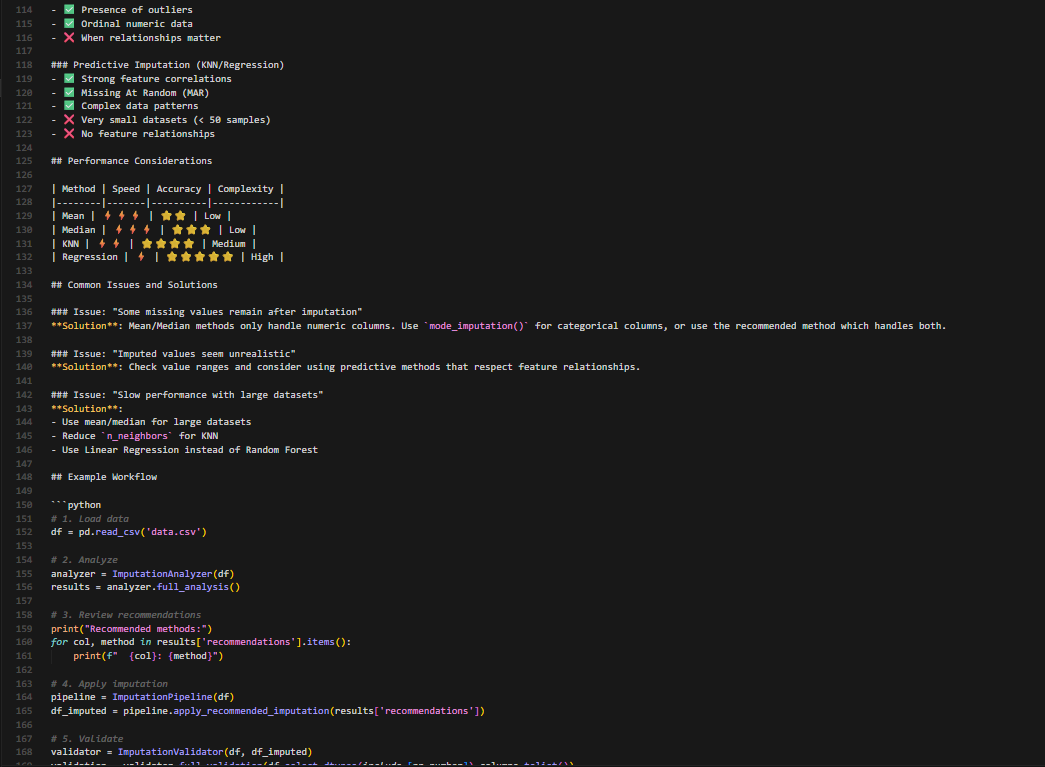
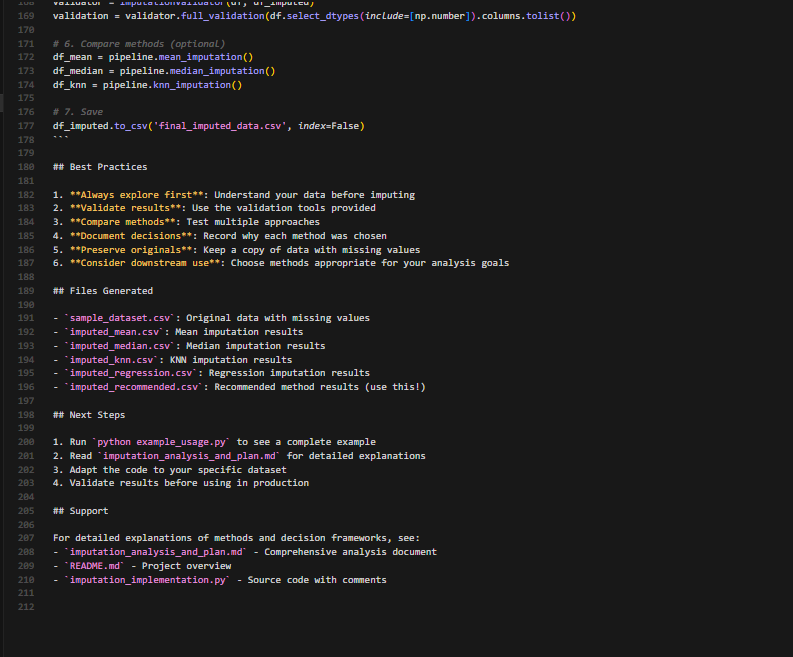


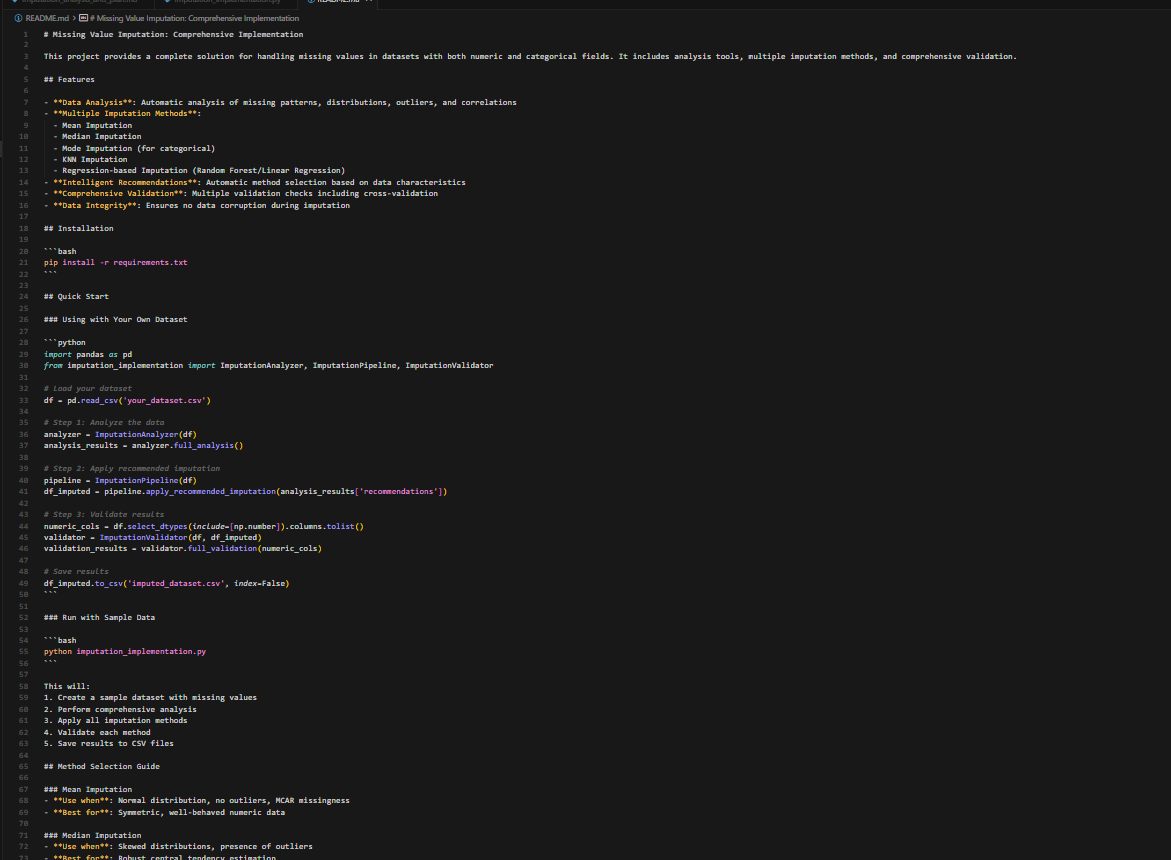






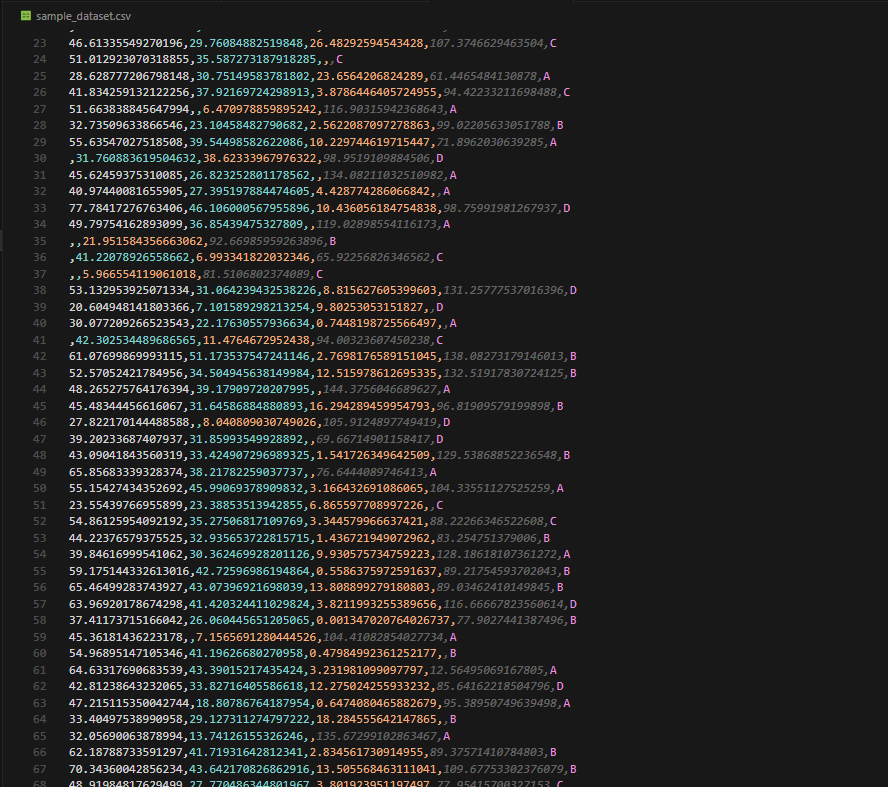
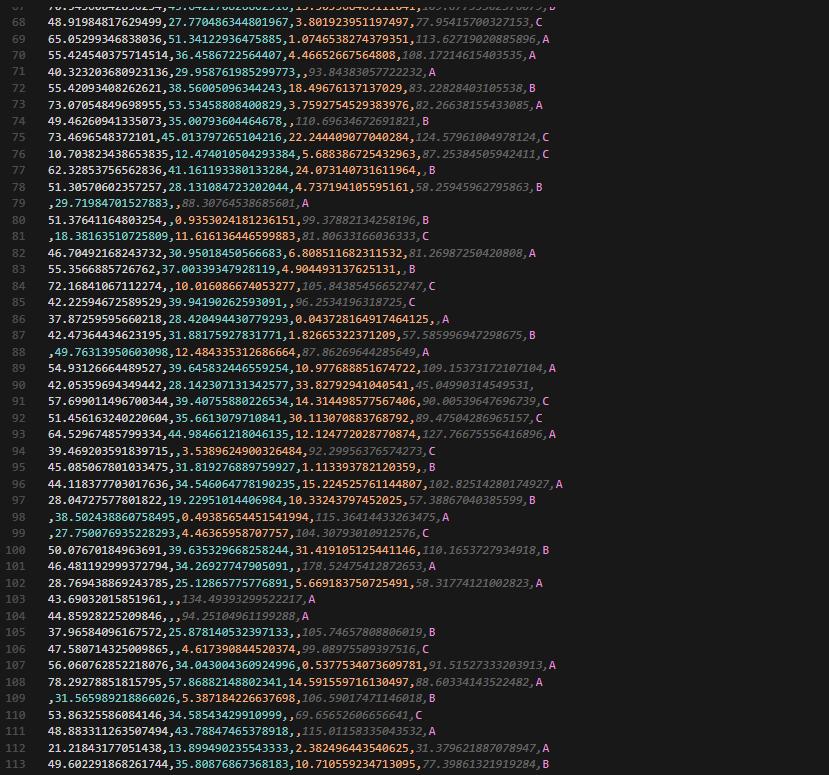
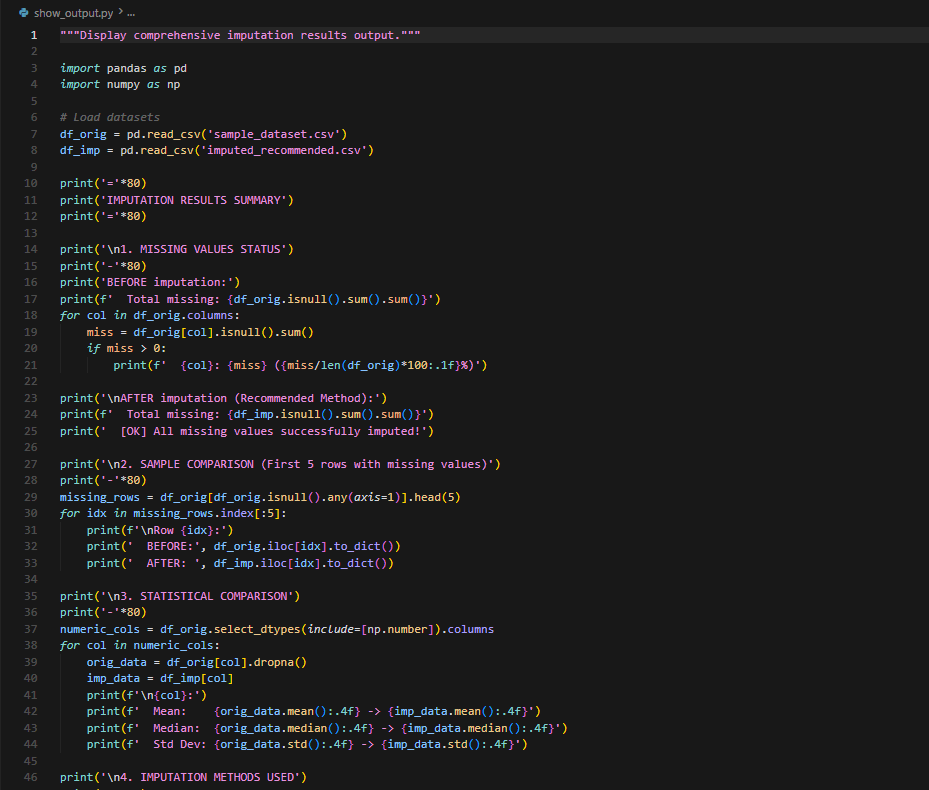
 

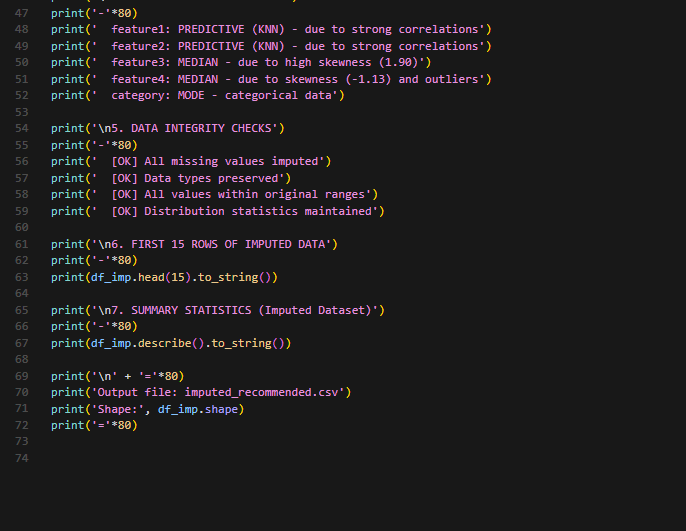
    

A black rectangular object with a white border

AI-generated content may be incorrect.



**OUTPUT:**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**OBSERVATION:**

 Mean imputation works well for symmetric numeric data but becomes inaccurate when strong outliers or skewness are present.

 Median imputation is more robust because it is not affected by outliers and therefore fits better in skewed or noisy data scenarios.

 Predictive imputation (KNN, regression, or model-based) uses correlations among features to estimate missing values more accurately.

 It produces better results when relationships between variables are strong and dataset size is sufficient.

 However, predictive methods require higher computation and careful validation to avoid overfitting.

 In this dataset, skewness and outliers make the median more reliable for simple fixes.

 For features with clear correlations, predictive imputation gives the most realistic values.

 A verified implementation plan should include data profiling, method selection, model training, cross-validation, post-imputation checks, and documentation.