# 🧹 AI Data Cleaner using Gemini

📁 Upload your dataset (CSV or Excel)

Drag and drop file here
Limit 200MB per file • CSV, XLSX

Browse files

UberDataset (1).csv 85.6KB ✕

## 🔍 Original Dataset Preview

| | ATE | END_DATE | CATEGORY | START | STOP | MILES | PURPOSE |
|---|---|---|---|---|---|---|---|
| 0 | 16 21:11 | 01-01-2016 21:17 | Business | Fort Pierce | Fort Pierce | 5.1 | Meal/Entertain |
| 1 | 16 01:25 | 01-02-2016 01:37 | Business | Fort Pierce | Fort Pierce | 5 | None |
| 2 | 16 20:25 | 01-02-2016 20:38 | Business | Fort Pierce | Fort Pierce | 4.8 | Errand/Supplies |
| 3 | 16 17:31 | 01-05-2016 17:45 | Business | Fort Pierce | Fort Pierce | 4.7 | Meeting |
| 4 | 16 14:42 | 01-06-2016 15:49 | Business | Fort Pierce | West Palm Beach | 63.7 | Customer Visit |
| 5 | 16 17:15 | 01-06-2016 17:19 | Business | West Palm Beach | West Palm Beach | 4.3 | Meal/Entertain |
| 6 | 16 17:30 | 01-06-2016 17:35 | Business | West Palm Beach | Palm Beach | 7.1 | Meeting |
| 7 | 16 13:27 | 01-07-2016 13:33 | Business | Cary | Cary | 0.8 | Meeting |
| 8 | 16 08:05 | 01-10-2016 08:25 | Business | Cary | Morrisville | 8.3 | Meeting |
| 9 | 16 12:17 | 01-10-2016 12:44 | Business | Jamaica | New York | 16.5 | Customer Visit |

**Original Shape:** (1156, 7)

## 📝 Cleaning Summary Report

Data Cleaning Report

This report summarizes the cleaning process applied to a dataset originally containing 1156 rows and 7 columns. After cleaning, the dataset now contains 1155 rows and 7 columns.

**Cleaning Actions:**

The cleaning process involved the following steps:

1. **Duplicate Row Removal:** One duplicate row was identified and removed to ensure data accuracy and avoid bias in analysis. This reduced the dataset size by one row.

2. **Missing Value Imputation:** Missing values were present in several columns. These were addressed using the mode (most frequent value) imputation technique. Specifically:

    - `END_DATE` : Missing values were filled with '01-01-2016 21:17', which is the most common date and time in this column.
    - `CATEGORY` : Missing values were filled with 'Business', the most frequent category.
    - `START` : Missing values were filled with 'Cary', the most frequent starting location.
    - `STOP` : Missing values were filled with 'Cary', the most frequent ending location.
    - `PURPOSE` : Missing values were filled with 'Meeting', the most frequent purpose.

**Rationale:**

The mode imputation was chosen for these columns because it's a simple and effective method for handling missing data, particularly when dealing with categorical variables like `CATEGORY` , `START` , `STOP` , and `PURPOSE` . For `END_DATE` , while more sophisticated imputation might be considered in other contexts, using the mode provides a reasonable approximation given the available data. The removal of the duplicate row ensured data integrity by eliminating redundant information.

This cleaning process resulted in a cleaner and more consistent dataset suitable for further analysis. The use of mode imputation may introduce a slight bias; however, given the context and the small number of missing values, this approach was deemed acceptable.

**Cleaned Shape:** (1155, 7)

⬇ Download Cleaned Dataset

⬇ Download Cleaning Report