# Data Cleaning Project Report

Internship Task Report: Data Cleaning Project
Project Overview
Date: 23-12-2025
Intern: Abhishek CD
Role: Data Analytics Intern
Task: Level 1 - Data Cleaning
Tools Used: Google Colab, Python, Pandas, NumPy, Matplotlib, Seaborn
Dataset: New York City Airbnb Open Data

Executive Summary
Successfully completed data cleaning operations on a dataset to ensure data quality and integrity. The project involved identifying and resolving multiple data quality issues including missing values, duplicates, inconsistent formatting, and potential outliers.

1. Dataset Information
Initial Dataset Status
Original Size: [X] rows × [Y] columns

File Format: CSV

Encoding Used: Latin-1 (resolved initial UnicodeDecodeError)

Memory Usage: [Z] MB

Data Structure
text
[Include df.info() output here]
2. Data Cleaning Process & Results
2.1 Missing Values Handling
Issues Identified:

[Number] columns contained missing values

Total missing values: [Count]

Most affected column: [Column Name] with [X]% missing

Actions Taken:

Numeric Columns: Imputed missing values with median

Categorical Columns: Imputed missing values with mode

Extreme Cases: [If any columns were dropped, mention here]

Results:

Missing values reduced from [Initial Count] to 0

Data completeness: 100%

2.2 Duplicate Removal
Issues Identified:

[Number] duplicate rows found

Duplicate rate: [X]% of total dataset

Actions Taken:

Removed all exact duplicate rows

Kept first occurrence of each duplicate set

Results:

Removed [X] duplicate rows

Unique rows preserved: [Y]

2.3 Data Type Standardization
Issues Identified:

[List any data type issues found]

Inconsistent date formats: [Yes/No]

Mixed data types in columns: [Columns if any]

Actions Taken:

Text columns: Stripped extra whitespace

Date columns: Converted to datetime format

Numeric columns: Ensured proper numeric types

Results:

Consistent data types across all columns

columns converted to optimal data types

2.4 Outlier Detection
Issues Identified:

[Number] numeric columns analyzed

Outliers detected in: [List columns with outliers]

Extreme values range: [Provide examples]

Actions Taken:

Used IQR method for outlier detection

Created outlier flag columns for each numeric column

[Optional: Mention if any capping was done]

Results:

Identified [X] outlier data points

Added [Y] new columns for outlier tracking

Data range standardized for analysis

3. Quality Metrics
Before Cleaning:

| Metric | Value |
|---|---|
| Total Rows | [X] |
| Total Columns | [Y] |
| Missing Values | [Z] |
| Duplicate Rows | [A] |
| Memory Usage | [B] MB |

After Cleaning:

| Metric | Value | Improvement |
|---|---|---|
| Total Rows | [X] | -[D]% |
| Total Columns | [Y] | +[E] (outlier flags) |
| Missing Values | 0 | 100% reduction |
| Duplicate Rows | 0 | 100% reduction |
| Memory Usage | [C] MB | -[F]% |

4. Visual Documentation
4.1 Missing Values Heatmap
https://link-if-you-saved-image
Description: Visual representation of missing data distribution before cleaning

4.2 Data Distribution Before & After
https://link-if-you-saved-image
Description: Comparison of key variable distributions before and after cleaning

4.3 Outlier Detection Boxplots
https://link-if-you-saved-image
Description: Boxplots showing outlier detection in numeric columns

5. Key Findings & Insights
5.1 Major Data Quality Issues:
[Issue 1]: Describe the most significant issue found

[Issue 2]: Describe secondary issues

[Issue 3]: Describe any unexpected findings

5.2 Impact on Analysis:
Without cleaning: Analysis would have been skewed by [X]%

Data reliability improved by: [Y]%

Key metrics affected: [List metrics that would be wrong without cleaning]

5.3 Recommendations:
For future data collection: [Suggestions]

For ongoing maintenance: [Suggestions]

For analysis: [Cautions or notes]

6. Code Implementation Summary

Techniques Applied:
Encoding Resolution: Fixed UnicodeDecodeError using Latin-1 encoding

Missing Data Imputation: Median for numeric, mode for categorical

Duplicate Management: Exact match removal

Outlier Handling: IQR method with flagging

Data Type Conversion: Automated type inference and correction

Key Functions Used:
python
# Sample of key functions implemented
- pd.read_csv() with encoding parameter
- df.isnull().sum() for missing value detection
- df.dropna() / df.fillna() for missing value handling
- df.drop_duplicates() for duplicate removal
- IQR calculation for outlier detection
- pd.to_datetime() for date conversion
7. Challenges & Solutions
Challenge 1: UnicodeDecodeError
Problem: Could not read CSV file with UTF-8 encoding
Solution: Used Latin-1 encoding which resolved the issue
Learning: Different systems use different default encodings

Challenge 2: [Other Challenge if any]
Problem: [Describe]
Solution: [Describe]
Learning: [What you learned]

8. Deliverables
Files Created:
cleaned_dataset.csv - Final cleaned dataset

data_cleaning_report.pdf - This report

cleaning_code.ipynb - Complete Google Colab notebook

Metrics Achieved:
✓ 100% missing values resolved

✓ 100% duplicates removed

✓ Data types standardized

✓ Outliers identified and flagged

✓ Data ready for analysis

9. Learnings & Skills Developed
Technical Skills:
Pandas Mastery: Data manipulation, cleaning, transformation

Encoding Handling: Understanding different text encodings

Quality Metrics: Calculating and tracking data quality

Visualization: Creating informative data quality visuals

Analytical Skills:
Problem-solving approach to data issues

Decision-making for handling missing data

Critical thinking for outlier management

Documentation and reporting skills

10. Next Steps
Immediate:
Share cleaned dataset with team

Present findings to supervisor

Begin exploratory data analysis on cleaned data

Future Improvements:
Automate cleaning pipeline

Create data quality dashboard

Implement validation rules for new data

Appendices
Appendix A: Complete Code Output
text
[Paste key outputs from your Colab notebook]
- df.info() before and after
- Missing value summary
- Duplicate count
- Outlier statistics
Appendix B: Column-wise Changes

| Column | Changes Made | Reason |
|---|---|---|
| [Col1] | [Changes] | [Reason] |
| [Col2] | [Changes] | [Reason] |
| ... | ... | ... |

Appendix C: Quality Check Results
text
Final Quality Check Results:
- Missing values: 0 ✓
- Duplicates: 0 ✓
- Data types: Consistent ✓
- Memory optimized: Yes ✓
- Ready for analysis: Yes ✓
Conclusion
Successfully transformed raw, messy data into a clean, analysis-ready dataset. All data quality issues were identified and resolved using appropriate statistical methods. The cleaned dataset now meets industry standards for data quality and is ready for further analysis and modeling.

Prepared by: Abhishek CD
Date: 23-12-2025