Dataset Analysis Report

Generated: 2026-02-16 13:47:33
Dataset: sample_dataset.csv
Total Records: 15

1. Executive Summary

This report provides a comprehensive analysis of the resume-job matching dataset. The dataset contains **15 records** with **3 columns**. The data quality assessment indicates **Excellent** grade with **100.0%** completeness. The dataset is designed for training machine learning models to match resumes with job descriptions.

2. Dataset Overview

Column Name	Data Type	Missing Values
resume_text	str	0
job_description	str	0
label	int64	0

Class Distribution:

Class	Count	Percentage
1	11	73.33%
0	4	26.67%

3. Data Quality Assessment

Quality Grade: Excellent
Completeness: 100.0%
Missing Cells: 0 out of 45

4. Detailed Text Analysis

resume_text:

- Average Length: 67 characters
- Min Length: 45 characters
- Max Length: 79 characters
- Average Words: 8 words

job_description:

- Average Length: 69 characters
- Min Length: 58 characters
- Max Length: 75 characters
- Average Words: 9 words

5. Storage Format Analysis: CSV vs PDF vs Parquet

CSV (Comma-Separated Values) - Current Format:

✓ Human-readable and easy to view in text editors
✓ Universally compatible across all platforms and tools
✓ Small file size for text-based data storage
✓ Easy to parse and process with any programming language
✓ Good for data exchange between different systems
✓ Efficient for streaming and incremental reads

✗ No native support for complex data types or hierarchies
✗ Delimiter conflicts when data contains commas
✗ No built-in compression - larger files than binary formats
✗ Slower performance for very large datasets (millions of rows)
✗ No support for formatting or metadata preservation
✗ Requires full rebuild to add columns or modify structure

PDF - Report Format (NOT for data storage):

✓ Professional, stable, formatted documents
✓ Preserves layout and typography consistently
✓ Excellent for distribution and sharing
✓ Supports multimedia (images, links, fonts)

✗ NOT suitable for data storage or analysis
✗ Very difficult to extract and parse structured data from
✗ Large file sizes compared to text formats
✗ Cannot efficiently query or filter data
✗ Data modification requires regeneration
✗ Not designed for machine learning workflows

Parquet - Recommended for Production:

✓ Columnar format optimized for analytics
✓ Excellent compression (50-80% size reduction)
✓ Fast read/write performance, especially for analytics
✓ Native support for complex data types
✓ Supports predicate pushdown for efficient filtering
✓ Industry standard in big data and ML ecosystems

✗ Not human-readable in text editors

✗ Slightly steeper learning curve

✗ Requires specialized libraries to read

✗ Less suitable for data exchange with non-technical users

Recommendation:

For this resume-job matching dataset:
- **CSV:** Keep for data import/export and sharing with non-technical stakeholders
- **PDF:** Use for reports, analysis documentation, and stakeholder presentations (current approach)
- **Parquet:** Recommended for production ML pipelines and large-scale data processing

6. Analysis Methodology

Data Collection & Preparation:
The dataset was loaded from CSV format and validated for completeness. All columns were analyzed for data types, missing values, and statistical properties.

Text Analysis:
For text columns (resume_text and job_description), we calculated:
- Average character length and word count
- Minimum and maximum text lengths
- Distribution analysis

Quality Metrics:
Data quality was assessed using:
- Completeness: (Total Cells - Missing Cells) / Total Cells × 100
- Class Balance: Ratio of majority class to minority class
- Missing Value Analysis: Identification of null/empty cells

Format Comparison:
Storage formats were evaluated based on use cases: data storage, analytics, performance, compatibility, and ease of use. Each format was rated for production ML workflows.