NETFLIX

Netflix Data Cleaning Preprocessing

A Complete Data Analysis Internship Task

8,807 Records | 12 Columns | Excel Processing

NETFLIX

Project Overview

OBJECTIVE

Clean and prepare Netflix dataset with nulls, duplicates, and inconsistent formats

DATASET SOURCE

Netflix Movies and TV Shows from Kaggle

PROCESSING TOOL

Microsoft Excel

DATASET SIZE

8,807 rows × 12 columns

DATA QUALITY ISSUES

4,307 missing values, inconsistent formats

DATA TYPES

Mixed (text, dates, numbers, categories)

Initial Data Analysis

NETFLIX

Comprehensive dataset quality assessment

8,807	12	4,307	0
TOTAL ROWS	Columns	MISSING VALUES	DUPLICATE ROWS

📊 Missing Values

Director: 29.91%, Cast: 9.37%, Country: 9.44%

📅 Date Formats

Inconsistent formatting across records

📝 Column Headers

Non-standardized naming conventions

⭐ Rating Values

Invalid entries mixed with duration data

STEP 1: DATA CLEANING

NETFLIX

Handling Missing Values

Comprehensive approach to resolve 4,307 missing data points

Director

2,634

→ Filled with "Not Available"

Cast

825

→ Filled with "Not Available"

Country

831

→ Filled with "Unknown"

Date Added

10

→ Filled with "Unknown"

Result: All 4,307 missing values successfully handled

STEP 2: DATA CLEANING

NETFLIX

Duplicate Removal

Advanced duplicate detection and removal process

Method Used

drop_duplicates()

Python function for duplicate detection

Duplicates Found

0

No duplicate rows detected

Action Taken

None

No rows needed removal

Final Row Count

8,807

All rows maintained

Result: Dataset integrity maintained - no duplicates found

STEP 3: DATA CLEANING

NETFLIX

Text Standardization

Standardizing categorical text fields for consistency

Type Column

Standardized to Title Case

Before: movie, tv show

After: Movie, TV Show

Country Names

Removed extra spaces and standardized separators

Before: "United States , UK , Canada"

After: "United States, UK, Canada"

Rating Values

Fixed invalid entries and consistent format

Before: "74 min", "84 min", "Not Rated"

After: "Not Rated", "Not Rated", "Not Rated"

Result

Clean, consistent categorical text across all fields

STEP 4: DATA TRANSFORMATION

NETFLIX

Date Format Conversion

Standardizing date formats for consistent analysis

Original Format

Month DD, YYYY

```
September 25, 2021
```

New Format

DD-MM-YYYY

```
25-09-2021
```

Method

datetime conversion with standardized format

Benefits

Enables reliable sorting, filtering, and time-based analysis

Result: Consistent date format for all 8,807 records

STEP 5: DATA CLEANING

NETFLIX

Column Header Standardization

Consistent naming conventions for improved data structure

Approach

Lowercase with underscores
Remove special characters and spaces

Examples

show_id	→	show_id
date_added	→	date_added
release_year	→	release_year

Result

Clean, uniform headers
Programming-friendly naming

Result: Clean, uniform headers with consistent naming conventions

STEP 6: DATA CLEANING

NETFLIX

Data Type Verification

Validating and correcting data types for reliable analysis

Release Year

Integer

Verified as whole number format

Date Added

Standardized

Converted to consistent date format

Text Columns

String

Enhanced string formatting and trimming

Outcome: All critical data types validated and corrected

EXCEL WORKBOOK STRUCTURE

NETFLI

Final Excel File Organization

Comprehensive 7-sheet workbook with complete project documentation

Project Info

Complete project documentation and methodology

1
SHEET

Original Data

Uncleaned dataset with all original values

8,807
ROWS

Cleaned Data

Final cleaned dataset ready for analysis

8,807
ROWS

Data Quality Report

Metrics and statistics for data quality assessment

1
REPORT

Missing Values Analysis

Before/after comparison of missing data

1
ANALYSIS

Summary of Changes

Detailed log of all data cleaning modifications

1
LOG

Column Information

Data types and metadata for all columns

1
METADATA

Total Sheets

Complete documentation in one workbook

7
SHEETS

DATA QUALITY IMPROVEMENTS

NETFLIX

Before vs After Comparison

Comprehensive transformation of data quality metrics

Before Cleaning

Missing Values	4,307
Inconsistent Formats	Multiple issues
Invalid Ratings	Present

After Cleaning

Missing Values	0
Standardized Formats	100%
Data Quality	Significantly Improved

Result: Complete transformation from messy to clean, analysis-ready dataset

KEY LEARNINGS

NETFLIX

Key Learnings

Critical insights from the data cleaning process

Missing Data - Requires clear, consistent policies for handling null values to maintain data integrity

Standardization - Ensures consistency across all fields, enabling reliable analysis and comparisons

Date Formatting - Critical for time-series analysis and chronological sorting of data

Column Naming - Improves readability and makes data more accessible for analysis and coding

Data Quality Checks - Essential to perform before any analysis or modeling to ensure reliable results

Data Validation - Critical step to verify data types and ensure data consistency across all records

TOOLS & TECHNIQUES

NETFLIX

Tools & Techniques Used

Professional data cleaning with Excel and Python

Tools

Microsoft Excel

via Python openpyxl library for Excel file creation

Techniques

- fillna() for missing values
- drop_duplicates() for duplicates
- String manipulation for standardization
- datetime conversion for dates
- Column renaming for consistency
- Data type validation

Result: Professional data cleaning with comprehensive documentation

PROJECT DELIVERABLES

NETFLIX

Netflix Data Cleaning Project Results

Comprehensive data processing and quality improvement

Cleaned Dataset

Complete dataset with all data quality issues resolved

Records: 8,807

Comprehensive Excel Workbook

Complete project documentation with detailed analysis

Sheets: 7

Data Quality Report

Detailed metrics and statistics for data quality assessment

Quality Score: 100%

Missing Values Analysis

Before/after comparison of missing data handling

Issues Resolved: 4,307

Complete Documentation

Professional documentation of all cleaning processes

Documentation: Complete

All deliverables completed successfully

NETFLIX

Project Completed Successfully

Netflix Data Cleaning & Preprocessing Project

Chandra Prakash Choudhary

Data Analyst

GitHub

github.com/ChandraPrakash6846

LinkedIn

linkedin.com/in/chandra-prakash-choudhary-17b96b212/