Task 1: Excel Basics – Data Cleaning & Formatting

Tool Used:

Microsoft Excel

Dataset Used:

Netflix Movies and TV Shows Dataset

Steps Performed

Data Import

Imported the dataset into Microsoft Excel.

Verified column headers and data types.

Handling Missing Values

Identified blank cells using filters.

Filled missing values where possible (e.g., Country, Rating).

Removed rows with critical missing data.

Removing Duplicates

Used Data → Remove Duplicates to delete duplicate records.

Text Cleaning

Removed extra spaces using TRIM() function.

Standardized text case (proper case for titles and names).

Date Formatting

Converted date columns to proper Date format.

Ensured consistency in date values.

Column Formatting

Renamed column headers for clarity.

Adjusted column width and alignment.

Applied bold formatting to headers.

Data Validation

Checked for incorrect or inconsistent values.

Ensured uniform data across all columns.

Final Formatting

Applied table formatting for better readability.

Used filters and freeze header row.

Outcome

The dataset is now clean, structured, and analysis-ready.

Missing values and duplicates were handled properly.

Data is consistently formatted and easy to understand.

Conclusion:

This task improved data quality by applying essential Excel data cleaning and formatting techniques, making the dataset suitable for further analysis and reporting.

If you want, I can also:

Make it shorter

Convert it to handwritten-style answer

Customize it for college / internship submission

CODE:

```
import pandas as pd


# Load dataset
df = pd.read_csv("netflix_titles.csv")
```

```python
# Display initial info
print("Initial Dataset Info:")
print(df.info())


# 1. Handle missing values
df['country'].fillna('Unknown', inplace=True)
df['rating'].fillna('Not Rated', inplace=True)
df.dropna(subset=['title', 'type'], inplace=True)


# 2. Remove duplicate rows
df.drop_duplicates(inplace=True)


# 3. Text cleaning
df['title'] = df['title'].str.strip()
df['director'] = df['director'].str.strip()


# 4. Date formatting
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')


# 5. Standardize text case
df['type'] = df['type'].str.title()
df['country'] = df['country'].str.title()


# 6. Rename columns for clarity
df.rename(columns={
    'release_year': 'Release_Year',
```

```python
    'date_added': 'Date_Added'
}, inplace=True)


# 7. Final dataset info
print("\nCleaned Dataset Info:")
print(df.info())


# Save cleaned data
df.to_csv("netflix_cleaned.csv", index=False)


print("\nData cleaning completed and saved as netflix_cleaned.csv")
```