Data Quality Assignment 1:

Messy Dataset Analysis

Assignment Overview

You work as a junior data engineer tasked with analyzing a messy dataset to identify and report data quality issues. Your goal is to write Python functions that can detect common data quality problems that decrease data veracity.

Learning Objectives

By completing this assignment, you will:

- Identify common data quality problems in real-world datasets
- Write Python functions to detect missing data, duplicates, and outdated records
- Create data quality reports for business stakeholders
- Practice data cleaning techniques using pandas

Dataset Selection

Choose ONE dataset from the provided "Messy Datasets / Dirty Datasets" resource file. Recommended datasets for beginners:

- **Cafe Sales - Dirty Data** (easiest, good for first-time data cleaning)
- **Dirty Sales Data Sample** (moderate difficulty)
- **Music Tours Dataset** (moderate difficulty)

Download your chosen dataset and save it in the same folder as your Python file.

Required Tasks

Choose ONE task:

Task 1: Missing Data Detection (30 points)

Write a function that identifies and counts missing data in your dataset.

Requirements:

- Load the dataset using pandas
- Identify all columns with missing values (empty strings, NaN, None, etc.)
- Count missing values per column
- Calculate percentage of missing data per column
- Identify rows with the most missing data

Task 2: Duplicate Detection

Write a function that finds and analyzes duplicate records.

Requirements:

- Find exact duplicate rows
- Find potential duplicates based on key identifying columns (e.g., name, email, ID)
- Count how many duplicates exist
- Show examples of duplicate records

Task 3: Outdated/Problematic Data Detection (35 points)

Write a function that identifies problematic data based on your dataset's context.

Requirements: Choose ONE approach based on your dataset:

- **If your dataset has dates**: Find outdated records (old dates, future dates where inappropriate)
- **If your dataset has numeric data**: Find outliers and abnormal values
- **If your dataset has categorical data**: Find inconsistent formatting or invalid categories

Final Deliverable: Data Quality Report

Create a comprehensive report summarizing your findings.

Requirements:

- Summary statistics (total records, columns analyzed)
- **Missing data summary (which columns, how much missing) [if applicable]**
- **Duplicate analysis (how many, examples) [if applicable]**
- **Problematic data findings (outliers, inconsistencies, etc.) [if applicable]**
- Overall data quality score (your own calculation)
- Recommendations for data cleaning

Submission Requirements

Files to Submit:

1. **`data_quality_analysis.py`** - your Python file with all functions
2. **`quality_report.(doc/pdf)`** - data quality report

Code Requirements:

- Use pandas for data manipulation
- Include clear comments explaining your logic
- Handle different types of missing data (NaN, empty strings, None)
- Use appropriate data types and methods
- Test your functions and show outputs

Report Requirements:

Your quality report should answer:

- What percentage of the data is missing?
- How many duplicate records exist?
- What are the main data quality issues?
- Which columns have the most problems?
- What would you recommend to fix these issues?

Tips for Success

Pandas Essentials:

```
# Check for missing data
df.isnull().sum()
df.info()

# Find duplicates
df.duplicated()
df.drop_duplicates()

# Basic statistics
df.describe()
df.value_counts()
```

```
# Data types
df.dtypes
df.astype()
```

Common Data Quality Checks:

- **Email validation**: Check for proper email format
- **Date validation**: Ensure dates are reasonable (not in future, not too old)
- **Numeric ranges**: Check if numbers fall within expected ranges
- **Text consistency**: Look for variations in spelling/formatting
- **Required fields**: Identify which fields should never be empty

---

Due Date: 12/8/2025
Dataset Resource: See "Messy Datasets/ Dirty Datasets" file
Questions? Contact zaid.momani@uop.edu.jo