Homework 1: Introduction to Data Processing

Points: 20 | **Due:** Sunday, February 2, 2026 @ 11pm Pacific

---

Learning Objectives

1. **Connect** to real-world data sources using Python
2. **Load and explore** datasets using Pandas
3. **Assess data quality** by identifying missing values, duplicates, and outliers
4. **Discover and communicate** an interesting finding from data

---

Grading

Component	Points	What We're Looking For
Data Connection	5	Load data from any source into a DataFrame
Data Exploration	5	Answer questions about dataset structure
Quality Assessment	5	Analyze missing values, duplicates, outliers
Interesting Finding	5	One insight with evidence and business relevance
Total	**20**	

Instructions

1. Open `MIS769_HW1_Data_Processing.ipynb` in Google Colab
2. Choose a data source (HuggingFace, Kaggle, or your own CSV)
3. Run the exploration code and answer questions in markdown cells
4. Complete the data quality assessment
5. Find something interesting and write up your insight

What Counts as "Interesting"?

- An unexpected distribution ("90% of reviews are 5-stars")
- A surprising correlation ("longer reviews tend to be more negative")
- A pattern ("complaints spike on Mondays")
- An anomaly ("one product has 10,000 reviews")

Submission

Upload to Canvas: - Your completed `.ipynb` notebook with all cells executed - A screenshot of your "Interesting Finding" cell