Ajijedidun Oluwaferanmi Assignment

Student Name: Ajijedidun Oluwaferanmi

Matric No: SCI/24/25/0231

Course: DTS 201

Lecturer: Dr. Sakinat Folorunso

Week 1: Exploring Sample Datasets and Data Types

STUDENT TASK 1: CSV Dataset Analysis

1. What does each row represent?

Each row represents a student's study session record showing the relationship between study hours per day and corresponding exam scores achieved.

2. Name at least three columns and their information type:

Since this dataset contains only 2 columns:

- **Hours** - Study hours per day (numerical - float type)
- **Scores** - Exam scores achieved (numerical - integer type)

STUDENT TASK 2: CSV Structure Analysis

1. How many rows and columns does your CSV dataset have?

- Rows: 25 rows
- Columns: 2 columns
- Shape: (25, 2)

2. Which columns are numerical?

- Hours: float64
- Scores: int64
- All columns are numerical

3. Which columns are categorical/text?

None - This dataset contains no categorical or text columns.

STUDENT TASK 3: Excel vs CSV Comparison

1. How are they similar?

- Both are structured data organized in tabular format
- Both can be loaded using pandas library
- Both have defined column headers

2. How are they different?

Aspect	CSV Dataset	Excel Dataset
Content	Student study data	E-commerce sales data
Complexity	Simple (2 columns)	Complex (16+ columns)
Data Types	Only numerical	Mixed types
Size	25 rows, 2 columns	1000+ rows, 16+ columns

STUDENT TASK 4: Excel Structure Analysis

1. Are the data types in the Excel dataset mostly similar to those in the CSV file?

No, they are very different. CSV dataset contains only numerical data types while Excel dataset contains mixed data types including text and numerical data.

2. Identify one column that is categorical and one that is numerical:

- **Categorical:** product_name (contains product descriptions)
- **Numerical:** rating (contains product ratings as decimal numbers)

STUDENT TASK 5: Data Type Classification

1. CSV dataset (student_score.csv)

- Type: STRUCTURED
- Justification: Organized in clear rows and columns with defined data types

2. Excel dataset (Sales_data.xlsx)

- Type: STRUCTURED
- Justification: Maintains tabular format with defined columns

3. Folder of .jpg images

- Type: UNSTRUCTURED
- Justification: No fixed format, binary image data

4. WhatsApp chat exports in .txt format

- Type: SEMI-STRUCTURED
- Justification: Contains timestamps and names but irregular text content

5. JSON file containing product data

- Type: SEMI-STRUCTURED

- Justification: Uses key-value pairs but not strictly tabular

FINAL REFLECTION

1. What was the most surprising thing you learned about data formats today?

The significant complexity difference between academic and commercial datasets. The student score CSV was clean with just 2 columns, but the sales Excel file had 16+ mixed-type columns, demonstrating the messiness of real-world business data.

2. Which format would you prefer for a beginner data scientist and why?

CSV format would be preferable for beginners because it's simpler, has fewer data type complications, and is easier to understand and manipulate. The straightforward structure helps focus on learning core concepts without getting overwhelmed by data complexity.

Week 2: Data Collection, Cleaning, and Quality Assessment

STUDENT TASK: Dataset Analysis (CSV, Excel, TXT)

For each dataset (CSV, Excel, TXT):

1. What the rows represent:

- **CSV Dataset:** Each row represents a product listing with customer reviews and ratings from an e-commerce platform
- **Excel Dataset:** Each row represents the same product data but with some formatting differences and potential data type variations

- **TXT Dataset:** Each row represents identical product information but stored in tab-delimited text format

2. At least 3 columns identified:

- **product_name:** Text/String - Contains product descriptions and titles
- **rating:** Numerical - Product ratings (decimal values like 4.2, 3.9)
- **discounted_price:** Text/String - Price with currency symbol (₹399, ₹199)
- **actual_price:** Text/String - Original price with currency and formatting
- **category:** Text/String - Product categories separated by pipes
- **rating_count:** Mixed - Number of ratings (some with commas, some without)

3. Cleanest vs Messiest Dataset:

- **Cleanest:** CSV Dataset - Most consistent data types and formatting
- **Messiest:** Excel Dataset - Contains mixed data types, some numerical columns stored as text, and inconsistent formatting in price columns (some show as decimals, others as text with currency symbols)

STUDENT REFLECTION: Data Quality Issues

2 Data Quality Issues Identified:

Issue 1: Missing Values in rating_count column

- Problem: 2 missing values found in the rating_count column
- Analysis Impact: Missing rating counts make it impossible to assess the reliability of product ratings. A 4.5-star rating based on 2 reviews is very different from one based on 10,000 reviews

Issue 2: Inconsistent Data Types for Price Columns

- Problem: Price columns (discounted_price, actual_price) stored as text with currency symbols and inconsistent formatting
- Analysis Impact: Cannot perform numerical operations like calculating average prices, finding price ranges, or computing actual discount amounts without extensive cleaning

STUDENT TASK: Data Cleaning Applied

Three Cleaning Techniques Applied:

1. Handling Missing Values

- Technique: Filled missing rating_count values with median
- Reason: Median is less affected by outliers than mean, providing a reasonable estimate for missing rating counts

2. Data Type Conversion

- Technique: Converted price columns from text to numeric by removing currency symbols and commas
- Reason: Enables mathematical operations and statistical analysis on price data

3. Standardizing Text Data

- Technique: Standardized category labels by converting to lowercase and removing extra spaces
- Reason: Ensures consistent categorization and prevents duplicate categories due to case differences

FINAL REFLECTION

Which cleaning step was most useful and which dataset needed most work?

Most Useful Cleaning Step: Converting price columns from text to numeric was essential as it enabled mathematical operations on previously unusable data. This transformation was crucial for conducting price analysis, discount calculations, and generating actionable business insights.

Dataset Needing Most Work: The Excel dataset required the most cleaning due to mixed data types where some numerical columns were stored as text, inconsistent number formatting, and data type confusion between different columns. The TXT dataset was actually cleaner in terms of consistent data types, even though it required delimiter specification.

Key Learning: Commercial datasets are inherently messy and demand extensive preprocessing. Different file formats of identical data can present unique quality challenges, highlighting data validation and cleaning as fundamental data science competencies.

Week 3: Exploratory Data Analysis (EDA) and Basic Data Visualization

STUDENT TASK 1: Dataset Structure Analysis

What does one row represent?

Each row represents a complete product listing from an e-commerce platform, including product details, pricing information, customer ratings, and associated customer reviews.

Five columns and their representations:

- **product_name:** Text description of the product being sold
- **rating:** Numerical customer rating score (1-5 scale)
- **discounted_price:** Current selling price with discount applied
- **actual_price:** Original price before discount
- **category:** Product classification hierarchy (e.g., Electronics|Accessories)

STUDENT TASK 2: Summary Statistics Analysis

Two numerical columns analyzed: rating and discount_percentage

Rating Column:

- Mean: 4.1

- Median: 4.2
- Standard Deviation: 0.3

Discount Percentage Column:

- Mean: 0.58 (58%)
- Median: 0.61 (61%)
- Standard Deviation: 0.22

Interpretation:

- **Rating:** Mean and median are very close (4.1 vs 4.2), indicating a symmetric distribution. Low standard deviation (0.3) shows ratings are clustered around 4.0-4.2, suggesting generally satisfied customers.
- **Discount:** Mean slightly lower than median suggests some products with very low discounts. Higher standard deviation (0.22) indicates more variability in discount strategies across products.

STUDENT TASK 3: Categorical Data Analysis

Category Column Analysis (Top 3 most frequent):

- **Computers&Accessories|Accessories&Peripherals|Cables:** 892 products
- **Electronics|Mobiles&Accessories:** 234 products
- **Home&Kitchen|Kitchen&HomeAppliances:** 156 products

Distribution Assessment:

The distribution is **highly imbalanced**. Computer accessories (especially cables) dominate with over 60% of all products, while other categories have much smaller representation. This imbalance suggests the dataset may be from a technology-focused retailer or represents a specific product category focus, which could bias any analysis toward tech product patterns.

STUDENT TASK 4: Bar Chart Analysis

Analysis of product categories bar chart:

- **Most Common:** Computer cables and accessories dominate the dataset with overwhelming frequency
- **Rare Categories:** Home appliances, books, and clothing categories appear very infrequently
- **Observation:** The extreme dominance of one category suggests this might be a specialized electronics retailer rather than a general marketplace

STUDENT TASK 5: Histogram Analysis

Rating distribution histogram analysis:

- **Distribution Shape:** Left-skewed (negatively skewed) with most ratings concentrated between 4.0-4.5
- **Peak:** Single peak around 4.2, indicating most products receive good ratings
- **Outliers:** Very few products with ratings below 3.0, suggesting either quality control or potential rating bias
- **Interpretation:** The concentration of high ratings might indicate customer satisfaction, but could also suggest rating inflation or selection bias in the dataset

STUDENT TASK 6: Scatter Plot and Correlation Analysis

Variables analyzed: discount_percentage vs rating

Correlation coefficient: -0.127

Interpretation:

- **Relationship:** Weak negative correlation
- **Meaning:** Products with higher discounts tend to have slightly lower ratings, but the relationship is very weak

- **Scatter plot support:** The plot shows a slight downward trend but with high variability, confirming the weak correlation
- **Business insight:** Heavy discounting might indicate products that are harder to sell due to lower quality or customer satisfaction

MINI EDA EXERCISE: E-commerce Sales Dataset Analysis

Summary Tables Generated:

- **Numerical Summary:** df.describe() revealed 1,465 products with ratings averaging 4.1 and discount percentages averaging 58%
- **Categorical Summary:** value_counts() showed extreme category imbalance with cables dominating 60% of listings

Three Plots Created:

- **Bar Chart:** Product categories showing overwhelming dominance of computer accessories
- **Histogram:** Rating distribution revealing left-skewed pattern with most products rated 4.0+
- **Scatter Plot:** Discount vs Rating showing weak negative relationship

Key Findings Summary:

The dataset represents a technology-focused e-commerce platform with high customer satisfaction (average 4.1 rating) and aggressive pricing strategies (average 58% discount). The extreme category imbalance suggests specialization in computer accessories, particularly cables. The weak negative correlation between discounts and ratings hints that heavily discounted items may have quality concerns, though the relationship is not strong enough to be definitive.

FINAL REFLECTION

Most Useful Plot Type:

The histogram was most useful because it revealed the true distribution shape of numerical variables, particularly exposing the left-skewed nature of ratings that wasn't apparent from summary statistics alone.

While the mean and median were close, the histogram showed the concentration of high ratings and scarcity of low ratings, providing crucial insights into customer behavior patterns.

Key Insight Discovered:

The most surprising discovery was the extreme category imbalance - over 60% of products being computer cables and accessories. This wasn't apparent from initial data loading but became obvious through bar chart visualization. This insight completely changed my understanding of the dataset from a general e-commerce platform to a specialized technology retailer, affecting how I interpret all other patterns.

EDA Skills for Future Modeling:

These EDA skills will be crucial for modeling because they reveal data quality issues (like category imbalance) that could bias machine learning models, identify which variables have predictive potential (like the discount-rating relationship), and help in feature engineering decisions. Understanding distribution shapes helps choose appropriate algorithms, while correlation analysis guides feature selection. Most importantly, EDA prevents building models on incorrect assumptions about the data structure and relationships.

Week 4: Distributions, Probability and Simulation

STUDENT TASK 1: Coin Toss Simulation Analysis

Simulation Results for Different Sample Sizes:

n_tosses = 50:

- Heads: 0.52 (26 tosses)
- Tails: 0.48 (24 tosses)

n_tosses = 5000:

- Heads: 0.4998 (2499 tosses)

- Tails: 0.5002 (2501 tosses)

Comparison with Theoretical Probabilities:

Yes, the relative frequencies get much closer to the theoretical probability of 0.5 as n_tosses increases. With 50 tosses, we see noticeable deviation (0.52 vs 0.48), but with 5000 tosses, the frequencies are extremely close to 0.5. This demonstrates the Law of Large Numbers - as sample size increases, empirical probabilities converge to theoretical probabilities.

STUDENT TASK 2: Dice Roll Analysis

Relative Frequencies vs Theoretical Probability (≈ 0.1667):

- Face 1: 0.164 (328 rolls)
- Face 2: 0.171 (342 rolls)
- Face 3: 0.159 (318 rolls)
- Face 4: 0.168 (336 rolls)
- Face 5: 0.173 (346 rolls)
- Face 6: 0.165 (330 rolls)

Analysis:

Yes, some faces are slightly more frequent than others in the simulation. This is completely expected due to random variation. Even with a fair die, we don't expect exactly equal frequencies in any finite sample. The variations we see (ranging from 0.159 to 0.173) are well within normal random fluctuation around the theoretical 0.1667. With more rolls, these frequencies would converge closer to 1/6.

STUDENT REFLECTION 3: Law of Large Numbers

Observation:

As the number of rolls increases, the estimated probability of rolling a 6 converges toward the theoretical probability of 1/6 (≈ 0.1667). With small samples (10-50 rolls), the estimated probability fluctuates significantly, but with larger samples (1000-5000 rolls), it stabilizes very close to 1/6.

Law of Large Numbers Support:

This perfectly demonstrates the Law of Large Numbers, which states that as the number of trials increases, the empirical probability approaches the theoretical probability. The plot shows the estimated probability oscillating around 1/6 with decreasing amplitude as sample size grows, eventually converging to the true value.

STUDENT TASK 4: Normal Distribution Analysis

Parameter Changes (mean = 10, std_dev = 2):

- **Position:** The entire curve shifted right from center 0 to center 10
- **Shape:** The curve became wider and flatter due to increased standard deviation
- **Height:** The peak became lower to maintain the same total area under the curve

Standard Deviation Effect:

Increasing the standard deviation from 1 to 2 makes the distribution more spread out. The curve becomes wider and flatter, meaning values are more dispersed around the mean. A larger standard deviation indicates greater variability in the data, while a smaller standard deviation indicates values cluster more tightly around the mean.

STUDENT TASK 5: Binomial Distribution Analysis

Distribution Shape Changes:

p_success = 0.2:

- Distribution is right-skewed (positively skewed)
- Peak occurs at low values (0-2 successes)
- Most experiments result in few successes

p_success = 0.8:

- Distribution is left-skewed (negatively skewed)
- Peak occurs at high values (8-10 successes)
- Most experiments result in many successes

Pattern:

As probability of success increases, the distribution shifts from right-skewed to left-skewed. When p < 0.5, the distribution peaks at low values; when p > 0.5, it peaks at high values. When p = 0.5, the distribution is symmetric around the middle.

STUDENT TASK 6: Uniform Distribution Analysis

Parameter Changes (low = -3, high = 5):

- The distribution shifted and expanded to cover the range -3 to 5
- The height decreased to maintain constant density across the wider interval
- The flat, rectangular shape remained the same

Comparison with Normal Distribution:

The main difference is shape: uniform distribution is completely flat (rectangular) with equal probability density across the entire interval, while normal distribution is bell-shaped with highest density at the center and decreasing density toward the tails. Uniform distribution has sharp cutoffs at the boundaries, while normal distribution extends infinitely in both directions with gradually decreasing probability.

MINI SIMULATION EXERCISE: Exam Score Modeling

Real-World Scenario:

Modeling final exam scores for a statistics course with 200 students, where the instructor expects an average score of 75 with a standard deviation of 12 points.

Distribution Choice Justification:

Normal distribution is appropriate because exam scores typically follow a bell curve - most students score around the average, with fewer students at the extremes (very high or very low scores). This reflects natural variation in student ability and preparation.

Simulation Parameters:

- Mean: 75 points
- Standard deviation: 12 points
- Sample size: 200 students

Results Interpretation:

The simulation generated realistic exam scores with most students scoring between 63-87 (within one standard deviation of the mean). The histogram shows the expected bell curve shape, with approximately 68% of scores falling within one standard deviation of 75. A few students scored exceptionally high (>95) or low (<55), which is realistic for any exam. This simulation could help instructors set grade boundaries and predict score distributions before administering the actual exam.

FINAL REFLECTION

Easiest vs Most Challenging:

The coin toss simulation was easiest to understand because it directly mirrors a familiar real-world activity with clear, intuitive outcomes. The normal distribution was most challenging initially because it involves continuous probability density rather than discrete counts, and the concept of probability density (rather than just probability) required more abstract thinking.

Simulation vs Formulas:

Simulation provides intuitive understanding by letting you "see" probability in action through repeated trials and visual patterns. While formulas give exact theoretical values, simulation shows how randomness actually behaves, demonstrates the Law of Large Numbers in practice, and makes abstract concepts like "probability density" tangible through histograms. The visual feedback from thousands of simulated trials builds intuition that pure mathematical formulas cannot provide.

Future Applications in Data Science:

These simulation skills are fundamental for data science modeling and hypothesis testing. Understanding distributions helps choose appropriate statistical tests, simulation enables bootstrap sampling and Monte

Carlo methods for uncertainty quantification, and probability concepts underpin machine learning algorithms. In hypothesis testing, we simulate null distributions to calculate p-values. For modeling, understanding how different distributions behave helps select appropriate models and interpret results correctly.