EDA Project Guidelines and Dataset Assignments for Day 7 Presentation

Prepared for Beginner Students

June 2025

Introduction

This document provides guidelines for your Day 7 Exploratory Data Analysis (EDA) project presentation. Each of the five teams (Team A, Team B, Team C, Team D, Team E) is assigned a beginner-friendly dataset to analyze. You will apply the EDA skills learned on Days 4–6 to explore your dataset, create visualizations, and present your findings in a 5–7 minute slide presentation. The document includes dataset descriptions, EDA instructions, and tips for creating appealing slides.

1 Dataset Assignments

Each team is assigned a dataset accessible via Python's `seaborn` or `sklearn.datasets` libraries. Below are the datasets and their key characteristics.

Team 1: Netflix Original Films and IMDB Scores Dataset

Access: https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores/data

Team 2: Superstore Sales Dataset

Access: https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting

Team 3: Data Science Salaries 2023 Dataset

Access: https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023

Team 4: Pokemon Dataset

Access: https://www.kaggle.com/datasets/abcsds/pokemon

Team 5: Video Game Sales Dataset

Access: https://www.kaggle.com/datasets/gregorut/videogamesales

Team 6: IBM HR Analytics Dataset

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Team 7: Anime Recommendation Dataset

https://www.kaggle.com/datasets/hernan4444/anime-recommendation-database-2020

2 EDA Guidelines

Your team should perform a complete EDA on your assigned dataset, using the skills from Days 4–6. Follow these steps and include them in your presentation.
Note!!!!!! This is just a guidelines you can add more steps on your own.

2.1 Step 1: Data Profiling

- Use `df.info()` to check columns, data types, and non-null counts.
- Use `df.head()` to view the first few rows.
- Use `df.describe()` for numerical feature statistics.
- Use `df.value_counts()` for categorical feature distributions.

2.2 Step 2: Data Quality Checks

- Check for missing values with `df.isnull().sum()`.
- Handle missing values (e.g., fill with median for numerical, mode for categorical, or drop columns with excessive missing data).
- Check for duplicates with `df.duplicated().sum()` and remove if necessary.
- Verify data types (e.g., convert strings to categories if needed).

2.3 Step 3: Visualizations

Create at least 3–4 visualizations to explore your data:

- **Univariate**: Histograms (`sns.histplot`) or count plots (`sns.countplot`) for single features.
- **Bivariate**: Scatter plots (`sns.scatterplot`) or box plots (`sns.boxplot`) to show relationships.
- **Correlation**: Heatmap (`sns.heatmap`) for numerical features.
- **Distribution**: KDE plots (`sns.kdeplot`) or violin plots (`sns.violinplot`) for comparisons.

2.4 Step 4: Outlier Detection and Handling

- Use box plots to identify outliers in numerical features.
- Handle outliers (e.g., remove using IQR method or cap values).
- Example:

```
q1 = df['column'].quantile(0.25)
q3 = df['column'].quantile(0.75)
iqr = q3 - q1
df = df[(df['column'] >= q1 - 1.5 * iqr) & (df['column'] <=
    q3 + 1.5 * iqr)]
```

2.5 Step 5: Feature Engineering

- Create at least one new feature (e.g., ratios, binning, or combining features).
- Visualize the new feature to show its impact.

2.6 Step 6: Summarize Findings

- Write 3–5 key insights (e.g., "Petal length strongly correlates with species in the Iris dataset").
- Include statistical summaries (e.g., mean, median) and visualization insights.

3 Presentation Guidelines

Your presentation should be 5–7 minutes long and use slides (e.g., PowerPoint, Google Slides). Follow these tips to create an appealing and effective presentation.

3.1 Slide Design Tips

- **Minimal Text**: Use bullet points or short sentences (3–5 per slide). Avoid paragraphs.
- **Clear Visuals**: Include 3–4 high-quality visualizations (e.g., histograms, scatter plots). Ensure plots have clear titles, labels, and legends.
- **Consistent Design**: Use a clean template with readable fonts (e.g., Arial, size 24+ for text, 28+ for titles). Choose a light background with dark text for readability.
- **Logical Structure**: Organize slides as follows:
  1. Title slide: Team name, dataset, and date.
  2. Introduction: Brief dataset description and EDA goals.
  3. Data Profiling: Key dataset characteristics (rows, columns, types).
  4. Data Quality: Missing values, duplicates, and how you handled them.
  5. Visualizations: Show and explain 3–4 plots.
  6. Feature Engineering: Describe new features and their insights.
  7. Key Findings: Summarize 3–5 insights.
  8. Conclusion: Recap and any challenges faced.
- **Engage the Audience**: Explain plots clearly, e.g., "This scatter plot shows a strong correlation between X and Y."

3.2 Presentation Delivery

- **Practice**: Rehearse to stay within 5–7 minutes.
- **Teamwork**: Each team member should present at least one section.
- **Clarity**: Speak clearly and avoid jargon. Explain technical terms (e.g., "Correlation measures how two variables move together").
- **Visual Focus**: Point to plots when discussing them to guide the audience.

4 Tips for Success

- **Collaborate**: Work as a team to divide tasks (e.g., one member handles cleaning, another visualizations).
- **Explore Creatively**: Try different visualizations to uncover unique insights.
- **Document Everything**: Keep a notebook with your code and findings to reference during the presentation.
- **Seek Help**: Ask the instructor if you encounter issues with the dataset or code.