Task 1: Comparing Visualizations

COURSE NAME

CSDS 413 Introduction to Data Analysis

Authors:

Jacob Anderson
Wiam Skakri

September 18, 2025

Contents

Context

In data science, the choice of visualization plays a critical role in shaping how insights are derived and communicated. Different visual encodings can highlight or obscure structure in data — including spread, skew, outliers, modality, or differences between categories. In this task, you will explore how three different types of visualizations can be used to compare distributions across categories, and evaluate which is most appropriate depending on the dataset context.

For this task, you are provided with three datasets, each containing categorical grouping variables and a numerical measurement. Each dataset comes with a research scenario/question. Your task is to clean the data, visualize the distribution across categories using multiple plotting techniques, and discuss which visualization is the most appropriate in addressing the research question for each dataset.

1 Best-Selling Albums Dataset

Attributes: Year, Ranking, Artist, Album, Genre, Worldwide Sales, Tracks, Album Length

Scenario: A media analytics firm is interested in understanding whether certain genres consistently produce top-selling albums or if success is more scattered across genres.

Research Question: How does the distribution of album sales vary across music genres for albums in the previous decade (released after 2015), and are high-sales outliers concentrated in certain genres?

1.1 Part A: Data Cleaning and Preprocessing

First, filter your dataset so that only the variables critical for your analysis remain. Then clean your data so that there is consistency in variable types, capitalization, and handle any missing or invalid values.

1.2 Part B: Generate Three Visualizations

Produce the following types of plots:

- **Error Bar Plot:** Show the mean and variability (e.g., standard error or 95% confidence intervals) of the numerical variable across each category.
- **Barcode Chart:** Also known as a strip plot or rug plot. Shows individual data points across categories.
- **Histogram:** Plot the distribution of the numerical variable, grouped by the categorical variable (using hue or facet).

1.3 Part C: Evaluate and Justify Visualization

For the dataset:

- Discuss the advantages and disadvantages of each visualization type.
- Decide which visualization is best for the research question.
- Support your answer with evidence from the plots and reasoning based on dataset size, shape, or structure.

2 Anime Dataset

Attributes: Rank, Name, Japanese_name, Type, Episodes, Studio, Release_season, Tags, Rating, Release_year, End_year, Description, Content_Warning, Related_Mange, Related_anime, Voice_actors, staff

Scenario: A streaming service is considering expanding its short anime series catalog ($<$ 25 episodes) and wants to understand how viewer ratings differ between anime TV series and movies released after 2015. The goal is to determine which format generally receives better audience reception to inform licensing and promotion strategies.

Research Question: How do audience ratings compare between anime TV series and movies released after 2015, and which format generally receives higher ratings?

2.1 Part A: Data Cleaning and Preprocessing

First, filter your dataset so that only the variables critical for your analysis remain. Then clean your data so that there is consistency in variable types, capitalization, and handle any missing or invalid values.

2.2 Part B: Generate Three Visualizations

Produce the following types of plots:

- **Error Bar Plot:** Show the mean and variability (e.g., standard error or 95% confidence intervals) of the numerical variable across each category.
- **Barcode Chart:** Also known as a strip plot or rug plot. Shows individual data points across categories.
- **Histogram:** Plot the distribution of the numerical variable, grouped by the categorical variable (using hue or facet).

2.3 Part C: Evaluate and Justify Visualization

For the dataset:

- Discuss the advantages and disadvantages of each visualization type.
- Decide which visualization is best for the research question.
- Support your answer with evidence from the plots and reasoning based on dataset size, shape, or structure.

3 Algorithm Performance Dataset

Attributes: Algorithm, Epoch, Accuracy, Trial Number

Scenario: You are testing two reinforcement learning (RL) algorithms on a sequential decision task. To avoid overfitting and simulate real-world noise, you shuffle the dataset for each trial and run 10 independent trials per algorithm. For each trial, you track the accuracy across 10 training epochs (one pass through a dataset). Due to how you shuffle your data and algorithmic stochasticity, accuracy results vary across trials.

Research Question: Which algorithm performs more accurately on average across epochs, and how does the use of a visualization help you assess reliability and variation of each algorithm?

3.1 Part A: Data Cleaning and Preprocessing

First, filter your dataset so that only the variables critical for your analysis remain. Then clean your data so that there is consistency in variable types, capitalization, and handle any missing or invalid values.

3.2 Part B: Generate Three Visualizations

Produce the following types of plots:

- **Error Bar Plot:** Show the mean and variability (e.g., standard error or 95% confidence intervals) of the numerical variable across each category.
- **Barcode Chart:** Also known as a strip plot or rug plot. Shows individual data points across categories.
- **Histogram:** Plot the distribution of the numerical variable, grouped by the categorical variable (using hue or facet).

3.3 Part C: Evaluate and Justify Visualization

For the dataset:

- Discuss the advantages and disadvantages of each visualization type.
- Decide which visualization is best for the research question.
- Support your answer with evidence from the plots and reasoning based on dataset size, shape, or structure.