Design Document: Anime Data Visualization

Introduction

The analysis focuses on the visualization of data collected from the raw_anime.csv database, which spans from 1970 to 2019 and contains score, genre, episodes, duration, and type information. The research aims to reveal which characteristics drive popularity and ratings in order to benefit both consumers and production teams as well as scientific researchers. The preprocessing Python script processes data to create seven CSV files from the initial 15,278 records and reduces them to 7,125 valid entries. The visualizations showcase a dashboard (dashboard.json) with heatmaps, bar charts, scatter plots, and boxplots together with separate charts like Genre popularity trends over the years.json, bar graph for epi duration.json, score by anime type.json, and genre ratings over time.json. The visualizations display genre performance alongside runtime-trends through encoded interaction elements that enable users to explore different patterns.

Preprocessing

The dataset cleanup and consolidation processes occur in the Python notebook document labeled assignment 2 in python.ipynb.

Assessment and numerical conversion, along with year extraction through regular expressions (regex), as well as genre string list processing and duration transformation from verbal statements into numeric values (such as "1 hr 30 min" converting to 90 min), complete the cleaning procedures. The total runtime amount is determined by multiplying the number of episodes by their corresponding duration.

The filter operation kept data points with score validity and time ranges between 1970 and 2020, which produced 7,125 records. The runtime visualizations have a limit of 5,000 minutes (6,878 records) to avoid the inclusion of extreme values.

•Aggregation: Produces seven CSVs:

- genre_trends_over_time.csv: Yearly percentages for top 10 genres.
- average_genre_ratings_over_time.csv: Yearly genre score averages.
- average_genre_ratings_by_decade.csv: Decadal genre score averages.
- score_by_episode_duration.csv: Scores by duration bins (<15, 15–24, 25–34, 35–60 min).
- score_by_anime_type.csv: Scores by type (TV, movie, OVA).
- score_distribution_by_episode_count.csv: Scores by episode bins (1, 2–6, etc.).
- runtime_vs_score_by_type.csv: Runtime vs. score data. The data preparation methods produce datasets that maintain reliability for visualization purposes.

Chart Descriptions and Encodings

Data encoding works effectively through Vega-Lite within the visualizations.

Dashboard (dashboard.json):

Heatmap (Genre Ratings by Decade): Encodes genres (y-axis, nominal), decades (x-axis, ordinal), and average score (color, quantitative, viridis). Tooltips show score/count.

This bar chart shows average score values through teal tealblues colors while using type as the nominal x-axis variable and the quantitative y-axis variable shows scores. Error bars indicate standard error.

The scatter plot utilizes runtime as its log-scale quantitative x-axis variable and score as its quantitative y-axis variable with a nominal type axis value that uses color as the element identifier. Includes regression line.

The boxplot displays episode ranges on the x-axis along with score distribution shown on the y-axis, which has quantitative values.

Interactions: Dropdowns for genre/type, min count slider, tooltips.

Standalone Charts:

The standalone chart shows the genre popularity trends over the years.json document with nominal category 10 colors at x-axis quantitative year points and y-axis quantitative percentage points. Supports zoom/pan and legend selection.

The bar graph for epi duration. json Presents duration categories in the X-axis and shows quantitative average scores through the Y-axis and viral color inflection. Error bars included.

- Score by Anime Type (score by anime type.json): Similar to the dashboard's type chart, with error bars.
- Genre Ratings Over Time (genre ratings over time.json): Year (x-axis, ordinal), average score (y-axis, quantitative), genre (color, nominal).

The heatmap visualizes two-dimensional categorical datasets, yet bar charts alongside error bars are best for categorical average data, scatter plots reveal quantitative connections, line charts monitor temporal changes, and boxplots reveal distribution patterns. The viridis and teal blue color scales provide good contrast along with interactive features that enable users to conduct dynamic analysis.

Alternative Designs and Pros/Cons

Heatmap:

Alternative: Bar chart (genres × decades).

A heatmap design offers compactness together with a clear presentation of two dataset dimensions. While offering approximate values, the heatmap solution maintains imprecision until tooltips are used. Bar Chart Cons: Cluttered with many categories, harder to compare.

Bar Charts (Type, Duration):

Alternative: Line chart.

The bar chart shows precise average amounts because it contains error bars that display reliability. Cons: Limited to categorical data. Line Chart Cons: Implies continuity, unsuitable for nominal categories.

Scatter Plot:

Alternative: Boxplot by runtime bins.

The log scale feature in scatter charts effectively displays individual points while handling data ranges of any scale. Cons: Overplotting (mitigated by opacity). Boxplot Cons: Loses granularity, less intuitive for runtime.

Line Chart (Popularity):

Alternative: Area chart.

The line chart provides users with strong pattern detection abilities along with interactive control features managed through the legend. The lines in this chart overlap with each other, which could be solved by selecting the lines to separate them. Area charts have a point against them because the stacking function adds values together even if the data points represent fractional quantities.

Interactions:

Alternative: Static charts.

The filtering system achieves two benefits by allowing users to establish their own field criteria ("Action," for instance). The addition of dropdowns simplifies the complexity that arises from this feature.

Conclusion

The anime trend analysis tool implements position and color as data encoding methods and includes filtering mechanisms along with interactive controls such as zoom function and tooltips. The preprocessing stage maintains data integrity by cleaning out incorrect records and by calculating confirmed metrics, which include average scores together with standard errors. Drama anime demonstrated stable high ratings (greater than 7.5) through all decades, while shounen anime reached its maximum popularity in the 2000s, when it accounted for a quarter of the series distribution and episodes between 15 and 24 minutes earned an average rating of 6.8. The visual display becomes more readable when designers focus on details like continuing Inter fonts combined with well-structured labels and properly formatted tooltips. The analysis only examines the top 10 genres and restricts the runtime to certain parameters, which might overlook unique niche trends in the data. These visual representations deliver important information that helps anime enthusiasts find top-quality shows while helping producers select profitable demographics and revealing societal patterns to researchers.