Visualization

Mounika Devabhaktuni

Nakshatra Reddy Lenkala

1. Describe 3 tasks for the users of this dataset. The same as the visualization proposal, the tasks should be challenging that can not be solved directly and the tasks should require interactive visualization.

1. **Release Year Trends:**

   Description: Plot the number of movies and TV shows released each year to identify trends over time.

   Method: Group the data by release year and type (movie or TV show), then plot the line chart showing the trends over the years.

   Visualization: Line plot

2. **Distribution of Content by Category and Type:**

   Description: Visualize the distribution of content by category and type (movie or TV show) using a sunburst chart.

   Method: Group the data by category and type, count the number of titles, and create a sunburst chart to represent the hierarchical structure.

   Visualization: Sunburst chart

3. **Distribution of Content by Country:**

   Description: Display the distribution of content by country using a choropleth map.

   Method: Group the data by country, count the number of titles, and create a choropleth map to visualize the distribution across different countries.

   Visualization: Choropleth map

4. **Content duration over time:**

   Description: Explore the relationship between the duration of movies and their release year using a scatter plot.

Method: Extract relevant columns (duration and release year), convert duration to numeric, and plot a scatter plot to visualize the relationship.

Visualization: Scatter plot

5. **Treemap of Content Categories:**

   Description: Visualize the distribution of content categories based on their popularity or number of titles using a treemap.

   Method: Group the data by category, count the number of titles, and create a treemap to represent the hierarchical structure of categories.

   Visualization: Treemap

6. **Rating Distribution:**

   Description: Create a pie chart to visualize the distribution of ratings among Netflix titles.

   Method: Count the occurrences of each rating, then create a pie chart to show the distribution.

   Visualization: Pie chart

7. **Distribution of Movie Durations:**

   Description: Visualize the distribution of movie durations using a histogram.

   Method: Filter the data to include only movies, extract the duration column, convert it to numeric, and create a histogram to show the distribution of durations.

   Visualization: Histogram

2a. Explanation of Visualization Library Methods:

1. **Release Year Trends:**

   Library Used: Matplotlib

   Method: Grouped the data by release year and type, then plotted the line chart using the plot method of Matplotlib's Axes object.

   Link to Original Example: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html/

2. **Distribution of Content by Category and Type:**

Library Used: Plotly Express

Method: Grouped the data by category and type, then used Plotly Express's sunburst function to create the sunburst chart.

Link to Original Example: https://plotly.com/python/sunburst-charts/

3. **Distribution of Content by Country:**

   Library Used: Plotly Express

   Method: Grouped the data by country, then used Plotly Express's choropleth function to create the choropleth map.

   Link to Original Example: https://plotly.com/python/choropleth-maps/

4. **Content duration over time:**

   Library Used: Matplotlib

   Method: Plotted a scatter plot using Matplotlib's scatter function to explore the relationship between movie duration and release year.

   Link to Original Example: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html/

5. **Treemap of Content Categories:**

   Library Used: Plotly Express

   Method: Used Plotly Express's treemap function to create the treemap visualization of content categories.

   Link to Original Example: https://plotly.com/python/treemaps/

6. **Rating Distribution:**

   Library Used: Plotly Express

   Method: Created a pie chart using Plotly Express's pie function to visualize the distribution of ratings among Netflix titles.

   Link to Original Example: https://plotly.com/python/pie-charts/

7. **Distribution of Movie Durations:**

Library Used: Matplotlib

Method: Created a histogram using Matplotlib's hist function to visualize the distribution of movie durations.

Link to Original Example: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html/

2b. For any customized visualization, describe your work.

This code is used to create a sunburst chart visualization representing the distribution of content by category and type (e.g., movie or TV show) using Plotly Express library in Python. Here's a breakdown of the code:

1. **Importing Libraries:**

   import pandas as pd: Imports the pandas library, which is used for data manipulation and analysis.

   import plotly.express as px: Imports the Plotly Express library, which provides easy-to-use functions for creating interactive plots.

2. **Loading the Data:**

   netflix_df = pd.read_csv("netflix_titles.csv"): Reads the Netflix dataset from a CSV file into a pandas DataFrame named netflix_df.

3. **Data Aggregation:**

   category_type_counts = netflix_df.groupby(['listed_in', 'type']). size(). Reset_index (name='count'): Groups the Netflix data by 'listed_in' (content category) and 'type' (movie or TV show). It then calculates the size of each group (i.e., the count of titles) and resets the index to convert the result into a DataFrame. The resulting DataFrame category_type_counts contains three columns: 'listed_in', 'type', and 'count'.

4. **Creating the Sunburst Chart:**

   fig = px.sunburst(category_type_counts, path=['type', 'listed_in'], values='count', title='Distribution of Content by Category and Type'): Uses Plotly Express to create a sunburst chart (px.sunburst). The category_type_counts DataFrame is passed as the data source. The path parameter specifies the hierarchical structure to be represented in the chart, with 'type' as the outer ring and 'listed_in' as the inner ring. The values parameter

specifies the numerical values to be represented by the size of each sector, which is the count of titles. Finally, the title parameter sets the title of the chart.

5. **Displaying the Chart:**

   fig.show(): Displays the sunburst chart in the output. The chart is interactive, allowing users to hover over segments to view additional information.

2c. Program to generate visualizations for your tasks, and explain the answers.

1. **Release Year Trends (Line Plot):**

   Description: Plot the number of movies and TV shows released each year to identify trends over time.

   Method: Group the data by release year and type, then plot the line chart showing the trends over the years.

   Visualization: Line plot

   Explanation:

   The data is loaded from the CSV file into a DataFrame df.

   The data is grouped by 'release_year' and 'type', and the count of titles for each group is computed.

   The grouped data is then unstacked to create a DataFrame where the index consists of release years, and the columns represent the counts of movies and TV shows.

   Finally, a line chart is plotted to visualize the trends over the years, with release year on the x-axis and the number of titles on the y-axis.

2. **Distribution of Content by Category and Type** (Sunburst Chart)**:**

   Description: Visualize the distribution of content by category and type (movie or TV show) using a sunburst chart.

   Method: Group the data by category and type, count the number of titles, and create a sunburst chart to represent the hierarchical structure.

   Visualization: Sunburst chart

   Explanation:

The data is grouped by 'listed_in' (content category) and 'type' (movie or TV show).

The number of titles for each category-type combination is counted.

A sunburst chart is created using Plotly Express, where the hierarchical structure is represented with 'type' as the outer ring and 'listed_in' as the inner ring.

3. **Distribution of Content by Country** (Choropleth Map)**:**

   Description: Display the distribution of content by country using a choropleth map.

   Method: Group the data by country, count the number of titles, and create a choropleth map to visualize the distribution across different countries.

   Visualization: Choropleth map

   Explanation:

   The data is grouped by 'country', and the number of titles for each country is counted.

   A choropleth map is created using Plotly Express, where the color intensity represents the count of titles for each country.

4. **Content duration over time (Scatter plot):**

   Description: Explore the relationship between the duration of movies and their release year using a scatter plot.

   Method: Extract relevant columns (duration and release year), convert duration to numeric, and plot a scatter plot to visualize the relationship.

   Visualization: Scatter plot

   Explanation:

   The data is loaded from the CSV file into a DataFrame netflix_df.

   Relevant columns 'duration' and 'release_year' are extracted.

   The 'duration' column is converted to numeric by extracting the numeric part of the string.

   A scatter plot is created to visualize the relationship between movie duration and release year.

5. **Treemap of Content Categories** (Treemap)**:**

Description: Visualize the distribution of content categories based on their popularity or number of titles using a treemap.

Method: Group the data by category, count the number of titles, and create a treemap to represent the hierarchical structure of categories.

Visualization: Treemap

Explanation:

The data is grouped by 'listed_in' (content category), and the count of titles for each category is computed.

A treemap is created using Plotly Express to visualize the distribution of content categories, where the size of each rectangle represents the number of titles in that category.

6. **Rating Distribution** (Pie Chart)**:**

   Description: Create a pie chart to visualize the distribution of ratings among Netflix titles.

   Method: Count the occurrences of each rating, then create a pie chart to show the distribution.

   Visualization: Pie chart

   Explanation:

   The ratings of Netflix titles are counted, and a pie chart is created to show the distribution of ratings among Netflix titles.

7. **Distribution of Movie Durations** (Histogram)**:**

   Description: Visualize the distribution of movie durations using a histogram.

   Method: Filter the data to include only movies, extract the duration column, convert it to numeric, and create a histogram to show the distribution of durations.

   Visualization: Histogram

   Explanation:

   The data is filtered to include only movies.

   The 'duration' column is extracted, and the numeric part of the string is converted to numeric.

A histogram is created to visualize the distribution of movie durations, with the x-axis representing duration (minutes) and the y-axis representing the number of movies.

2d.Overall evaluation of the visualization methods that are appropriate for the project

Evaluation of Visualization Methods:

1. **Release Year Trends (Line Chart):**

   Pros: Clearly shows the trend of movies and TV shows released each year. Suitable for comparing continuous data over time.

   Cons: May become cluttered with too many data points, making it hard to discern individual trends.

2. **Distribution of Content by Category and Type (Sunburst Chart):**

   Pros: Provides a hierarchical view of content categories and types. Interactive and visually appealing.

   Cons: Limited to categorical data and may not be suitable for comparing numerical values.

3. **Distribution of Content by Country (Choropleth Map):**

   Pros: Clearly visualizes the geographic distribution of content by country. Suitable for exploring spatial patterns.

   Cons: May not be suitable for comparing data across different countries due to variations in population and internet penetration.

4. **Content Duration over Time (Scatter plot):**

   Pros: Shows the relationship between two continuous variables. Helps identify trends or correlations.

   Cons: May not clearly show patterns if data points overlap or are too dense.

5. **Treemap of Content Categories (Treemap):**

   Pros: Provides a hierarchical view of content categories based on popularity. Easy to understand and visually appealing.

   Cons: Limited to categorical data and may not effectively compare numerical values.

6. **Rating Distribution (Pie Chart):**

Pros: Clearly shows the distribution of content ratings. Easy to understand and visually appealing.

Cons: May not effectively compare categories with many small values. Limited to categorical data.

7. **Distribution of Movie Durations (Histogram):**

   Pros: Clearly visualizes the distribution of movie durations. Suitable for exploring data distribution.

   Cons: May not be suitable for comparing specific duration ranges or identifying trends over time.

Roles -

Mounika Devabhaktuni -
Documentation and plot for Release Year Trends (Line Plot), Content Duration over Time (Scatter Plot), Treemap of Content Categories, Rating Distribution (Pie Chart).

Nakshatra Reddy Lenkala -
Documentation and plot for Distribution of Content by Category and Type (Sunburst Chart), Distribution of Content by Country (Choropleth Map), Distribution of Movie Durations (Histogram).