**Project - Phase II: Decision Making – Team 23**

**Kartheek Panyam & Gokuleswar Reddy Mooli**

**Arun Pal Bommaji & Maneesh Reddy Devara**

MS IN INFORMATION TECHNOLOGY, ARIZONA STATE UNIVERSITY

**IFT 533: Data Visualization & Reporting for IT**

**PROF. ASMAA ELBADRAWY**

APRIL 27, 2025

**Section 1: Used Visualization Tools**

For our project, we have selected **Tableau, MySQL Workbench, Python (Pandas library), and Mural** as our primary visualization tools. Tableau will be used for the main dashboard visualizations, while MySQL Workbench and Python will be used for the initial querying, cleaning, and preprocessing of the data. Mural will be used for planning and design phases.

We utilized Tableau because it is easy to use and straightforward to build sophisticated visualizations such as scatter plots, maps, tree maps, and histograms with little coding. It also comes with interactivity out of the box, which consists of filters, parameters, and dashboard actions required for dynamic, user-facing dashboards. Additionally, Tableau generates clean, professional-quality visuals ideal for presenting to a large crowd. Since we already have experience using Tableau during Phase I, it will make our project run more smoothly.

Python and MySQL Workbench are utilized owing to their capabilities in data manipulation. MySQL Workbench allows it to perform the data preprocessing functions such as frequent cleaning, filtering, joining, and exporting the datasets at affordable costs before being inserted into Tableau. Python, particularly through the Pandas library, proves useful in more complex data preprocessing functions such as dealing with missing values, aggregating data, derivation of fields that are derived, and normalization of the dataset for viewing.

Mural is an integral component of our collaboration process. It allows us to organize brainstorming ideas, finish dashboard designs, and align team members on the design plan. It provides us with an intuitive space to make rough-stage sketches and map user questions to the appropriate visualization concepts, enabling us to have a clear and well-organized planning stage.

We considered working with other software like Power BI and Excel but eliminated them. Power BI, although excellent, is more complex with licensing issues and less freedom than Tableau Public. Excel, while fantastic for basic graphing, is not interactive in dynamics as needed for our dashboard. While Python packages Matplotlib and Seaborn provide great flexibility, developing an entire interactive dashboard in Python would require other packages like Dash or Streamlit, which would require time and be less convenient within our project timeline.

**Section 2: Explanation of Required Data Pre-processing, if any**

The data for our project includes Netflix titles with attributes title, type (Movie/TV Show), director, cast, country, date added, release year, rating, duration, and genres. As a precursor to visualization, there are a few preprocessing tasks to be done on the data.

First, we have to handle missing values. Columns like director, cast, and country contain missing values. These missing values will either be imputed with 'Unknown' or handled through exclusion depending on the requirements of visualization.

Second, we will extract useful time-based features. From the date\_added column, we will extract the year and month to facilitate temporal analysis, like the trend in new content addition over time.

Third, we will perform data cleaning on categories. The listed\_in (genres) column typically contains several genres per title separated by commas. We'll split those into separate genre categories so we can facilitate genre-based filtering and aggregation.

Fourth, we will classify content based on the rating attribute. Ratings like "TV-MA," "PG," and "R" will be rolled up to higher-level categories like "Family," "Teens," and "Adults" to simplify visualizations.

Fifth, normalization of fields like duration will be completed. Duration will be recorded in minutes for films and the number of seasons for television shows. This formatting will ensure proper comparisons can be made easily.

Finally, all the fields will be cleaned, normalized, and exported into a structured format (CSV) through Python (Pandas). MySQL will assist in querying any needed specific subsets. The preprocessed datasets will then be imported into Tableau for dashboarding.

**title**: No changes were applied. The title field is used directly for labeling and identification purposes.

**type**: Retained as it is. This attribute was used to differentiate between Movies and TV Shows and to create filters in the dashboard.

**director**: Missing values in the director field were filled with the placeholder 'Unknown' to maintain consistency. This attribute was utilized to identify the top directors with the most titles.

**cast**: Missing values were filled with 'Unknown'. The cast field was exploded, meaning that entries with multiple actors were split into individual records, allowing us to calculate the frequency of actor appearances across titles.

**country**: Missing country values were filled with 'Unknown'. If multiple countries were listed for a title, they were split into separate records to allow for more accurate country-based analyses.

**date\_added**: This attribute was converted from a string to a datetime format. From this, separate fields for year and month were extracted, enabling temporal analyses such as the number of titles added per year.

**release\_year**: Used as-is to examine trends in the historical release of content. This attribute allowed us to create filters and timelines for visualizations.

**rating**: Used to group titles into broader maturity rating categories. No major cleaning was necessary, but standardization ensured consistent formatting for filtering and analysis.

**duration**: For movies, the duration field was parsed to extract the numeric value representing minutes. For TV shows, the number of seasons was extracted. This allowed accurate visualizations of movie durations and TV show longevity.

**listed\_in**: The genres listed in the 'listed\_in' field were split into multiple genres wherever multiple categories were listed. This enabled accurate analysis of genre distribution and preferences.

**description**: Although not directly visualized in the dashboard, the description field was retained for possible future sentiment analysis or thematic studies.

**Section 3: List of Final Set of Questions**

The dashboard is designed to allow users to explore and answer the following questions based on the Netflix Titles dataset:

1. How many movies vs. TV shows are currently available on Netflix?
2. What are the most popular content ratings (e.g., TV-MA, PG-13, R) on Netflix?
3. Which countries produce the most content available on Netflix?
4. How has the number of new titles added to Netflix changed over time (by year or month)?
5. What are the most common genres or genre combinations across all titles?
6. Which years had the highest number of releases?
7. What is the average duration of movies on Netflix? Are there noticeable trends by year or by content rating?
8. Which directors have the most titles featured on Netflix?
9. Which actors or actresses appear most frequently across Netflix content?
10. Are there genre preferences based on country (e.g., India vs. US vs. Japan)?
11. How are titles distributed across different maturity ratings by type (Movie vs. TV Show)?
12. Are more international titles being added to Netflix in recent years compared to earlier years?
13. How many one-season vs. multi-season TV shows are available on Netflix?
14. What genres are most common among TV shows compared to movies?
15. Top 10 most represented countries in each genre?

**Section 4: Dashboard Plot Drafts**

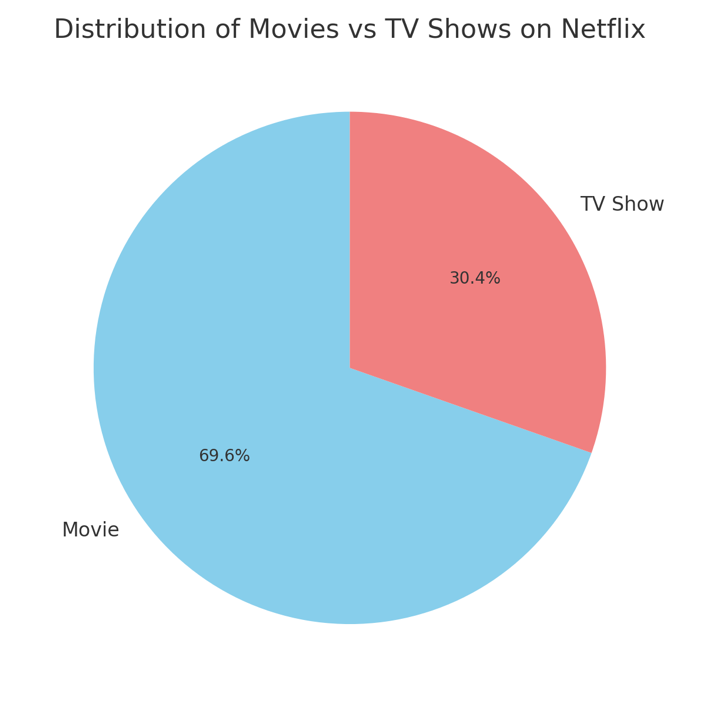
1. **Movies vs. TV Shows Distribution**

**Plot Type:** Pie Chart

**Description:** Visualizes the proportion of Movies and TV Shows available on Netflix, helping users quickly understand content focus.

**Pre-attentive Attributes:** Color (Blue for Movies, Orange for TV Shows), Shape (slice size).

**Sample Plot:**



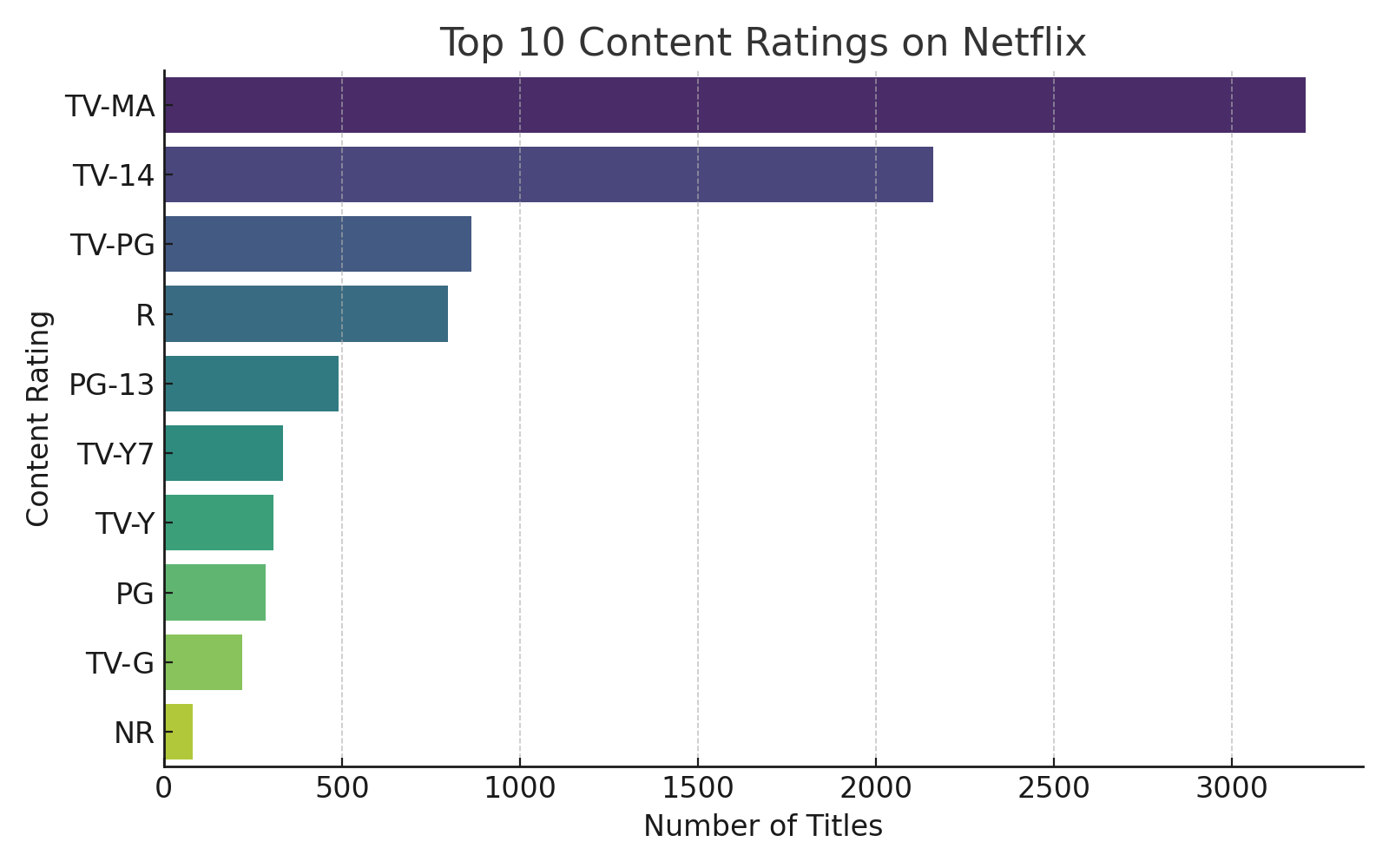
1. **Content Ratings Popularity**

**Plot Type:** Horizontal Bar Chart

**Description:** Displays the number of titles per content rating (e.g., TV-MA, PG-13), identifying popular target audiences.

**Pre-attentive Attributes:** Bar length, Darker colors for higher counts.

**Sample Plot:**



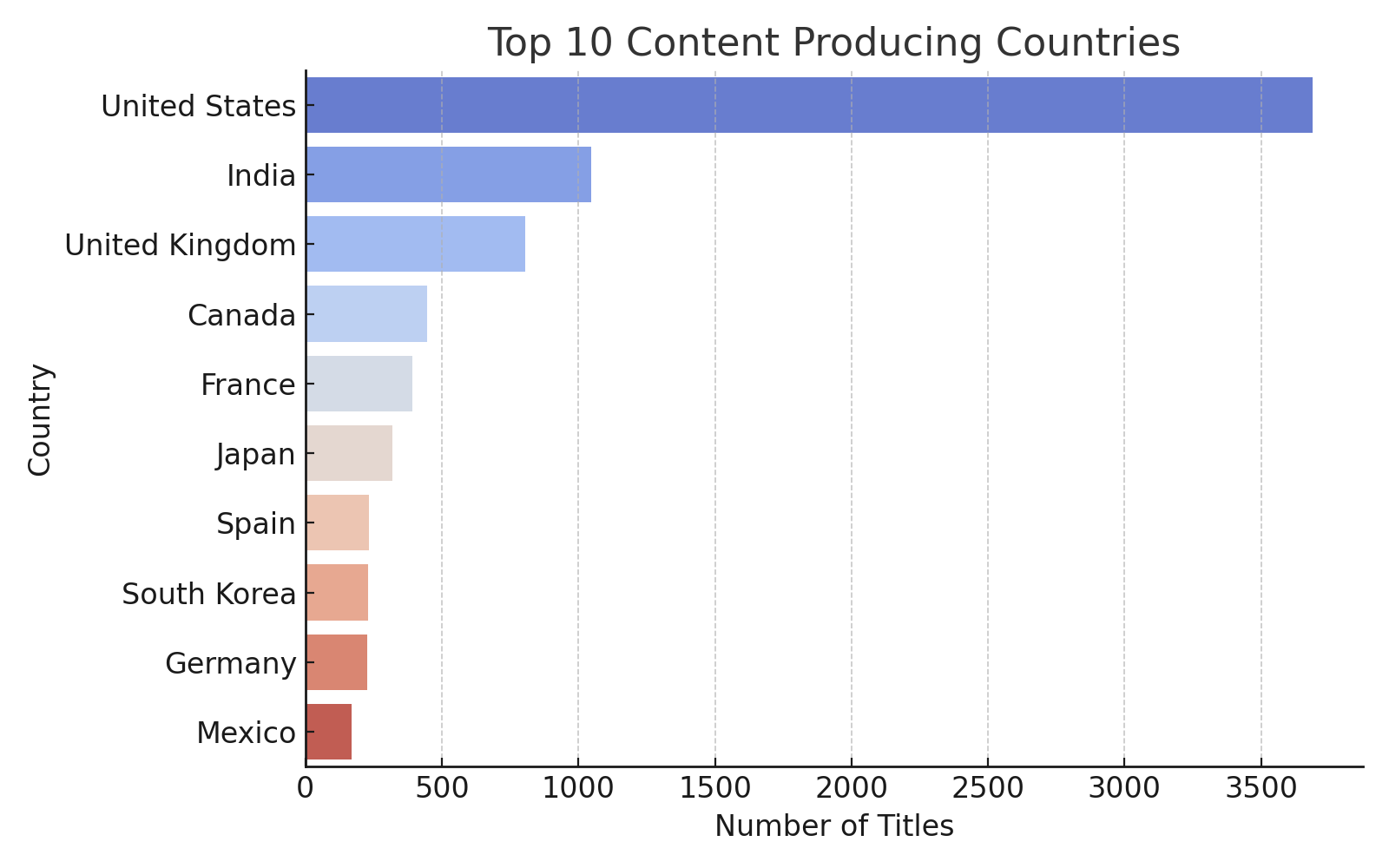
1. **Top Content Producing Countries**

**Plot Type:** Vertical Bar Chart

**Description:** Highlights countries with the highest contributions to Netflix's catalog.

**Pre-attentive Attributes:** Color intensity, Bar height.

**Sample Plot:**



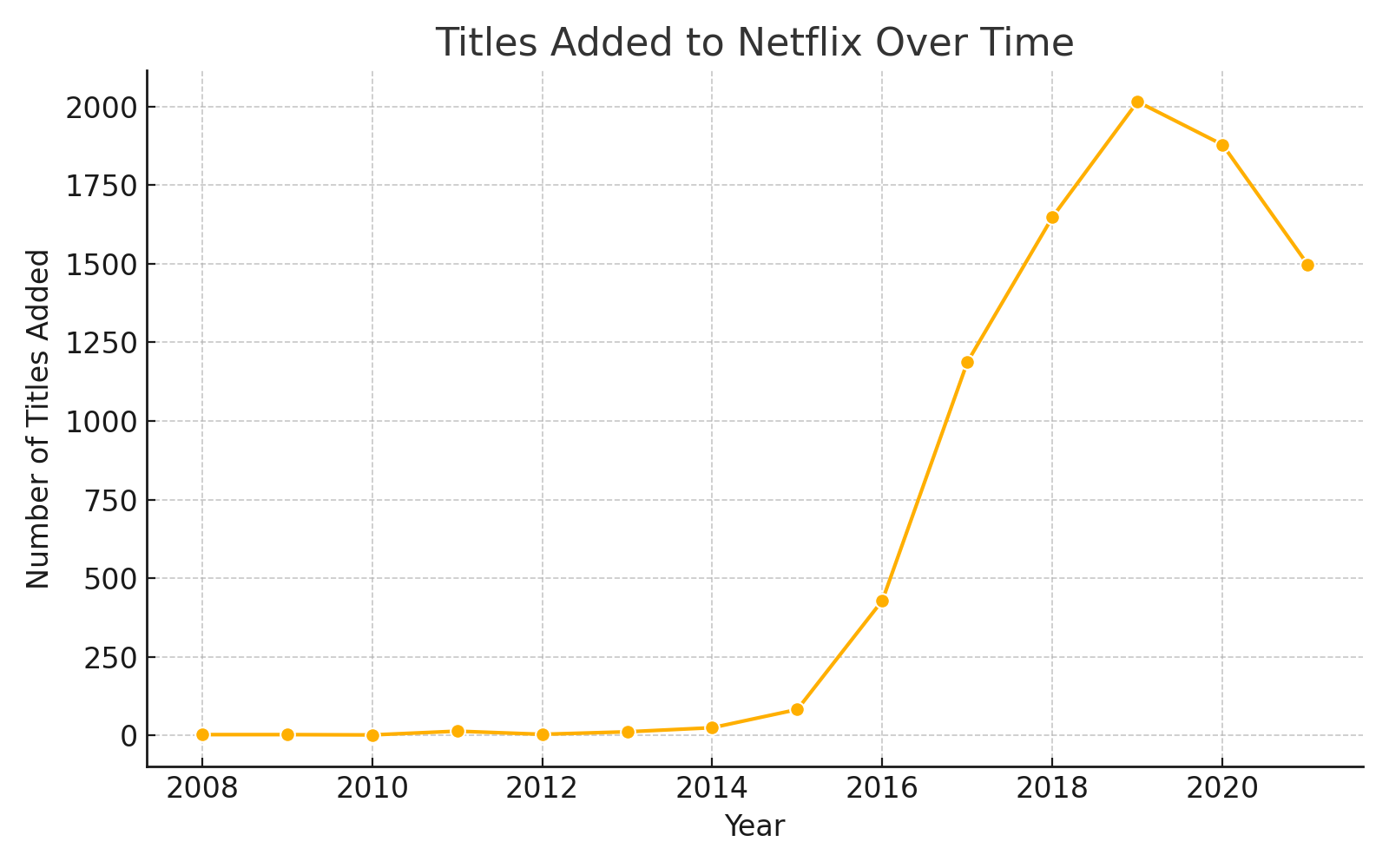
1. **Titles Added Over Time**

**Plot Type:** Line Chart

**Description:** Depicts how many titles were added over the years or months, providing insights into Netflix's content growth trends.

**Pre-attentive Attributes:** Line slope, Position.

**Sample Generated Plot:**

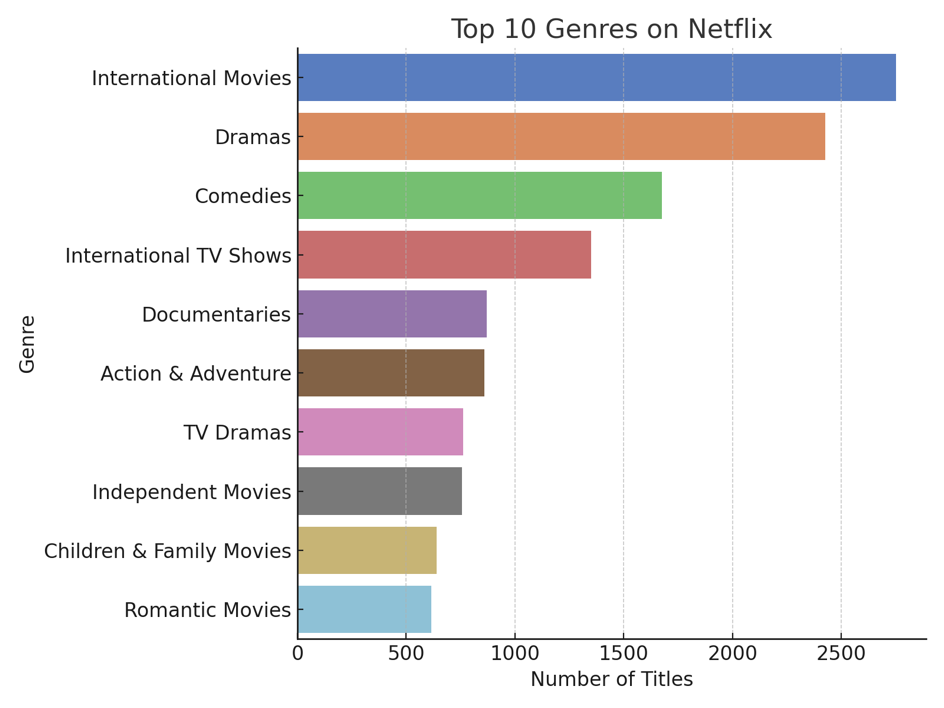


1. **Genre Distribution**

**Plot Type:** Bar chart

**Description: Represents genres based on the number of titles, offering a high-level view of genre popularity.**

**Pre-attentive Attributes: Bar length, Color coding by genre.**



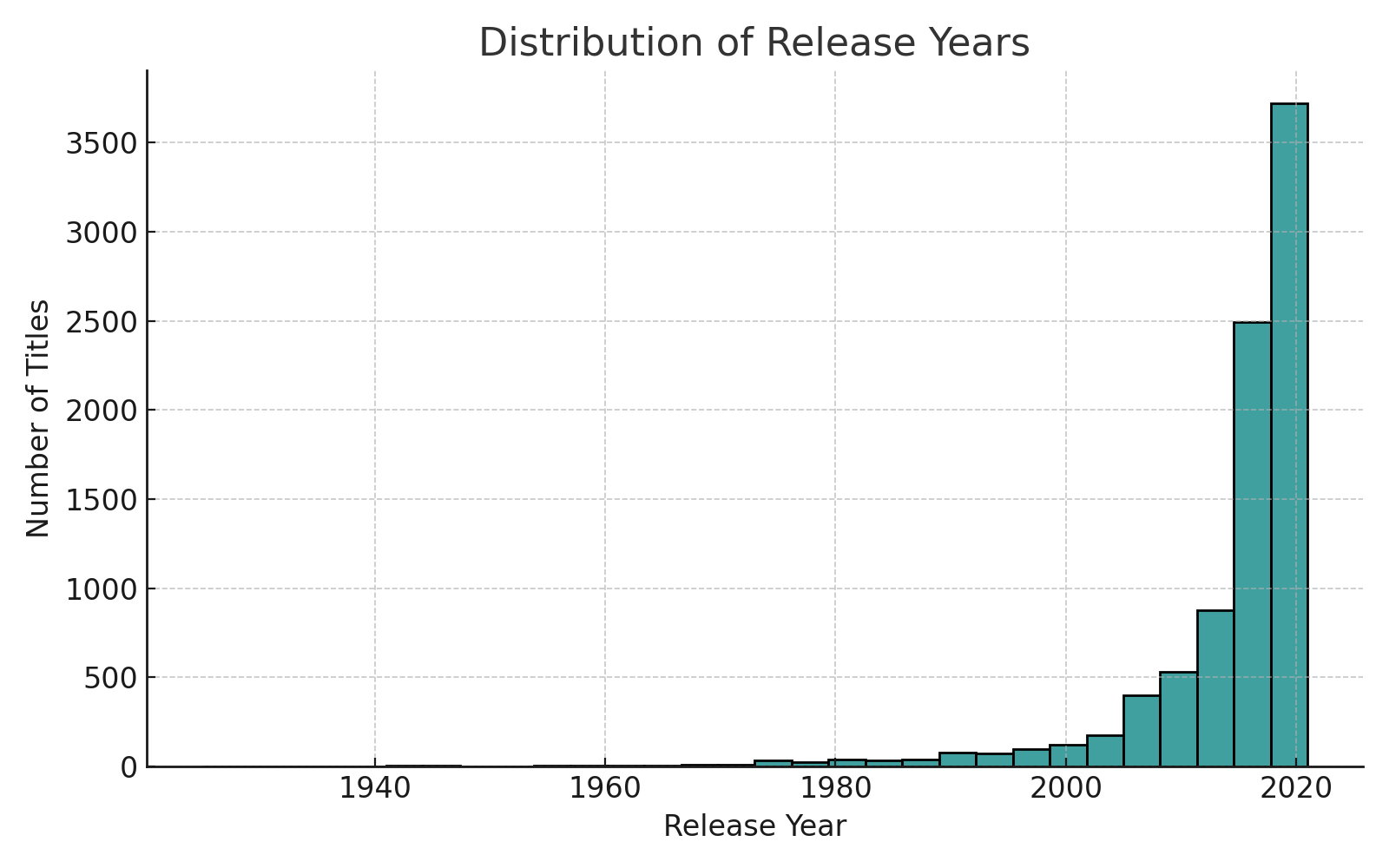
1. **Release Year Trends**

**Plot Type:** Histogram

**Description:** Analyzes the frequency of title releases over different years to detect content production peaks.

**Pre-attentive Attributes:** Bar height, X-axis alignment.

**Sample Plot:**



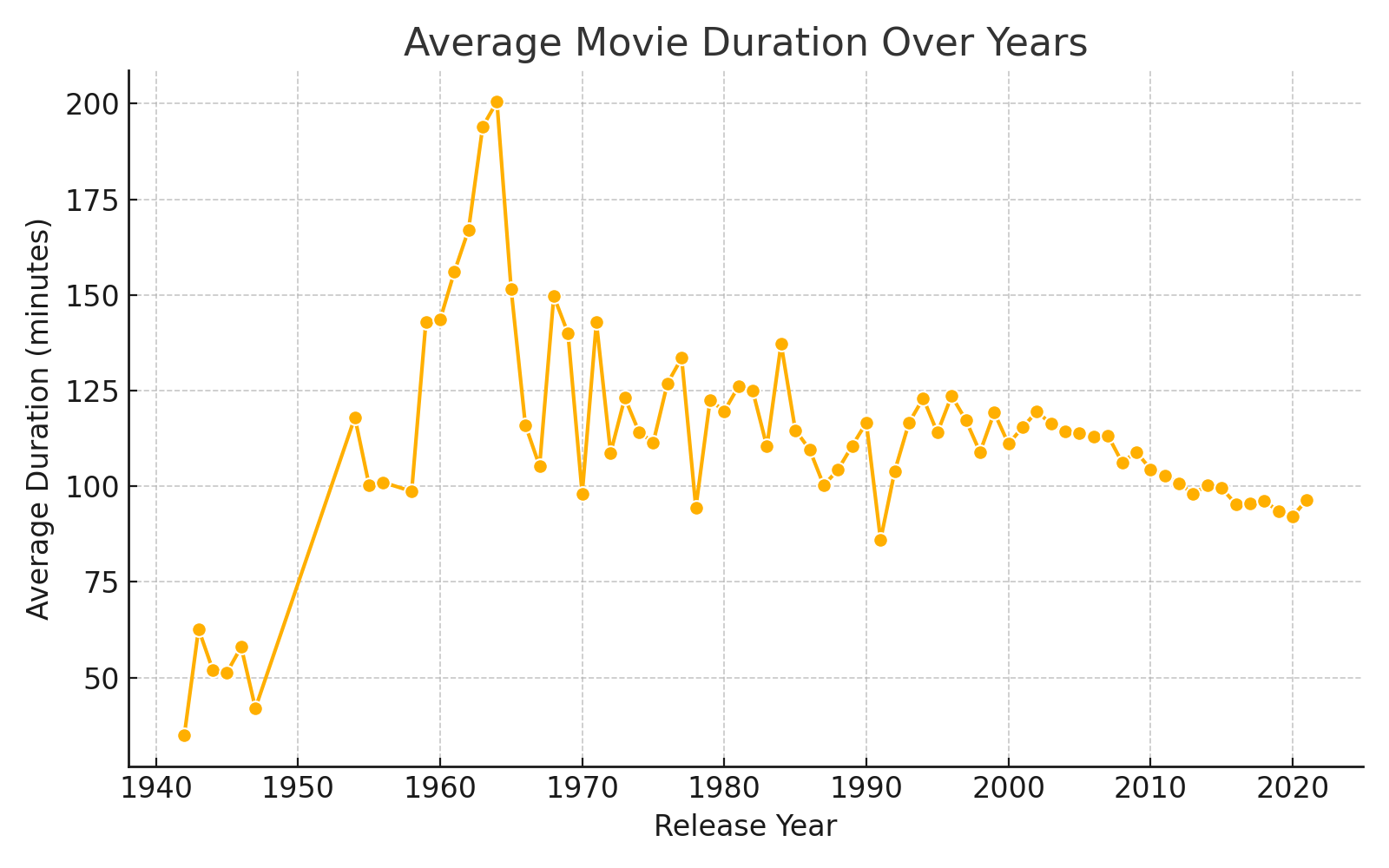
1. **Average Movie Duration Over Time**

**Plot Type:** Line Chart

**Description:** Shows how the average movie duration changed across release years, indicating shifts in content length.

**Pre-attentive Attributes:** Line slope and trend movement.

**Sample Plot:**



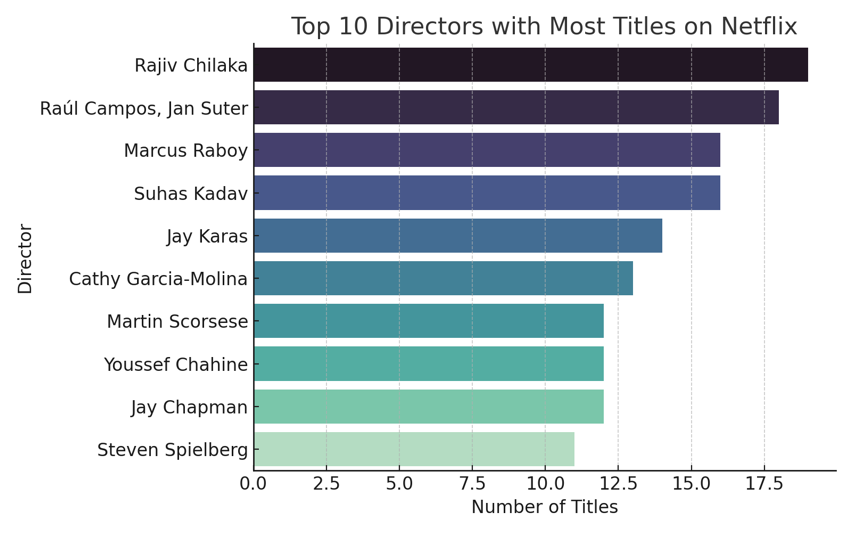
1. **Top Directors**

**Plot Type:** Vertical Bar Chart

**Description:** Displays directors with the highest number of Netflix titles, highlighting prominent contributors.

**Pre-attentive Attributes:** Bar length, Label clarity.

**Sample Plot:**



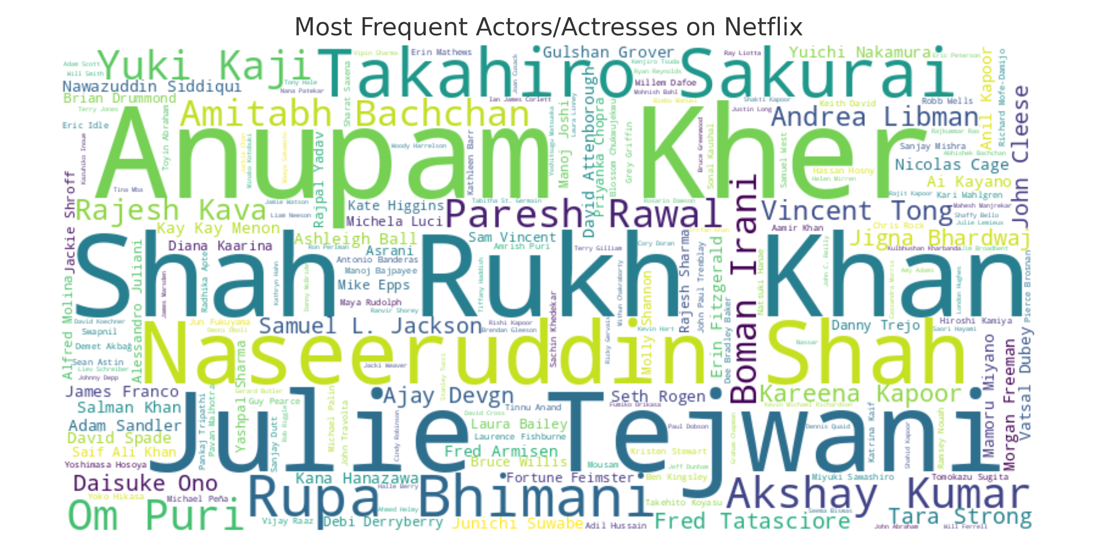
1. **Most Frequent Actors/Actresses**

**Plot Type:** Word Cloud

**Description:** Shows which actors/actresses appear most often in Netflix titles, based on title frequency.

**Pre-attentive Attributes:** Font size.

**Sample Generated Plot:**



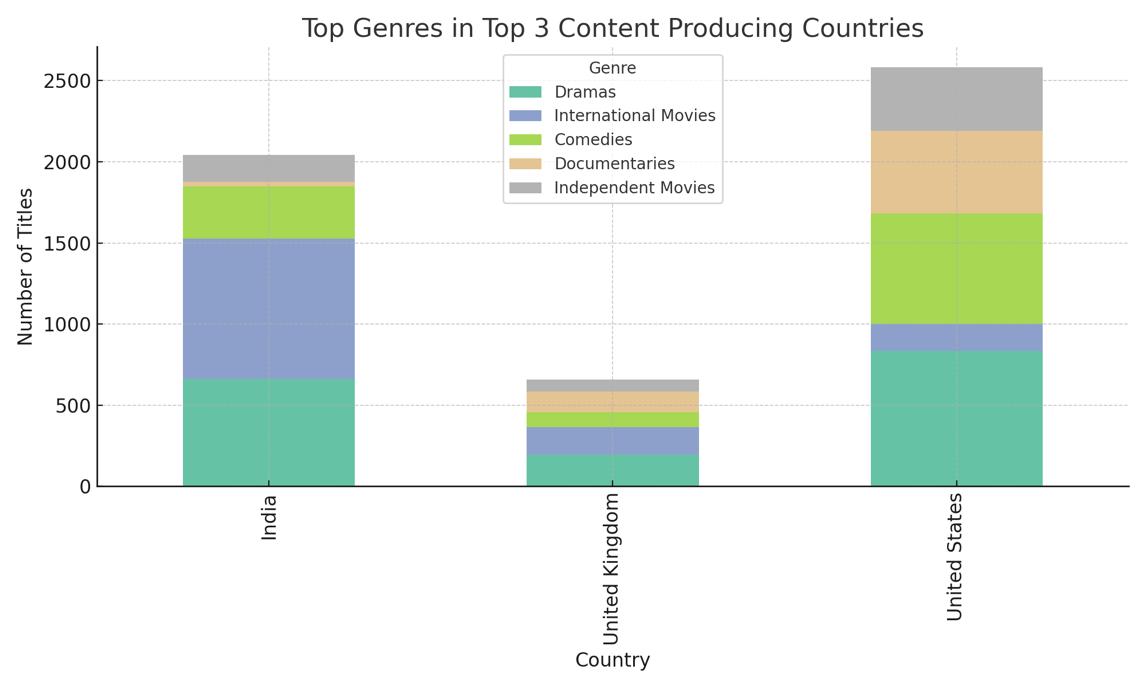
1. **Genre Preferences by Country**

**Plot Type:** Stacked Bar Chart

**Description:** Compares genre popularity across different countries to reveal cultural content trends.

**Pre-attentive Attributes:** Stack height, Color differentiation for genres.

**Sample Generated Plot:**



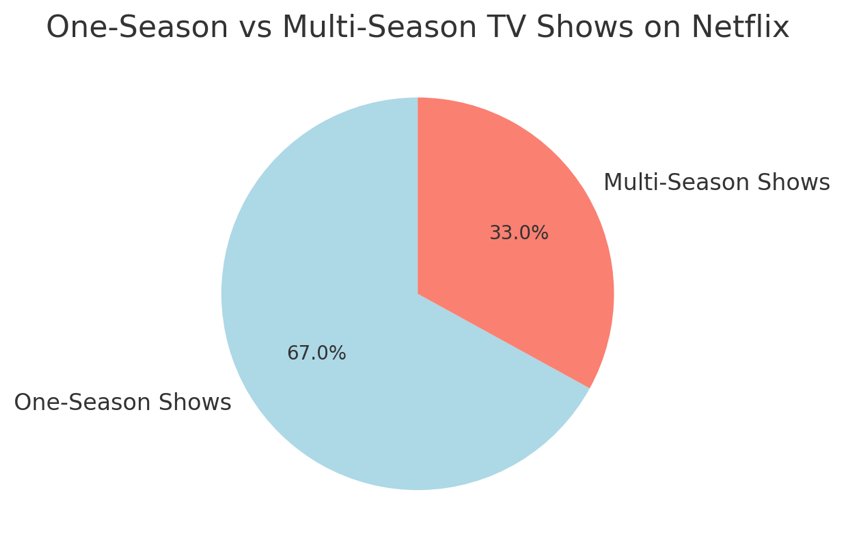
**11. One-Season vs Multi-Season TV Shows**

**Plot Type:** Pie Chart

**Description:** Shows the distribution of TV Shows based on the number of seasons. Helps understand how many are single season limited series versus multi-season productions.

**Pre-attentive Attributes:** Slice size, Color coding for 1 season vs multiple seasons.

**Sample plot:**



**Section 5: Dashboard Interactivity**

Our dashboard will present various interactive controls to allow users to dynamically traverse Netflix content and tackle the salient questions specified above. Interactive capability is indispensable in order to make user experience richer, since it assists users to filter, tailor, and cut the visualized data as per their preferences. The interactive elements specified below will be implemented within our dashboard:

**Content Type Filter (Movies vs TV Shows):** A dropdown menu will enable users to filter the plots based on content type, either Movies or TV Shows. This will be linked directly to the Movies vs TV Shows Distribution, Average Movie Duration Over Time, and Genre Distribution plots. By switching between content types, users are able to more clearly comprehend distinctions between TV shows and movies in Netflix's library. Values for this control are derived from the 'type' column within the dataset and provide the 'Movie' and 'TV Show' options.

**Release Year Range Slider:** A range slider will enable the titles released within a specified period to be highlighted, say 2000-2020. The control will be associated with the Titles Added Over Time, Release Year Trends, and Average Movie Duration Over Time charts. It offers temporal analysis of Netflix content strategy and growth trends over time. The range of slider values is between 1925 and 2021, according to information in the 'release\_year' column.

**Genre Multi-Select Filter:** A multi-select dropdown widget will allow users to select one or more than one genre of interest. The option will be associated with the Genre Distribution and Genre Preferences by Country plots so that users can zoom in on specific types of content like Comedies, Dramas, or Documentaries. The available values will be dynamically drawn from the unique genre entries in the 'listed\_in' column of the data.

**Country Filter:** A dropdown search will offer the ability to select one or more countries and view the content from them. This filter corresponds to the Top Content Producing Countries and Genre Preferences by Country charts. Through this control, users will be able to explore regional content production patterns and establish geographical difference in what Netflix provides. The available values are taken from the 'country' column.

**Content Rating Selector:** A drop-down control will allow users to choose titles according to their maturity rating, e.g., TV-MA, PG-13, or R. It will be linked with the Content Ratings Popularity graph. This control allows users to study how the different audience age segments and levels of maturity are served by the Netflix services. The selectable values are loaded from the 'rating' column.

These tools can be used alone or in combination, providing multi-dimensional analysis and enabling exploration to uncover additional insights. The interactivity makes the dashboard dynamic and adaptive to different user interests and exploration orders, transforming it from a static tool into an interactive and dynamic tool.

**Section 6: References**

**1.** **Netflix Movies and shows Dataset:** [https://www.kaggle.com/datasets/shivamb/netflix-shows](1.%20Netflix%20Movies%20and%20shows%20Dataset:%20https:/www.kaggle.com/datasets/shivamb/netflix-shows%20%202.%20Mural%20link:%20https:/app.mural.co/t/ift5338876/m/ift5338876/1744948618212/1841962e5e3801a8ff05bc1b200746c861dcb20b?sender=u028cea11dd94b51431478940)

**2. Mural link:** [https://app.mural.co/t/ift5338876/m/ift5338876/1744948618212/1841962e5e3801a8ff05bc1b200746c861dcb20b?sender=u028cea11dd94b51431478940](1.%20Netflix%20Movies%20and%20shows%20Dataset:%20https:/www.kaggle.com/datasets/shivamb/netflix-shows%20%202.%20Mural%20link:%20https:/app.mural.co/t/ift5338876/m/ift5338876/1744948618212/1841962e5e3801a8ff05bc1b200746c861dcb20b?sender=u028cea11dd94b51431478940)