DATA ANALYST-INTERNSHIP PROJECT

UNIFIED MENTOR

Topics Covered:

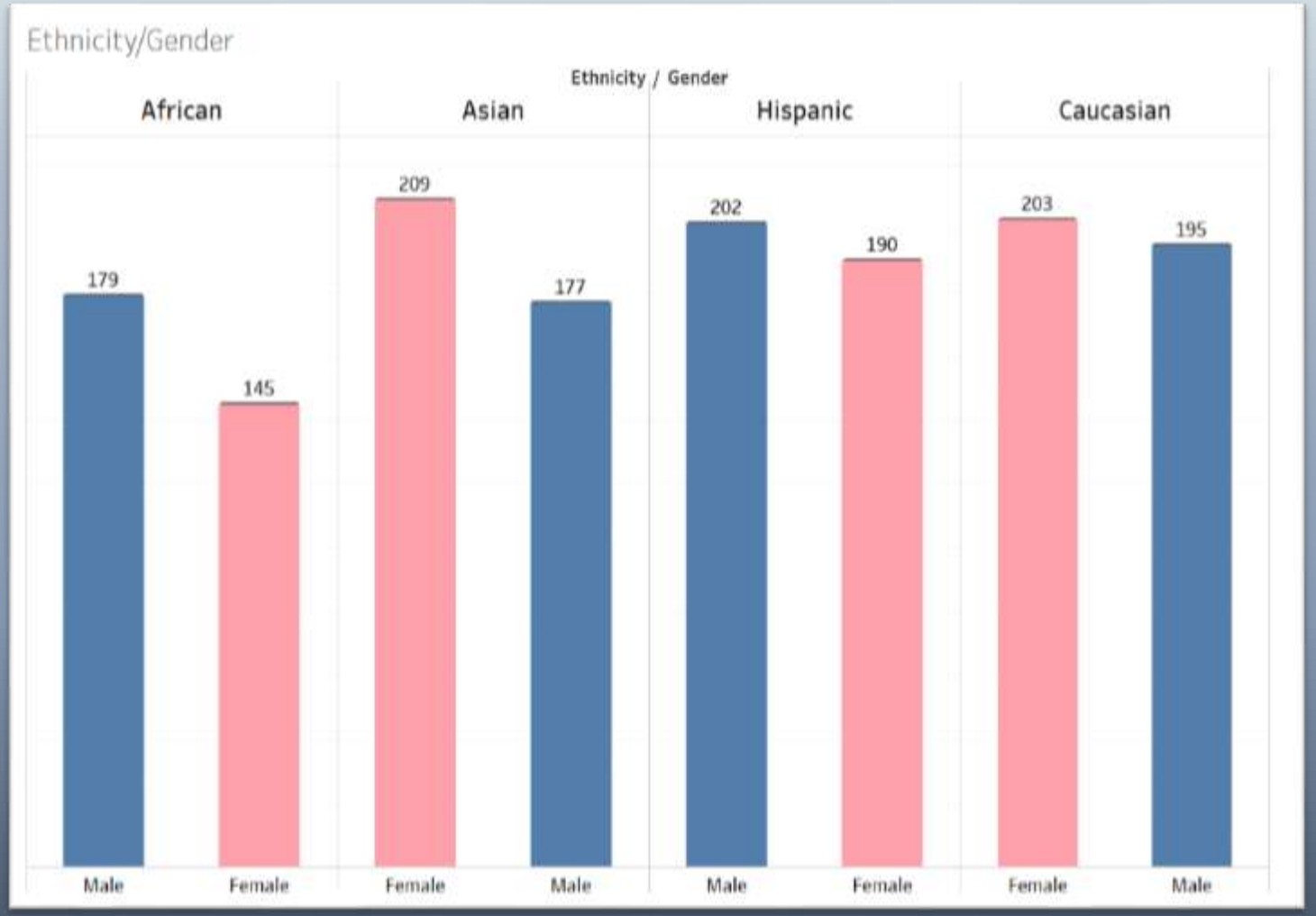
1. OCD Patient Dataset: Demographics & Clinical Data
2. Netflix Data: Cleaning, Analysis and Visualization
3. Supermart Grocery Sales - Retail Analytics Dataset

Name : **Suraj S G Dhanva**
Intern ID : **UMIP 26552**
Batch : **Nov 1st (2 months)**

1. OCD Patient Dataset: Demographics & Clinical Data

The **OCD Patient Dataset: Demographics & Clinical Data** project focuses on analyzing a comprehensive dataset of individuals diagnosed with Obsessive-Compulsive Disorder (OCD). This dataset contains key demographic details like age, gender, and ethnicity, alongside clinical data such as diagnosis date, symptom duration, and Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) scores.

The project aims to uncover patterns and correlations within the data, offering insights into symptom severity, co-occurring conditions, and treatment approaches. Through visualizations and statistical analysis using tools like Python and Tableau, this study seeks to enhance understanding of OCD and support informed decision-making in mental health research and clinical practices.



This graph compares the distribution of OCD patients based on **Ethnicity** and **Gender**. Each bar represents the count of Male (blue) and Female (pink) patients across ethnic groups: African, Asian, Hispanic, and Caucasian. It highlights variations in patient representation among different ethnicities and genders, with notable trends like a higher number of Asian females and nearly equal male-female counts for Caucasians. This data can help understand demographic trends in OCD prevalence.

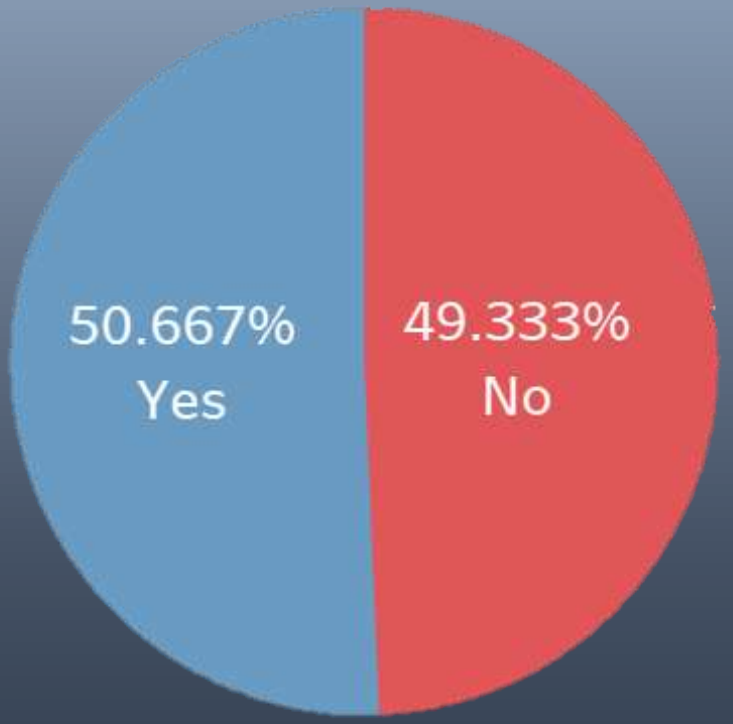
Marital Status

The percentages for marital status categories suggest that OCD prevalence is relatively evenly distributed among single, divorced, and married individuals. However:

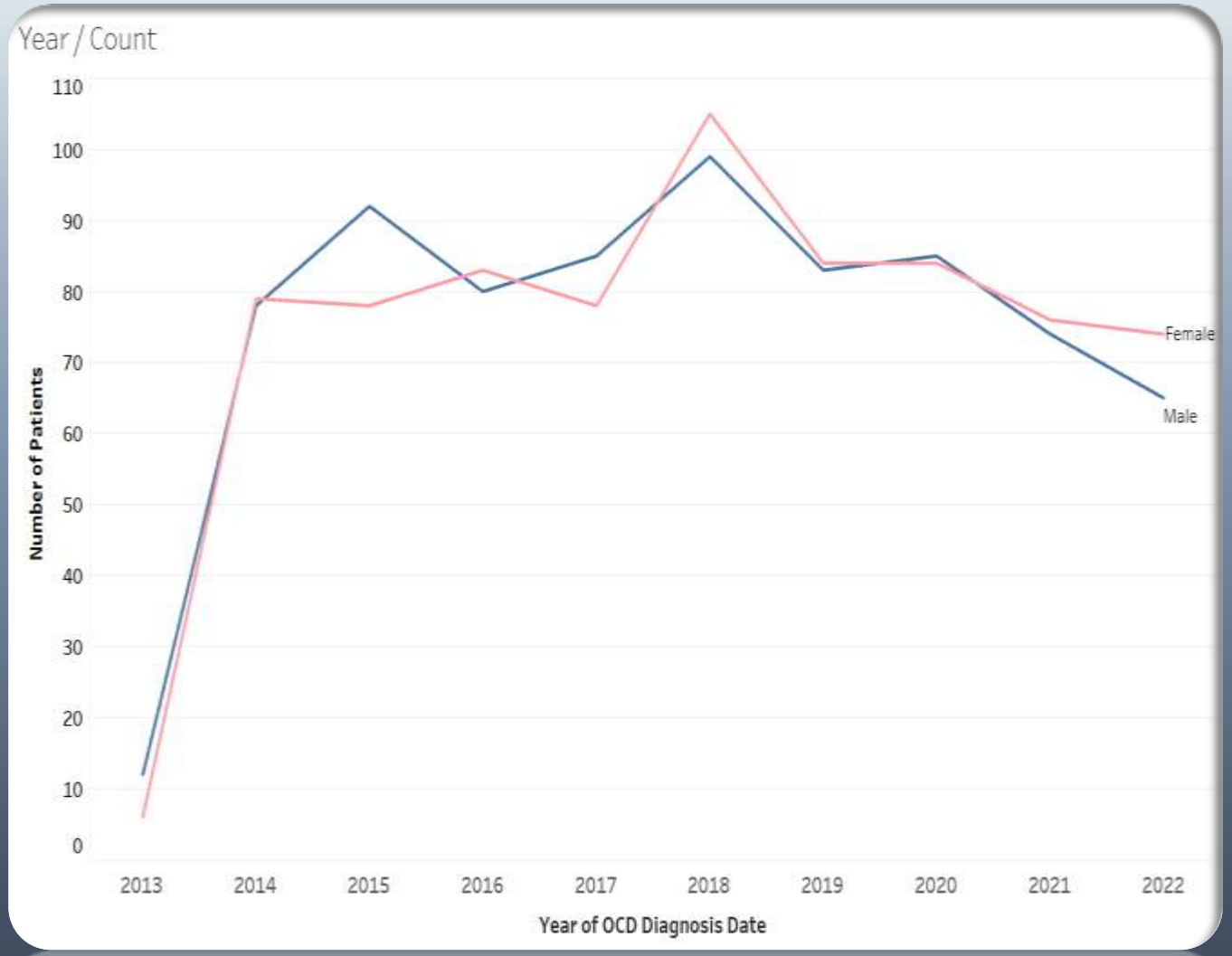
- **Slight Majority**: Single individuals form the largest portion of the OCD population.
- Divorced and Married Close: Similar proportions suggest marital status may not strongly influence OCD.

The pie chart shows that 50.667% of OCD patients have a family history, while the remaining 49.333% do not. This indicates a possible genetic link but also highlights that environmental or other factors could contribute to OCD in many cases.

- **Equal Distribution: Family history of OCD shows a near-equal split, with 50.667% having a positive family history.**
- **Genetic Link: This suggests a possible genetic or hereditary component in OCD prevalence..**



Family History of OCD



This graph shows the average number of OCD patients diagnosed each year, divided by gender:

1.Initial Increase (2013-2017): The number of diagnoses for both males and females increased significantly, suggesting rising awareness or better diagnosis during this period.

1.Peak Diagnoses (2018): The highest number of patients were diagnosed in 2018 for both genders, with females slightly surpassing males.

1.Decline Post-2018: Diagnoses steadily decreased after 2018, indicating either improved treatment, a decline in reporting, or external factors like limited healthcare access during events like the pandemic.

1.Gender Trends: Female patients consistently had higher diagnosis rates compared to males across most years, highlighting possible gender-specific factors in OCD occurrence or diagnosis rates.

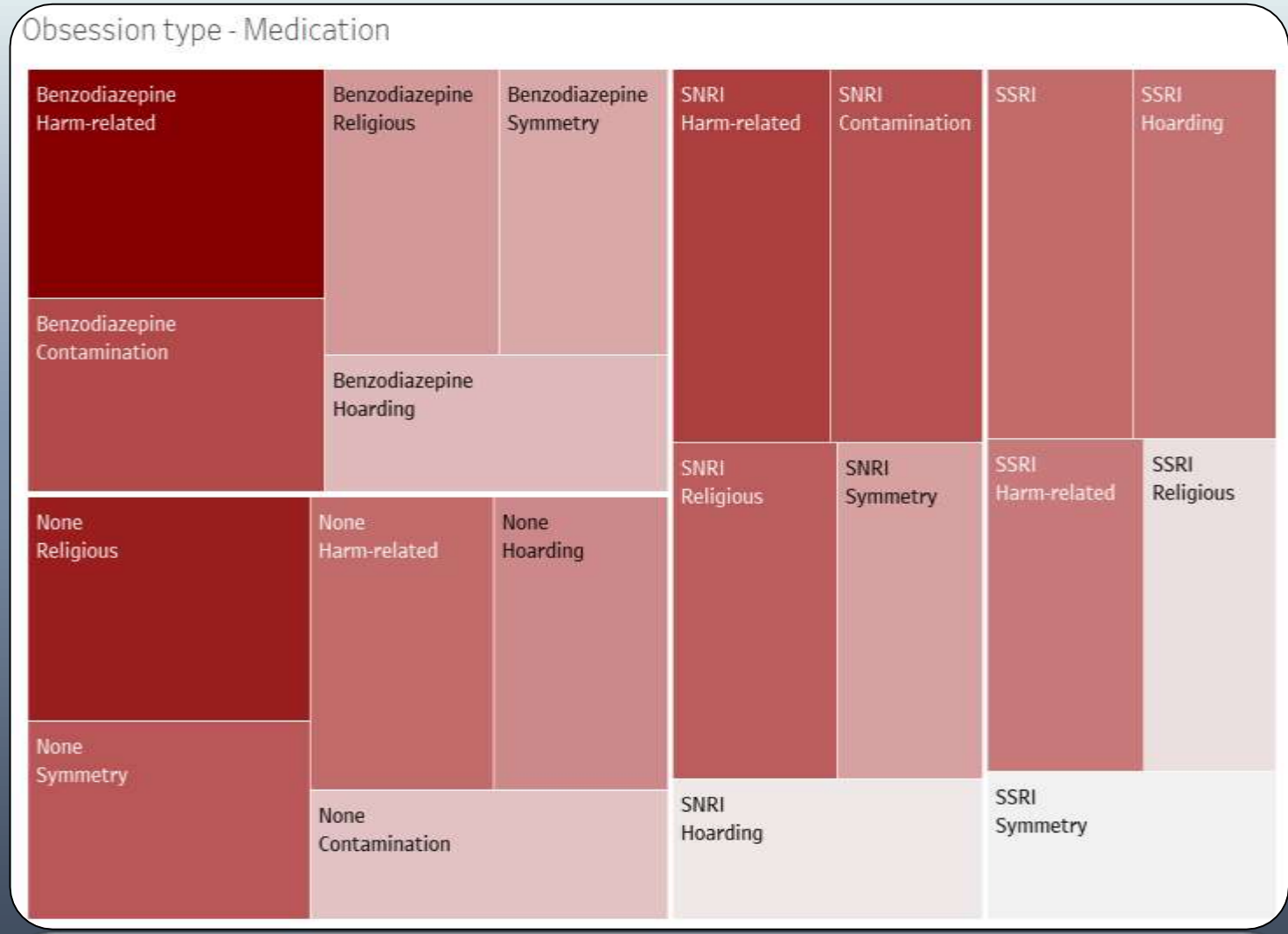
The treemap visualizes the distribution of obsessive-compulsive disorder (OCD) obsessions and associated medication types. The categories include different obsession types (e.g., harm-related, contamination, religious) and corresponding medication types (e.g., benzodiazepine, SSRI, SNRI, and no medication).

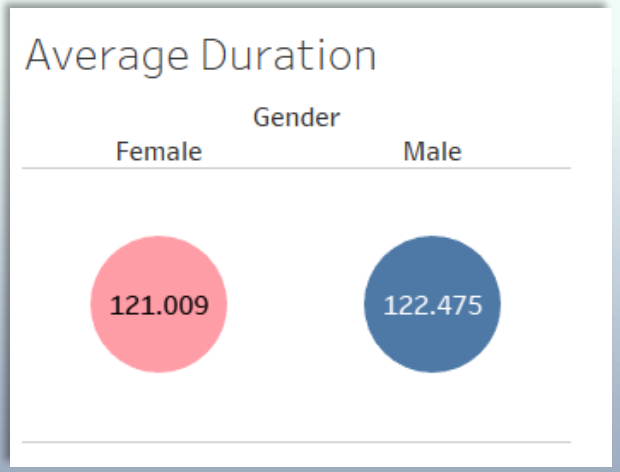
Prevalence Across Medication Types:

- Benzodiazepines are associated with all the obsession types, but harm-related and religious obsessions appear larger in the visualization, indicating a higher frequency.
- SSRIs also span all obsession types, with harm-related and religious obsessions being prominent.
- SNRIs show a broader distribution but tend to have smaller areas overall compared to SSRIs and benzodiazepines, suggesting a less frequent association.
- "None" (no medication) is a notable category, suggesting many individuals with OCD are not using medications or are managed differently.

Obsession Types:

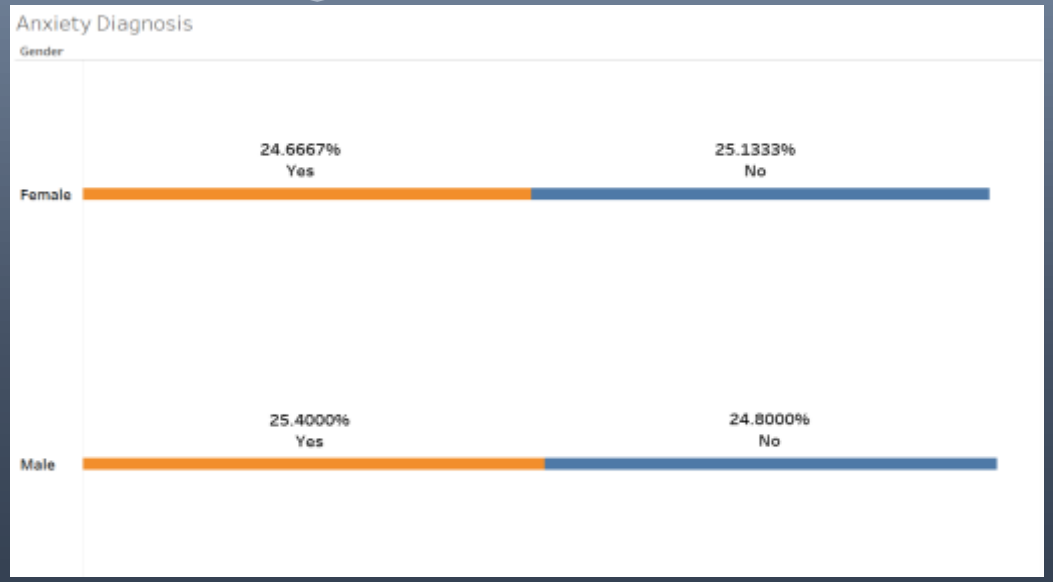
- Harm-related obsessions are prominent across multiple medication types, suggesting these are a common OCD subtype.
- Religious obsessions are also significant, especially with benzodiazepines and SSRIs.
- Contamination, symmetry, and hoarding obsessions have smaller but consistent representations, showing their presence in the data but likely less frequent compared to harm-related and religious obsessions.





From this data, we can conclude that, on average, males tend to experience slightly longer episodes of OCD than females, though the difference is relatively small. This could suggest that there may be underlying factors influencing the duration of OCD episodes differently across genders, potentially related to biological, psychological, or social aspects.

- The numbers indicate that, on average, females have a slightly higher "No" response compared to "Yes" responses, suggesting that slightly more females might not be diagnosed with anxiety compared to those who are.
- For males, the "Yes" responses are slightly higher than the "No" responses, suggesting that a marginally higher number of males are diagnosed with anxiety.
- Based on this data, it appears that males are slightly more likely to be diagnosed with anxiety compared to females, although the difference is relatively small.



- For females, the "Yes" responses are slightly higher than the "No" responses, indicating that slightly more females are diagnosed with depression than those who are not.
- For males, the "Yes" responses are higher than the "No" responses , suggesting that males are also more likely to be diagnosed with depression than not.
- The data suggests that both males and females have a similar likelihood of being diagnosed with depression, with females having a slightly higher rate.

2.Netflix Data: Cleaning, Analysis and Visualization

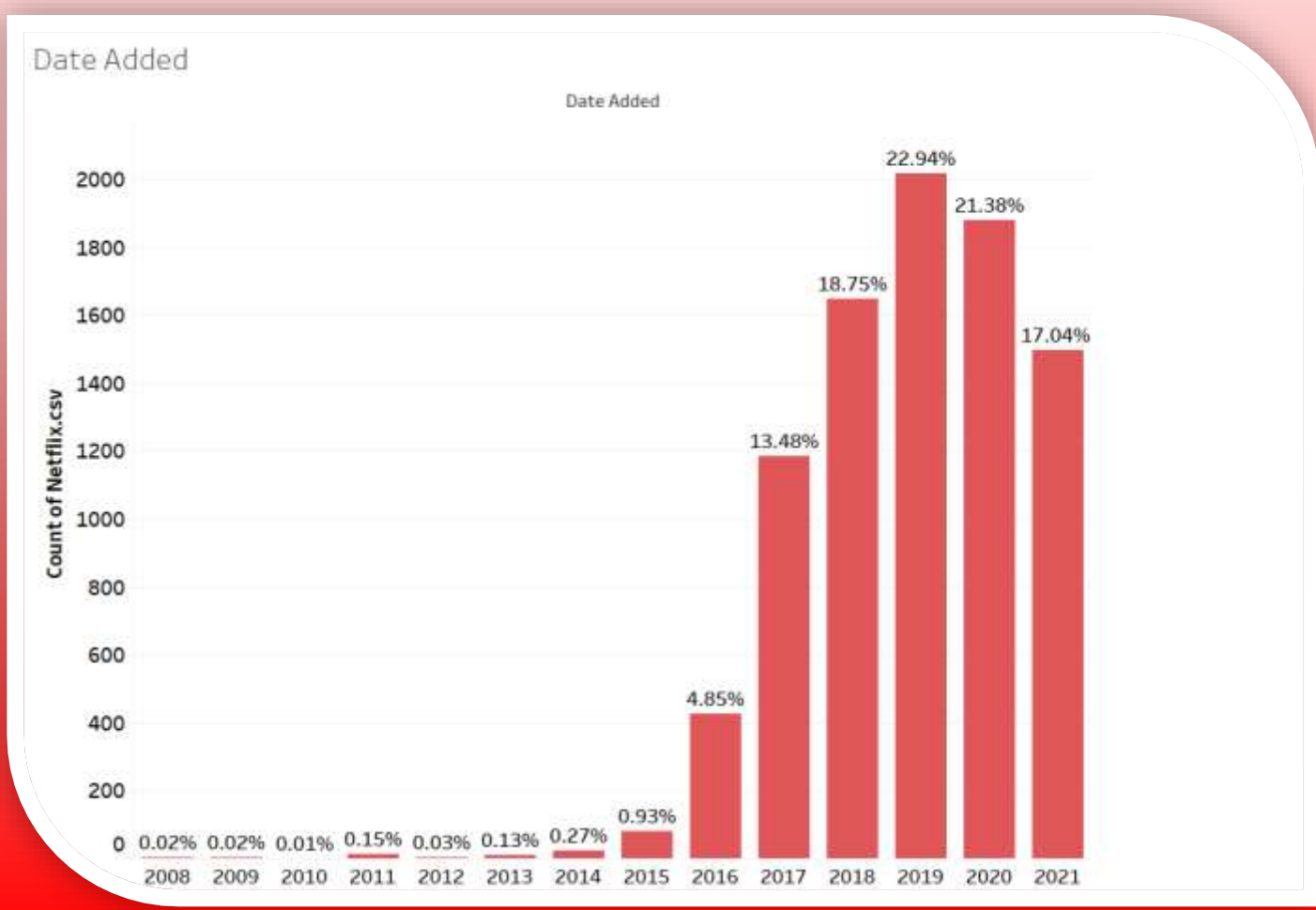
Netflix is a global streaming service that offers a vast library of TV shows, movies, documentaries, and original content. It began as a DVD rental service in 1997 and transitioned to online streaming in 2007. Over the years, Netflix has become a dominant player in the entertainment industry, producing award-winning original content and expanding to over 190 countries.

The graph represents the number of titles added to Netflix each year and their percentage contributions.

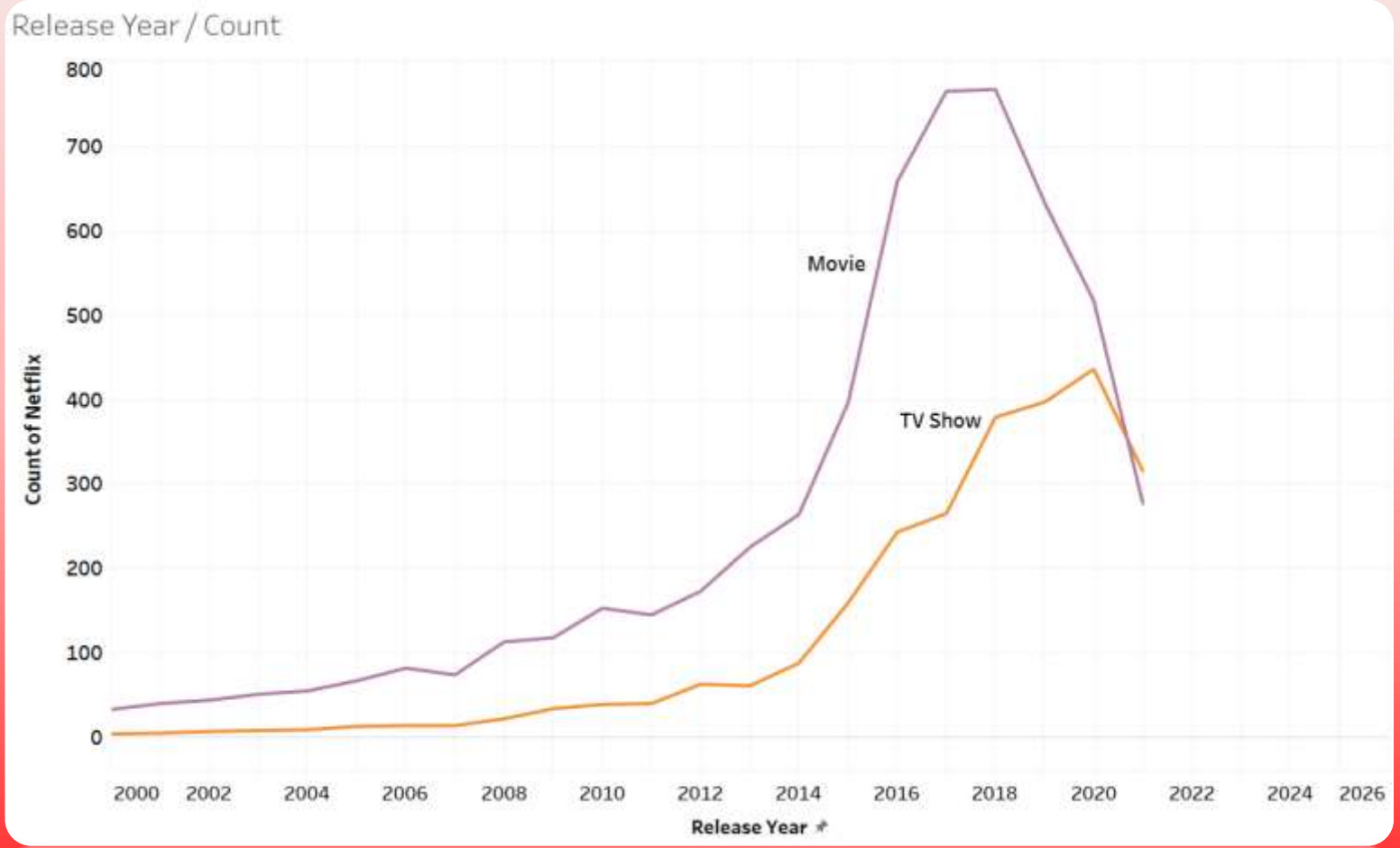
1. **Netflix experienced significant growth in its content additions starting in 2016, with a steady increase in titles added yearly. The peak years were 2018 and 2019, where the largest percentages (22.94% and 21.38%) of titles were added.**

2. **After 2019, there was a decline in new additions, likely due to factors like market saturation or external disruptions, such as the COVID-19 pandemic in 2020.**

3. **Very few titles were added between 2008 and 2014, reflecting Netflix's initial focus on building its infrastructure and transitioning to streaming before expanding its library aggressively.**



This graph displays the count of movies and TV shows on Netflix by their release year. Below are key observations:



1

DOMINANCE OF MOVIES:

Movies consistently outnumber TV shows on Netflix, showing the platform's focus on offering a robust collection of films. The movie count experienced a sharp rise from 2015 to 2018 before declining post-2019.

2

GRADUAL GROWTH OF TV SHOWS:

TV shows saw a steady increase in availability starting in the mid-2010s, peaking around 2020. This aligns with Netflix's investment in producing original series and licensing popular shows.

3

DECLINE IN RECENT ADDITIONS:

Both movies and TV shows exhibit a decline in titles released after 2020. This could be attributed to production slowdowns during the COVID-19 pandemic or a strategic shift in Netflix's content acquisition strategy.

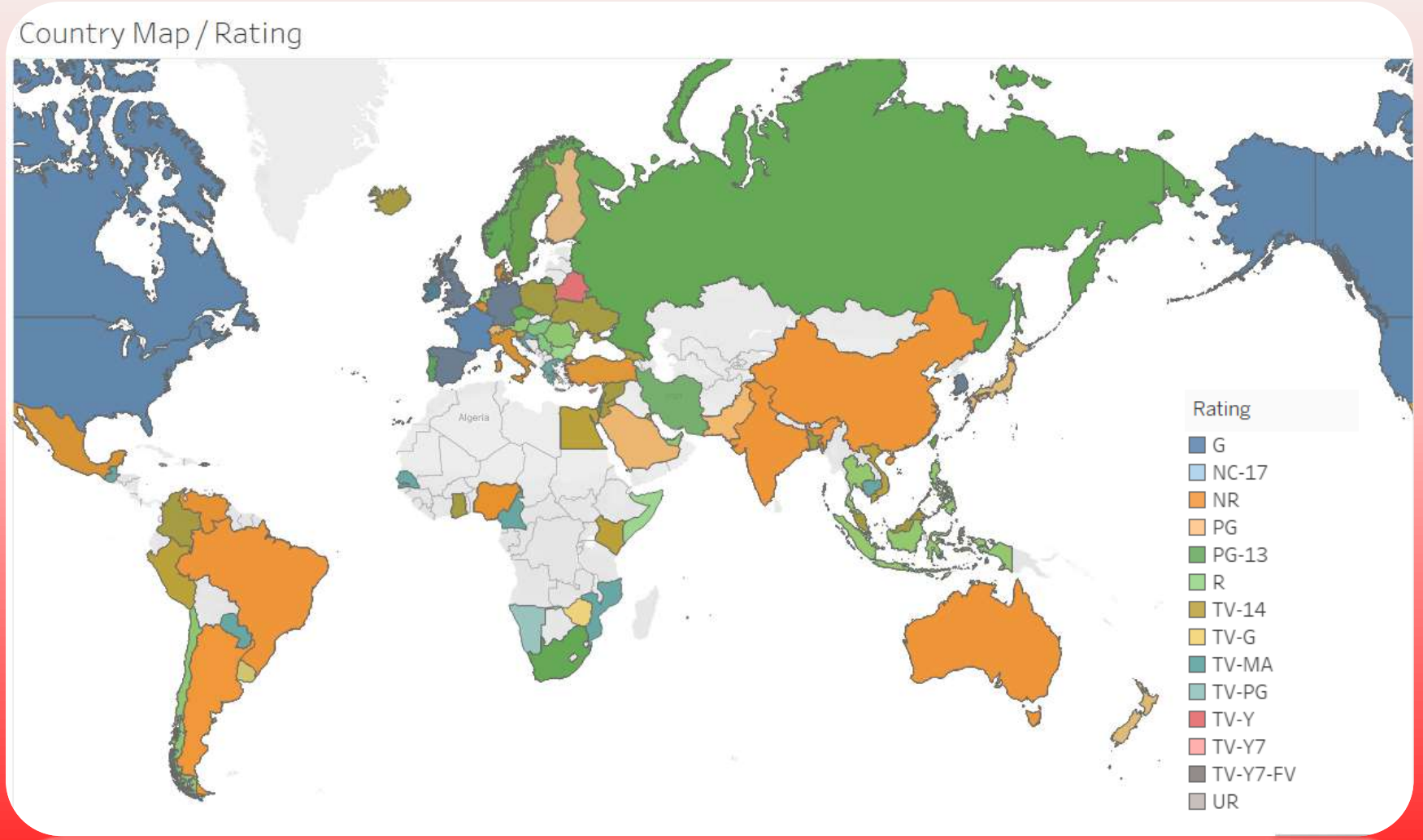
4

PEAK PRODUCTION PERIOD (2015–2020):

The graph highlights a surge in content during this period, reflecting Netflix's aggressive expansion and investment in building its catalog to attract a global audience.

This map illustrates the global distribution of Netflix content ratings, reflecting its strategy to cater to diverse audiences. Regions like the United States and Canada predominantly feature MPAA ratings such as G, PG, PG-13, and R, while Europe and Asia show a mix of TV and movie ratings like TV-MA and TV-14, tailored to local preferences. South America and Africa exhibit a balance between family-friendly (PG) and mature content (TV-MA). Netflix's adaptation to regional standards, such as using age-appropriate labels (e.g., TV-Y7, TV-G), highlights its efforts to meet cultural, regulatory, and viewer-specific demands, ensuring a globally appealing library.

This map visualizes the distribution of Netflix content ratings across different countries.



The treemap visualizes the distribution of Netflix shows across different countries based on percentages. Here's an analysis:

United States Dominates:The United States holds the largest share of Netflix shows, with 36.86%. This indicates the dominance of American content on the platform.

India as the Second Largest Contributor:India contributes 12.03%, showing a significant presence of Indian shows on Netflix, likely driven by regional demand and content production.

United Kingdom:The United Kingdom ranks third with 7.26%, highlighting its notable contribution to Netflix's global catalog.

Other Significant Contributors:Pakistan (4.79%), Canada (3.08%), and Japan (2.95%) have a smaller, yet noticeable share of shows.

Smaller Contributions:Countries like South Korea (2.43%), France (2.42%), Spain (2.07%), and Mexico (1.57%) show modest representation, likely due to regional preferences or production scales.

Less Represented Countries:Countries such as Brazil (1.00%), Thailand (0.75%), and Indonesia (0.98%) have minimal representation, indicating either lower production volume or lesser global reach.

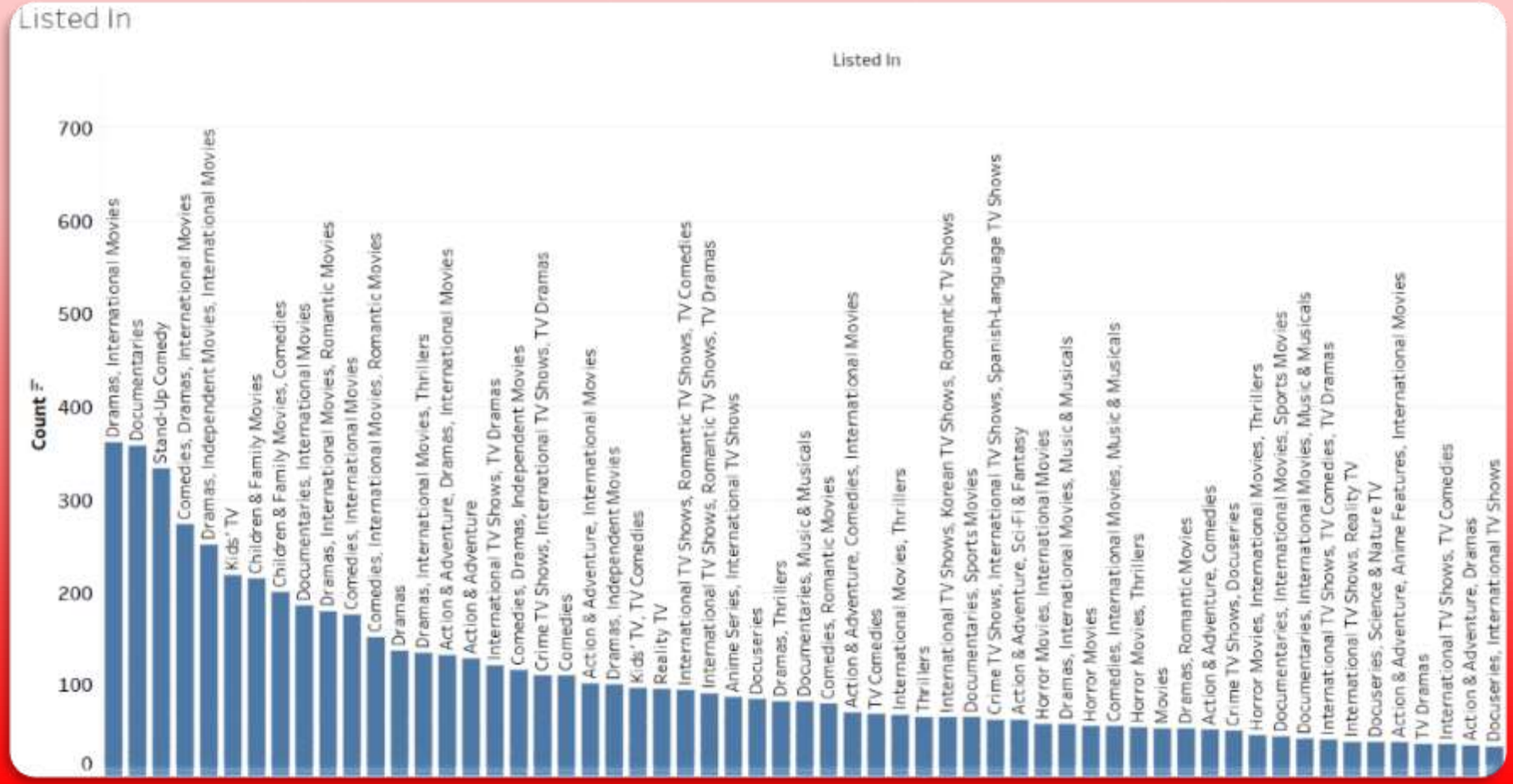
In conclusion, the treemap shows Netflix’s catalog is dominated by the United States (36.86%), followed by India (12.03%) and the UK (7.26%). Countries like Pakistan, Canada, and Japan contribute smaller shares, while regions like Brazil and Thailand have minimal representation, suggesting opportunities for regional growth. The "Not Given" category (3.27%) indicates some unspecified content origins.

The bar chart represents the distribution of Netflix content across various genres/categories. Here's an analysis with 3 key points:

Top Categories:The most popular category is **Dramas, International Movies**, followed by **Documentaries** and **Stand-Up Comedy**, indicating a strong preference for diverse and engaging content.

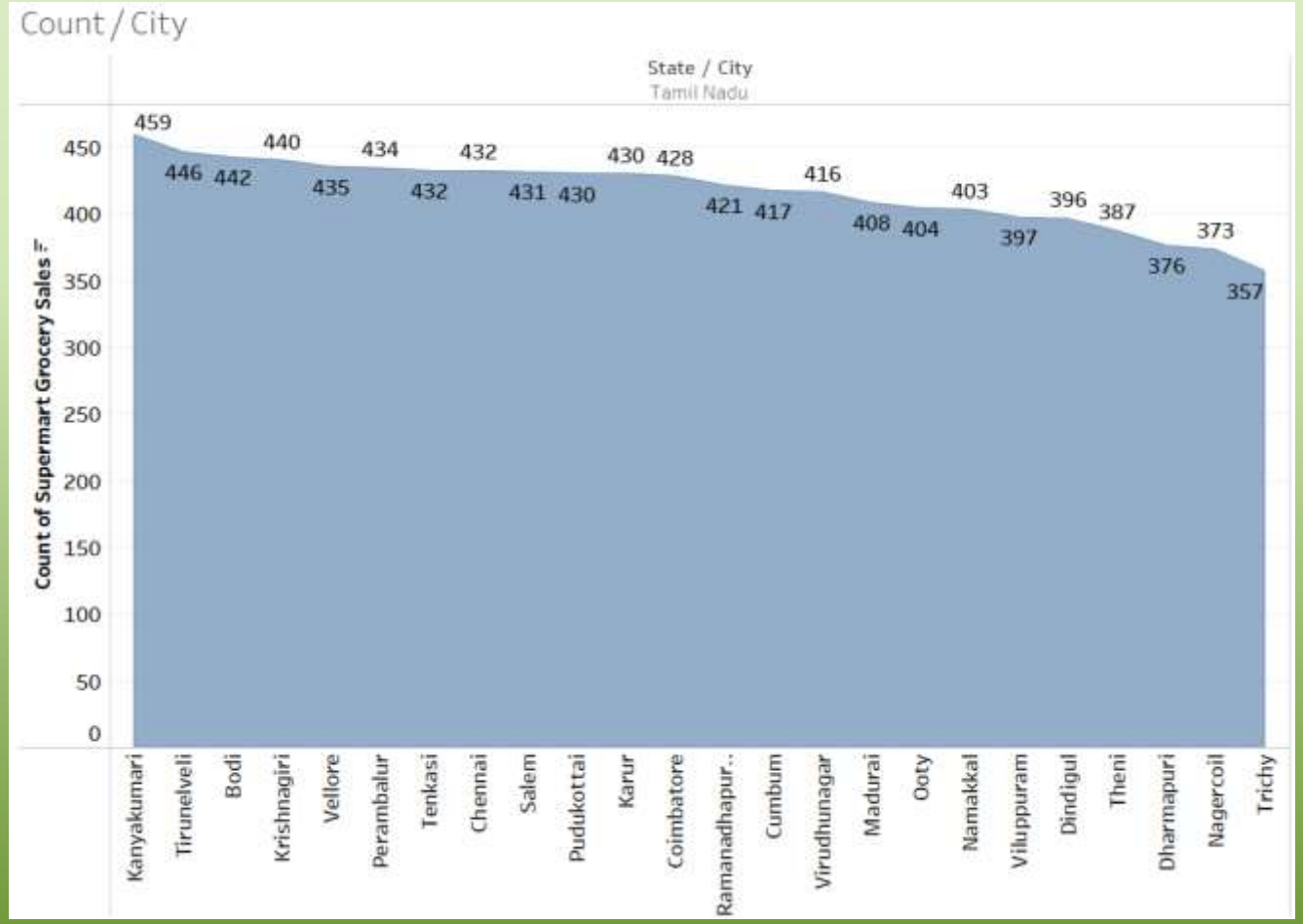
Children and Family-Oriented Content:Categories like **Kids' TV** and **Children & Family Movies** also have significant representation, showcasing Netflix's focus on catering to family audiences.

Niche Genres:Genres like **Anime Series, Spanish-Language TV Shows**, and **Korean TV Shows** are represented, though their counts are relatively lower, suggesting targeted content for specific audience groups.



3. Supermart Grocery Sales - Retail Analytics Dataset

The **Supermart Grocery Sales - Retail Analytics Dataset** captures sales performance across various cities, regions, and product categories. It is typically used to analyze trends, identify high-demand areas, and optimize supply chains for retail businesses. Insights derived from this dataset can guide marketing strategies, inventory management, and customer targeting efforts.

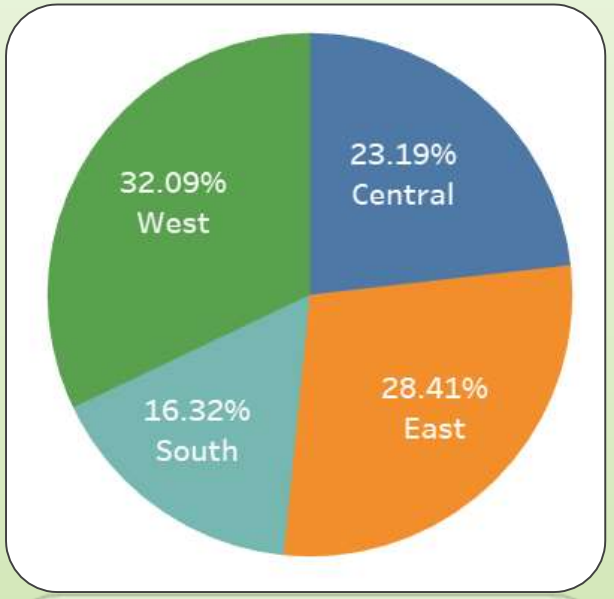


The chart showcases the count of Supermart Grocery sales in different cities across Tamil Nadu. Key observations include:

- **Top-Performing Cities**: Kanyakumari leads with the highest sales count (459), followed by Tirunelveli (446) and Bodi (442).
- **Moderate Sales**: Cities like Chennai (432) and Salem (431) demonstrate steady sales performance.
- **Low-Performing Cities**: Trichy (357) has the lowest sales, followed by Nagercoil (373).

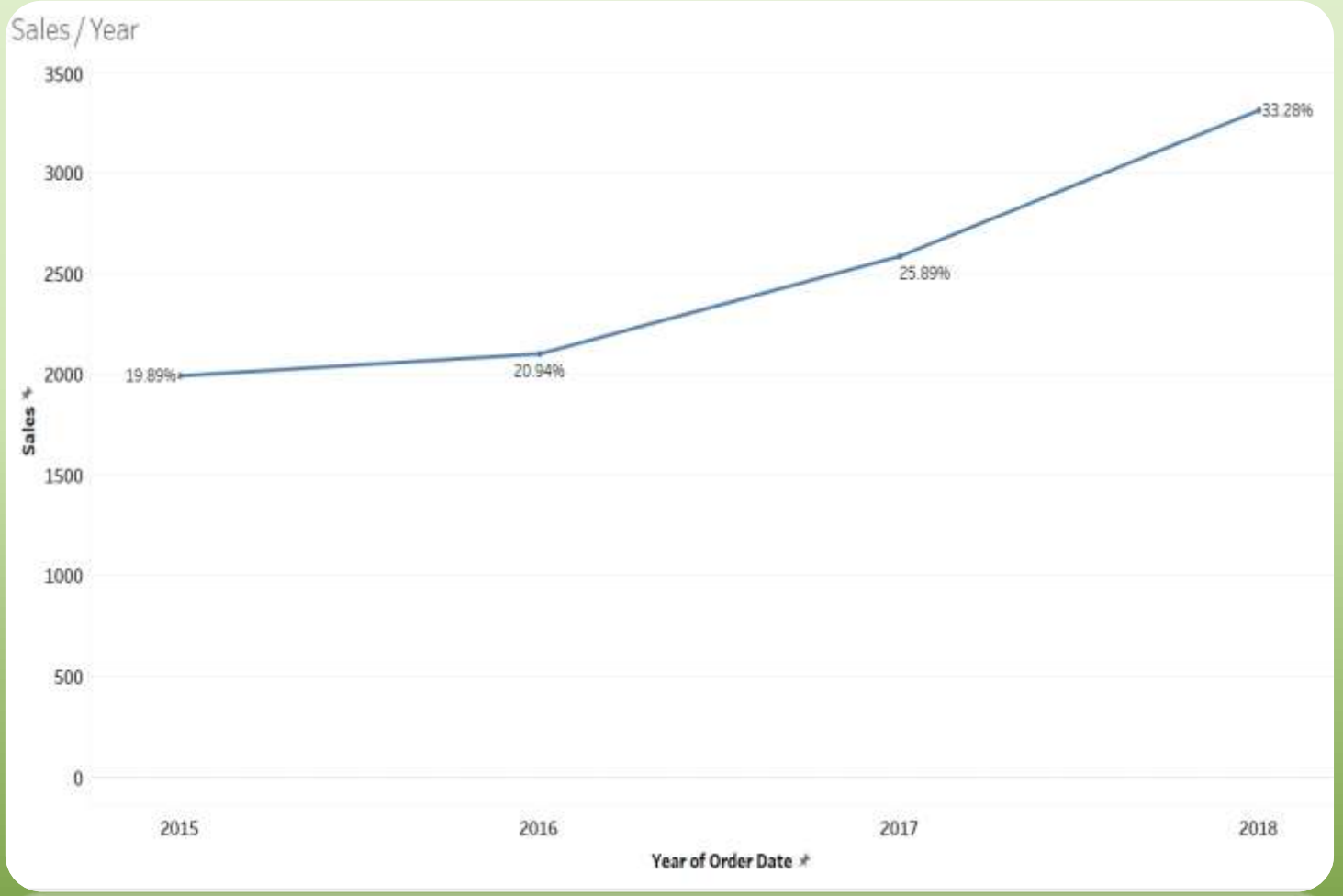
This distribution indicates a varying demand for groceries across different cities in Tamil Nadu, with smaller cities performing competitively compared to larger ones.

The pie chart highlights the distribution of sales across different regions, showing which regions perform better and identifying areas with lower sales potential.



1. **West Region**: Highest sales at 32.09%, indicating strong demand and effective retail operations.
2. **East Region**: Accounts for 28.41% of sales, showing significant contribution.
3. **Central Region**: Moderate sales at 23.19%, reflecting steady performance.
4. **South Region**: Lower sales at 16.32%, highlighting potential challenges or reduced demand.

This is a line graph showing sales performance over the years, with the percentage growth annotated for each year. Here are some observations:



•1. Overall Trend

•Consistent Growth: The sales exhibit a steady upward trend from **2015 to 2018**, indicating a positive performance trajectory.

2. Growth Rate

Increasing Annual Growth Rate:

2015: 19.89% **2016:** 20.94%

2017: 25.89% **2018:** 33.28%

Observation : accelerating growth

•3.Sales Volume

•Sales Progression:

•Started **below 2,000** in 2015. Rose to **exceed 3,000** by 2018.

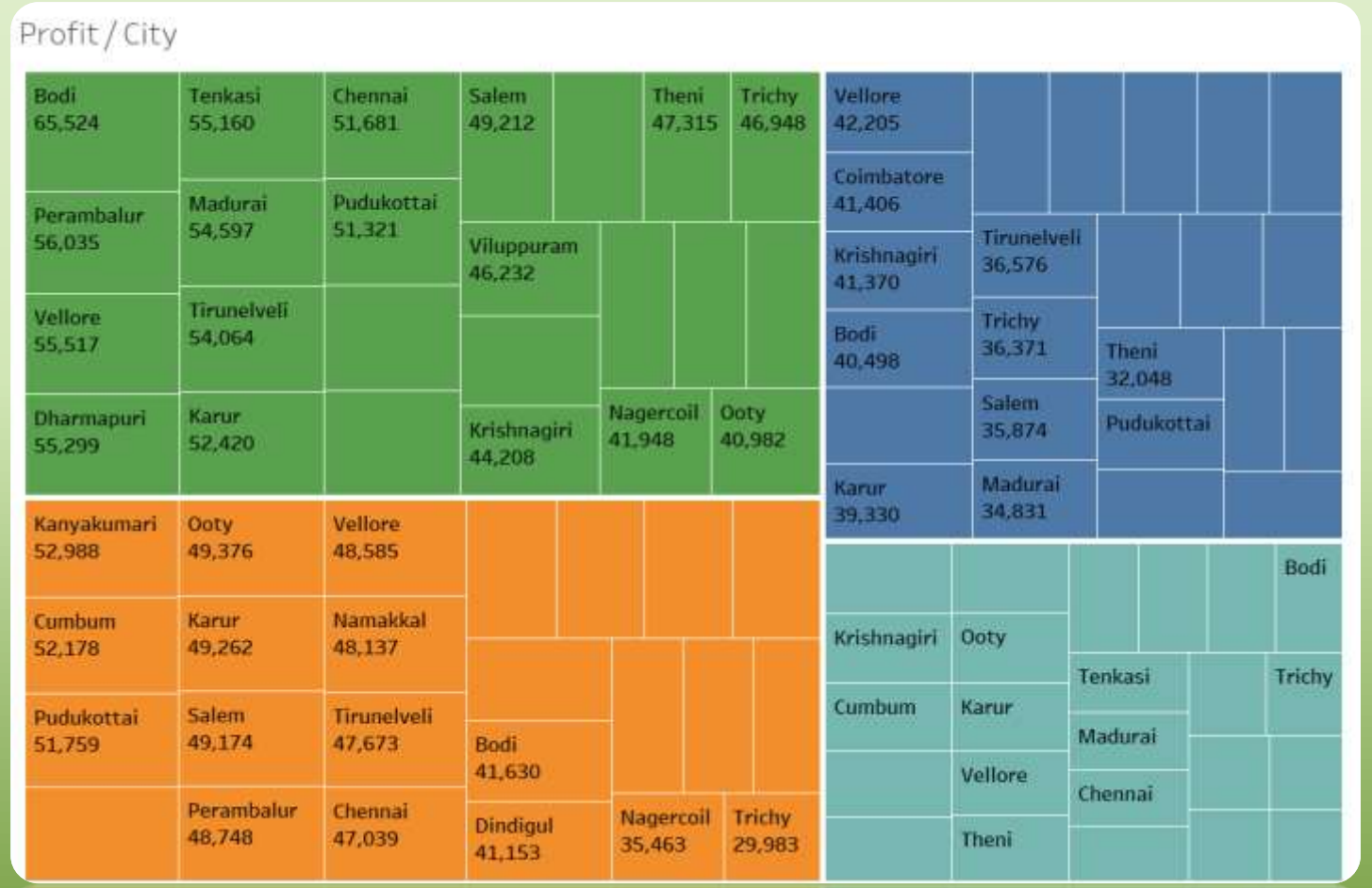
•Observation: Sales volume shows significant improvement, more than doubling over the four years.

•4. Performance Highlights

•Highest Growth Rate: The peak growth rate occurred in **2018** (33.28%), indicating that external factors (market conditions, strategy changes) may have influenced the business's performance that year.

•Lowest Growth Rate: 2015 had the lowest growth rate (19.89%), though still substantial.

This treemap shows the profit distribution across various cities. Each block represents a city, with the size corresponding to the profit amount, and the colour scheme possibly indicating different ranges of profit. Here's an analysis:



1

Top-Performing Cities:

- Bodi (65,524): Achieves the highest profit among all cities.
- Tenkasi (55,160) and Vellore (55,517): Also significant contributors to profit, indicating strong performance in these regions.

2

Moderately Performing Cities:

- Cities like Chennai (51,681), Pudukottai (51,321), and Kanyakumari (52,988) maintain a balanced profit, contributing consistently to overall earnings.

3

Low-Performing Cities

- Cities such as Theni (32,048), Nagapattinam (35,463), and Trichy (29,983) show the lowest profit figures, indicating potential challenges in these markets.

4

Profit Distribution:

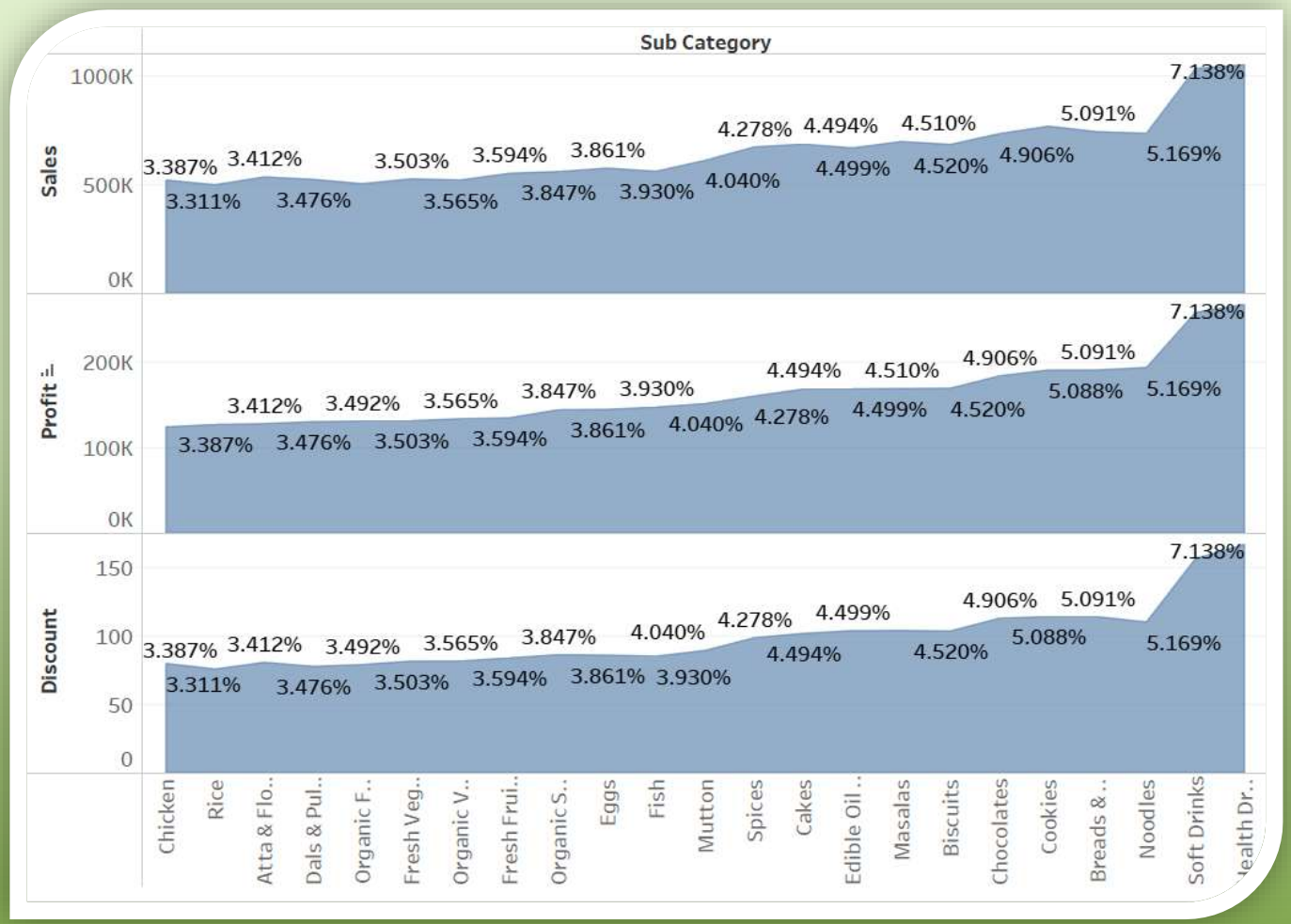
- Profits appear evenly distributed across many cities, with a few cities like Bodi standing.
- Smaller block sizes in the lower range may indicate markets that require strategic focus to boost performance.

This graph illustrates Sales, Profit, and Discounts across various Sub-Categories, with percentage contributions indicated for each metric. Here's an analysis:

1. **High-Performing Sub-Categories** : Soft Drinks: Leads in Sales (7.138%), Profit (7.138%), and Discount (7.138%), indicating strong consumer demand but possibly higher promotional efforts. Breads & Noodles and Cookies: Perform well across all metrics but slightly trail Soft Drinks.

2.**Moderate Performers** : Edible Oil, Masalas, and Biscuits maintain relatively balanced contributions in sales, profit, and discounts, indicating consistent demand without excessive discounting.

3. **Low-Performing Sub-Categories** : Chicken, Rice, and Atta & Flours have lower percentages in sales and profit, reflecting lower consumer demand or possibly smaller market segments.



THANK YOU...