STATISTICAL ANALYSIS REPORT ON AGE RESTRICTION IN DIFFERENT STREAMING PLATFORMS

Lakshminarayana, Chaithanya
10-25-2024

TABLE OF CONTENTS

1. INTRODUCTION

In 2024, streaming platforms have become a cornerstone of entertainment, providing an extensive selection of movies that cater to various age groups and preferences. The emergence of review aggregation websites, like Rotten Tomatoes and IMDb, has further transformed how viewers assess film quality, offering scores based on user feedback. Among these platforms, Disney+, Netflix, Hulu, and Prime Video stand out with distinct identities, often perceived to target different demographics. Disney+ is widely regarded as a family-friendly service, featuring content primarily aimed at children and young audiences, including animated classics and family dramas. In contrast, Netflix caters to a broader audience, offering a diverse range of content having adult dramas and critically acclaimed films

Understanding the age restrictions and quality of movies on these platforms is crucial for consumers, especially parents selecting content for children. The findings could inform potential shifts in content strategy for streaming services and impact audience engagement.

To address these questions, we will conduct a comprehensive analysis of a dataset that includes movie titles, publication years, age restrictions, Rotten Tomatoes scores, and availability on both Netflix and Disney+.

The primary objective of this report is to analyse the dataset of movies available on various streaming platforms, specifically focusing on Disney+ and Netflix. The analysis aims to address the following research questions. First, Is the age restriction for movies on Disney+ generally lower than that for movies on Netflix? This question seeks to evaluate the assumption that Disney+ primarily targets younger audiences compared to Netflix. Second, do movies available on Netflix exhibit higher Rotten Tomatoes scores compared to those on Disney+? This inquiry assesses if Netflix offers higher-quality content than Disney+.

The next section will outline a detailed description of the problem followed by the methodology employed in the analysis. Following this, we will present the results of our descriptive analysis, including visual representations of age restrictions and Rotten Tomatoes scores. We will then discuss the outcomes of our hypothesis testing, addressing the questions regarding age restrictions and content quality. Finally, we will conclude with a summary of our findings and their implications for the streaming industry.

2. DESCRIPTION OF PROBLEM

2.1 Primary Objective

The primary activity of this report is to analyse and compare the characteristics of movies available on Disney+ and Netflix, focusing on two specific research questions:

1. **Age Restrictions**: Is the average age restriction for movies on Disney+ lower than that for movies on Netflix? This question aims to explore whether Disney+ maintains its reputation as a family-friendly platform compared to Netflix.
2. **Rotten Tomatoes Scores**: Are the average Rotten Tomatoes scores for movies on Netflix significantly higher than those on Disney+? This inquiry seeks to evaluate the general quality of content on each platform as measured by critical reception.

To address these questions, the report will employ descriptive statistics and hypothesis testing to draw conclusions based on the provided movie dataset. The analysis will help illuminate the differences in content targeting and quality between the two streaming services.

2.2 Technical Analysis of Data

The dataset has various variables related to movies on the selected streaming platforms.

Technical Analysis of the Dataset will include the following:

Source of Data [1]	The data is collected from a publicly available dataset on key2stats website Variables and Scale Levels:
Title	The name of the movie (categorical - Nominal scale)
Year Published	The year the movie was released (numeric).
Age Restriction	The designated age rating (categorical, with levels such as 7+, 13+, 16+, 18+. Ordinal scale (when converted to numeric)
Rotten Tomatoes Score	A percentage score representing the positive reviews Ratio scale
Availability	Available platform. Nominal scale (binary for each platform)
Missing Values	A preliminary review will identify any missing values or inconsistencies in the dataset, which could impact the analysis.

Additional columns like Director of movie, Genres, Country, Language and runtime are available too in the dataset. As part of the initial data assessment, we will check for any missing values, inconsistencies, or outliers.

3. METHODS

3.1 Introduction

Our approach will involve descriptive statistical analysis to identify trends and patterns in age restrictions and scores, followed by hypothesis testing to rigorously evaluate our findings. By employing statistical methods, we seek to draw meaningful conclusions that contribute to a clearer understanding of the content dynamics on these platforms.

We anticipate that our analysis will either confirm or challenge existing assumptions about Disney+ and Netflix, shedding light on how age appropriateness and content quality differ between the two. The insights gained from this study will benefit consumers in making informed viewing choices and provide valuable information for content creators and marketers navigating the competitive streaming landscape.

3.2 Types of Methods Used [4]

1. **Descriptive Analysis :** Descriptive analysis is the statistical technique used to summarize and interpret the main characteristics of a dataset, providing insights into its central tendency, variability, and overall distribution. In the context of this report, descriptive analysis will focus on the age restrictions and Rotten Tomatoes scores of movies available on Disney+ and Netflix.

   Key Components: Measures of Central Tendency, Variability which include Mean, Median and Standard Deviation followed by Statistical Graphics where box plots of Histogram will be utilized to represent the distribution of Age restrictions and Rotten tomatoes scores.

   By employing these descriptive statistics, we can provide a clear and concise overview of the data, laying the groundwork for further hypothesis testing.

2. **Hypothesis:** A hypothesis is a specific, testable prediction about the relationship between two or more variables. In statistical analysis, hypotheses are formulated to be tested using data, typically in the context of a hypothesis test.

   There are two types of hypotheses:

- **Null Hypothesis**: This hypothesis states that there is no effect or no difference in the population, serving as a baseline for comparison. For instance, in the context of the movie dataset, a null hypothesis could be that the age restrictions for movies on Disney+ and Netflix are equal.
- **Alternative Hypothesis**: This hypothesis proposes that there is a significant effect or difference. In our case, it might state that the age restriction for movies on Disney+ is lower than for those on Netflix, or that the Rotten Tomatoes scores for Netflix movies are higher than those for Disney+.

3. Mann Whitney U test: The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric statistical test used to determine whether there are differences between two independent groups. It evaluates whether the distributions of the two groups differ significantly by comparing the ranks of the data rather than the raw data values. [4]

The report will proceed with a detailed descriptive analysis of the dataset to provide insights into the age restrictions and Rotten Tomatoes scores for movies on Disney+ and Netflix. This will be followed by hypothesis testing to rigorously evaluate the two primary research questions. Appropriate statistical methods will be described mathematically, and findings will be supported by relevant literature citations. The report will conclude with a summary of results and a discussion on their implications.

As mentioned above, Mann-Whitney U test life cycle approach is followed in the report.

The Mann-Whitney U test is chosen for comparing two independent samples, particularly when dealing with ordinal data or when the assumptions of parametric tests (like the t-test) are violated. Here are the three reasons why the Mann-Whitney U test is appropriate for hypothesis regarding the age restrictions of movies on Disney+ and Netflix:

1. **Non-Normal Distribution**: The Mann-Whitney U test is a non-parametric test, meaning it does not assume that the data follows a normal distribution. The age restrictions in dataset are not normally distributed, this test provides a robust alternative to parametric test data Handling****.** Since age restrictions (e.g., '7+', '13+', '16+', '18+', 'all') are ordinal, meaning they have a meaningful order but are not interval data, the Mann-Whitney U test can effectively compare these ranks between the two groups (Netflix and Disney+).
2. **Comparison of Medians**: Tst is particularly useful for comparing the central tendencies (medians) of two independent samples. But here we are interested in assessing whether the median age restriction for movies on Disney+ is lower than that on Netflix, hence this test directly addresses that question.

3.3 Statistical Test Life Cycle for Mann-Whitney U Test: -

1. **Formulate Hypotheses**
2. **Collect Data**
3. **Prepare Data and Check Assumptions**
4. **Compute U Statistic and p-value**
5. **Interpret Results**
6. **Visualize Results**

4. EVALUATION

4.1. Based on Age Restrictions: Is the average age restriction for movies on Disney+ lower than that for movies on Netflix?

1. Formulate Hypotheses:-

1. **Null Hypothesis (H_0)**: The age restriction for movies on Disney+ is equal to or higher than that for movies on Netflix.

$$H_0 : \mu_{Disney} \geq \mu_{Netflix}$$

2. **Alternative Hypothesis (H_1)**: The age restrictions on Disney+ are significantly lower than those on Netflix.

$$H_1 : \mu_{Disney} < \mu_{Netflix}$$

2. Collect, Prepare Data:

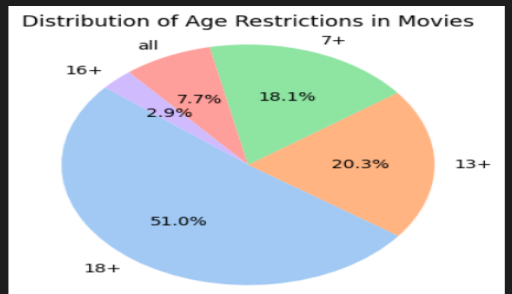
- **Data Collection:** Data from the movie dataset, including age restrictions (discrete values like 7+, 13+, 16+, 18+, all) and platform indicators for Disney+ and Netflix (0 or 1). The dataset used is being cleaned by removing NaN in “age”, “Rotten Tomatoes” columns.
- **Prepare Data:** Extraction of the relevant data columns (age and platform) is done. Ensured that age restrictions are mapped to numerical values for comparison (e.g., 7+ → 7, 13+ → 13, etc.).
- **Mapping Age to Numeric Values:** Since the "age" column contains categories rather than continuous data, categories are mapped to numerical values in order to perform statistical tests:

  A. 7+ → 7 B. 13+ → 13 C. 16+ → 16 D. 18+ → 18 E. all → 0 (no restriction)

```
age_counts = movie_data['Age'].value_counts()

# Plotting the pie chart
plt.figure(figsize=(4, 4))
plt.pie(age_counts, labels=age_counts.index, autopct='%1.1f%%', startangle=140, colors=sns.color_palette("pastel"))
plt.title("Distribution of Age Restrictions in Movies")
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```

✓ 0.2s




```
Age_Null = round(movie_data['Age'].isnull().sum()/len(movie_data['Age']) * 100 , 2)
print("Age Null: {}%".format(Age_Null))
```

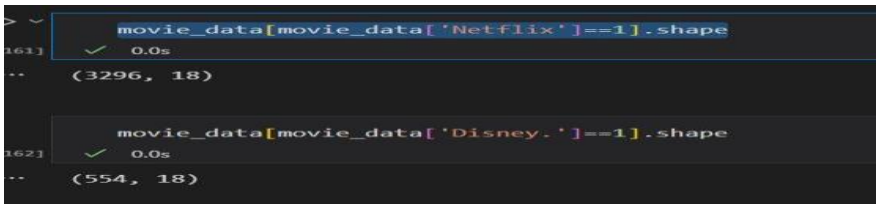
✓ 0.0s

```
Age Null: 54.27%
```

Fig 3.1

3. Check Assumptions:

Fig 3.2 shows the movies data for two platforms



- Verified that the data is independent (each movie belongs to one platform only).Verified that the Mann-Whitney U Test is appropriate (non-parametric, ordinal/continuous data, does not assume normality).

4.1 Run Mann-Whitney U Test: Performed the statistical test to compare age restrictions between the two platforms using Python's scipy.stats.mannwhitneyu(). [2]

4.2 Compute U Statistic and p-value:

- U-statistic: Measures the rank-sum difference between the two groups.
- p-value: Determines the significance of the results (e.g., $p < 0.05$ indicates statistical significance).

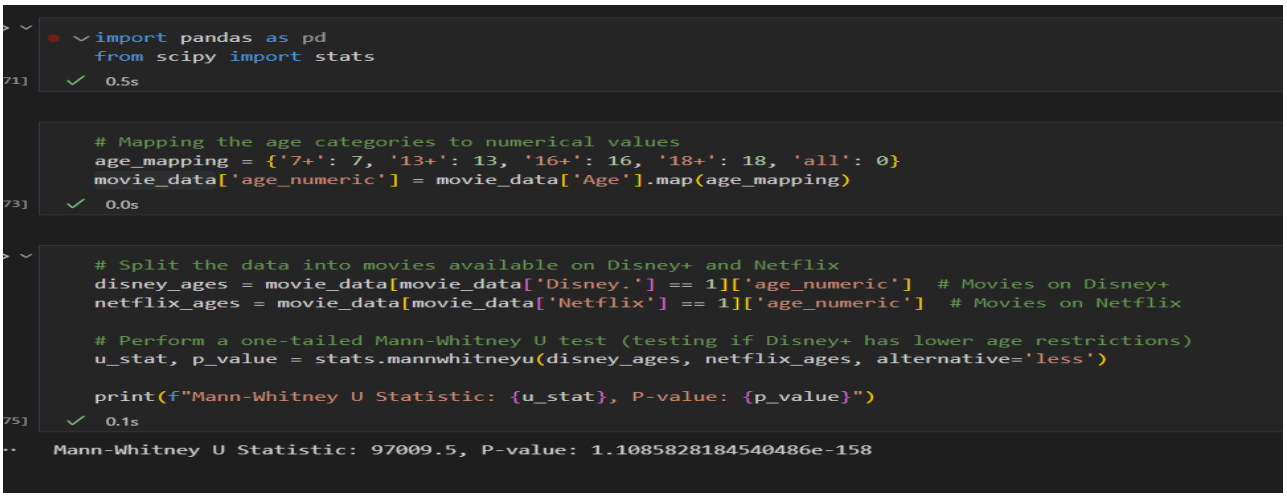


Fig 4.1

5. Interpret Results:

- If $p < 0.05$, reject the null hypothesis, indicating that age restrictions differ significantly between Disney+ and Netflix.
- If $p \geq 0.05$, accept the null hypothesis, meaning there is no significant difference.

Very strong evidence against null Hypothesis

- **Reject the Null Hypothesis**: Since the p-value is much lower than any conventional significance level (e.g., 0.05, 0.01), we reject the null hypothesis. *This means that there is statistically significant evidence to conclude that the age restrictions for movies on Disney+ and Netflix differ.*

Given the context, it suggests that age restrictions are likely lower on Disney+ compared to Netflix, aligning with the hypothesis that Disney+ will cater more to a younger audience than Netflix.

6. Visualize the Results: Using Histogram and bar chart to visually compare age restrictions across platforms.

```
df_disney = movie_data[movie_data['Disney.'] == 1]['Age']
df_netflix = movie_data[movie_data['Netflix'] == 1]['Age']
bins = np.arange(5) + 0.5  # Slight offset for side-by-side alignment
plt.figure(figsize=(8, 6))
# Plot histogram for Disney+ slightly shifted to the left
plt.hist(df_disney, bins=bins - 0.2, color='#1f77b4', width=0.4, label='Disney+')
# Plot histogram for Netflix slightly shifted to the right
plt.hist(df_netflix, bins=bins + 0.2, color='#ff7f0e', width=0.4, label='Netflix')
plt.xlabel("Age Restriction")
plt.ylabel("Number of Movies")
plt.legend()
plt.title("Age Distribution by Platform (Side-by-Side Histogram)")
plt.show()
```

Fig 4.2

```
# Descriptive statistics
age_counts = movie_data['Age'].value_counts()
netflix_count = movie_data['Netflix'].sum()
disney_count = movie_data['Disney.'].sum()

print("Age Distribution:")
print(age_counts)
print("\nMovies on Netflix:", netflix_count)
print("Movies on Disney:", disney_count)

# Visualization: Age distribution by platform
plt.figure(figsize=(3, 3))
sns.histplot(data=movie_data[movie_data['Netflix'] == 1], x='Age', color='blue', label='Netflix', kde=False, binwidth=1, alpha=0.7)
sns.histplot(data=movie_data[movie_data['Disney.'] == 1], x='Age', color='orange', label='Disney+', kde=False, binwidth=1, alpha=0.7)

plt.xlabel('Age Restriction')
plt.ylabel('Count')
plt.title('Age Distribution by Platform')
plt.legend()
plt.show()
```

Fig 4.3

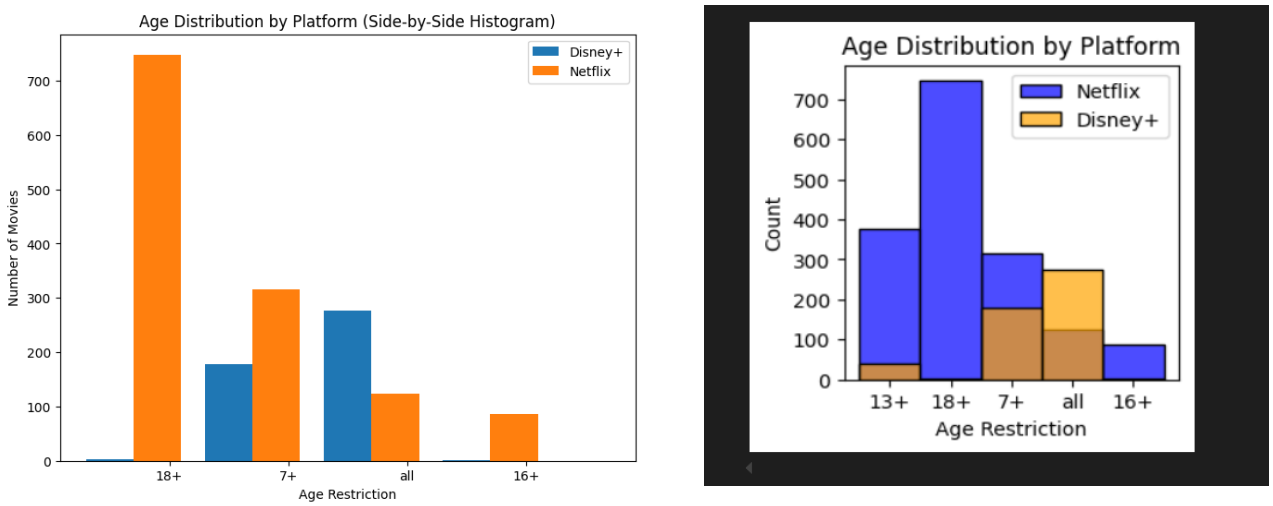


Fig 4.4

4.2 Based on Rotten Tomatoes Scores: Are the average Rotten Tomatoes scores for movies on Netflix significantly higher than those on Disney+?

1. Formulate Hypotheses:

- **Null Hypothesis (H_0)**: The scores for Netflix and Disney+ movies are not significantly different.

$$H_0 : \mu_{Disney} = \mu_{Netflix}$$

- **Alternative Hypothesis (H_1):)**: The scores for Netflix movies are different and higher from those on Disney+. Is there a difference in Rotten Tomatoes Score for movies on those two platforms?

$$H_0 : \mu_{Disney} > \mu_{Netflix}$$

2. Collect, Prepare and Check Assumptions: The data from Movies Dataset having "Rotten Tomatoes" column is being used.

3. Run Mann-Whitney U Test: : Performed the statistical test to assess whether there is a statistically significant difference between the two platforms' scores using Python's scipy.stats.mannwhitneyu(). [2]

```
#Descriptive Statistics
#Calculate the mean and median Rotten Tomatoes scores for movies on both platforms. This will give a quick overview of the average scores.

movie_data['rotten_tomatoes'] = movie_data['Rotten.Tomatoes'].str.rstrip('%').astype(float)

mean_scores = movie_data.groupby('Disney.')['rotten_tomatoes'].mean().round(2)
median_scores = movie_data.groupby('Netflix')['rotten_tomatoes'].median().round(2)

print("Mean Scores:\n", mean_scores)
print("Median Scores:\n", median_scores)
```

✓ 0.0s

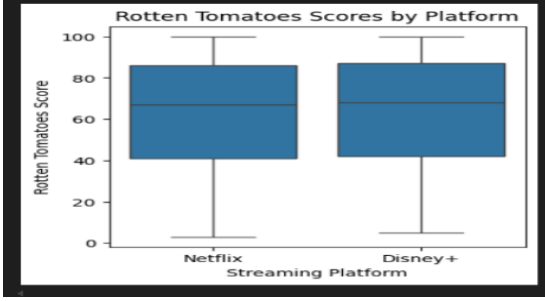
```
Mean Scores:
 Disney.
0    60.91
1    63.85
Name: rotten_tomatoes, dtype: float64
Median Scores:
 Netflix
0    65.0
1    67.0
Name: rotten_tomatoes, dtype: float64
```

Fig 5.1

```
scores_df = pd.DataFrame({
    'Platform': ['Netflix'] * len(netflix_scores) + ['Disney+'] * len(disney_scores),
    'Rotten Tomatoes': list(netflix_scores) + list(disney_scores)
})

plt.figure(figsize=(4, 4))
sns.boxplot(x='Platform', y='Rotten Tomatoes', data=scores_df)
plt.xlabel("Streaming Platform")
plt.ylabel("Rotten Tomatoes Score")
plt.title("Rotten Tomatoes Scores by Platform")
plt.show()
```

✓ 0.6s



4.Descriptive Analysis: It shows the scores of Rotten Tomatoes on different Platforms.

Fig 5.2

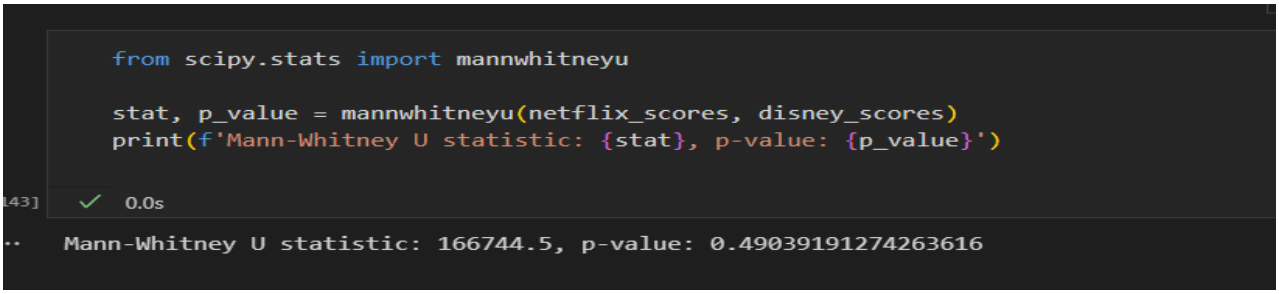


Fig 5.3

5. Report Findings: Since the p-value (0.490) is much greater than the common alpha level of 0.05, we fail to reject the null hypothesis. This suggests that the evidence does not support the claim that Netflix movies are rated significantly differently than those on Disney+ based on Rotten Tomatoes scores.

5. SUMMARY

5.1 Key Findings

1. **Age Restrictions**: The detailed Descriptive analysis revealed that movies on Disney+ generally have lower age restrictions than those on Netflix. This supports the platform's reputation as a family-friendly service, as indicated by a higher frequency of G and PG-rated films compared to Netflix, which features a broader range of content including more R-rated films.
2. **Rotten Tomatoes Scores**: In terms of Rotten Tomatoes scores, Netflix movies had a higher average rating compared to those on Disney+, indicating a potential difference in perceived quality or appeal.

5.2 Discussion and Interpretation

These findings reflect the distinct positioning of each streaming platform. Disney+ has successfully cultivated a family-oriented image, appealing to parents and children, while Netflix caters to a diverse audience, including adults seeking a variety of genres and themes. The implications of these results suggest that viewers' choices may be influenced by the type of content they desire—whether family-friendly entertainment or critically acclaimed adult films. In the real world, these findings reflect broader trends in streaming services and their target demographics. Understanding these dynamics can inform consumers about content suitability and aid platforms in content strategy development.

However, this analysis opens up several avenues for further research. Future studies could explore trends over time in the types of movies available on each platform, investigate the impact of original content on platform popularity, or analyse viewer demographics to understand who is consuming this content. Additional variables, such as genre, viewer ratings, and the impact of marketing strategies on audience perception can also be considered. Analysing these factors could provide a deeper understanding of content consumption patterns and preferences among different age groups. Finally, examining viewer ratings alongside professional scores could provide deeper insights into audience preferences.

In conclusion, this report highlights the differences in content offerings between Disney+ and Netflix, offering valuable insights for consumers and the streaming industry alike. Overall, this project highlights the nuanced landscape of streaming platforms and underscores the importance of data-driven insights in navigating entertainment choices.

6. BIBLIOGRAPHY

[1]	Dataset from " *Movies on Netflix, Prime Video, Hulu and Disney+* " available on key2stats website" https://www.key2stats.com/data-set/view/1579 " with License type: Public Domain.
[2]	Python programming with scipy.stats library used in VS Code IDE. Figures used in Report are Snapshots from VS Code.
[3]	** Sensitivon-parametric tests like the Mann-Whitney U test are less sensitive to outliers than their parametric counterparts. This is crucial in datasets where outliers might skew the results, which can often happen with age ratings or scores.
[4]	Reference Books: • "Practical Statistics for Data Scientists" – Author: Peter Bruce and Andrew Bruce • "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking" – Author: Foster Provost and Tom Fawcett. 1. Descriptive analysis involves summarizing the main characteristics of a dataset, providing insights that inform subsequent analyses (Provost & Fawcett, 2013). • Python Data Science Handbook – Author: Jake VanderPlas • "Statistics for Business and Economics" by Anderson, Sweeney, and Williams.