Character Personality Type (MBTI) and its Impact on Film Popularity

---

Team Members: Tanuja Voruganti, Janice Tran, Muhammad Faizan Raza, Animesh Agrawal, Sahil Batav

Project Goal: Determine if there is any relationship between the first lead and second lead's personality types and the film's popularity.

Research Question: How does the MBTIs of the first and the second lead impact the success of films?

Introduction

Films are more than just visual storytelling. They are works of art that represent the people, represent society, and they're woven from compelling characters, narratives, and emotions. The primary driving force in firms, however, are the characters and the plot that they're written into. They're all distinct, with their own characteristics, mannerisms, and expressions that help the audience see them as real people. These traits and mannerisms of the characters, along with their thoughts and emotions and motivation be analyzed using psychological tools. One such tool is called MBTI.

The Myers-Briggs Type Indicator is a psychological tool that classifies individuals into sixteen distinct personalities. The personality is a combination of four personality traits.

- (I)ntrovert vs. (E)xtrovert: If people gain energy from being alone or with others.
- (S)ensing vs. I(N)tuitive: Whether people prefer facts and specifics above big ideas and possibilities.
- (T)hinking vs. (F)eeling:Whether people make decisions based on logic and reason or on their own feelings and beliefs.
- (J)udging vs. (P)erceiving: Whether individuals like to be organized and planned, or spontaneous and adaptable.

Classifying film characters according to these MBTI personality types, we can uncover trends that shed light on how these traits affect the narrative development and the way the audience resonates with the film. Understanding these dynamics is important, especially when these characters can impact how the films are received by the audience. This affects the popularity and success of the film, which is why it is important to study it.

Significance of Study

Films are crucial to societal impact. This is because they can influence public opinion and attitude. Furthermore, films can deliver messages in many ways, such as teaching life lessons, bringing attention to emerging issues, or providing introspection to other cultures, lifestyles, and more. While films are influential, so are the characters that appear in them, specifically the protagonist and their respective antagonist.

The main character and the rival are key characters that can shape the likability of a film. This is because these key characters can be relatable, and therefore, easy to empathize with. Main characters typically are present for the entirety of the film, meaning viewers are able to see their development, which is essential to most film storylines. Antagonists can greatly influence the main character's development, therefore, they are just as pivotal to the film. Character development serves to engage the audience, and can play a major role in the story overall.

Exploring the relation between character personality and film popularity can provide insights into the values and goals that resonate with the public. This is because films shape morals, depict cultures and lifestyles, and spread awareness of issues that need more attention. Characters in film can represent these ideologies. Using film popularity and character personality, it can be determined what is important to society and ourselves.

Hypotheses and Expected Results

When it comes to general results, we anticipate that the models we create will give us insights into the relationship between the personality types of the First Lead and Second Lead, and the popularity of the movies. There are a couple of hypotheses that we propose when it comes to the results of this project.

1. Considering the genres of the movie, we speculate that there will be a different personality type that dominates for each.
2. When taking movie timelines into perspective, we hypothesize that over time, the dominating personality types for First Leads and Second Leads might change.
3. Since the relationship between the First Lead and Second Lead are a crucial element of movies, we theorize that they will either have opposite or identical personality traits, depending on the genre.
4. There may be certain generic personality types that will be present in movies regardless of the genre or other features.

Data Collection and Preprocessing

Data Sources

Name	Reference
IMDB Movies Dataset (*date accessed: 1st September 2024*)	https://datasets.imdbws.com/
Personality Database (*MBTI of characters*)	https://www.personality-database.com/

- **IMDB**: In order to create the dataset needed for the objective, first we incorporated the dataset provided by IMDB, a website that showcases ratings and reviews for movies and media alike. The dataset consists of t-const, the primary title of the movie, the average rating, the number of ratings, and the genre of the movie.
- **PDB**: Next, we then scraped the profile id, character name, movie name, personality type, and the character votes from PDB, a website that displays character MBTI types from various types of media.

Scraping

1. **Selenium**: A python module used in conjunction with chrome webdriver to automate web browsing.
2. **Randomizing User Agent**: To prevent blocking or scraping detection, five random user agents were created to simulate user interaction.

```python
# Function to randomly generate a User-Agent
def random_user_agent():
    """Generate a random user-agent to mimic different browsers and avoid detection."""
    user_agents = [
        "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.3",
        "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36",
        "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.157 Safari/537.36",
        "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/75.0.3770.142 Safari/537.36"
    ]
    return random.choice(user_agents)
```

3. **Meta keyword**: On the website, there is no visible label for what type of media the character belongs to. In the website source, there are keywords for what category the

character belongs to. Since we are searching for movie characters, the keyword is "Movies."

```python
# Function to check if the profile contains 'Movies' in the meta keywords
def check_if_movie_category(browser, profile_id, skip_list):
    """Check if the profile contains 'Movies' in the meta keywords and skip if it contains other unwanted keywords."""
    try:
        # Initial wait for 5 seconds (increased timeout)
        WebDriverWait(browser, 5).until(EC.presence_of_element_located((By.XPATH, '//meta[@name="keywords"]')))

        # Extract the meta keywords tag content
        meta_keywords = browser.find_element(By.XPATH, '//meta[@name="keywords"]').get_attribute('content')

        # Skip if any unwanted keyword is found
        for keyword in skip_keywords:
            if keyword in meta_keywords:
                print(f"Profile ID {profile_id} contains '{keyword}'. Skipping.")
                skip_list.append(profile_id)  # Add to the skip list
                return False

        # Check if the keyword list contains 'Movies'
        if "Movies" in meta_keywords:
            print(f"Movie keyword found. Scraping profile ID {profile_id}.")
            return True
        else:
            print(f"Profile ID {profile_id} does not contain 'Movies'. Skipping.")
            skip_list.append(profile_id)  # Add to the skip list
            return False
```

4. **RegExpression**: A python module that matches the desired strings. It searched for profile id, character name, movie name, and number of four letter votes of the MBTI.

```python
# Regex extraction for name, personality type, show/movie, and votes
name_match = re.search(r'<h1 class="profile-name">(.*?)</h1>', page_source)

if name_match:
    character_name = name_match.group(1).strip()

    # **Check if the character has already been scraped (lead or second lead)**
    if is_character_scraped(characters_df, character_name):
        print(f"Character '{character_name}' has already been scraped. Skipping.")
        skip_list.append(profile_id)  # Add to skip list
        return None  # Skip this character if already scraped

    character_data['name'] = character_name

personality_match = re.search(r'<div class="profile-personality">Personality Type: (.*?)</div>', page_source)
show_match = re.search(r'<h1>(.*?)</h1>', page_source)
votes_match = re.search(r'<label class="personality-vote-title">Four Letter</label>.*?<label class="personality-vot
```

5. **Randomizing Delays**: Incorporated random time delays with an interval between 5 and 15 seconds for each character to stimulate web browsing.

```
# Introduce a random delay between 5 and 15 seconds after each profile is scraped
delay = random.uniform(5, 7)
print(f"Sleeping for {delay:.2f} seconds to avoid overloading the website.")
time.sleep(delay)
```

6. **Batches**: The code scrapes in batches set by the user for efficiency and sake of computational power. Once a batch is completed, it is appended to the csv and the next batch starts.

```
if __name__ == '__main__':
    # Specify the profile range you would like to start with
    start_id = 80000
    end_id = 90000
    # Amount per batch, keep this low for smoother performance, 10 is also fine
    batch_size = 5
    # File name to save scraped data
    file_name = 'profiles80000to90000.csv'
```

Example Output:

```
Scraping profile ID: 89003
Movie keyword found. Scraping profile ID 89003.
{'profile_id': 89003, 'name': 'The Monster', 'personality_type': 'INFP', 'show_movie': 'Frankenstein', 'four_letter_votes'
Sleeping for 6.25 seconds to avoid overloading the website.
Scraping profile ID: 89004
Movie keyword found. Scraping profile ID 89004.
{'profile_id': 89004, 'name': 'Henry Frankenstein', 'personality_type': 'INTP', 'show_movie': 'Frankenstein', 'four_letter
Sleeping for 6.46 seconds to avoid overloading the website.
Batch 89000 to 89004 completed and appended to CSV.
Skipped profiles appended to skip_profiles.csv after batch 89000 to 89004.
Starting batch: 89005 to 89009
Scraping profile ID: 89005
Movie keyword found. Scraping profile ID 89005.
{'profile_id': 89005, 'name': 'Elizabeth Lavenza', 'personality_type': 'ISFJ', 'show_movie': 'Frankenstein', 'four_letter_
Sleeping for 6.98 seconds to avoid overloading the website.
```

Merging and the Final Dataset

Once completed, we combined both datasets to create our final dataset which consists of the following columns:

tconst	titletype	primarytitle	originaltitle	isadult	startyear	runtimeminutes	genres	tconst-2	averagerating	numvotes	lead_character	lead_MBTI	lead_votes	second_lead	secondlead_MBTI	secondlead_votes	relationship
tt0111161	movie	The Shawshank Redemption	The Shawshank Redemption	0	1994	142	Drama	tt0111161	9.3	2934011	Andy Dufresne	INTJ	1155	Warden Norton	ESTJ	215	Enemy
tt0468569	movie	The Dark Knight	The Dark Knight	0	2008	152	Action,Crime,Drama	tt0468569	9.0	2914377	Bruce Wayne	INTJ	1404	The Joker	ENTP	1767	Enemy
tt1375666	movie	Inception	Inception	0	2010	148	Action,Adventure,Sci-Fi	tt1375666	8.8	2587898	Dominick Cobb	ISTP	321	Arthur	ISTJ	209	Friend
tt0137523	movie	Fight Club	Fight Club	0	1999	139	Drama	tt0137523	8.8	2365490	The Narrator	ISFJ	42	Tyler Durden	ENTP	86	Rival

Our approach with the analysis was to create two different datasets with the exact schema:

1. **MBTI_top_1000:** Consists of the top 1000 movies from the IMDB dataset based on the number of votes given to the rating of the movie.

2. **MBTI_all_4200:** Consists of a random sample of 3200 movies from the IMDB dataset where the number of votes was at least 10,000.

We chose the number of votes as a minimum of 10,000 movies because the movies with less than 10,000 votes did not have enough data on PDB for the characters' MBTI. Additionally, the random sampling was done by generating a random number for each of the movies, and then sorting them in descending order to pick the top 4000 movies. We chose 4000 as the number because of the time constraint with scraping the data from the PDB website. Later on, due to missing or erroneous values, the 4000 movies dropped down to a final list of 3200 movies.

Preliminary Analysis

After finalizing the datasets and doing the preprocessing, we performed basic statistical analysis to get an idea of what kind of data we are working with and how to proceed with the analysis.

We used df.describe() method to get the following results for MBTI_top_1000 and MBTI_all_4200:

	averagerating	numvotes	lead_votes	secondlead_votes
count	1000.000000	1.000000e+03	1000.000000	1000.000000
mean	7.119600	3.795006e+05	483.828000	343.988000
std	0.873601	3.789258e+05	689.808571	493.300877
min	1.900000	1.185960e+05	3.000000	3.000000
25%	6.500000	1.592780e+05	124.000000	101.750000
50%	7.150000	2.299345e+05	213.500000	179.500000
75%	7.725000	3.875650e+05	454.000000	360.000000
max	9.300000	2.934011e+06	5340.000000	5942.000000

df.describe() for MTBI_top_1000 dataset

The range of numvotes is interesting, ranging from almost 3 million votes to 118 thousand votes, despite the fact that this is the top 1000 movies from IMDB.

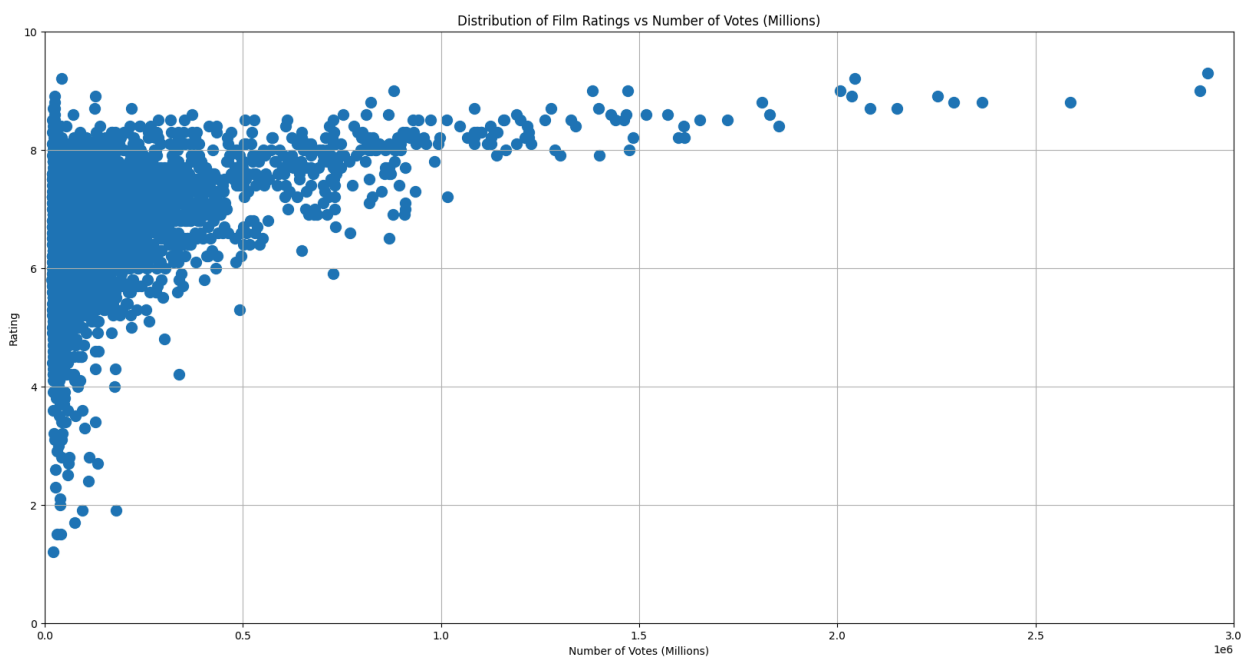
One particular distinction is the maximum *secondlead_votes* is larger than the maximum *lead_votes*, but this is because the character "Joker" from the movie "The Dark Knight" is extremely popular, and the incredible acting by the late actor Heath Ledger made the character iconic and loved by the audience.

Another interesting thing is the minimum for *averagerating* being 1.9/10 for the movie "Radhe" despite 180 thousand people voting for it, which shows that a movie can be popular despite being despised by the audience.

	averagerating	numvotes	lead_votes	secondlead_votes
count	4124.000000	4.124000e+03	4124.000000	4124.000000
mean	6.708584	1.558755e+05	203.461203	155.689622
std	0.982478	2.423878e+05	384.921096	275.201282
min	1.200000	1.728100e+04	3.000000	3.000000
25%	6.100000	3.461850e+04	70.000000	53.000000
50%	6.800000	6.931750e+04	102.000000	87.000000
75%	7.400000	1.672615e+05	165.000000	146.000000
max	9.300000	2.934011e+06	5340.000000	5942.000000

df.describe() for MTBI_all_4200 dataset

Our final dataset is made up of 4,200 films, which shows a variety of voting trends. The collection appears to contain films with varying degrees of popularity, as evidenced by the votes, which range from 17,000 to almost 3 million. Furthermore, both positive and negative popular films are represented by the movie ratings, which vary from 1.2 to 9.3. The dataset's mix of popular and less popular characters is reflected in the lead and second lead character votes, which range from 3 to 5,000.

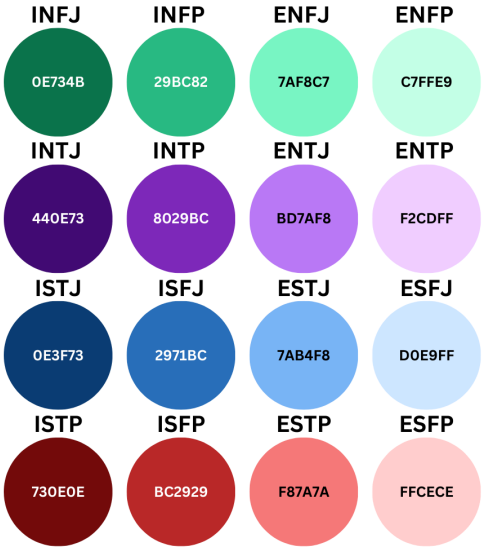


After observing the above number of votes vs rating graph carefully we can conclude that the most of our movies' votes are varying from 0 to 500K. Also for the rating most of the movie ratings are varying from 4 to 9.

Visualizations

Color Assignment and Classification:

When it comes to the visualizations, for ease of interpretation and understanding, we have divided the 16 MBTI personality types into four groups based off of the website 16Personalities (https://www.16personalities.com/personality-types) and we assigned colors to them.



color classification of MBTI types

<u>The Diplomats (Green):</u>

- They possess the N + F personality traits and are usually empathetic, passionate, and idealistic people.
- Example: Frodo Baggins from The Lord of The Rings.

<u>The Analysts (Purple):</u>

- They possess the N + T personality traits and are usually rational, intellectual, and righteous.
- Example: Bruce Banner (the Hulk) from The Avengers.

<u>The Sentinels (Blue):</u>

- They possess the S + J personality traits and are usually practical, stable, and prioritize security.
- Example: Mikasa Ackerman from Attack on Titan.

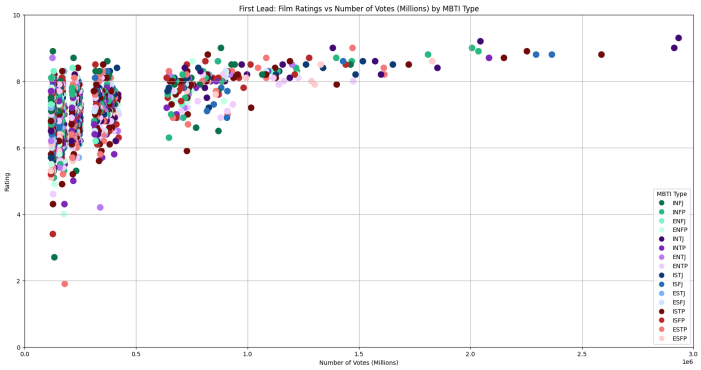
The Explorers (Red):

- They possess the S + P personality traits and are usually ingenious, flexible, and spontaneous people.
- Example: Monkey D. Luffy from One Piece.

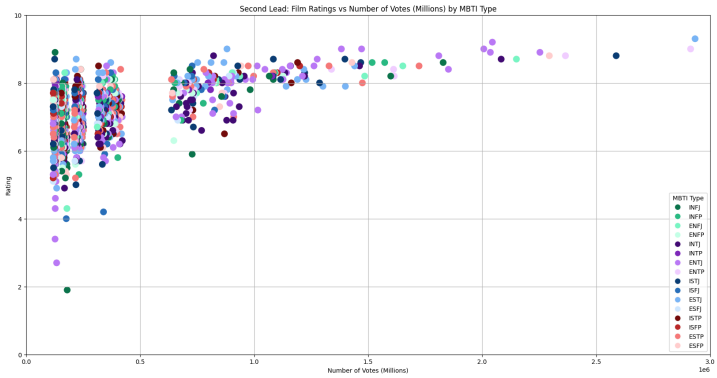
Another thing to observe about the classification and color assignment is that the darker colors are an indicator of Introversion whereas the lighter colors are for Extroversion. This makes it easier to visually analyze if the distribution has a presence of extroverts or introverts. Furthermore, the domination of a particular color is easier to understand because of the dominant traits that the personality types share.

Popularity Analysis:

Moving on, we tried analyzing the MBTI types based on the film popularity to see if there are any prominent trends. We plotted a singular scatter plot for First Lead and Second Lead characters against the Number of Votes (x-axis) and the Average Rating (y-axis), and obtained the following results.



MBTI_top_1000: First Lead MBTI distribution

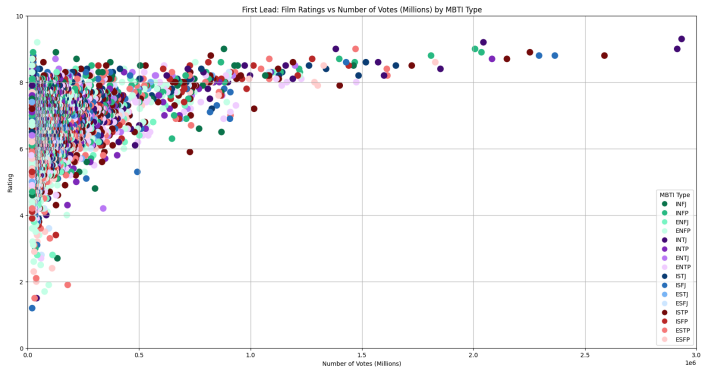


MBTI_top_1000: Second Lead MBTI distribution

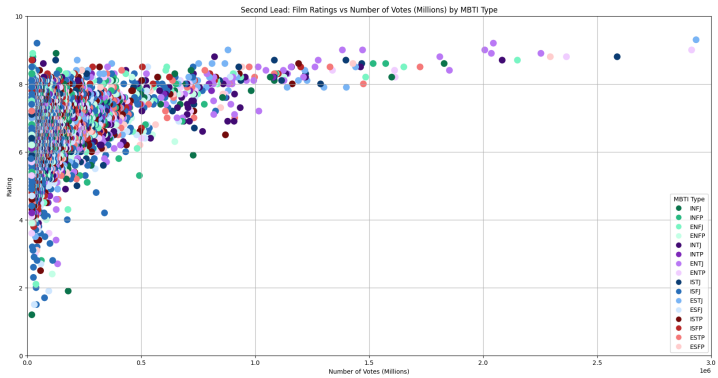
The above scatterplots show a diverse color graph of the MBTI's. However, it is important to note that for the First Lead, there is a dominating presence of Red in the overall movies; especially the darker reds. This indicates that for the top 1000 movies, a significant amount of Lead Characters are introverted, free-spirited, spontaneous, and ingenious. Additionally, for the First Lead, the top 2 movies—Shawshank Redemption and The Dark Knight—the personality type is INTJ. The domination of introverts can be explained by the genres of the top movies which will be analyzed in the further sections.

Coming to the Second Lead, it is imperative to note the presence of lighter colors, especially the purples. This indicates that while the Lead Characters are expected to be free spirited, the Second Leads are more rational and put together in order to balance out the energies and drive the plot forward.

We repeated the same analysis for the entire dataset of 4200 movies.



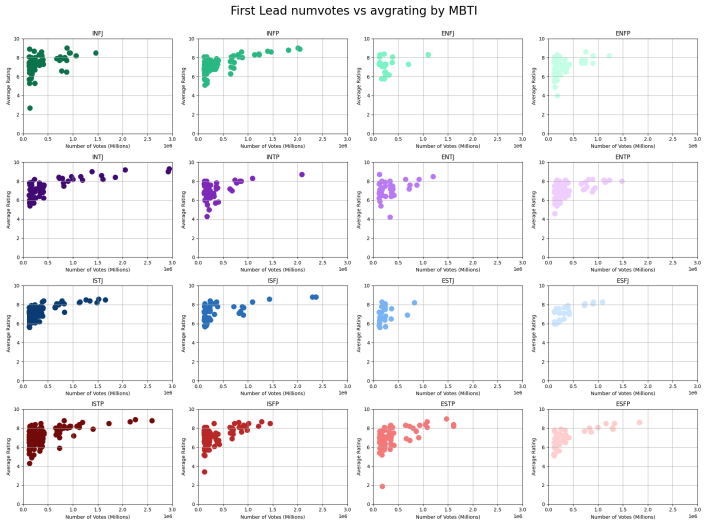
MBTI_all_4200: First Lead MBTI distribution

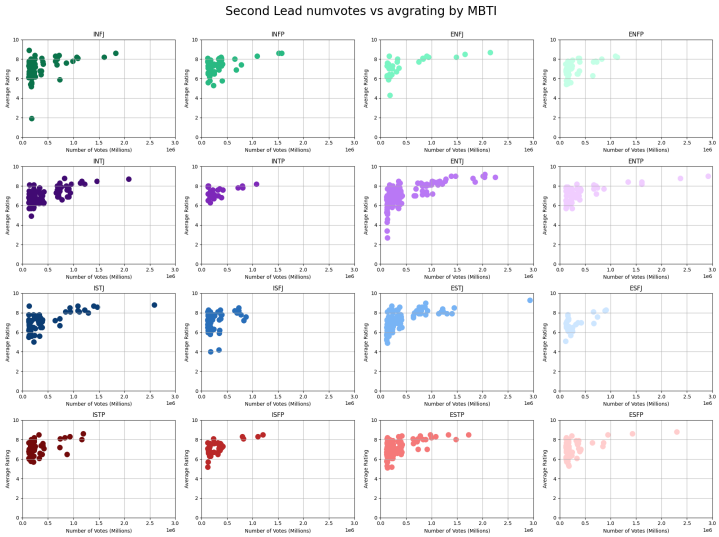


MBTI_all_4200: Second Lead MBTI distribution

It is intriguing to see the change in colors from the top 1000 movies to the entire dataset. For the First Lead, there are all sorts of MBTI types with a little dominance of Green.

Since the concentration of data points makes it harder to perform a more thorough analysis, we broke it down into separate scatter plots by the MBTI to visualize the trends more distinctly. The following sets of scatter plots show the distribution for the top 1000 movies.





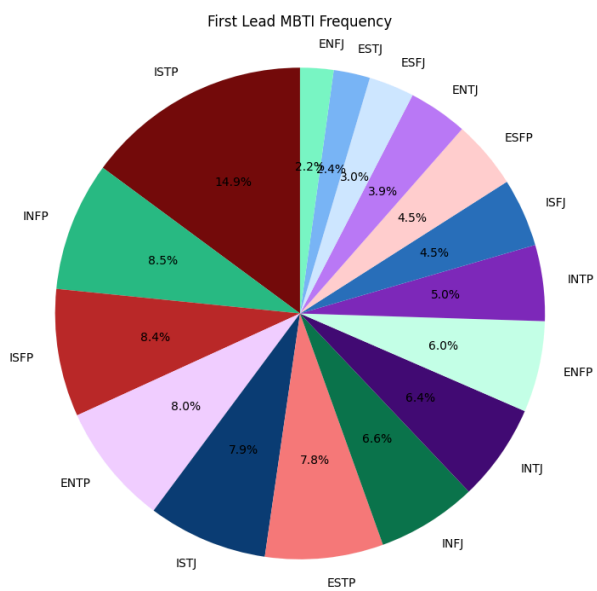
It is interesting to see that certain MBTI types for the First Leads are more scattered about while others are concentrated to form a cluster. The Introverted MBTI types for the First Lead are usually more spread out, with higher number of votes and a higher average rating. Whereas for the lighter colors (the Extroverts) are more clustered together with lower number of votes and a higher range in the ratings.

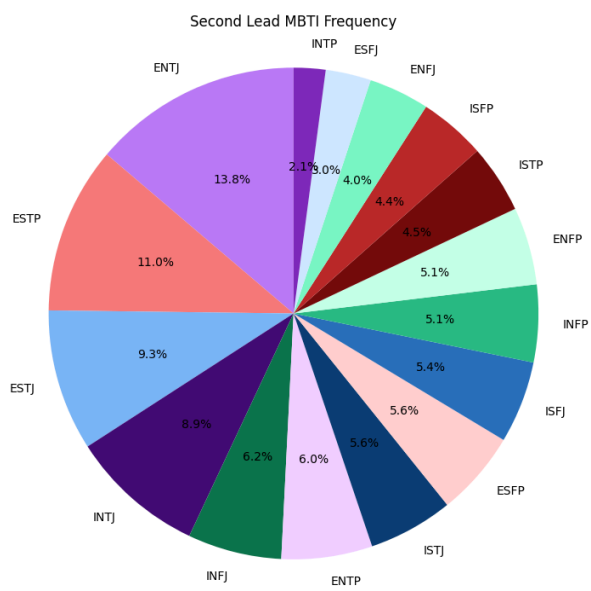
As for the Second Lead, it is the exact opposite. The Extroverted MBTI types are more spaced out into the higher ratings and higher votes, whereas the Introverted MBTI types are more clustered together. This once again backs up the idea of the First Lead and Second Lead being opposite in order to balance each other out.

In conclusion, the popularity analysis shows distinct trends in MBTI types among First and Second Leads based on film ratings and votes. In the top 1000 movies, First Leads are mainly Introverted and spontaneous, with INTJ types prominent in top films. Second Leads are more Extroverted and rational, providing balance. In the full dataset of 4200 movies, there's more variety among First Leads, with some Green dominance. Scatter plots reveal that Introverted First Leads generally receive higher ratings and votes, while Extroverted Second Leads tend to achieve higher ratings, highlighting their complementary roles in stories.

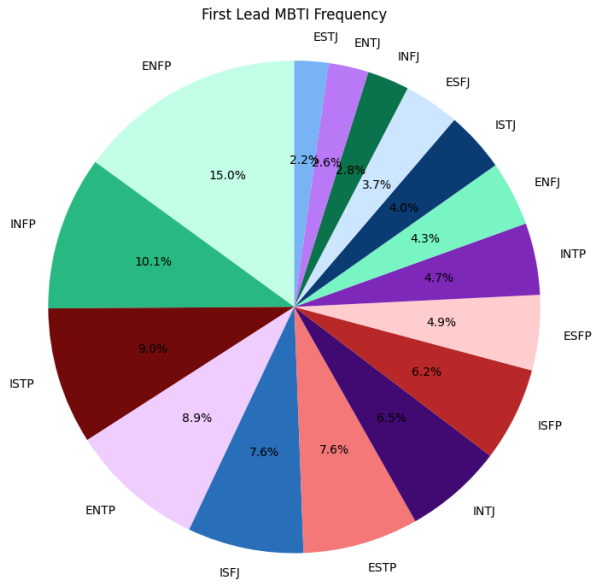
MBTI Distribution:

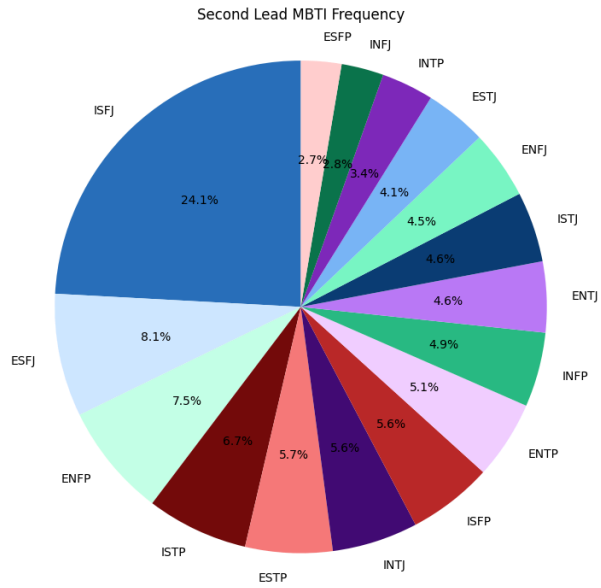
Moving on, we analyzed the distribution of each MBTI types and visualized them as Pie Charts for better understanding.





In the top 1000 movies, the dominating personality type for First Leads is ISTP, whereas for Second Leads it is ENTJ. The only common personality type they have is that they both share the Thinking (T) personality trait. This indicates that popular movies have a tendency of featuring intelligent and logical characters.





Coming to the entire dataset, it is interesting to see how the dominating personality types have changed. The First Lead has a Green dominance which indicates that films are featuring a lot of protagonists who are creative, diplomatic, empathetic, and in-touch with their emotions. It's also interesting to see that both ENFP and INFP share the N + F + P personality traits. As for the Second Lead, ISFJ is an extremely dominating personality type. This is also the first time we are seeing a Blue dominance in the analysis. This indicates that second leads are more likely to be logical, calculated, and practical about things, which balances out the creative and empathetic First leads.

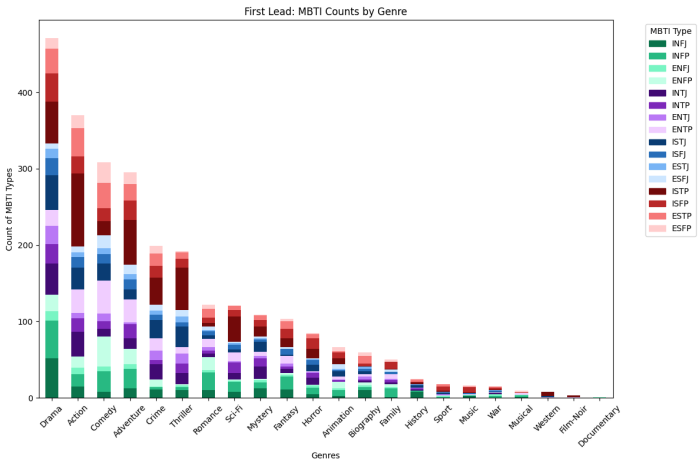
Another interesting observation is how both the First Lead and Second Lead's dominating personality types just differ on the Extroversion and Introversion aspect (ENFP + INFP and ISFJ + ESFJ).

In conclusion, the MBTI distribution analysis shows that ISTP is the dominant personality type for First Leads in the top 1000 movies, while ENTJ leads among Second Leads, both sharing the Thinking (T) trait. In the entire dataset, First Leads shift to a Green dominance, indicating more creative and empathetic characters, with ENFP and INFP types prevalent. For Second Leads, ISFJ is dominant, introducing a Blue characteristic that suggests practicality. Overall, the key difference between the leads is their Extroversion and Introversion, balancing creativity with logic.

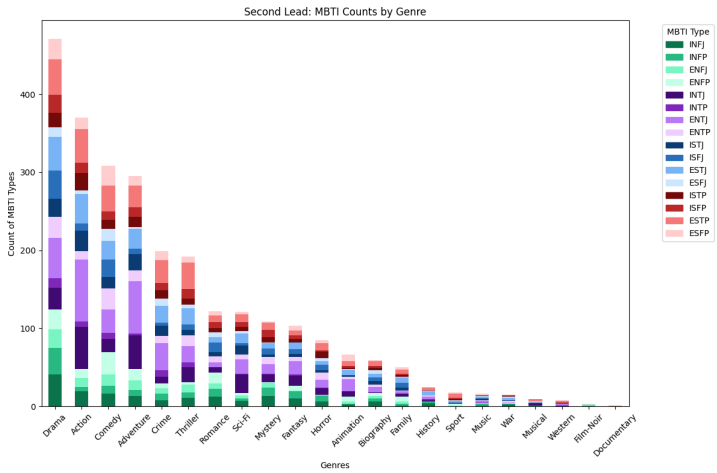
Genre Analysis:

One of the determining factors that affects the personality of a character when it comes to stories is the genre of the story being written. A protagonist from a Thriller movie would feel out of place in a Romance movie because their backstory, their drive, and the crux of their character would not fit in with the requirements of the genre.

Therefore, we decided to analyze the MBTI types based on the genres of the movies in our dataset, and the results are as displayed below.

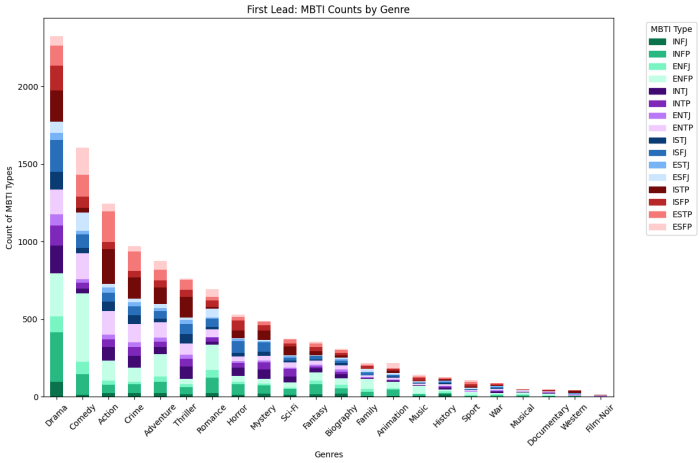


MBTI_top_1000: First Lead MBTI count by genre

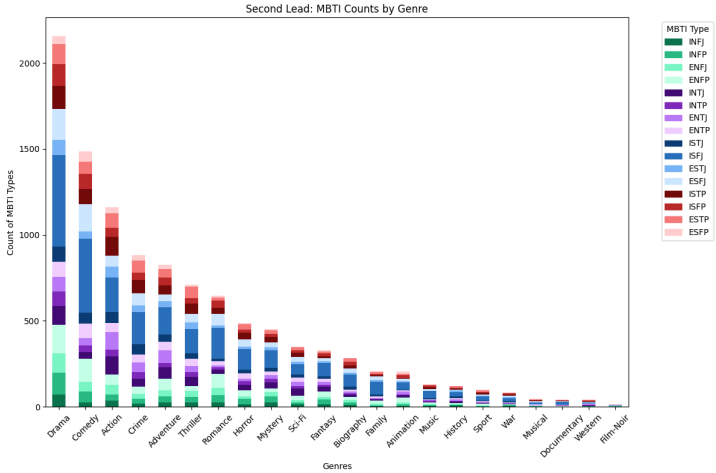


MBTI_top_1000: Second Lead MBTI count by genre

Despite the relatively even distribution of the four color groups in all the genres, it is fascinating to note that for the First Lead, darker colors are more dominating in all genres except for Romance, indicating that there is a preference for Introversion. However, it is the opposite for the second lead where there is a dominance of Extroversion as can be interpreted from the lighter colors.



MBTI_all_4200: First Lead MBTI count by genre

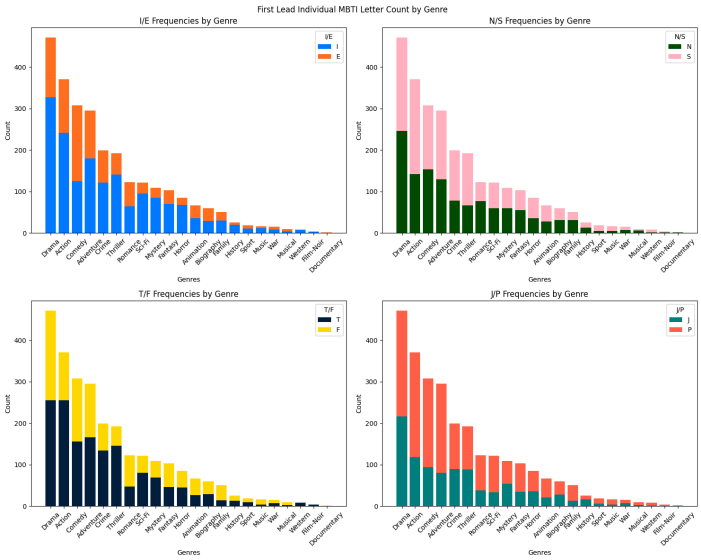


MBTI_all_4200: Second Lead MBTI count by genre

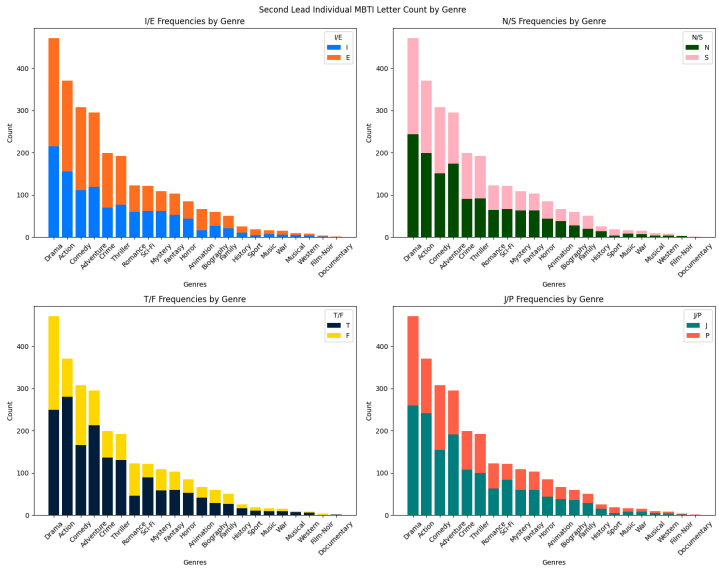
The charts reveal a noticeable dominance of Green MBTI for First Leads, which varies significantly by genre, as anticipated. In contrast, the Second Lead shows a clear preference for the ISFJ type and overall Blue characteristics.

This suggests that Second Leads are often written to be practical and analytical, providing a balance to their counterparts. Their roles either complement or contrast with the First Lead, enhancing the dynamics between the characters. This difference between personality types highlights the intentional choices writers make to create engaging and nuanced relationships within the story.

In order to be able to properly understand the underlying patterns, further analysis was done for the individual personality traits for both First Lead and Second Lead characters. First we look at the top 1000 movies.



MBTI_top_1000: First Lead Analysis by Traits

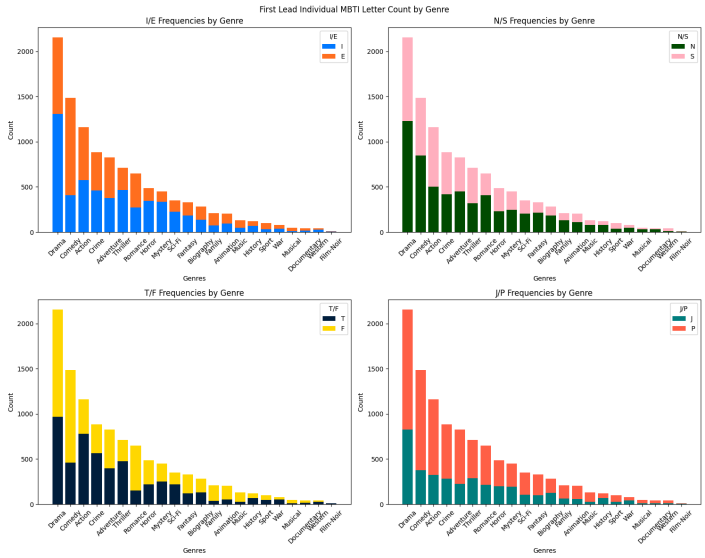


MBTI_top_1000: Second Lead Analysis by Traits

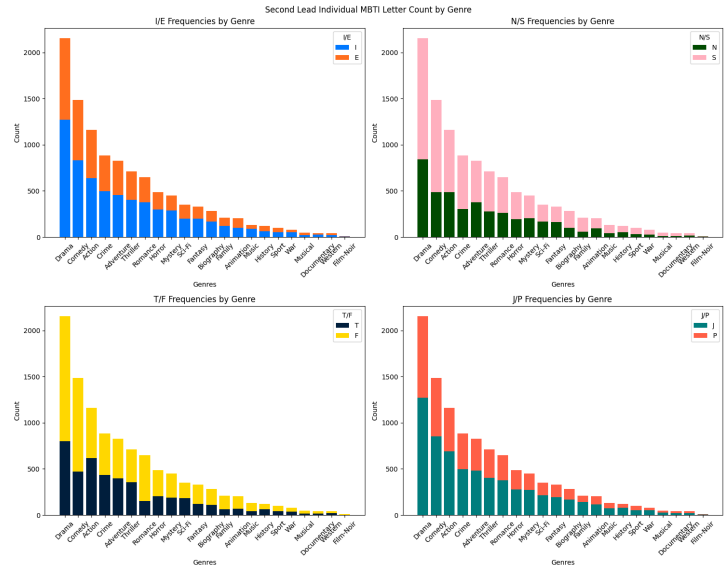
The analysis shows a clear preference for Introversion among First Leads, along with a notable inclination toward Thinking types in both leads. This trend shifts in Romance, where Feelers and Thinkers are more balanced, reflecting the genre's emphasis on emotional connections. First Leads also tend to Perceive, indicating a tendency to live in the moment, which aligns with the dominance of Red in the personality charts.

In contrast, for the top movies, Second Leads are usually Judging types, making them more rooted in the past and grounded, as evidenced by the prevalence of Purple in the charts. These distinctions highlight the intentional choices made in character development to enhance storytelling dynamics.

Now we compare the results of the top 1000 movies to the rest of the dataset.



MBTI_all_4200: First Lead Analysis by Traits



MBTI_all_4200: Second Lead Analysis by Traits

When it comes to Comedy, First Leads lean towards Extroversion, while Second Leads reflect a more balanced distribution of personality types. The balance between Thinking and Perceiving types in Mystery is intriguing, as one might expect the lead to be more practical; this could be explained by the innocence often associated with the character.

For Romance, it's amusing to note the scarcity of Thinking traits, reinforcing the genre's focus on emotional depth. Interestingly, the distribution of Judging and Perceiving types remains consistent across both the top 1000 and all 4200 films, suggesting a stable trend in character dynamics across different datasets.

In conclusion, the analysis shows that genre significantly shapes the personality traits of First and Second Leads in films. In the top 1000 movies, First Leads typically lean towards darker, more Introverted characteristics, while Second Leads lean towards Extroversion, highlighting their complementary roles. The preference for Thinking types in both leads, along with Judging traits in Second Leads, reflects intentional choices by writers to enhance character dynamics. Observations in genres like Comedy, Mystery, and Romance further reinforce how personality traits align with genre conventions, underscoring the intricate relationship between character development and storytelling.

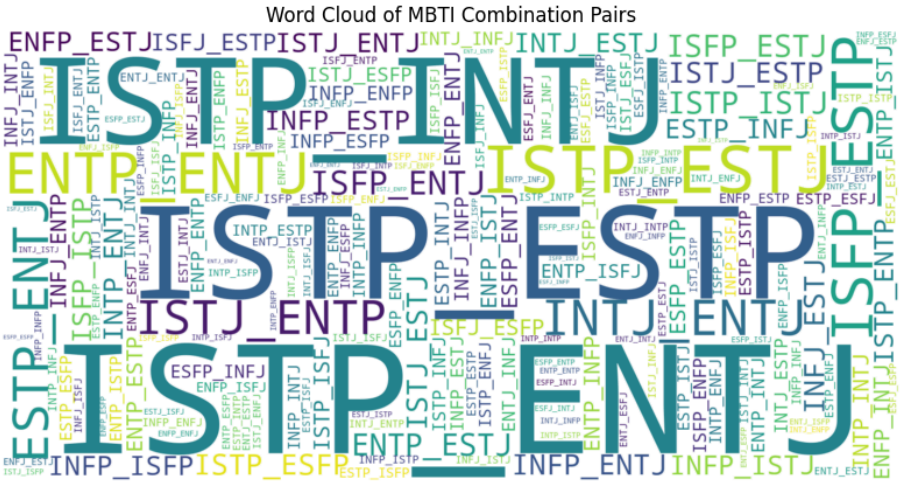
Character Pairs Analysis:

When creating the First and Second Lead characters, they are often made to be parallels or opposites of each other. The type of relationship they have can greatly influence the chemistry and the flow of the story. Rivals tend to push each other to their limits and they generally have

similar personalities. On the other hand, Friends usually have different personalities but they get along well with each other since they balance the other person's traits.

That is why it is important to see what kind of personality types the First Lead and Second Lead have in movies, so that we can better understand the nuances of the characters and what the audience finds most enjoyable.

We performed this analysis by creating word clouds in order to visualize the frequency of occurrence of the pairs. The following word cloud is for the top 1000 movies.

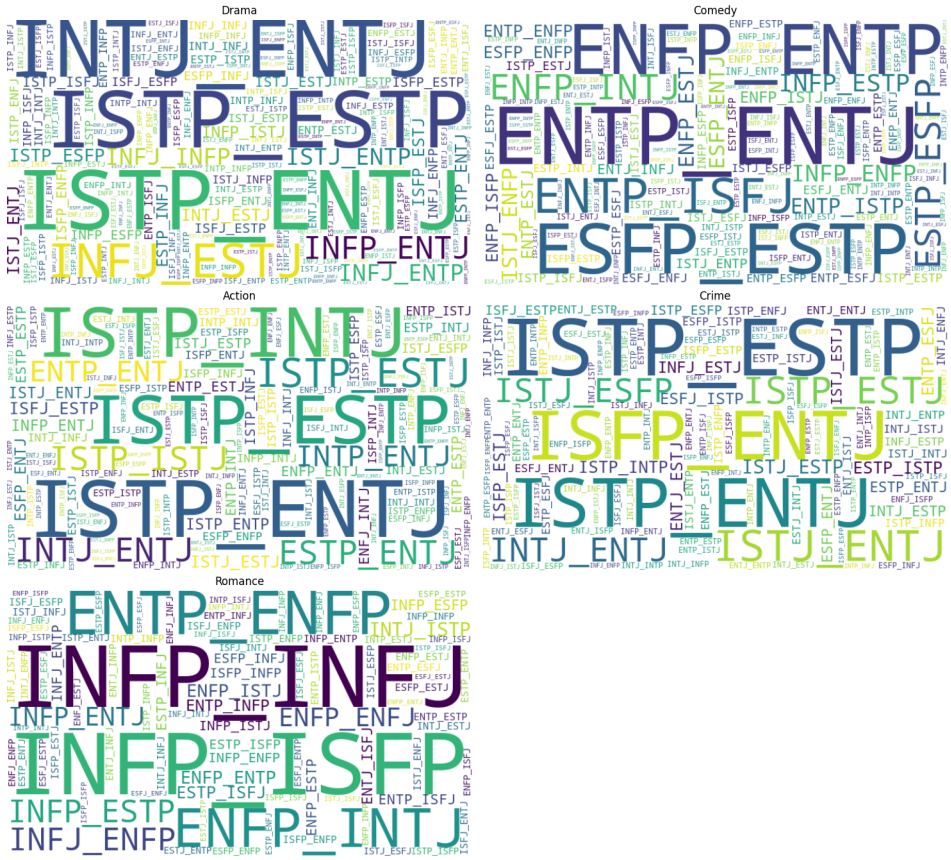


MBTI_top_1000: Word Cloud

It is extremely intriguing to see the dominating presence of ISTP for First Lead with different pairs of the Second Lead. This indicates that the ISTP personality type is compatible with a wide variety of personalities, which fits in with their nature of being innovative, practical, and resourceful.

One thing to take note of is ISTP paired with ESTP. It is the type of dynamic where they are very direct about the way they approach problems, and communicate effectively with each other. Considering how the top movies tend to have a lot of Action, Crime, and darker topics, this explains such a relationship.

We then analyzed the pairs based on the genres of the movies to get a better insight into the different dynamics.



MBTI_top_1000: Word Cloud distribution in Top 5 Genres

<u>Drama</u>:

- There is a dominance of Introvert First Lead and Extrovert Second Lead pairs.
- Pairs such as INTJ_ENTJ and ISTP_ESTP have only a single personality trait that differs, which indicates that characters who are very similar to each other are often present in these stories.
- On the other hand, we have ISTP_ENTJ where they only share a singular personality trait of T (Thinking).

<u>Comedy</u>:

- This is the only genre where there is a significant presence of Extroverted First Lead and Extroverted Second Lead.
- It is compatible because of the fast paced and exciting nature of comedy movies where a lot of energy is required in order to engage the audience.
- In other genres, it would be more difficult to pull off such a pair.

Action:

- Although the Second Lead has diversity, the First Lead is dominantly ISTP.
- As explained previously, ISTP characters tend to be very innovative and spontaneous, which are traits that are expected from a genre such as Action.
- There is also a strong presence of Thinking in the personality types, which adheres to the expectations of the genre.

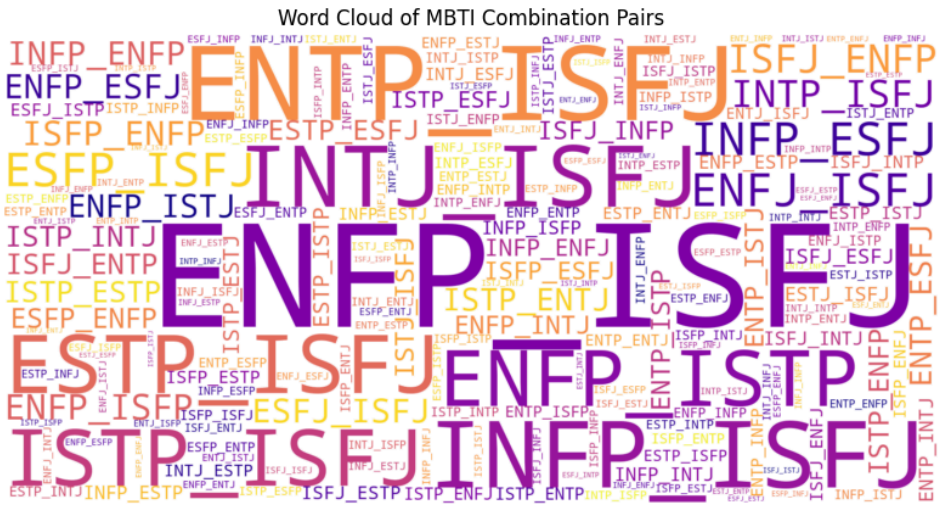
Crime:

- The prominent pairs are either very similar to each other (ISTP_ESTP) or very opposite to each other (ISFP_ENTJ and ISTP_ENTJ).
- This is interesting because it shows the different approaches screenwriters can take when building characters in a crime movie—either making them parallels or opposites of each other.

Romance:

- For all three frequent pairs, the First Lead and Second Lead have only a single personality trait that differs from each other.
- This ties in with the concept of Soulmates that exists in the Romance genre, where two people are said to be made for each other, often called “two halves of a whole”.
- It indicates that the most popular Romance movies feature two characters having such a rich and familiar chemistry between them, which the audience enjoys.

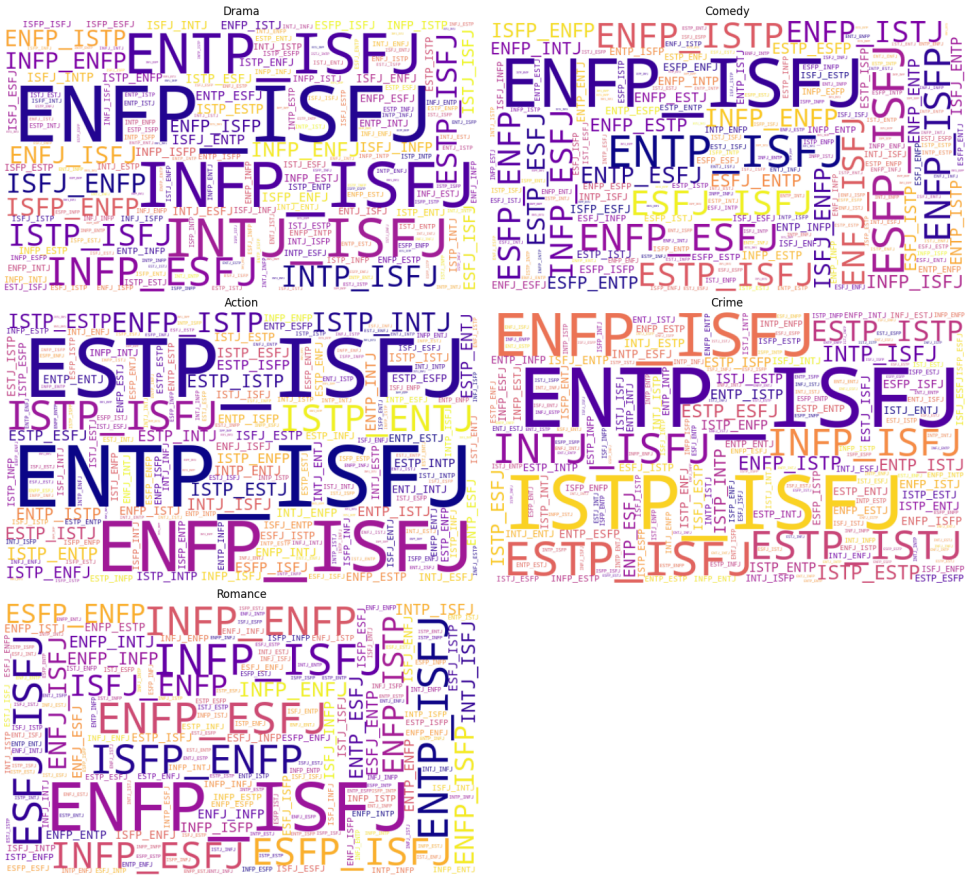
We then performed the same analysis on the MBTI_all_4200 dataset to see how the results compare.



MBTI_all_4200: Word Cloud

The first thing to note is that there is a presence of both Extroverted-Introverted pairs, and Introverted-Introverted pairs, but no Introverted-Extroverted pairs. This is an interesting result since it is starkly different from the top 1000 movies. Additionally, there is a significant presence of ISFJ as the Second Lead regardless of the pair.

This can be understood when broken down based on the genres.



MBTI_all_4200: Word Cloud distribution in Top 5 Genres

<u>Drama</u>:

- It is the exact opposite of the top 1000, where the First Lead is Extroverted and the Second Lead is Introverted.
- There is also a significant presence of ISFJ in the Second Lead.

<u>Comedy</u>:

- There is a lot more randomness in the pairs, with only a single predominant personality type unlike the other genres.
- This indicates that there is a lot of diversity in the dynamics between the First Lead and Second Lead characters in this genre.

<u>Action</u>:

- There is a presence of Extroverted First Lead and Introverted Second Lead in this genre.

- Considering that the relationship is mostly Enemy or Rival in this genre, it would explain the opposite personality types in most of the traits.
- Additionally, there is a significant percentage of the Second Lead that is ISFJ, which makes the character cold, practical, and in this context, also ruthless.

Crime:

- This is similar to the top 1000 movies where the First Lead and Second Lead are either very similar or very opposite to each other.

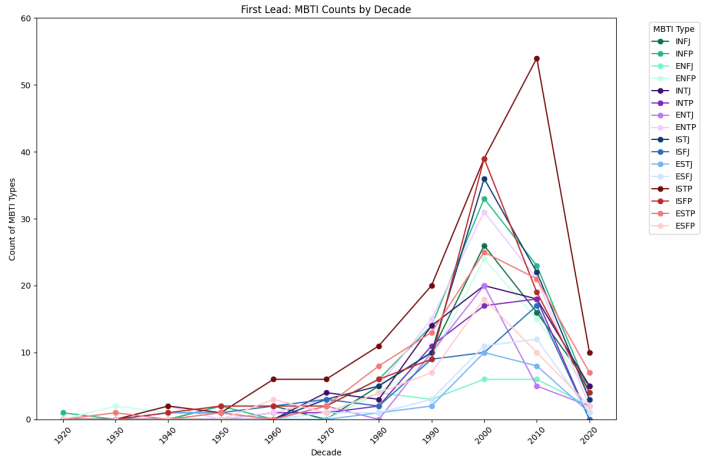
Romance:

- Interestingly enough, there are a lot of different pairs with all sorts of combinations, which indicates that the dynamic between characters has a lot of variety too.
- This implies that Romance as a genre has been thoroughly explored in terms of personality types of the characters.

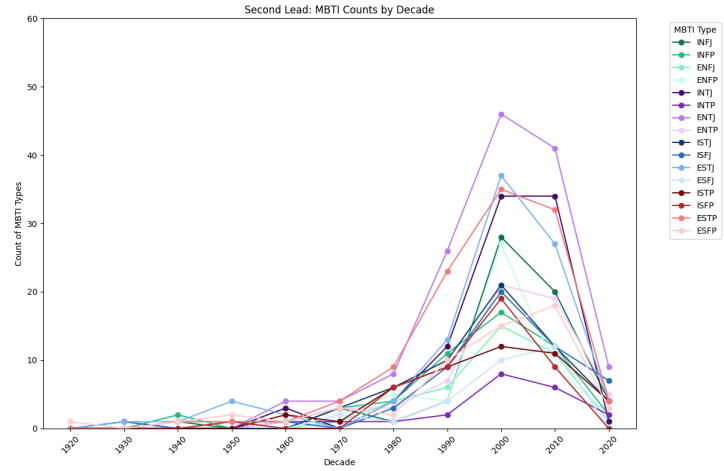
In conclusion, the analysis of character pairs across different film genres reveals how personality dynamics shape the storytelling experience. The relationships between First and Second Leads—whether as rivals, friends, or romantic interests—play a crucial role in establishing chemistry and engaging the audience. Understanding the character dynamics, especially in the context of the genres in which they're written into, can help us explore the rich storytelling possibilities in movies.

Time Series Analysis:

Analyzing the trend of popular MBTI of the movie characters across the decades for pattern recognition.

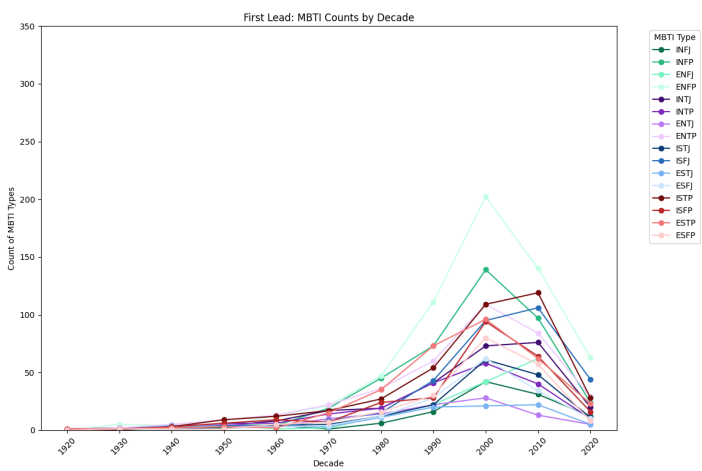


MBTI_top_1000: First Lead Time Series

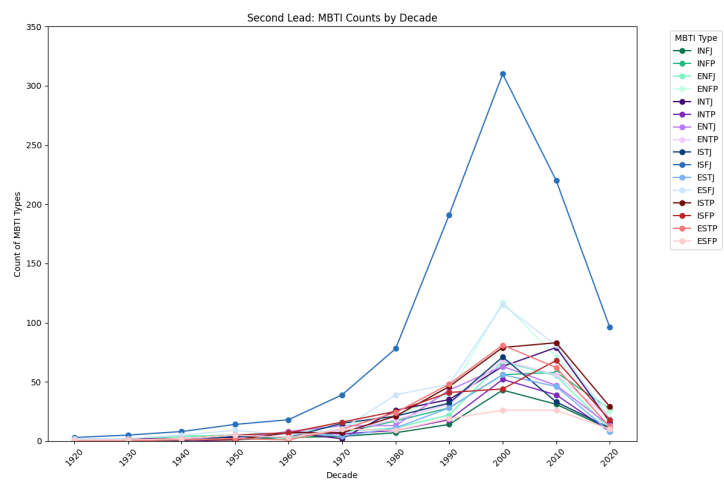


MBTI_top_1000: Second Lead Time Series

The trend in film production began to rise sharply from the 1990s, as more films were made due to technological developments, as can be seen from all the graphs. ISTP is the most prevalent personality type in the graph that displays the MBTI types of the main characters from a sample of the top 1,000 films. The second graph from the left, which shows the top 1,000 films, confirms that ENTJ is the most popular MBTI type.



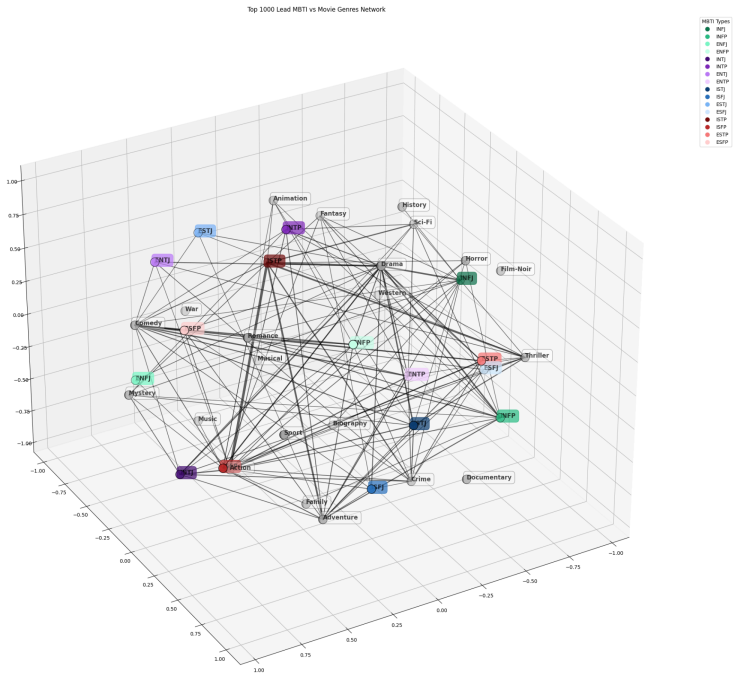
MBTI_all_4200: Second Lead Time Series



MBTI_all_4200: Second Lead Time Series

The 4,200-film dataset indicates that the ENFP personality type is more commonly linked to the first lead character, whereas the ISFJ personality type is more commonly linked to the second lead character. This could be as a result of the dataset's inclusion of both well-known and obscure movies, which is altering earlier patterns.

Network Analysis



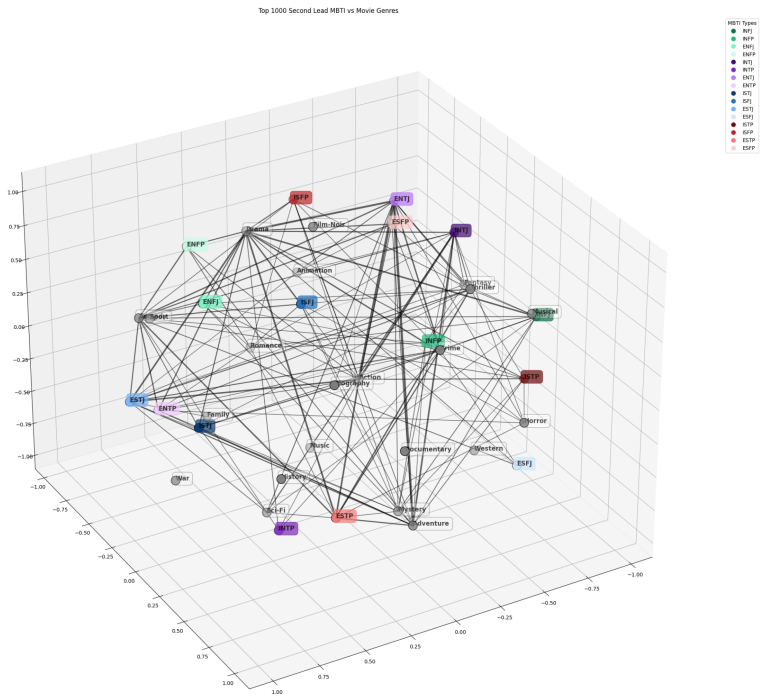
MBTI_top_1000: Lead MBTI vs. Movie Genres

Using network analysis, the most heavy edges depict the strongest connections. This is based on the mean of the MBTIs - providing a more clear network - and how frequently they appear in conjunction with the genres.

For the first lead, the strongest node from this visual is ENFP, which is located near the middle, as it is connected to several genres. This suggests that it is the most frequent across movie genres. Following ENFP are ISTJ and ESFP, again, suggesting that these types are more frequent compared to other types.

The heaviest genres are Action, Drama, Sci-Fi, and Comedy, meaning that they have the most variety of MBTI types.

In terms of edge thickness, ISTP and Action, and ENFP and Romance have strong links.



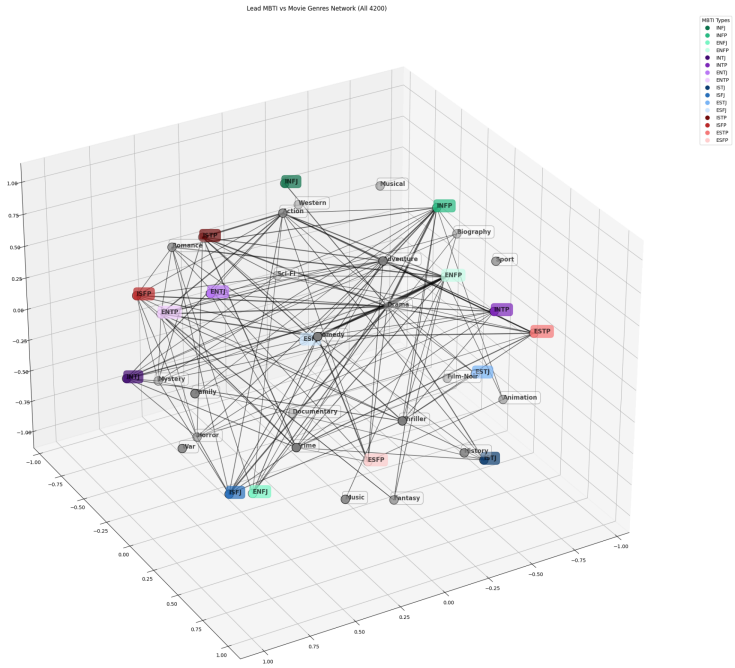
MBTI_top_1000: Second Lead MBTI vs. Movie Genres

For the second lead, the strongest nodes are ESFP, ENTJ, and ISTJ as they have the most connections to other genres. This suggests that these MBTI types frequently appear across many types of movies.

The thickest edges consist of Romance, Drama, and Crime. The MBTI types they are linked most to are ISFP and ESFP. This suggests that these genres have a majority of ISFP and ESFP as their second leads.

Finally, similar to the First Lead, Drama, Action, Romance, and Crime have the most connections overall with a variety of personality types, suggesting that the personality types for the second lead are more broad.

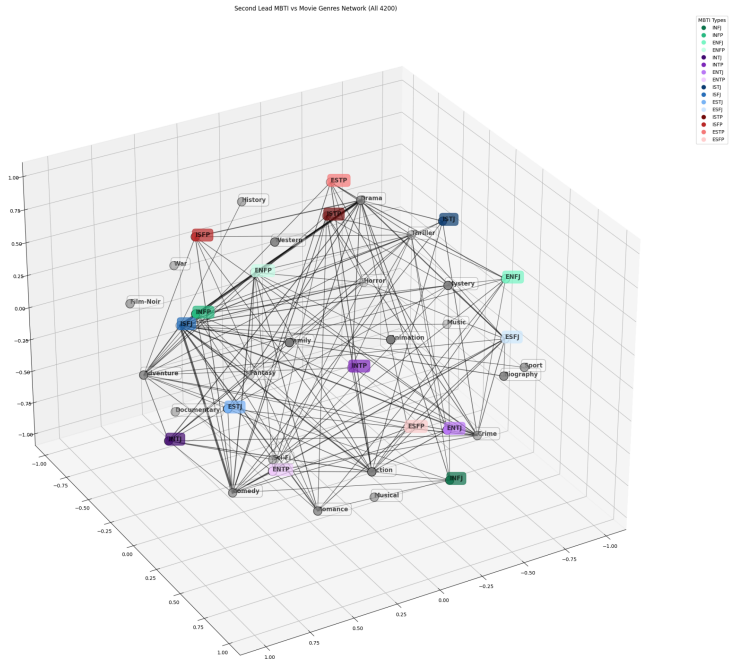
Compared to the First Lead, the network nodes are more connected with less outliers, which indicates that the second leads have a higher range of personality types across genres.



MBTI_all_4200: First Lead MBTI vs Movie Genres

For all 4200 movies, ENFP, ESTP, INFJ, INTP, are the most influential nodes, having a wide range of connections. This suggests that these personality types are widely represented across the network of genres.

Drama, Crime, and Action contain a variety of personality types, which is why they have thicker edges.



MBTI_all_4200: Second Lead MBTI vs Movie Genres

ESTP is again a frequent MBTI, connected to many different kinds of movies. ENFP is also well-connected, showing strong links to various movie genres.

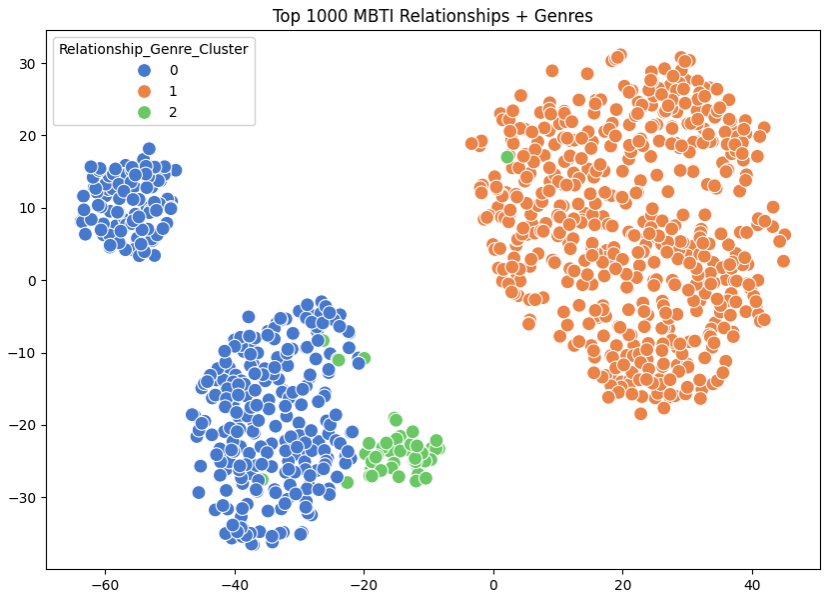
The most substantial edges are Drama, connected with ESTP and ENFP. ESFP and Romance also have a strong connection.

Drama, Thriller, and Mystery are at the center, showcasing that they have the strongest connections to a variety of MBTI types. Again, this reveals that these genres have the most variety of types for second lead characters.

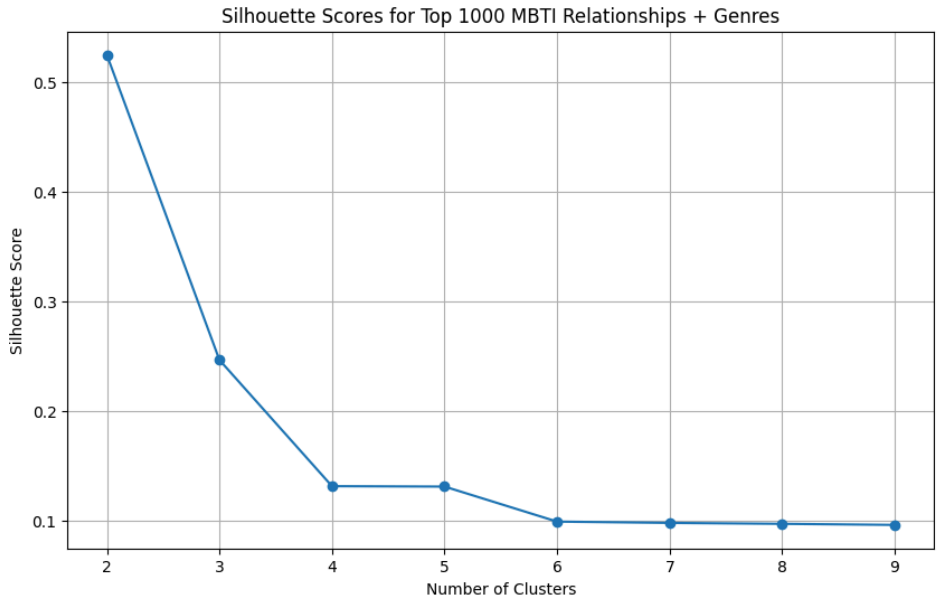
Clustering and PCA Analysis

Clustering:

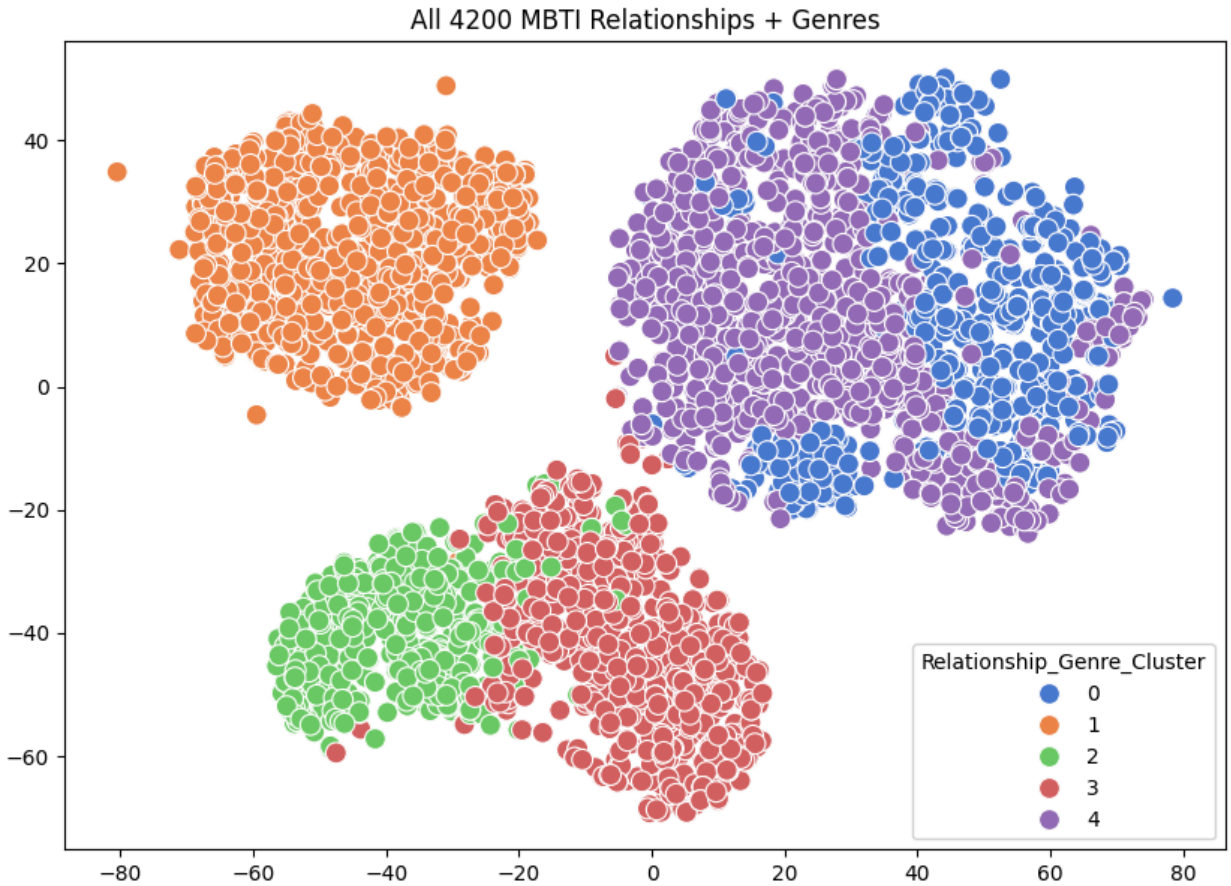
Before any clustering, one-hot encoding was done for both the first and the second lead personality types. This converted the categorical data into binary. The same was applied to genres. K-means clustering was then applied to define a total of three clusters for the relationship between MBTI and genre.



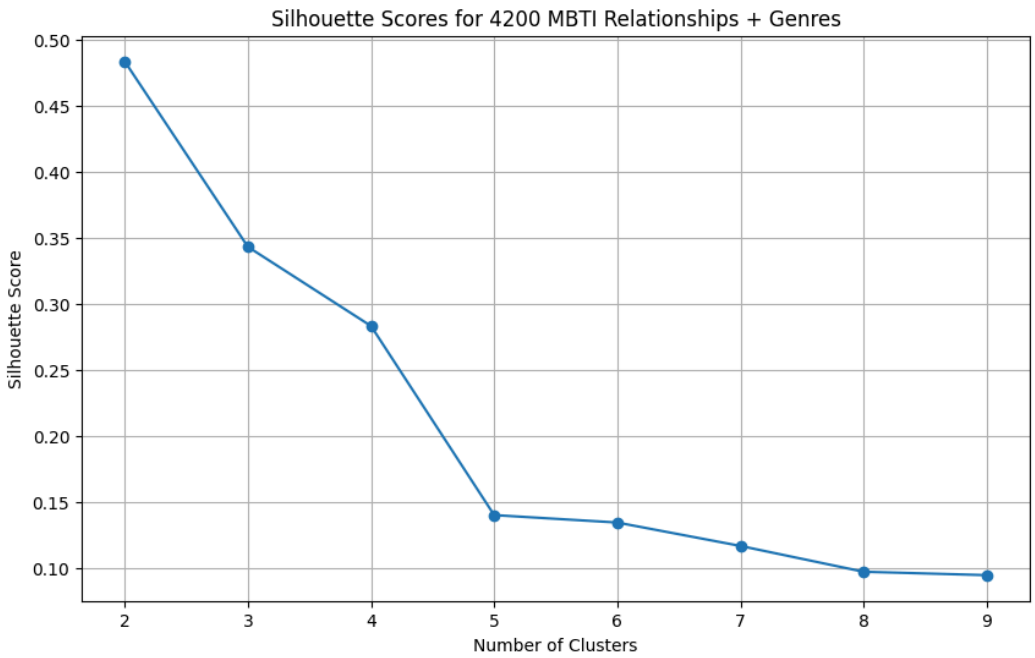
From this, t-SNE was applied to reduce the data to two dimensions, especially since the data is not linear, therefore this application is optimal. Because t-SNE was used, three distinct clusters were clearly separated, which implies that certain MBTIs are grouped into specific genres. Blue and green have separate clusters, yet are neighbors, with slight overlapping. This suggests that while the MBTIs are grouped by genre, these clusters have similarities between each other, such as having similar single letter combinations out of the four letter personality type. The most significant finding is that the orange grouping is the largest, and easily separated against the blue and green clusters. This reveals that a significant portion of the relationships between MBTIs and genres share similarities, meaning there is a strong association in this group. It could represent a large number of MBTIs that gravitate toward certain genres.



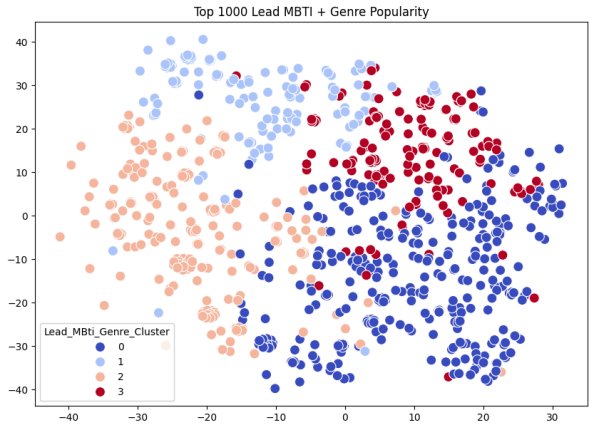
Upon completing the initial clustering, the silhouette score was then found. The y-axis displays the level of clustering, while the x-axis shows the number of clusters overall. The elbow point is the most optimal, which is 4, as shown in the above clustering.

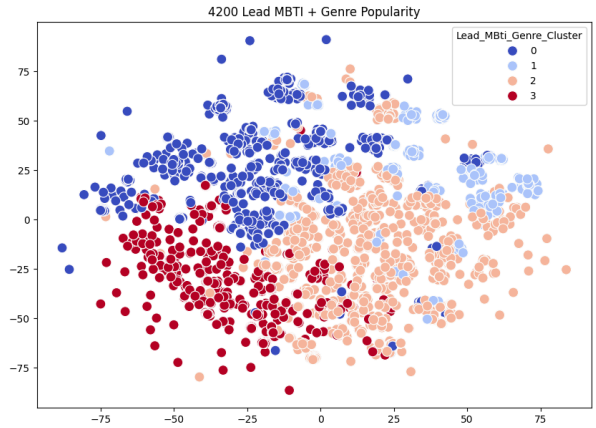


In the 4200 movies dataset, the clusters are once again distinct, reinforcing the separability of clusters in the top 1000 movies clustering. There is less overlapping in this instance. Because the clusters are large, more distinct, and tightly packed, this may suggest recurring trends between MBTI and genre relationships.



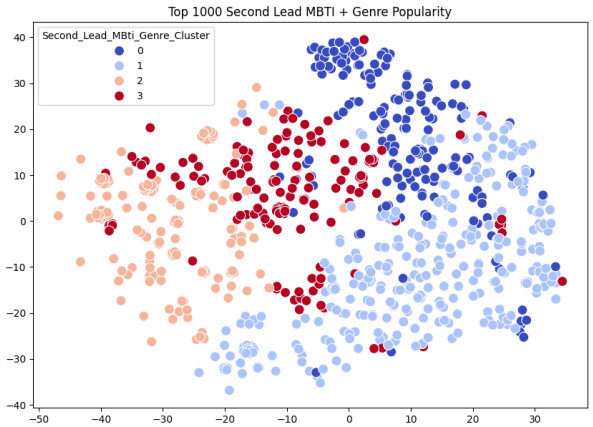
For 4200 movies, five were assigned with k-means, as suggested by the silhouette score, showing more distinct groupings compared to the initial analysis done.

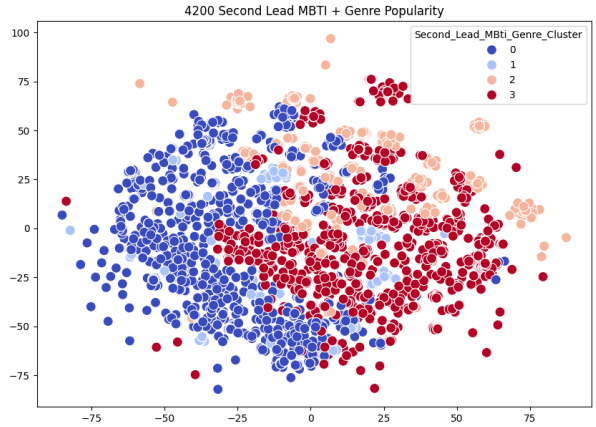




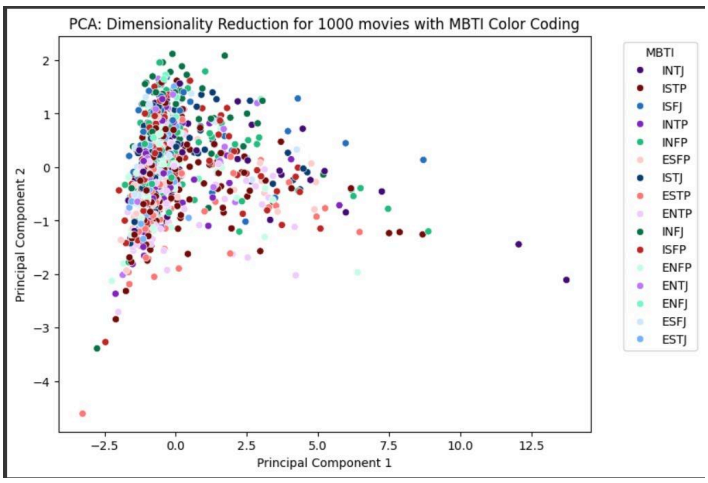
The clusters that relate Lead MBTI with Genre popularity are separated into four, with some degree of overlap. For the top 1000, the sizes of all the clusters are by and large similar to each

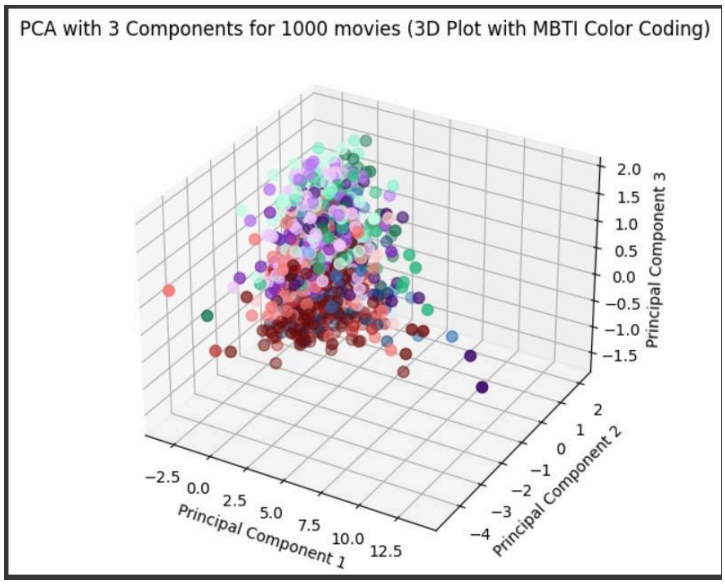
other, which illustrates that there are groups in which the Lead MBTI is very synergistic with particular genres. This finding is reinforced in our visualizations where certain MBTI pairs for Lead Character dominate in certain genres. For the 4200 movies dataset, three clusters are evenly sized with the light blue one being relatively smaller than the rest. The scatter points are dense, and cover up the space between each other, which indicates that for more data points, the clusters are well defined. For this dataset as well, the Lead Character's MBTIs are concentrated more towards the center with the genres which means that there are a lot of Lead MBTIs that go well with some genres.

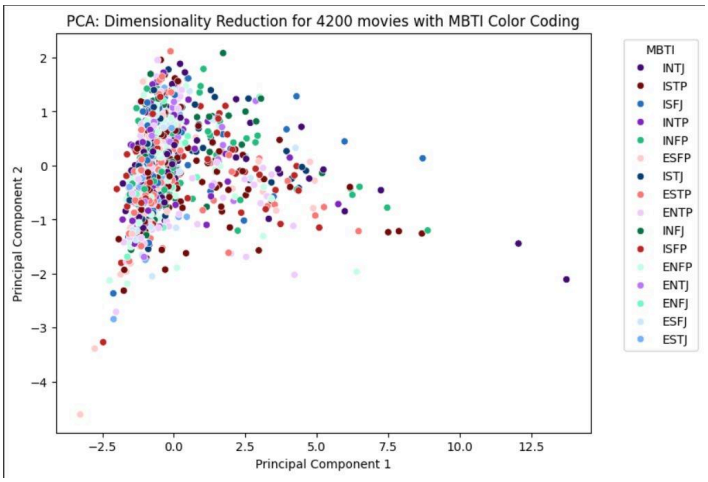


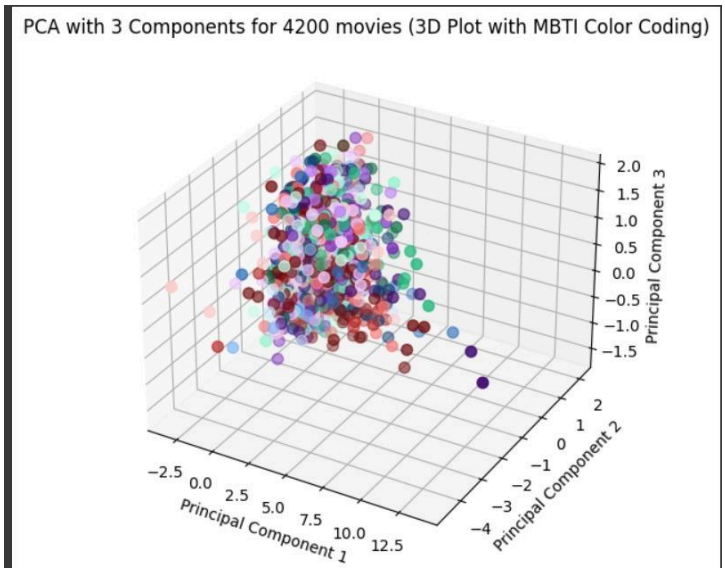


The Second Lead MBTI / Genres clusters also show a similar trend as the Lead MBTIs with the 1000 movies dataset. However, an interesting thing to witness in the 4200 dataset is two major clusters (red and deep blue) dominating the rest. This shows that for the Second Lead MBTI, a lot of personality types are paired well with some genres in two regions, and the combinations are not that frequent for the remaining two. Another inference that we can draw from this is that the Lead Character's MBTI goes well with a diverse range of genres as compared to the Second Lead Character's MBTI personality type.


PCA: Dimensionality Reduction for 1000 movies with MBTI Color Coding
Principal Component 1
Principal Component 2
MBTI
INTJ
ISTP
ISFJ
INTP
INFP
ESFP
ISTJ
ESTP
ENTP
INFJ
ISFP
ENFP
ENTJ
ENFJ
ESFJ
ESTJ


PCA with 3 Components for 1000 movies (3D Plot with MBTI Color Coding)
Principal Component 1
Principal Component 2
Principal Component 3


PCA: Dimensionality Reduction for 4200 movies with MBTI Color Coding
Principal Component 1
Principal Component 2
MBTI
INTJ
ISTP
ISFJ
INTP
INFP
ESFP
ISTJ
ESTP
ENTP
INFJ
ISFP
ENFP
ENTJ
ENFJ
ESFJ
ESTJ


PCA with 3 Components for 4200 movies (3D Plot with MBTI Color Coding)
Principal Component 1
Principal Component 2
Principal Component 3

A lot of films, regardless of genre, may have comparable personality representations since PC2 probably captures qualities and characteristics that are common to films with specific MBTI types or genres. For example, films might cluster along PC2 if several of them had similar character characteristics (for example, a high prevalence of introverted characters like lead_I).

The concentration in PC2's upper left corner raises the possibility that the sample may not be as diverse as one might anticipate in terms of personality types and specific film characteristics, such genres. This clustering suggests that there may be a bias or trend in the film industry where some MBTI traits are more prevalent or favored, which would reduce the diversity of character personality types in various films.

Therefore, the close clustering raises the possibility that future studies will examine whether specific genres are more likely to have specific MBTI personality distributions, which could expose biases in character development that affect how an audience reacts to them. This investigation demonstrates how PCA can be helpful in revealing underlying structures in intricate datasets, such as personality types and movie attributes.

Predictive Analysis

Data Analysis

In order to make our data suitable for the analysis, we encoded all the categorical variables that we had. This included Lead MBTI, Second Lead MBTI, MBTI Pairs, Genres, and Relationship.

Lead and Second Lead MBTI:

For the MBTI personality types of both the lead and the second lead, we split the personality types into 8 characters I, E, N, S, F, T, P, J to intricately analyze the impact each personality indicator has, on the movie popularity. The columns that we generated for the lead and second lead MBTI encoded personality types are the following: lead_I, lead_E, lead_N, lead_S, lead_F, lead_T, lead_P, lead_J and secondlead_I, secondlead_E, secondlead_N, secondlead_S, secondlead_F, secondlead_T, secondlead_P, secondlead_J. A .head() overview of this is illustrated below:

MBTI Pairs, Genres, and Relationship:

We also created MBTI pairs based on combining the personality types of both the lead character and the second lead character (ISTP-ENTJ, ISTJ-ENFJ etc.) We one-hot encoded these MBTI pairs in order to witness their predictive power in the analysis, because it is ultimately the combination of different MBTI types that causes the synergistic effect which translates into an interest from the audience. In addition to this, we also one-hot encoded the Genres and the Relationships (Enemy, Rival etc.) to assess their influence as well.

Lead and Second Lead Popularity:

In order to extend our analysis to include a very important factor that can influence the popularity of the movie, and which also relates to the budget as well as the marketing efforts done to promote the movie, are the popularity of the Lead and the Second Lead. We were effectively able to quantify this in our dataset, through the number of votes that the characters have on Pdb. Through this, our project included the majority of variables that influence movie popularity.

Model Selection

For model selection we considered Linear Regression, Logistic Regression, Random Forest, and XGBoost to narrow down for the best ML model suitable for the type of data we had.

Random Forest

We chose the Random Forest Regressor model because of its ability to handle complex interactions in non-linear data, which is the case with our features. In addition, the algorithm also gives us the feature importance, along with their importance scores, through which we were exactly able to figure out which features contributed the most towards movie popularity, and had the most predictive power. The importance scores rank the features in order of their contribution towards predicting the dependent variable (numvotes), which also provide us with which features relatively contribute the most. For data with very high dimensions, similar to the one we had, there is always a risk of overfitting. The algorithm reduces this by taking the average of the prediction of multiple decision trees, which also added to its alignment with our analysis. All of these reasons made RF Algorithm a perfect choice for our analysis.

Model Set Up and Evaluation:

We did an exploratory analysis on whether each of our final encoded categorical variables have any correlation with each other. Since there was a very little correlation between each of them, and any of their combinations (<0.002), it allowed us to consider them in our model without facing any multicollinearity. Moreover, since our data did not involve any time series analysis, we naturally avoided autocorrelation between independent variables. The final features that we selected to include in our analysis are the following:

- Lead MBTI, Second Lead MBTI, MBTI Pairs, Lead Popularity, Second Lead Popularity, Genre, and Relationship, as X
- Number of votes for the movie on IMDB, as Y

In order to assess the model performance we calculated the R-squared, Root Mean Squared, and Feature Importance to quantify the predictive impact that our model features had. We tried different combinations of features to assess the impact on results, and have addressed each of them in the results section.

The train_test_split for our dataset (1000 top, and 3200 randomly sampled movies) was 80% for training and 20% for testing.

Results and Interpretation

When considering all of the final features in X, and running the Random Forest Regressor Model, we got the following results:

```
Random Forest Root Mean Squared Error: 77475.24107219727
Random Forest R-squared: 0.8849764086492489

Feature Importance:
                                    Feature    Importance
16                               lead_votes  7.091581e-01
17                           secondlead_votes  2.168319e-01
295          genres_Action,Adventure,Sci-Fi  3.162203e-03
250                     mbti_pairs_ISTJ-INFP  2.569598e-03
212                     mbti_pairs_ISFJ-ENTP  1.977390e-03
596            genres_Drama,Sci-Fi,Thriller  1.510762e-03
283                 genres_Action,Adventure  1.485250e-03
288           genres_Action,Adventure,Drama  1.456615e-03
231                     mbti_pairs_ISFP-ESTJ  1.408437e-03
396         genres_Adventure,Family,Fantasy  1.383145e-03
1                                    lead_E  1.310609e-03
0                                    lead_I  1.204183e-03
279                       relationship_Other  1.203591e-03
260                     mbti_pairs_ISTP-ENTP  1.181285e-03
227                     mbti_pairs_ISFP-ENTJ  1.138890e-03
353                     genres_Action,Sci-Fi  1.130461e-03
68                      mbti_pairs_ENTP-ENTJ  1.034558e-03
275                       relationship_Enemy  1.019435e-03
200                     mbti_pairs_INTP-ESTP  1.015979e-03
106                     mbti_pairs_ESFP-INFJ  1.011374e-03
544                             genres_Drama  9.557346e-04
280                       relationship_Rival  8.895824e-04
361       genres_Adventure,Animation,Comedy  8.789993e-04
533                   genres_Crime,Thriller  8.210207e-04
601                        genres_Drama,War  7.981673e-04
```

Random Forest Algorithm (All Features x Number of Votes)

The Root Mean Squared Error is 77475.241 which is low considering the range of our num_votes variable (17k - 2.5M). The model was able to make predictions that were generally closer to the actual values. The R-squared value is 0.885 which indicates that our model features were able to explain 88.5% of the variance in movie popularity (as measured by numvotes). This is a very significant result, leading to our model being a good fit overall and being able to explain a significant variance in the data.

Interesting Findings:

- One of the major predictors of movie popularity are the popularity of Lead Character followed by Second Lead Character, measured by lead_votes and secondlead_votes respectively. These features dominate the results by having the highest importance scores. This is very true to a large extent, and is also evident as one of the major reasons why

popular actors create a significant buzz in the media which leads to a lot of viewership for the movie.

- Followed by these features, are the Genres that significantly impact movie popularity. From our findings, the most predictive ones are Action,Adventure,Sci-Fi & Drama,Sci-Fi,Thriller & Action,Adventure.

- In addition to these, we found out very interesting results for the MBTI pairs that influence film popularity. The most significant pairs that had the most predictive power are the following:

    - ISTJ-INFP

    - ISFJ-ENTP

    - ISFP-ESTJ

These results are in line with what our visualizations initially pointed towards. Looking at the results, we find a lot of I-E pairs in our findings which corroborate the fact that movies with an introverted lead and an extroverted second lead tend to perform better on the box office. For the romance genre however, similar traits are more synergistic which tend to reinforce the perfect match hypothesis for that category of movies.

- In terms of the relationships, we found out that the Enemy and Rival relationships contribute the most towards making the film popular. This finding is strengthened by the fact that it relates to the popular genres (Action) that we mentioned above, in which the Lead Character and the Second Lead are usually enemies or rivals of each other.

- When it comes to the most influential individual character traits for the lead character, E and I dominated the findings. The result relates to all our previous findings in which all personality letters do not make much difference individually, but E and I are the ones that have the most frequency.

The above mentioned findings capture all of the essential variables that have a lot of predictive power, and are very much representative of what actually matters when it comes to the movie creating a lot of waves. This result, combined with some additional variables, can act as a solid blueprint for production companies to have a strong sense of what is going to eventually be popular.

XGBoost

The second model that was in line with the types of features that we had in our data, was XGBoost.. Linear Regression was not the ideal model to use here due to the non-linear nature of our dataset. Moreover, Logistic Regression was also ruled out as it is suitable for the dependent variable that has a binary outcome. Since our dependent variable (numvotes) was not a binary variable and instead had interval data, it also did not suit our analysis. XGBoost, on the other hand, builds decision trees in a very structured way that corrects the error of the previous one as it moves forward. It is especially a very powerful model to use when it comes to dealing with both numeric and categorical data, which was the case with our dataset. It is also able to effectively handle complex relationships between different features through regularization, making it even better than Random Forest in this aspect. Finally, it offers more robust hyperparameters, which make model fine tuning even more powerful when trying to optimize model performance.

Model set up and evaluation:

The features that we evaluated XGBoost on, were the same through which we set up Random Forest Algorithm. The encoding of categorical variables was also the same as RF. In terms of the performance metrics that we calculated, we focused on finding the RMSE, R-squared, as well as features with their relative importance scores. We kept the train test split to be 80% for the training data and 20% for test data. The number of estimators were 200, maximum depth was 10, and the random state was 42.

Results and Interpretation:

When considering all of the final features in X, and running the XGB Regressor Model, we got the following results:

```
XGBoost Root Mean Squared Error: 82372.52153166883
XGBoost R-squared: 0.8699753284454346

Feature Importance:
                                    Feature  Importance
16                               lead_votes    0.066933
172                     mbti_pairs_INFP-INTJ    0.039520
17                          secondlead_votes    0.029664
231                     mbti_pairs_ISFP-ESTJ    0.027340
68                      mbti_pairs_ENTP-ENTJ    0.024944
260                     mbti_pairs_ISTP-ENTP    0.023966
227                     mbti_pairs_ISFP-ENTJ    0.019016
225                     mbti_pairs_ISFP-ENFJ    0.016581
250                     mbti_pairs_ISTJ-INFP    0.013272
283                 genres_Action,Adventure    0.013265
54                      mbti_pairs_ENTJ-ESFJ    0.012592
25                      mbti_pairs_ENFJ-ESTP    0.012506
200                     mbti_pairs_INTP-ESTP    0.012137
202                     mbti_pairs_INTP-INFP    0.011841
583                     genres_Drama,Mystery    0.011795
596            genres_Drama,Sci-Fi,Thriller    0.011478
173                     mbti_pairs_INFP-INTP    0.011084
36                      mbti_pairs_ENFP-ENTJ    0.010034
256                     mbti_pairs_ISTJ-ISTP    0.009907
75                      mbti_pairs_ENTP-INFP    0.009803
396         genres_Adventure,Family,Fantasy    0.008278
148                     mbti_pairs_INFJ-ENTJ    0.007841
295          genres_Action,Adventure,Sci-Fi    0.007487
84                      mbti_pairs_ESFJ-ENTJ    0.007155
108                     mbti_pairs_ESFP-INTJ    0.007081
62                      mbti_pairs_ENTJ-ISFJ    0.006628
```

XGB Regressor Algorithm (All Features x Number of Votes)

The Root Mean Squared Error is 82372.52 which is low considering the range of our num_votes variable (17k - 2.5M). The R-squared value is 0.870 which indicates that our model features were able to explain 87% of the variance in movie popularity (as measured by numvotes). The results are very similar to what we had for the Random Forest algorithm, which further reinstates the predictive power of all features contributing towards movie popularity.

Interesting Findings:

- As stated by our RF model, the popularity of the lead character and the second lead character were two of the major determinants when it comes to how popular the movie would eventually be.
- However, this time around, there is one MBTI pair that passed the second lead votes feature to have the second highest feature importance. This pair was INFP-INTJ.
- In addition to these, we found out very interesting results for the MBTI pairs through this ML algorithm as well. The most significant pairs that had the most predictive power in this case, are the following:

    - ISFP-ESTJ
    - ENTP-ENTJ
    - ISTP-ENTP
    - ISFP-ENTJ
    - ISFP-ENFJ
    - ISTJ-INFP

  These results are in line with what our RF algorithm pointed towards. Looking at the results, we find a lot of I-E pairs in our findings which corroborate the fact that movies with an introverted lead and an extroverted second lead tend to become more popular. Interestingly, this algorithm placed a lot of IS pairs for the Lead Character, and EN pairs for the second lead, on top. These findings are a bit more interesting because they are able to find out a lot of predictive power for two letters in MBTI as compared to one (I and E) in our previous algorithm.

- Followed by these features, are the Genres that significantly impact movie popularity. From our findings through XGB, the most predictive ones are Action,Adventure & Drama,Sci-Fi,Thriller.

Analysis Conclusion

Summarizing our findings from both models (RF, XGB), we conclude that the MBTI personality types of Lead and the Second Lead characters do have an impact on how popular the movie becomes. There are some pairs of personality types that have a lot of predictive power, such as ISFP-ESTJ, and 7 other dominant pairs. Overall, the I pair for the Lead character, and E pair for the Second Lead, influences movie popularity. In terms of pairs, IS for Lead and EN for Second Lead are the ones that have the major impact. Other than MBTIs, the popularity of the Lead and Second Lead Character also influences the popularity, which was an expected outcome. In terms of the genres and relationships, Action, Adventure, and Enemy/Rival were the most detrimental ones respectively. Overall, the study is a thorough blueprint for production companies to have a strong sense of what is going to eventually be popular.

Research Evidence Supporting Data

MBTI can influence film in numerous ways. This is because characters in film offer traits that are attractive and ideal, as well as traits that are relatable that may not be as desired. Because characters are the central point to the storylines, it can highly affect the film's success based on audience resonance (Hall, 2005).

One's personality can denote media preferences, including film. For example, extroverts require external stimulation when it comes to media. Movie genres such as action, comedy, and thrillers can offer that. Looking at our data, a majority of the leads in these films tend to be extroverted. From this, we can see that many of the leads being extroverts offer not only the attributes of being attractive to society, but for extroverts, it also offers the perspective of relatability, which could draw the viewer in even more (Debling, 2020).

For those that are introverted, yet enjoy films that extroverts naturally would, this goes back to the denotation of society valuing extraversion as culturally ideal. While having extroverted leads may not be relatable to introverts, they exhibit traits that are deemed as attractive or desirable to society, thus it could be attractive to introverts. The same can be said vice versa for extroverts that enjoy films that are more likely to appeal to introverts.

Introverts tend to require internal simulation. They seek more tranquility, and like to examine their thoughts and feelings (NeuroLaunch, n.d.), which Drama, Romance and Sci-fi films offer, as they are intellectually stimulating and feature introspective characters (BetterLYF, n.d.). Movies with introverted leads tend to have more complexity and depth, again, adding to the introspect of relatability, which can impact film success.

There also exist films that offer a variety of personality profiles. This increases the range of relatability, which can greatly impact film success. For example, the movie *The Avengers* features many superheroes of different backgrounds. There exist characters that are extroverted and introverted, with other MBTI letter traits. This allows for a wide range of audiences to recognize the feeling of familiarity, thus drawing in more viewers, and offering the possibility of maximizing success. With both extroverted and introverted representation in films, along with other defining personality traits, there is the element of inclusivity, enhancing success (Nuwer, 2013), (Murphy, 2023).

Impact of Data Analysis

Studying the relationship between MBTI personality types and its impact on film success is crucial as films are highly influential in society. By analyzing the most frequent personality types, it can help us understand what society connects to, desires, and values. Furthermore, this can assist filmmakers in creating storylines that resonate with the audience. This can lead to

more inclusivity and representation in films, drawing in a larger audience and further shaping cultural ideals.

This can also be helpful to authors and tv-show writers as well, as books and tv-shows can offer characters that are highly relatable or valued to the audience too.

Additionally, it can showcase what traits appear to be the most likable, allowing us to understand why certain politicians, public speakers, social media influencers, and other notable figures can greatly attract listeners.

Conclusion

The detailed analysis using the Random Forest model was able to prove the impact that characters and their personality types have on the popularity of the film, which addresses the primary focus of this study. Furthermore, the literature review backs up the findings to prove that the manner in which the audience connects with the character depends on aspects such as relatability and socially desirable personality traits. This connection with the character influences whether or not the audience resonates with the film, which ends up boosting the film's popularity.

Additionally, the exploratory analysis and clustering have given insights which can help verify the original hypotheses that were proposed.

Our findings confirm that different genres indeed favor specific MBTI personality types. For example, Darker, more Introverted characters dominate in genres like Drama and Action, while more Extroverted characters are prevalent in Comedy. This aligns with our hypothesis that genre influences personality type.

When it comes to changes over time, analyzing the entire dataset revealed a shift in the dominant personality types for both First Lead and Second Lead characters, especially with the introduction of various technologies. There are also dominating personality types that have been distinctly revealed over the decades, especially between the years 2000 and 2010.

The relationships between First Lead and Second Lead often display either opposing or similar traits, especially with Introversion and Extroversion. For instance, rivals in Crime films often share similar traits while Romance in the top 1000 movies features characters with minimal difference. This aligns with our hypothesis regarding character dynamics and the chemistry between the two lead characters.

Finally, when it comes to generic personality types across genres, the analysis identified certain personality types like ISFJ for Second Leads, appearing consistently across genres. This supports the hypothesis that specific traits remain prevalent regardless of genre.

In conclusion, our investigation reveals that films are powerful instruments of societal influence, shaping public opinion and fostering empathy through their protagonists and antagonists. These characters not only drive narratives but also embody the values and struggles that resonate with audiences, allowing viewers to reflect on their own beliefs and experiences. As we explore the relationship between character personality and film popularity, it becomes evident that the stories we tell are not merely for entertainment; they serve as mirrors to our society, illuminating what truly matters to us and inspiring change in our world.

Future Scope

One of things we could to further our project is we can observe the yearly or every five years, mbti frequency. This would be interesting because every year or every few years certain current events arise that we turn our attention to and it causes us to reflect on what we value as we watch world leaders and our community handle it, (i.e elections, natural disasters, covid, etc). We could observe the mbti frequency and see what values at the time did society want to hold true to and see if it reflected the major event that occurred at that time.

Other aspects of the analysis that can be explored are:

1. Analyzing the Directors' preferences in their characters' MBTI.
2. Considering International Movie Databases.
3. Extend the analysis to include TV Shows, Anime, and Web-Series.

All this analysis would make the analysis and project more interesting, and it has the possibility of providing interesting insight into the relationship between films and characters.

References

Dembling, S. (2020, April 20). *Let's talk about extroverts for a change*. Psychology Today. https://www.psychologytoday.com/us/blog/the-introverts-corner/202004/lets-talk-about-extroverts-change

Knobloch-Westerwick, S. (2008). Audience personality and the selection of media and media genres. Media Psychology, 11(3), 388-417. https://doi.org/10.1080/15213260802204396

Manolika, M. (2023). The Big Five and beyond: Which personality traits do predict movie and reading preferences? Psychology of Popular Media, 12(2), 197–206.

BetterLYF. (n.d.). Introvert meaning and its signs. BetterLYF. Retrieved October 15, 2024, from https://www.betterlyf.com/articles/mental-health/introvert-meaning-and-its-signs

Murphy, M. (2023, October 4). The craft of character creation: Building believable personalities in film Medium. https://medium.com/@markmurphydirector/the-craft-of-character-creation-building-believable-personalities-in-film-bec2edce183e

Nuwer, R. (2013, July 15). The psychology of character bonding: Why we feel a real connection to actors. Motion Picture Association. https://www.motionpictures.org/2013/07/the-psychology-of-character-bonding-why-we-feel-a-real-connection-to-actors/

SciTechDaily. (2023, October 4). When fiction feels real: Scientists discover that lonely brains can't differentiate between fictional characters and real-life friends. SciTechDaily. https://scitechdaily.com/when-fiction-feels-real-scientists-discover-that-lonely-brains-cant-differentiate-between-fictional-characters-and-real-life-friends/

van Krieken, K., Hoeken, H., & Sanders, J. (2017). Evoking and measuring identification with narrative characters: A linguistic cues framework. Frontiers in Psychology, 8, 1190. https://doi.org/10.3389/fpsyg.2017.01190

NeuroLaunch. (n.d.). *Unveiling the psychology of introverts: Quiet minds explored*. NeuroLaunch. Retrieved October 15, 2024, from https://neurolaunch.com/psychological-facts-about-introverts/

Hall, Alice. (2005). Audience Personality and the Selection of Media and Media Genres. Media Psychology - MEDIA PSYCHOL. 7. 377-398. 10.1207/S1532785XMEP0704_4.