



STAT W5243: Applied Data Science - Project 1: Korean Drama Analysis

Group 12: Mengyan Li (ml4779), Zishun Shen (zs2695), Zhisheng Yang (zy2675), Shayan Chowdhury (sc4040)

Spring 2025 - February 19, 2025

Link to Github: <https://github.com/My990813/Applied-Data-Science-Project-One>

Introduction and Data

The global popularity of Korean dramas (K-dramas) has grown significantly in recent years, presenting an opportunity to analyze the factors contributing to their success and audience reception. Our study examines a comprehensive dataset of K-dramas from 2015 to 2023, integrating information from multiple sources to understand the relationships between production characteristics, viewer ratings, and commercial success. The project began with team members proposing different datasets, with the Korean Drama dataset (proposed by Mengyan Li) ultimately being selected due to its complexity, data quality challenges, and interpretability potential. We began with the [Korean Drama from 2015-2023 with Actors & Reviews by Chanon Charuchinda - Kaggle](#) dataset, originally compiled by data expert Chanon Charuchinda through web scraping from [MyDramaList.com](#).

With that dataset as our starting point, we worked with five distinct but related datasets:

1. **Korean Dramas from 2015-2023** (1,752 records, 17 columns):
 - Source: [Chanon Charuchinda on Kaggle](#)
 - Contains comprehensive drama information including:
 - Basic metadata (ID, name, release year)
 - Production details (director, screenwriter)
 - Broadcast information (episodes, duration, air dates, network)
 - Content details (rating, synopsis)
 - Performance metrics (website ranking, popularity ranking)
 - Raw data location: [korean_drama.csv](#)

2. Top Korean Drama List (~1500)

- Source: [Top Korean Drama List \(~1500\) by Noor Rizki - Kaggle](#)
- Similar to Dataset 1 but has a larger number of dramas (1500)
- Raw data location: [kdrama_list.csv](#)

3. Top 100 Korean Dramas (2023)

- Source: [Top 100 KDrama 2023 by Gianina-Maria Petrascu - Kaggle](#)
- Similar to Dataset 1 but has more detailed information on the best dramas in 2023 only
- Raw data location: [top_100_kdrama.csv](#)

We merged datasets 1-3 to get a dataset with... [finish this]

4. Movie Industry Sales Revenue in South Korea (2014-2023)

- Source: [Movie Industry Sales Revenue in South Korea 2014-2023 by ID831717 - Statista](#)
- Contains revenue data for the Korean film industry from 2014 to 2023, which we used to
- Raw data location: [statistic_id831717_movie-industry-sales-revenue-in-south-korea-2014-2023.xlsx](#)

5. Reviews for Korean Dramas from 2015-2023 (Dataset 1)

- Source: [Korean Drama from 2015-2023 with Actors & Reviews by Chanon Charuchinda - Kaggle](#)
- Same as Dataset 1 but with review data
- Raw data location: [reviews.csv](#)

In our project, we looked at Korean Drama information data and review data for Korean Drama from 2015 to 2023. In the beginning of the project, each of us brought the data we are interested in to the meeting and voted for the best data. Mengyan Li brought this Korean Drama Data because of her interest in Korean drama. In the end of the meeting, Korean Drama Data got the highest vote because of its complexity and interpretability. The data is from Kaggle. <https://www.kaggle.com/datasets/chanoncharuchinda/korean-drama-2015-23-actor-and-reviewmydramalist?select=reviews.csv> It was made by Chanon

Charuchinda, a data expert, for educational purpose.

As explained by Chanon Charuchinda, the data was taken from https://mydramalist.com/shows/top_korean_dramas?page=1 through web scraping. Chanon Charuchinda shared four csv data in Kaggle. After discussion, we chose two csv data—korean_drama.csv which included 1752 Korean drama's information and review.csv which included 10625 reviews given to the drama from users on the website. We chose these two because the other two csv did not contain any numerical data.

There are 10 columns in review.csv including user ID, drama name, Score for Story, Score for acting, Score for music, Score for rewatch value, Overall Score, Review, Number of episode that the reviewer watched, and Number of people on the website that find this comment helpful.

Research Objectives

Our analysis focused on three key relationships:

1. the connection between popularity, ranking and viewer ratings
2. the relationship between popularity and textual review sentiment (using NLP methods)
3. the correlation between drama popularity and industry revenue

Methodology

Initial Data Collection

Because of interest in Korean drama, we searched Korean drama on Kaggle and found the files. The data was taken from https://mydramalist.com/shows/top_korean_dramas?page=1 through web scraping by Chanon Charuchinda—a data expert on Kaggle. The data is structured data with data quality problem such as date format, missing value, and outliers. The date contains mismatched date. Some are month-date-year but some are date-month-year. There are many missing values and outliers. For example, in the `pop` column in korean_drama.csv, there are many 99999 which are different from other Popularity Ranking

values and are obviously outliers. And for missing values: there are many missing values in the categorical variable for example director name. And we cannot use KNN or other imputation methods for it.

Our study focus is popularity of Korean dramas. Many variables may contribute to the popularity such as the length of drama, music, and acting. Thus we focus on what score the users give to the dramas in many different sectors—music, story, acting, etc. And we also look at the correlation between different variables such as rank and popularity. We used multiple statistical methods such as t-test, linear regression to approach this problem and found meaningful results.

Data Cleaning

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R		
1	kdrama_id	drama_name,year	director	screenwriter	country	type	tot_eps	duration	start_dt	end_dt	aired_on	org_net	content_rt	synopsis	rank	pop			
2	6614413931	Sing My Crus	2023 [So Joon Moon]		South Korea	Drama	8	1500	2-Aug-23	8-Aug-23	Wednesday	Not Yet Rate	Follow the st	1484	2238				
3	5ffccbea171	D. P. Season :	2023 [Kim Bo Tor] South Korea	Drama	6	3000	28-Jul-23	28-Jul-23	Friday		Netflix	15+ - Teens	This unfoldin	164	1084				
4	4000000000	My Love At	2023 [Han Dong H] [Song Jung H]	South Korea	Drama	8	3300	6/17/23	7/12/23	Wednesday	Disney+	Hulu	Follow the st	2443	6915				
5	#0f0ca5bfb1	To Be Tones	2023		South Korea	Drama	3	600	6/17/23	7/14/23	Friday	Not Yet Rate	Don't w	48885	5969				
6	04c1ef41948	Celebrity	2023 [Kim Chul G] [Kim Yi Your] South Korea	Drama	12	2700	30-Jun-23	30-Jun-23	Friday		Netflix	18+ - Restrict	Fame, Move	826	547				
7	2edf03f02c6	Blue Temper	2023		South Korea	Drama	4	420	6/28/23	7/19/23	Wednesday	Naver TV	Car	Not Yet Rate Set in the	47245	15405			
8	e352b17628	South Korean	2023 [Kim Chil Be] [Oh Hye Sec] South Korea	Drama	12	2600	6/23/23	7/29/23	Friday	Satur	MBC	15+ - Teens	Ho Woo is th	2686	1546				
9	4000000000	Secrets of the Restaurant	2023 [Lee Jung R] [Kim Eun He]	South Korea	Drama	12	4200	6/23/23	7/29/23	Friday	Netflix	15+ - Teens	Secrets of the	326	769				
10	4a5b176552	Adult Kids	2023		South Korea	Drama	8	600	6/19/23	6/27/23	Wednesday	Not Yet Rate	A hyper-	47813	10649				
11	565dbdb0582	Iles Hidden I	2023 [Jung Ji Hya] [Ji Ah Nee]	South Korea	Drama	8	3600	6/19/23	7/11/23	Monday	Tue ENA	Genre T	15+ - Teens	Moon Joo Ra	3381	1843			
12	24ec5e1ed6	King the Law	2023 [Im Hyun W] [Choi Rom]	South Korea	Drama	16	4800	6/17/23	7/24/23	Wednesday	Su [B]C Net	15+ - Teens	Go Go Woo	605	237				
13	4000000000	Secrets in the Land	2023 [Lee Na Jun] [Han Ah Re]	South Korea	Drama	12	4200	6/17/23	7/24/23	Wednesday	Netflix	15+ - Teens	Ban Ji Eun	461	207				
14	6dc4e88d84	Bloodhounds	2023		South Korea	Drama	8	3600	9-Jun-23	9-Jun-23	Friday	Netflix	18+ - Restrict	White, Snow	261	444			
15	9213b6aa40	Love Tractress	2023 [Yang Kyung Hee]	South Korea	Drama	8	1500	6/7/23	6/21/23	Wednesday	15+ - Teens	Seon Yul, a b	2342	1099					
16	e1e394b6d0	Romance by	2023		South Korea	Drama	10	900	6/7/23	6/28/23	Wednesday	15+ - Teens	Having never	6232	7118				
17	b5b3d87821	The Villain o	2023		South Korea	Drama	10	2700	6/9/23	6/27/23	Monday	Tue MBC	Drama	15+ - Teens	A coming-of	5035	7766		
18	4000000000	Secrets of the Restaurant	2023 [Kang Hee Ju]	South Korea	Drama	10	900	6/19/23	7/29/23	Friday	Netflix	15+ - Teens	Secrets of the	54893	8975				
19	ed6e6428d8	Bitch X Ruth	2023		South Korea	Drama	10	2100	5/31/23	6/28/23	Wednesday	TVING	15+ - Teens	Baek Je Na II	5659	1731			
20	4a4b68e6d0	Battle for Ha	2023 [Kim Yoon C] [Joo Young]	South Korea	Drama	16	4200	5/31/23	7/20/23	Wednesday	Amazon	Print	Not Yet Rate A	Suspense	2952	3767			
21	147a5a5383	King of the Land I	2023 [Lee Soo Ho] [Kim Jc]	South Korea	Drama	16	4200	5/29/23	7/18/23	Monday	Tue K	15+ - Teens	Clementon ar	3486	1302				
22	4000000000	Devil Off	2023 [Park Jin H] [Park M]	South Korea	Drama	8	1500	5/29/23	5/29/23	Monday	Netflix	15+ - Teens	Rate in the	1227	3710				
23	4796544845	Star Struck	2023 [Park Sun Ja] [Jung Hyun S]	South Korea	Drama	8	1000	5/18/23	5/26/23	Thursday	15+ - Teens	When love is	8567	1323					
24	213b7476e5	Ohi Youngsir	2023 [Kim Eun Ky] [Jeon Seon]	South Korea	Drama	10	3000	5/15/23	6/13/23	Monday	Tue ENA	Genre T	Not Yet Rate	Oh Young Sir	8020	2632			
25	e27845d516	Black Knight	2023		South Korea	Drama	6	2520	12-May-23	12-May-23	Friday	Netflix	15+ - Teens	In 2021, toxi	1422	608			
26	13189476d8	Red Rain	2023 [Lee Dong Y] [Kim Roe Ri]	South Korea	Drama	12	3480	5/29/23	6/13/23	Wednesday	TVING	15+ - Teens	Rate in the	4025	4178				
27	4000000000	Secrets of the Restaurant	2023 [Park Joon Sik]	South Korea	Drama	8	1000	5/29/23	6/13/23	Monday	Netflix	15+ - Teens	Rate in the	55545	12325				
28	2ee687a950	Wise Spring	2023 [Kang Shin I]	South Korea	Drama	8	1200	5/6/23	6/11/23	Saturday	Su K	15+ - Teens	An unexpected	101	546				
29	1470331a	All That We	2023 [Kim Jin Sun] [Kang Yoon]	South Korea	Drama	8	2100	5/5/23	5/26/23	Friday	TVING	15+ - Teens	Depicts the s	3866	2030				
30	756ab26055	Love Mate	2023 [So Joon Moon]	South Korea	Drama	8	1200	5/4/23	5/25/23	Thursday	15+ - Teens	As a team	5695	1135					
31	39740947c2	My Perfect S	2023 [Yang Soo Y] [Baek So Ye]	South Korea	Drama	18	4200	5/1/23	6/20/23	Monday	Tue KBS2	ViU	15+ - Teens	Yoon Hee	310	981			
32	67218f8634	Finland Papa	2023 [Yang Soo Y] [Baek So Ye]	South Korea	Drama	6	1800	4/29/23	5/24/23	Saturday, Sunday	Not Yet Rate	Lee Yu Ri is i	6394	5609					

	A	B	C	D	E	F	G	H	I	J	K
1	user_id	title	story_score	acting_cast	music_score	rewatch_val	overall_score	review_text	ep_watched	n_helpful	
2	c8ffdab3f2a3	Sing My Crus	9	9	10	9	9	the Best Son	8 of 8 episod	23	
3	c8ffdab3f2a3	Happy Merry	5	7	9	4	6.5	I'm Happy ar	8 of 8 episod	31	
4	c8ffdab3f2a3	Duty After Sc	4	9	3	1	4	This PART 2	4 of 4 episod	121	
5	c8ffdab3f2a3	Our Dating S	9	9.5	9	9	9	I want to pla	8 of 8 episod	79	
6	c8ffdab3f2a3	The Director	7.5	8.5	7	6	7	Half-Cooked,	10 of 10 epis	66	
7	c8ffdab3f2a3	Unlock My B	9	9.5	7	8	8.5	Satisfying se	12 of 12 epis	26	
8	c8ffdab3f2a3	Roommates	8	9	9	9	9	MATES of Po	8 of 8 episod	72	
9	c8ffdab3f2a3	The Golden S	8	9	6	8	8	RUSTED SPO	16 of 16 epis	37	
10	c8ffdab3f2a3	Big Mouth	8.5	10	8	8.5	8.5	RUSHED EN	16 of 16 epis	101	
11	c8ffdab3f2a3	Blueming	8	9.5	7	7	8.5	FULLY BLOO	11 of 11 epis	76	
12	c8ffdab3f2a3	Grid	6.5	9	5	6	6.5	AFTER EVER	10 of 10 epis	52	
13	c8ffdab3f2a3	Semantic Err	7	9.5	9	9.5	9	Illogically ad	8 of 8 episod	172	
14	c8ffdab3f2a3	Twenty-Five	9	10	9	7	9	BUT SERIOU	16 of 16 epis	212	
15	c8ffdab3f2a3	All of Us Are	6	7.5	5	5	6.5	You need to	12 of 12 epis	263	
16	c8ffdab3f2a3	Ghost Doctor	9	9.5	6	6	8.5	It started out	16 of 16 epis	50	
17	c8ffdab3f2a3	Bad and Craz	9	9	8	7	8	CRAZY GOO	12 of 12 epis	54	
18	c8ffdab3f2a3	Happiness	7.5	9	6	8.5	8	WARNING, T	12 of 12 epis	69	
19	c8ffdab3f2a3	My Sweet De	7	8.5	5.5	6	7	TOO SHORT!	8 of 8 episod	49	
20	c8ffdab3f2a3	Dali and the	8	9	7.5	7.5	8	WE NEED M	16 of 16 epis	46	
21	c8ffdab3f2a3	Lovers of the	7	7.5	8	4	6.5	Charming bu	16 of 16 epis	55	
22	c8ffdab3f2a3	Wish You: Yo	7.5	7.5	9	7	8	I WANT MO	8 of 8 episod	56	
23	c8ffdab3f2a3	The School N	9	9	6.5	8	8	NETFLIX, GIV	6 of 6 episod	47	
24	c8ffdab3f2a3	Where Your	8.5	9.5	9.5	9	9	MORE MORE	8 of 8 episod	41	
25	c8ffdab3f2a3	Itaewon Clas	7.5	9.5	6	6	8	Is SEO-JOON	16 of 16 epis	46	
26	01b015525e3	Sing My Crus	8.5	9	10	10	9	Second Winc	8 of 8 episod	37	
27	01b015525e3	Star Struck	8	8.5	7.5	6.5	7.5	More like a g	8 of 8 episod	24	
28	01b015525e3	Happy Merry	6	7	7.5	3.5	6	They got the	8 of 8 episod	25	
29	01b015525e3	The Eighth S	10	10	10	10	10	Blooming lik	10 of 10 epis	12	
30	01b015525e3	Individual Cir	9	9.5	8	8.5	8	Why did the	8 of 8 episod	4	
31	01b015525e3	The Director	6.5	7.5	7.5	5	6.5	Just . . . Bad.	10 of 10 epis	8	
32	01b015525e3	Happy Ending	8	9	9	8	8	An Empty Sp	8 of 8 episod	3	
33	01b015525e3	Summer Stri	7.5	8.5	8.5	6	7.5	Tumbling do	12 of 12 epis	17	
34	01b015525e3	Roommates	8	8.5	8	9	8.5	Throwback t	8 of 8 episod	51	

The above are the two tables we started for the data cleaning process. The following are our approaches:

In the first step, we focused on **removing outliers**, particularly in the popularity ('pop') and overall score columns, as these would serve as dependent variables in our linear regression analyses. Using boxplot visualization with interquartile range calculations, we identified and removed data points that fell outside the upper and lower limits. This process resulted in the removal of 609 outlier entries, significantly improving the data's statistical validity.

The second step **addressed inconsistencies in date formatting** across the dataset. We encountered various date formats, with some entries following a month-date-year format while others used date-month-year. Using Python's `datetime` library, we standardized all dates to a consistent year-month-day format. During this process, we made the decision to exclude entries that contained only month and year information, as these incomplete date entries could potentially skew our temporal analysis.

The third step involved **handling missing values** in the dataset. Our analysis revealed that the majority of missing data occurred in three main categories: director names, screenwriter names, and network information. While our initial plan was to complete these missing entries through web scraping from various drama databases, time constraints of our two-week project deadline made this approach impractical. We did end up writing a script for web scraping using Selenium and BeautifulSoup (can be found under [scripts/scrape_mydramalist.py](#)), but due to rate limiting, we were unable to scrape all the necessary data in time. Theoretically with more time, we could have scraped all the missing data and filled in the missing values. Instead, we opted to replace missing categorical values with "Others" as a placeholder and removed entries with missing numerical values to maintain data integrity for our quantitative analyses.

Data Merging

Merging the Korean Drama Datasets (Datasets 1-3)

To combine the three Korean drama datasets spanning 2015-2023, since we noticed that all the datasets have the drama names (with different column names), we decided to merge the datasets using that column.

First, datasets 1 (1,752 dramas) and 2 (1,647 dramas) were merged using an outer join on the title column, resulting in 2,302 unique entries. Missing values in overlapping columns were filled using corresponding values from the other dataset, with special handling for episode counts that required string parsing.

Second, we incorporated dataset 3 (100 top dramas from 2023) using another outer join on common columns plus the title field, expanding the final dataset to 2,402 entries. Duplicate entries were removed by keeping only the first occurrence when grouping by title and year. The merged dataset preserved the most complete information from each dataset while standardizing formats and handling overlapping data fields systematically.

The following is the final merged and cleaned Korean Dramas dataset, spanning 3 other datasets:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	title	year	kdrama_id	director	screenwriter	tot_eps	duration	start_dt	end_dt	aired_on	org_net	content_rt	synopsis	rank	pop	music_score	story_score	acting_cast	rewatch_vali	overall_score_n helpful	
2	The Family	2015	9170c31a02c	['Joo Dong M	['Kim Shin H	20	3600	1/3/15	3/15/15	Saturday, Su	SBS	15+	Teens : A grandma h	7995	4289	6	7.8	8.2	6	8	6
3	The Lover	2015	2dc53191bc1	['Kim Min Se	12	3000	4/2/15	6/25/15	Thursday	Mnet	18+	Restrict The series ta	5486	527	7.7	8.5	9	8	8.7	15	
4	Love of Eve	2015	832b2dcd1ac	['Lee Gye Joc	Others	120	2400	5/18/15	10/30/15	Monday, Tue	MBC	15+	Teens : Three wome	54646	12387	1	3	3	1	3	
5	Love on a Ro	2015	5eda227cf63	['Choi Ji Year	['Choi Min Ki	101	2100	4/6/15	8/21/15	Monday, Tue	KBS2	15+	Teens : A drama abe	8341	7469	5	5.4	5.1	3	5.1	
6	Kill Me, Heal	2015	e8c09f07d0	['Kim Jin Ma	['Jin Soo Wa	20	3720	1/7/15	3/12/15	Wednesday,	MBC	15+	Teens : A traumatic	232	29	8.8	8.5	9	8	8.7	
7	Delicious Lov	2015	a38a58a6c6	Others	Others	3	1800	11/5/15	11/10/15	Monday, Tue	Naver TV	Can Not Yet Rate	A teen roma	9163	4826	5.2	2.8	3.8	1	3.5	
8	Love Detecti	2015	f5f5006b7e	Others	Others	10	600	11/11/15	11/20/15	Monday, Tue	Naver TV	Can Not Yet Rate	Sherlock K Is	8964	4942	3.5	3.5	7	1	4	
9	The Time Wi	2015	12208c74d5d	['Jo Soo Won	['Jung Do Yeo	16	3840	6/27/15	8/16/15	Saturday, Su	SBS	15+	Teens : Choi Won is	5880	670	6.4	7	7.6	4.3	7.2	
10	The Dearest	2015	729e5a41c	['Choi Chang	['Seo Hyun J	116	2100	12/7/15	5/20/16	Monday, Tue	MBC	15+	Teens : An upbeat fa	53280	11354	3	4.2	5	1	3.5	
11	The Eccentric	2015	c11a722be1	['Lee Duk Gu	['Yoo Nam K	12	4500	8/17/15	9/22/15	Monday, Tue	KBS2	Not Yet Rate	Oh In Yeong	7125	2769	7.8	7.3	7.8	6.2	7.9	
12	Aw!	2015	b546cd4d98	['Kim Seok Y	['Kim Soo Jin	12	3900	10/24/15	11/29/15	Saturday, Su	JTBC	15+	Teens : Soo In has a	3410	3958	5.6	6.6	7.4	5.8	7.1	
13	Bubblegum	2015	f8ecb7e038	['Kim Byung	['Lee Mi Na'	16	3600	10/26/15	12/15/15	Monday, Tue	tvN	15+	Teens : Park Ri Hwar	7163	996	7	7.5	8.8	6.1	7.3	
14	Noble, My Lc	2015	e2328f790de	['Kim Yang H	Others	20	900	8/23/15	9/16/15	Monday, Tue	Naver TV	Can Not Yet Rate	A teen roma	5090	310	6.4	6.7	7.5	6.5	7.4	
15	My Mother Is	2015	65e265355	['Ho Heung S	['Lee Geun Y	136	2400	6/22/15	12/31/15	Monday, We	SBS	15+	Teens : Kang Hoon is	44090	14364	10	10	10	10	0	
16	Love Cells Se	2015	3be7f6025c1	Others	Others	12	600	9/14/15	10/15/15	Monday, Tue	Naver TV	Can Not Yet Rate	In this bitt	8646	3225	6	4.6	7.8	3.9	5.9	
17	Flower of the	2015	62d72045415	['Kim Min Sh	['Park Hyun J	50	3900	3/14/15	8/30/15	Saturday, Su	MBC	15+	Teens : South Korean	6954	4493	1	4	4.5	1	4	
18	Six Flying Dri	2015	61790e3827	['Shin Kyung	['Kim Young	50	3600	10/5/15	3/22/16	Monday, Tue	SBS	15+	Teens : A fictional, h	162	531	9.7	9.8	9.9	8.9	9.8	
19	Missing Noir	2015	64309e69d8	['Lee Seung J	['Lee Yoo Jin	10	4200	3/28/15	5/30/15	Saturday	OCN	15+	Teens : Gil Su Hyeon	1040	1413	9	8.3	9.2	6.9	8.7	
20	Jumping Girl	2015	5e0761a4d	Others	Others	15	600	4/27/15	4/28/15	Monday, Tue	Daum Kakao	15+	Teens : Follows the l	9145	3662	6.5	6.4	7.6	6.8	3	
21	My Beautiful	2015	e265f560507	['Kim Chul G	['Yoo Sung Y	16	3600	6/20/15	8/9/15	Saturday, Su	OCN	15+	Teens : Kim Do Hyun	2542	1538	8.9	8.8	8.9	7.1	8.7	
22	Shine or Go I	2015	45a1025119d	['Song Hyung	['Kim Sun Mi	24	3900	1/19/15	4/7/15	Monday, Tue	MBC	15+	Teens : Wang So, a	3958	1632	8	7.3	8.6	5.1	7.8	
23	Missing Kore	2015	0f989ec2490	['Min Doo Sh	Others	6	600	11/3/15	11/12/15	Tuesday, We	Naver TV	Can Not Yet Rate	Á»Missing K	8905	4337	3.9	6.4	3.4	6.4	1	
24	She Was Pre	2015	1985f185a51	['Jeong Dae	['Jo Sung He	16	3600	9/16/15	11/11/15	Wednesday, MBC	15+	Teens : As a young g	1933	32	7.9	7.6	8.6	6	7.8		
25	She Is 200 Yea	2015	f3ae6267d9	Others	Others	5	600	10/27/15	10/27/15	Others	Naver TV	Can Not Yet Rate	200-year-old	3872	2678	6	6.1	6.1	4.2	6.4	
26	Warm and C	2015	81cf1d83d7	['Park Hong J	['Hong Jung I	16	3600	5/13/15	7/2/15	Wednesday, MBC	15+	Teens : A man and a	5707	635	8.2	7.6	8.9	7.1	8.5		
27	Girl of OAM	2015	634519fd5e7	Others	Others	8	900	5/4/15	5/14/15	Others	MBC every1	Not Yet Rate	Gong Ji Dan,	7755	2845	7.4	7	8.1	5.5	7.6	
28	Girls' Love St	2015	8a57b42f2	Others	Others	50	540	6/16/15	8/15/15	Wednesday	Daum Kakao	Not Yet Rate	Four women	52369	8459	5	4.8	6.2	5.2	5.5	
29	Assembly	2015	56f7a1b9c72	['Hwang In H	['Jung Hyun f	20	3720	7/15/15	9/17/15	Wednesday,	KBS2	Not Yet Rate	Jin Sang Pil i	4927	5731	8.5	9.8	9.5	9	9.5	
30	We Broke Up	2015	dc6fe763f2	Others	['Jeon Seon	10	900	6/29/15	7/17/15	Monday, We	Naver TV	Can Not Yet Rate	Ji Won Yeon	6979	1712	8.5	7.3	8.2	7.1	8.1	
31	Splendid Poli	2015	75c1221d1d	['Kim Sang H	['Kim Yi Your	50	3600	4/13/15	9/29/15	Monday, Tue	MBC	15+	Teens : Princess Jung	4965	3091	7.1	6.9	7.4	4.7	7.2	
32	Spy	2015	14fb3a5d0c	['Park Hyun S	['Han Sang V	16	3000	1/9/15	3/6/15	Friday	KBS2	Not Yet Rate	Hye Rim is a	7174	2132	6.7	6.3	8.2	4.4	6.9	
33	To Be Contin	2015	05be094a11	Others	Others	12	900	8/18/15	9/3/15	Thursday	MBC every1	G - All Ages	This web dra	5890	911	8.8	6.6	7.6	7.2	7.6	
34	Late Night Ri	2015	1ed9e98879	['Hwang In P	Others	20	1800	7/4/15	9/5/15	Saturday	SBS	Not Yet Rate	A late night i	3327	3914	8.9	9.4	9.4	9.3	9.4	
35	Sweet, Sava	2015	53d882ba9	['Kang Dae S	['Son Geun J	16	3780	11/18/15	1/14/16	Wednesday,	MBC	15+	Teens : At home, Tai	7123	4165	5.2	6.8	7.5	4.5	6.8	

Merging the (Merged) Korean Drama Dataset with Review Data (Datasets 5)

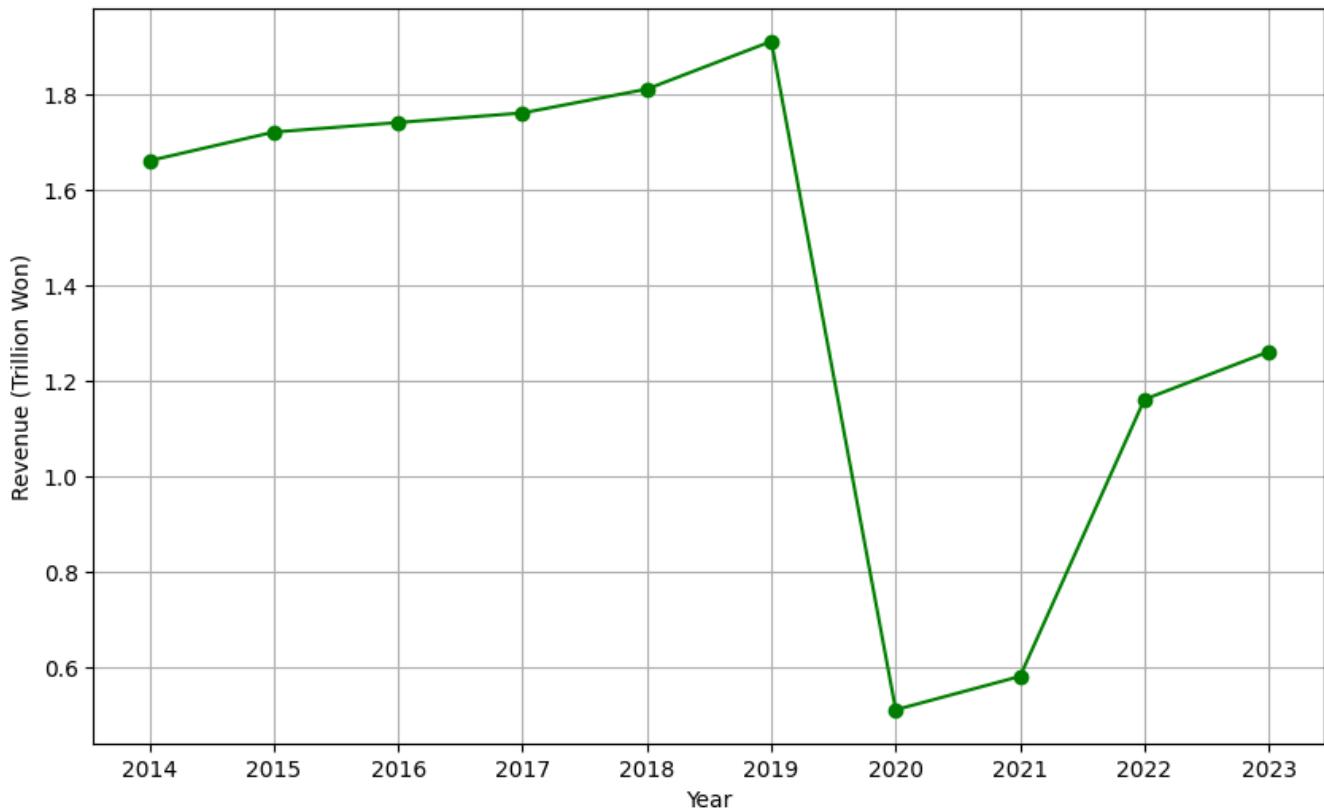
For merging the merged Korean Drama Dataset with the Review dataset (Dataset 5), we similarly merged the two tables on their respective drama title columns, and then dropped any rows where numerical values were missing. This resulted in a final dataset of 11,776 rows with 32 columns.

Merging the (Merged) Korean Drama Dataset with Yearly Revenue Data (Datasets 4)

Since this merging process is a little bit different from the previous two merges, we will explain it in detail.

Just to demonstrate, the revenue data is yearly:

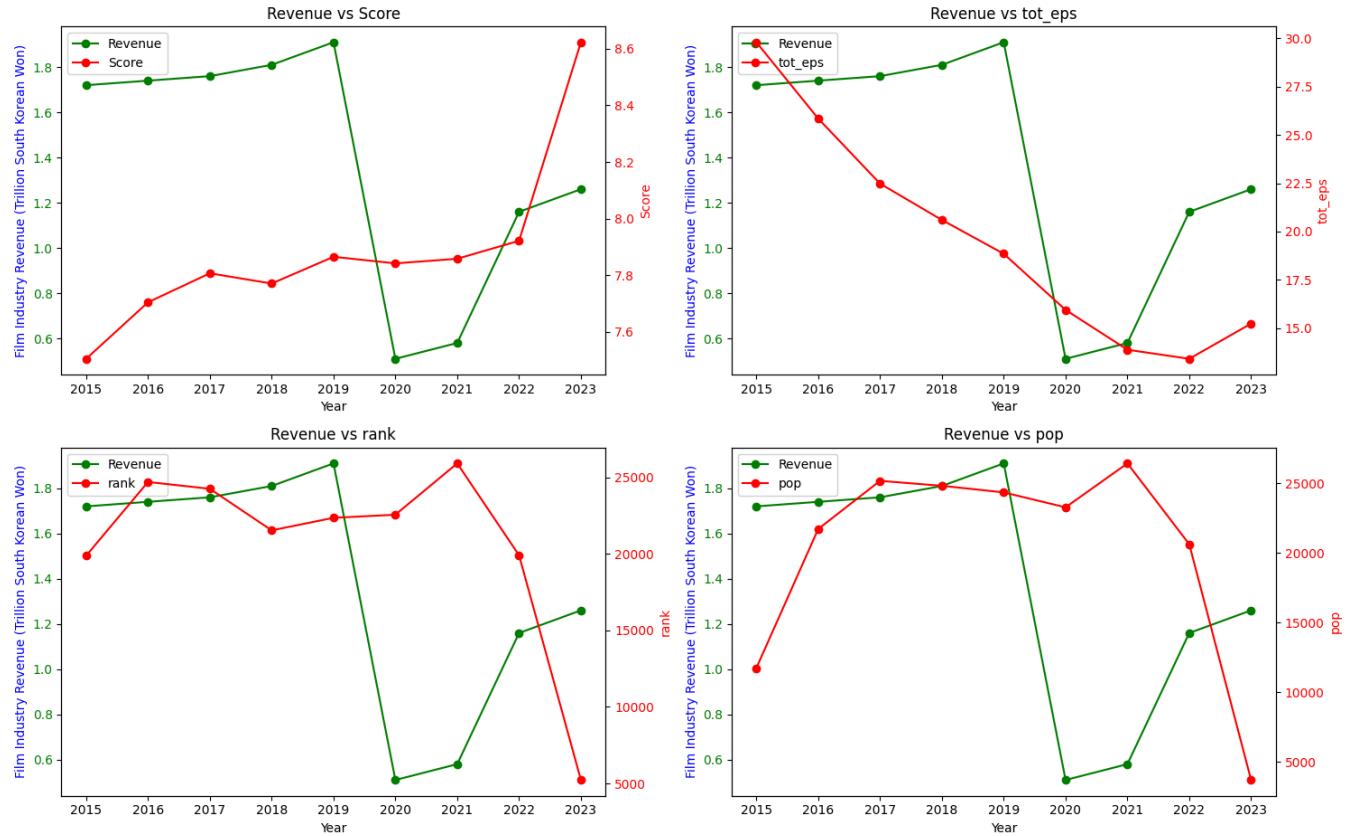
Yearly Revenue of Korean Film Industry
(in trillions of South Korean Won)



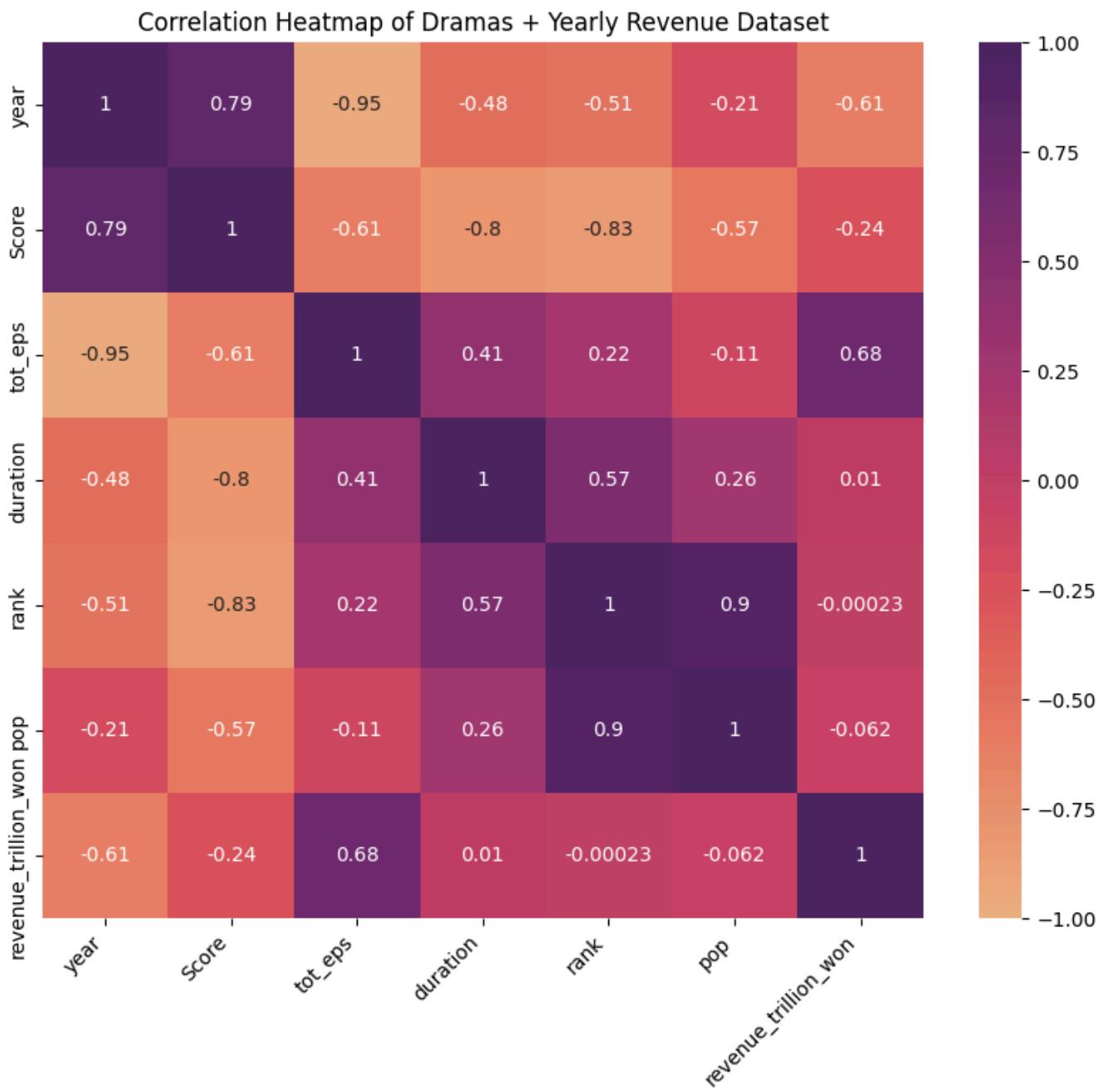
Note that this revenue variable is for the entire South Korean film industry, not just dramas. However, Korean dramas do indeed contribute a significant portion of South Korea's economic output, including revenue from exports and employment. This analysis is still useful to see how the drama industry is doing relative to the rest of the film industry.

Furthermore, we merge the revenue data with the drama data by year. To do so, we grouped the drama data by year and then merged the revenue data with the drama data by year. With our resulting dataset, we can compare the trends of revenue by year alongside the other numerical variables:

Yearly Revenue of South Korean Film Industry
vs Numerical Variables of South Korean Dramas by Year (TV Shows Only)



We also generated a correlation heatmap of the revenue data with the other numerical variables:



Here are some notable findings from the line plots and correlation matrix above:

- Pre-Pandemic Trends (2015-2019)
 - `revenue` of the South Korean Film Industry showed steady growth from 2015 to 2019, while scores gradually improved
 - `tot_eps` (total episodes) was decreasing steadily, suggesting a trend toward shorter series formats while maintaining higher revenue
 - `pop` (popularity) and `rank` maintained relatively stable patterns, indicating consistent audience engagement

- Pandemic Impact (2020)
 - `revenue` dropped sharply in 2020, falling to about 0.5 trillion won from 1.9 trillion won in 2019
 - Surprisingly, `score` remained stable during this period, suggesting that while the industry's financial performance suffered, content quality was maintained. Or that during the pandemic, shows that were filmed prior were being released. Or that the audience was more forgiving during the pandemic.
 - `tot_eps` continued its downward trend, possibly accelerated by production challenges during the pandemic
- Post-Pandemic Recovery (2021-2023)
 - `revenue` shows a gradual recovery pattern but hasn't returned to pre-pandemic levels
 - `score` reached its highest point (8.6) in 2023, showing significant improvement
 - `pop` and `rank` show volatile patterns during the recovery period

Other Notable Findings:

- The strong negative correlation between `year` and `tot_eps` (total number of episodes) (-0.95) suggests a clear industry shift toward shorter formats—whether due to the pandemic or other factors
- The positive correlation between `tot_eps` and `revenue` (0.68) implies that longer series historically generated more revenue, though this trend might be changing
- The relatively weak correlation between `revenue` and `popularity/rank` (-0.062/-0.00023) suggests that commercial success isn't necessarily tied to traditional popularity metrics

Exploratory Data Analysis

Below is the summary statistics of variables: We can see the mean for `music_score`, `story_score`, `acting_cast_score`, `rewatch_value_score`, and `overall_score` are very close to the median compared to other variables. Although they are close, the data itself is still skewed. The standard deviation of the popularity ranking is the highest, meaning there is a lot of variation in the data.

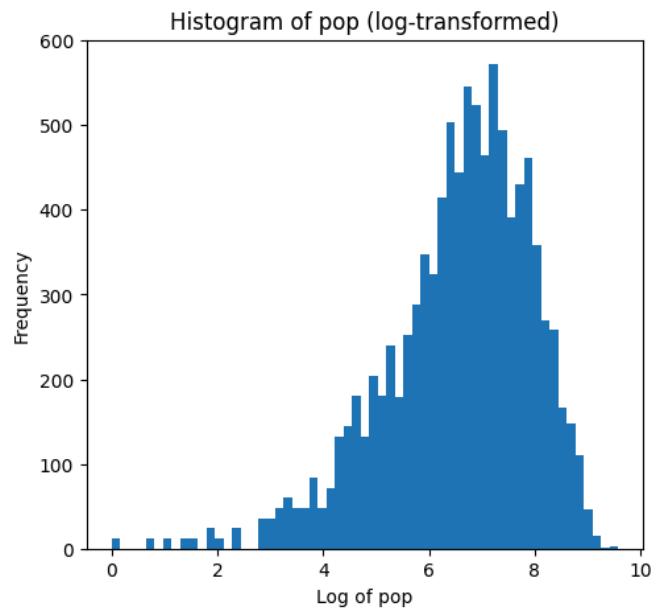
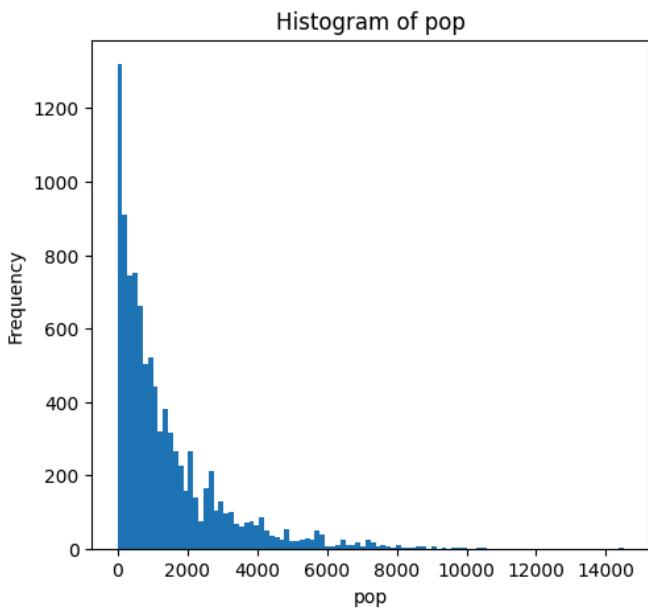
	year	tot_eps	duration	rank	pop	story_score	acting_cast_score
count	9823	9823	9823	9823	9823	9823	9823
mean	2019.11	18.7209	3020.11	3196.22	1401.66	7.56179	8.50000
std	2.29225	17.0432	1333.96	2816.44	1547.09	2.18148	1.68000
min	2015	1	52	8	1	1	
25%	2017	12	1980	985	331	6.5	
50%	2019	16	3600	2506	875	8	
75%	2021	16	4200	5090	1893.5	9	
max	2023	150	6000	56186	14507	10	

Definition of each variable (column header):

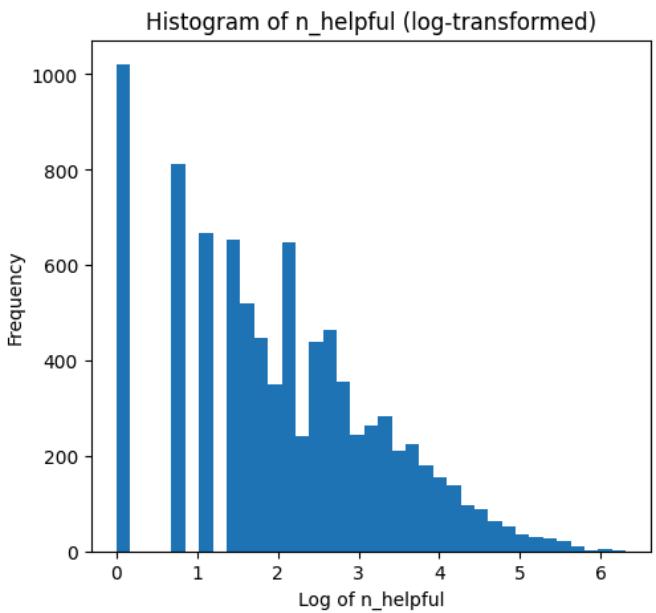
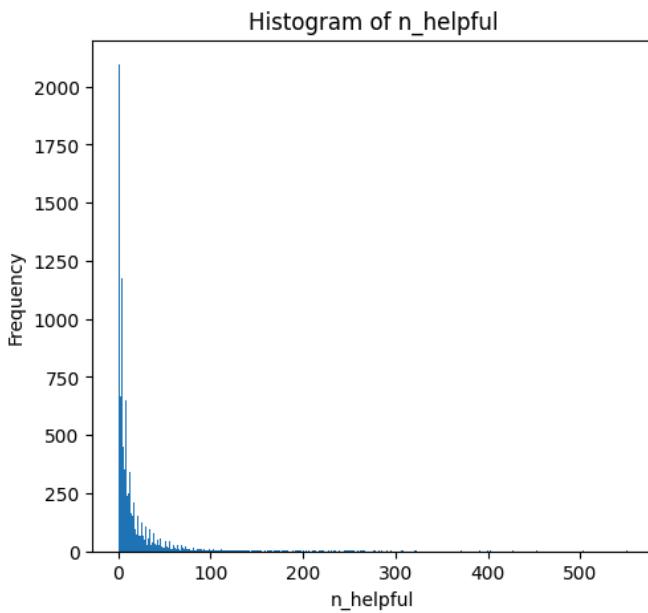
- `year` : Year of the drama (2015-2023)
- `tot_eps` : Total number of episodes
- `duration` : Duration of each episode
- `rank` : Ranking of the drama (lower is more popular)
- `pop` : Popularity ranking of the drama (lower is more popular)
- `story_score` : User-submitted score for story (from reviews)
- `acting_cast_score` : User-submitted score for acting cast (from reviews)
- `music_score` : User-submitted score for music (from reviews)
- `rewatch_value_score` : User-submitted score for rewatch value (from reviews)
- `overall_score` : User-submitted overall score (from reviews)
- `n_helpful` : Number of people that found review helpful (from reviews)

Distribution of Numerical Variables (and Log Transformed)

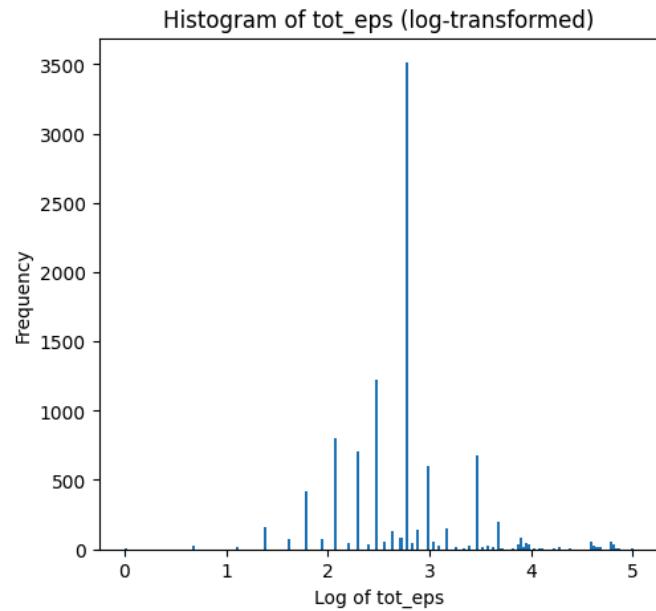
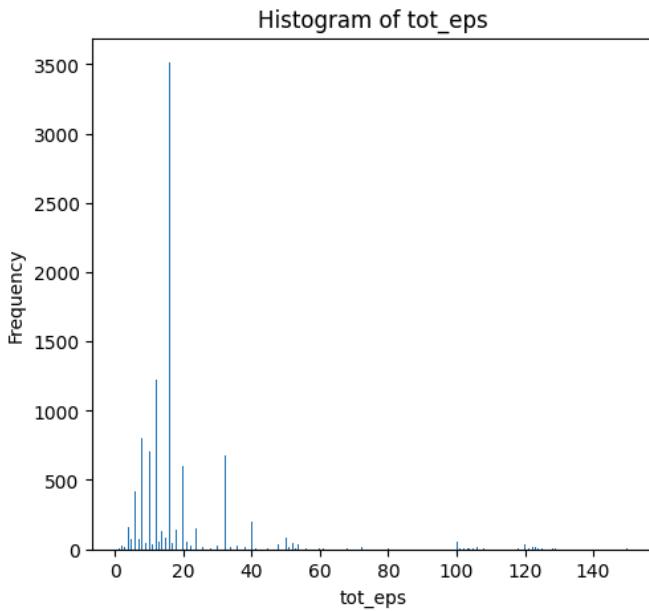
To study the distribution of numerical variables, we first plotted the histograms of all numeric variables, along with log transformations of those distributions as done in class. Below are the histograms of greatest interest, but all the distributions can be found in our [final_report.ipynb](#) Jupyter notebook.



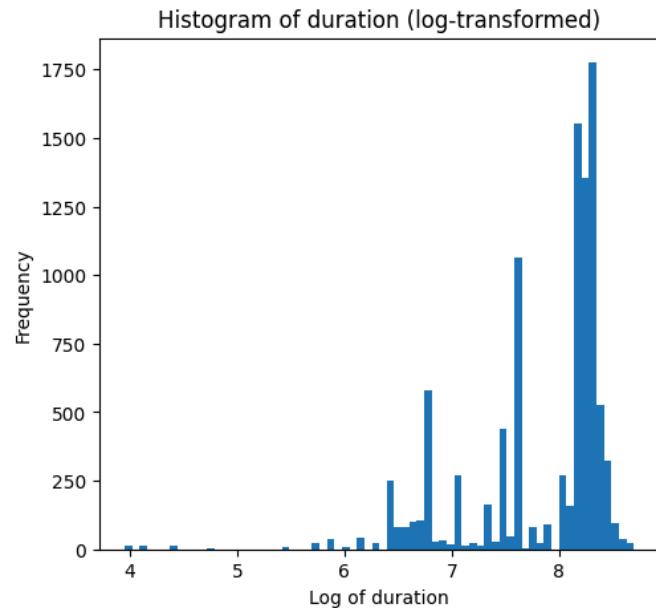
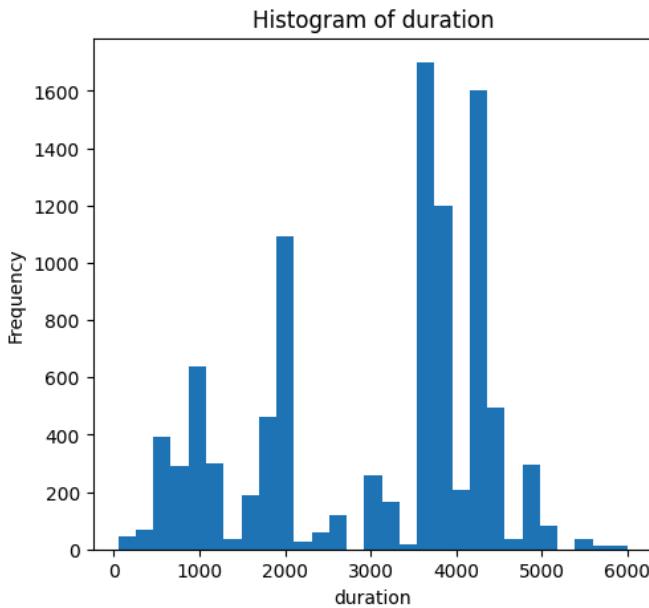
- `pop` (popularity ranking) remains heavily right-skewed, with most dramas having low popularity scores (<2000) and a few outliers reaching 12000+. As we saw in class, a log transformation makes the distribution more normal, suggesting a multiplicative effect in popularity growth



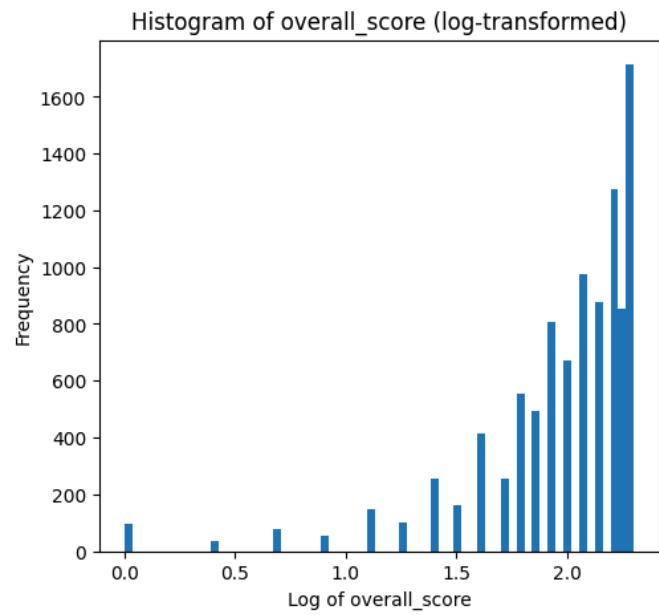
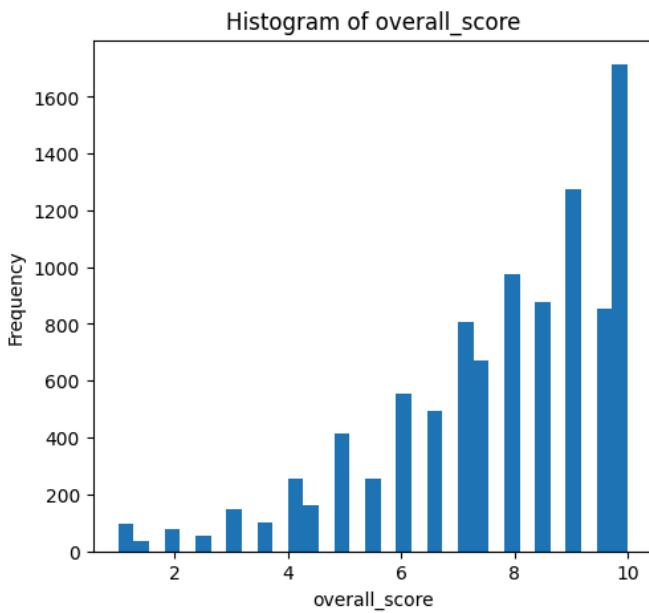
- `n_helpful` shows an extreme right skew, with most reviews getting few helpful votes and a rapid drop-off



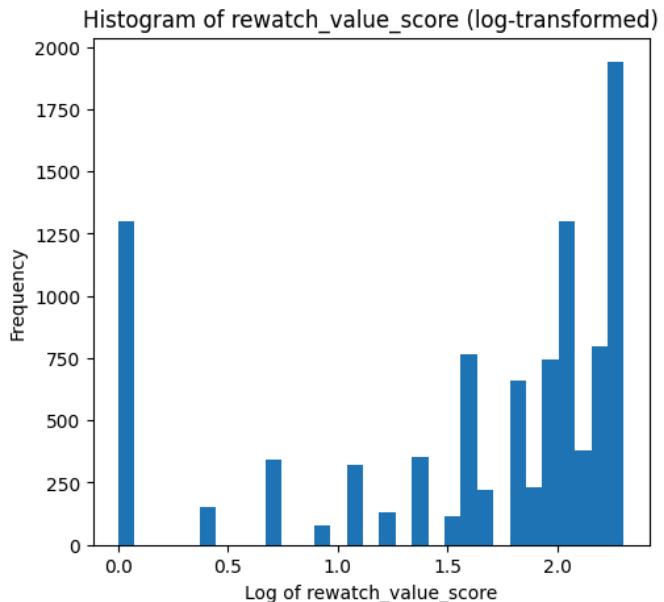
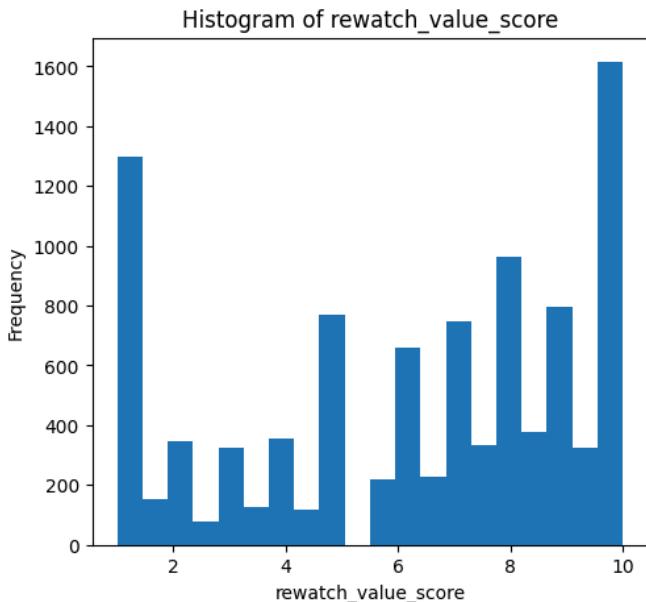
- `tot_eps` shows distinct clustering around certain episode counts (16, 32, or 50 episode dramas across multiple seasons), reflecting common show format lengths



`duration` also exhibits multiple distinct peaks, likely corresponding to standard episode lengths for different drama formats

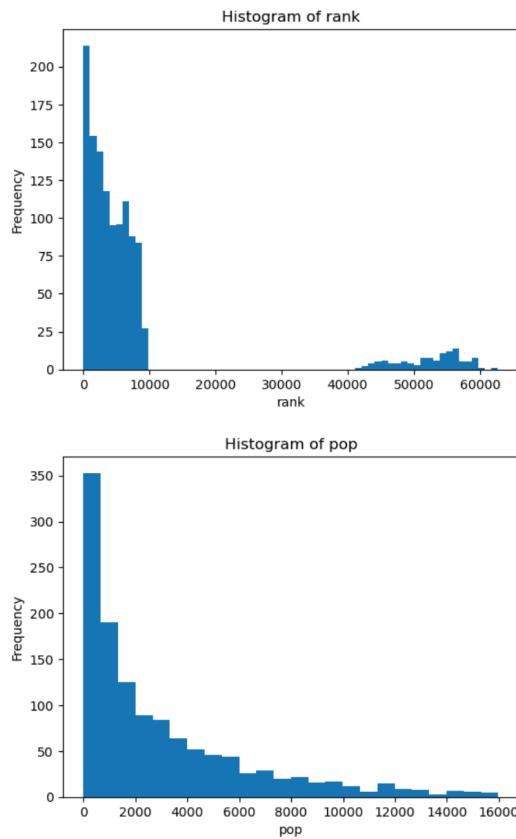


- Most score related columns: `story_score`, `acting_cast_score`, `music_score`, `overall_score` are left-skewed, meaning that most viewers gave high scores to most dramas, while a few have very low score (e.g. 1-2). As we saw in class, a log transformation of left-skewed data does not make the distribution more normal but in fact skews it more to the left.
- Only `overall_score` is shown above for brevity, but once again, all the distributions can be found in [final_report.ipynb](#)

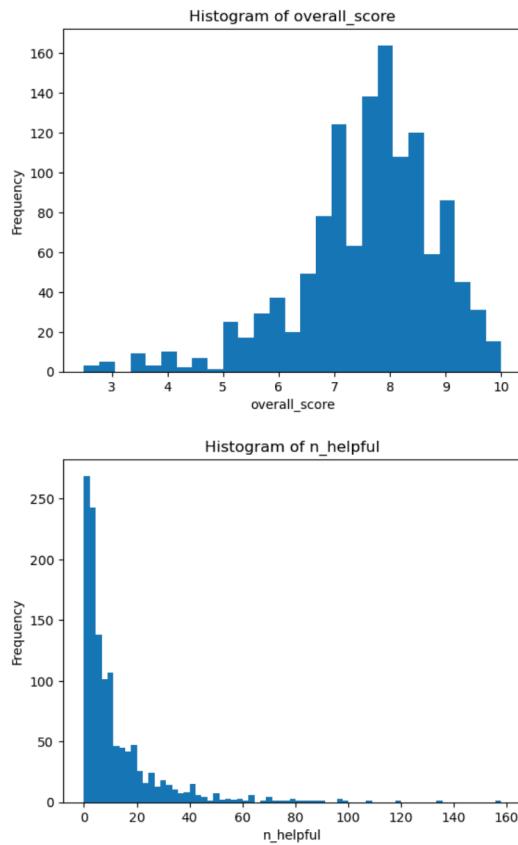


- The only potential exception is `rewatch_value_score`, which is bi-modal with a slight left skew, suggesting that most dramas fall into two categories: those that

are highly rewritable (9-10) and likely one-time watches (1-2)



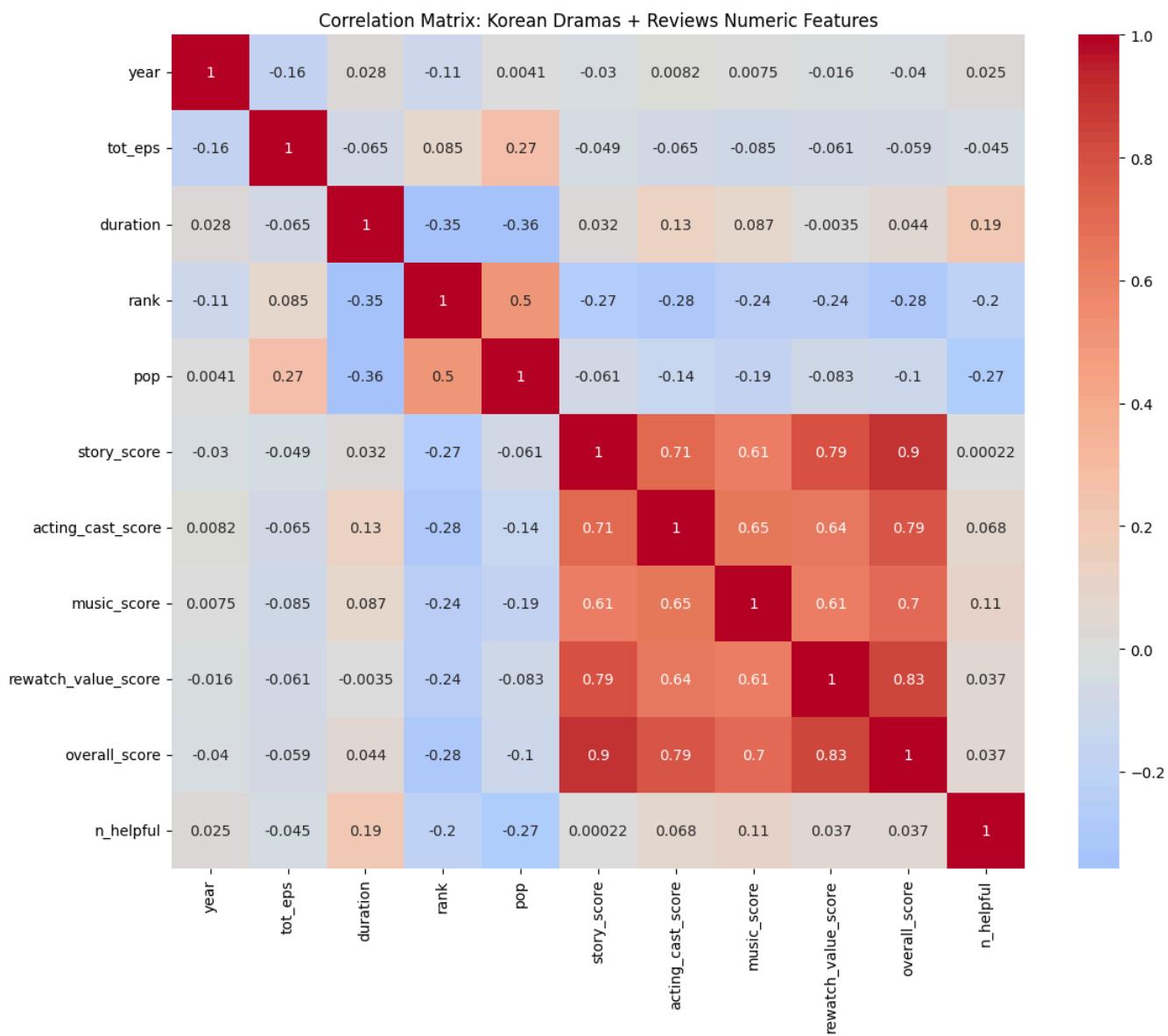
- `rank` and `pop` (popularity ranking) are also right skewed. Since higher number in ranking means lower popularity, most drama ranked high in the ranking and only a few have low popularity.



- `overall_score` is left skewed, meaning most audience gave high scores to the drama overall. The data for the `n_helpful` (number of people that found review helpful) is right skewed, meaning the review are very personalized so that not so many people hold same opinion.

Correlation Heatmap and Plots of Numerical Variables

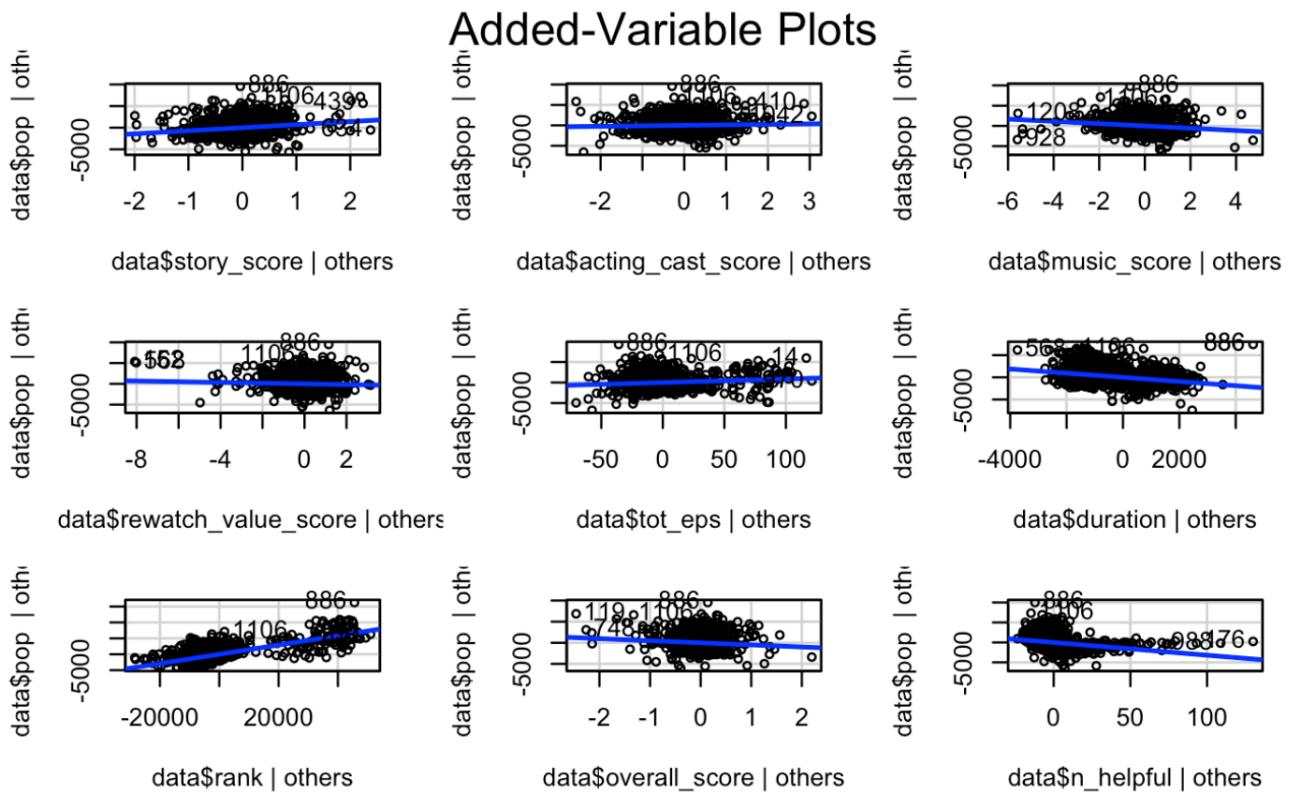
Below is the correlation heatmap of all numerical values and some notable findings:



- overall_score has very strong **positive correlations** with story_score (0.9), rewatch_value_score (0.83), acting_cast_score (0.79), and music_score (0.7)—which is to be expected, people who enjoy the story, acting, and music are more likely to enjoy the drama overall
- pop has a **moderate positive correlation** with rank (0.5), suggesting that more popular dramas might not always have the best rankings
- duration has **negative correlations** with rank (-0.40) and pop (-0.38)
- year exhibits very weak correlations with most variables, suggesting quality and sentiment metrics aren't strongly tied to when the drama was released

Digging deeper into individual variable correlations, we can see that pop (popularity) and rank are highly correlated. And music_score, story_score, acting_cast_score,

`rewatch_value_score`, and `overall_score` are highly correlated:

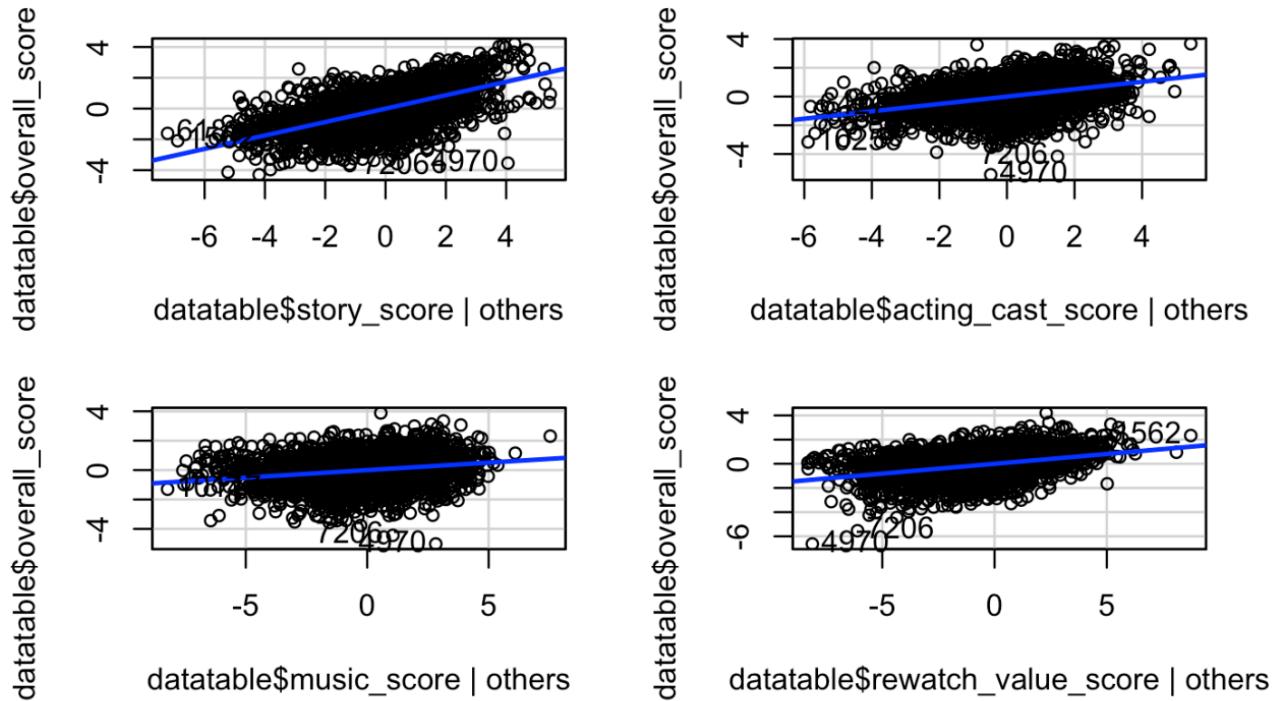


Linear Regression Analysis

Same thing happened when we look at the linear regression plot. We can see that there is a positive relationship between `pop` (popularity rank) and `rank`. We also found that there is a negative relationship between `duration` and `pop`, meaning longer drama has higher popularity. Interesting thing to see is that there is a negative relationship between `n_helpful` (number of people found the review helpful) and `pop`, meaning when there is more people sharing same opinion, the drama is more popular.

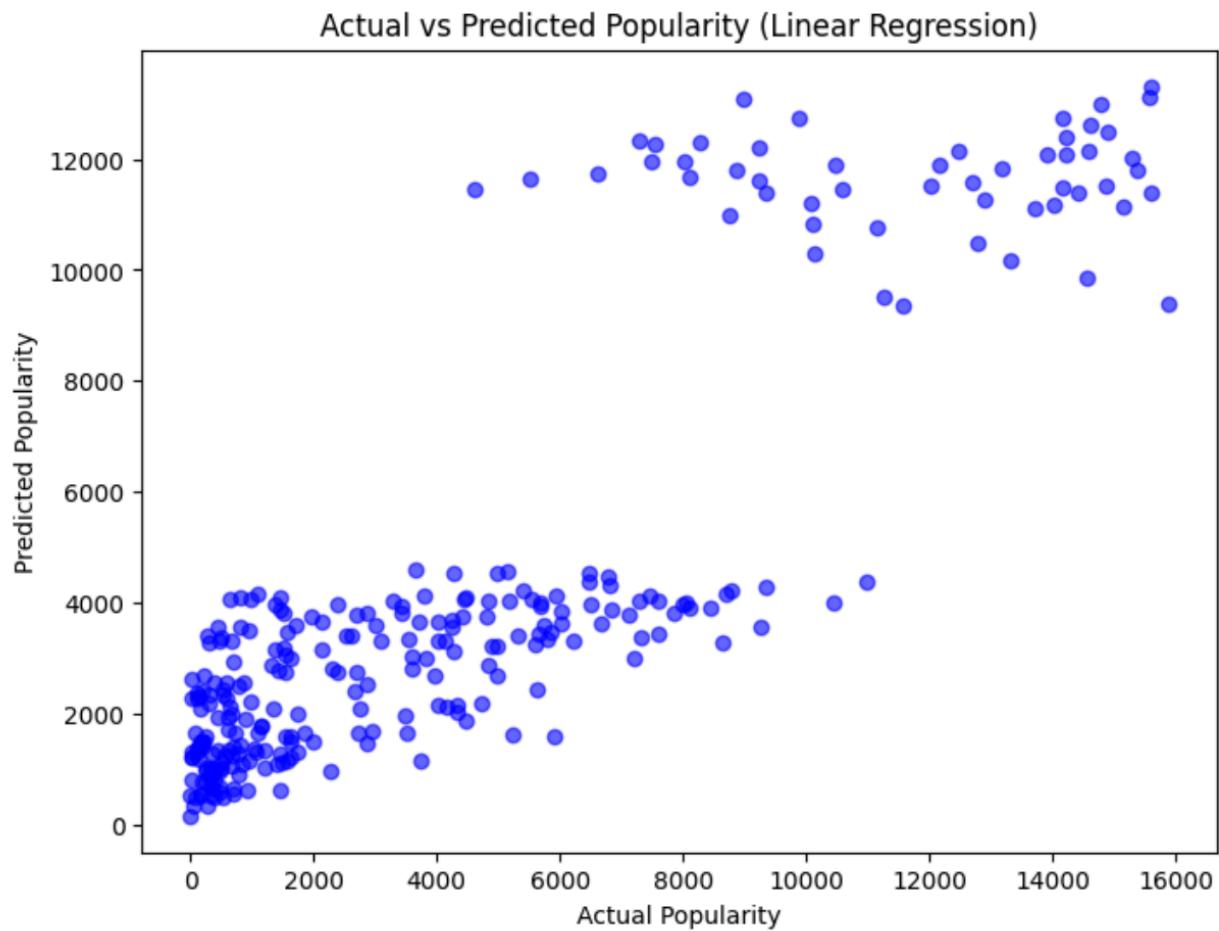
By looking at the drama information and the review separately, we did the following EDA and analysis:

Added-Variable Plots



This is the linear regression analysis between the `overall_score` and `story_score`, `acting_cast_score`, `music_score`, and `rewatch_value_score`. We can see that these are all positive relationship, meaning the `story_score`, `acting_cast_score`, `music_score`, and `rewatch_value_score` all have a positive impact on the `overall_score`.

By looking at the drama information without the impact of review, we found the following thing:



This is a simple linear regression without standardization and normalization.

Comparison: Without standardization and normalization and with standardization and normalization:

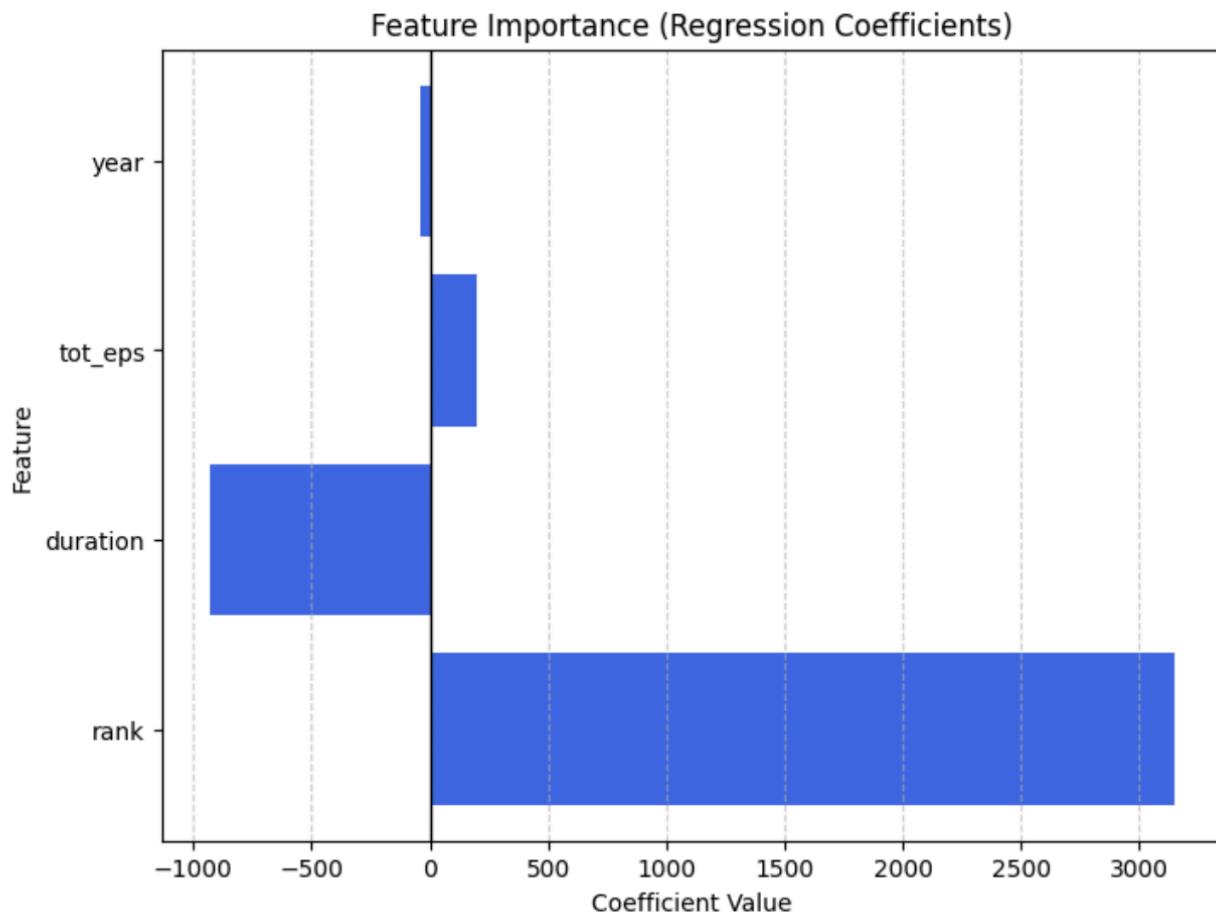
Feature	Coefficient
year	-16
tot_eps	7
duration	-0.6
rank	0.2

Feature	Coefficient
year	-38

Feature	Coefficient
tot_eps	195
duration	-929
rank	3146

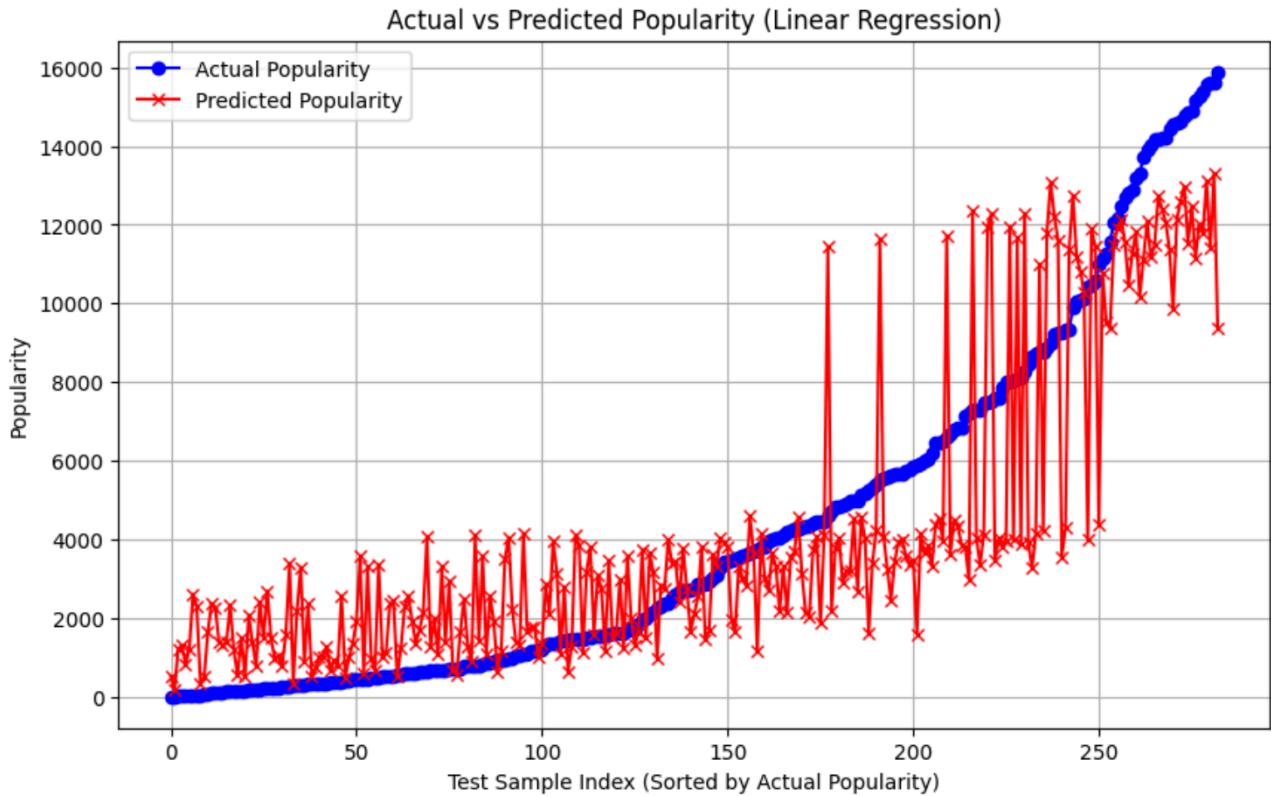
This is with standardization and normalization.

`rank` was already strongly correlated with `pop`, but because it had a larger scale originally, it may have had a smaller coefficient in the unstandardized model. After standardization, its impact is clearer, leading to a larger coefficient, indicating that it has the strongest influence in predicting popularity.



From the bar chart, `rank` has the strongest positive impact, suggesting that higher-ranked dramas are significantly more popular. `duration`, on the other hand, has a substantial negative effect, indicating that dramas aired for a longer period tend to be less popular,

possibly due to audiences losing interest over time. `tot_eps` has a smaller but still positive influence, implying that dramas with more episodes may attract greater engagement. Finally, the `year` of release has the least impact, showing that a drama's release year is not a major factor in determining its popularity.



The actual vs. predicted popularity of K-dramas can tell us some insights about the model fit.

The blue line represents the actual popularity values, sorted in increasing order. The red crosses represent the predicted popularity values from the linear regression model.

The predicted values roughly follow the increasing trend of actual popularity, indicating that the model is capturing some underlying patterns.

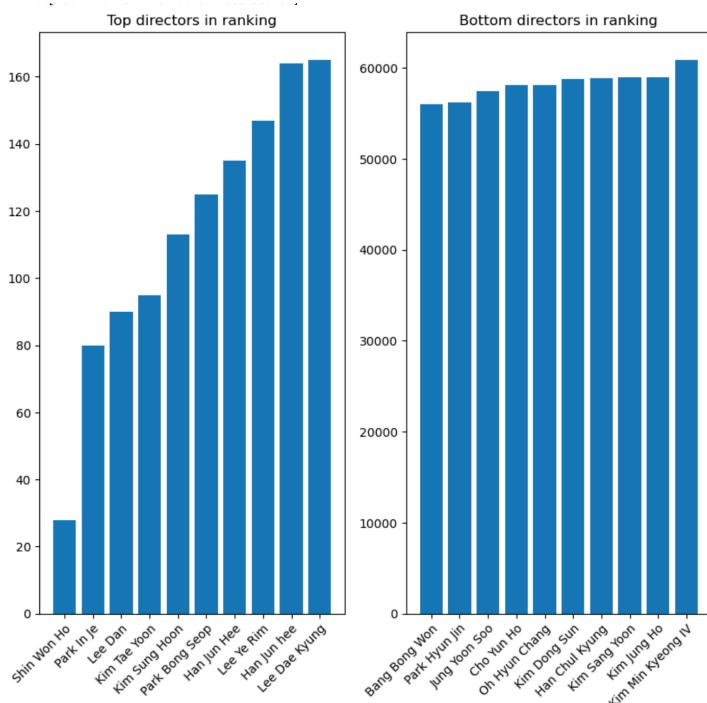
There is significant scatter in the predicted values, especially in the higher popularity range. This suggests that the model struggles to accurately predict very popular dramas. The linear regression model seems to perform better for less popular dramas (left side of the graph), where the red points are closer to the blue line. The variance increases significantly as popularity rises.

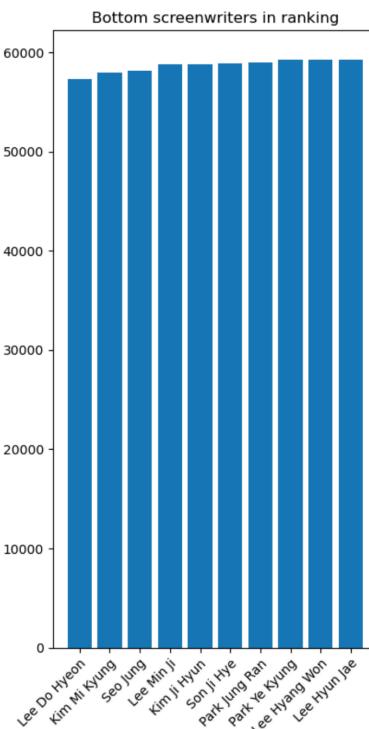
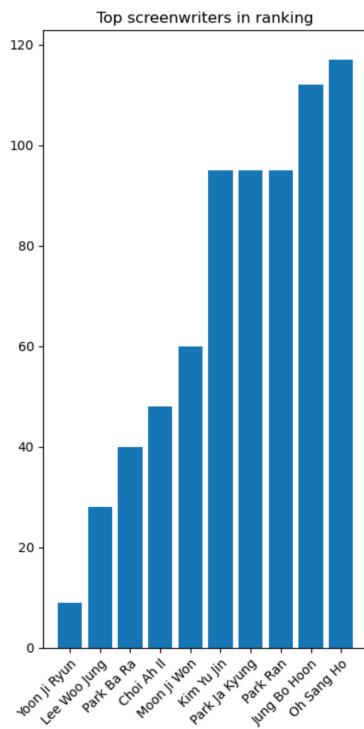
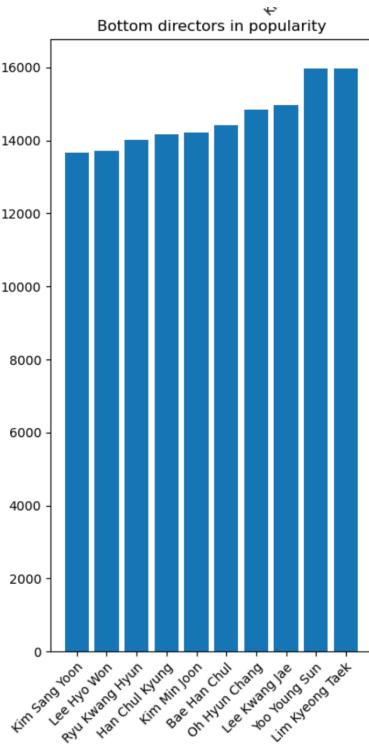
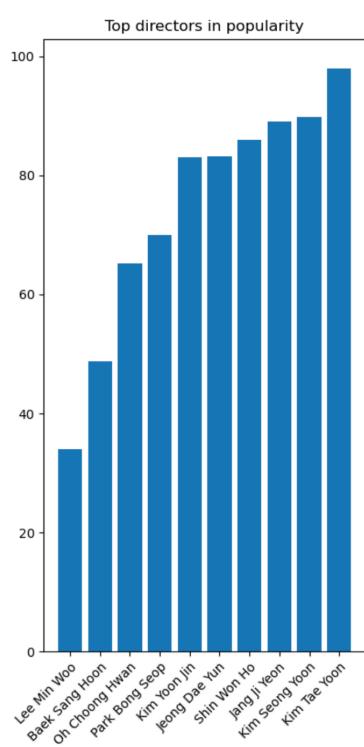
	T statistics	p value
Year	-0.03	0.98
Total number of episodes	0.30	0.76
Duration of each episode	2.65	0.01
rank	-0.79	0.43

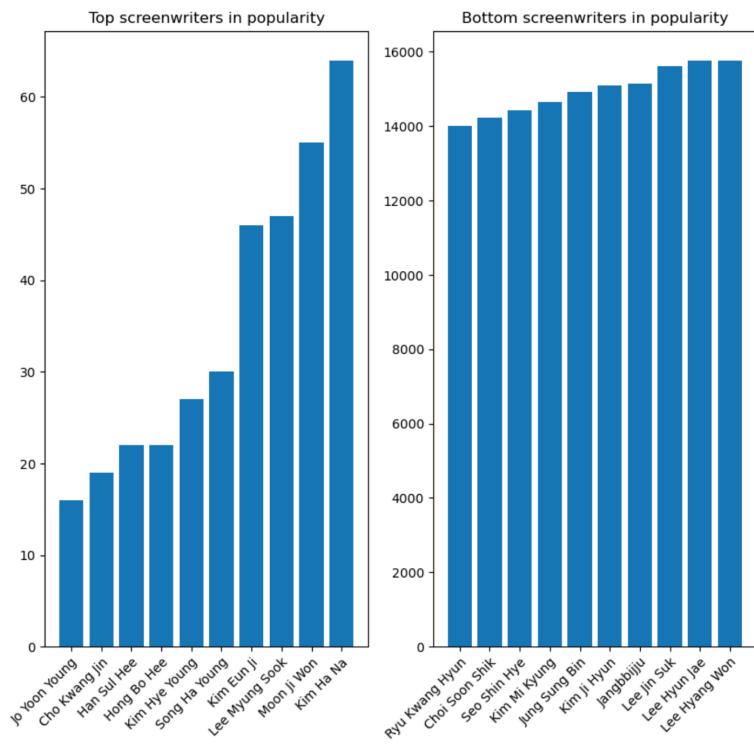
T-test on Standardized Variables

In this T-test on standardized variables, the null hypothesis (H_0) states that there is no significant difference in the mean values of each feature between the training and test sets. The results indicate that most features, including `year`, `tot_eps` (total number of episodes), and `rank`, are well-balanced between the training and test sets, as their high p-values suggest no significant difference in their distributions. However, `duration` stands out with a statistically significant p-value (0.008), implying a potential distribution shift between the two sets. This suggests that the model might face inconsistencies when predicting dramas with extreme duration values, possibly affecting overall performance.

Besides, we also did analysis on the directors and screenwriters of Korean drama:



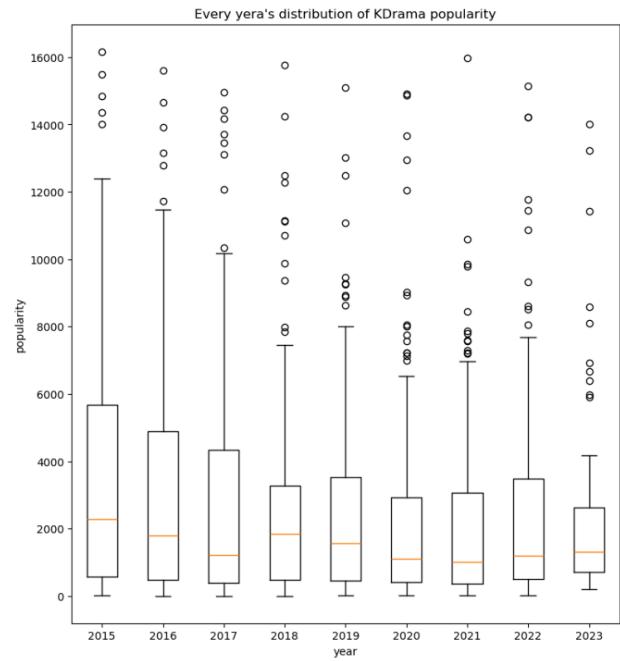
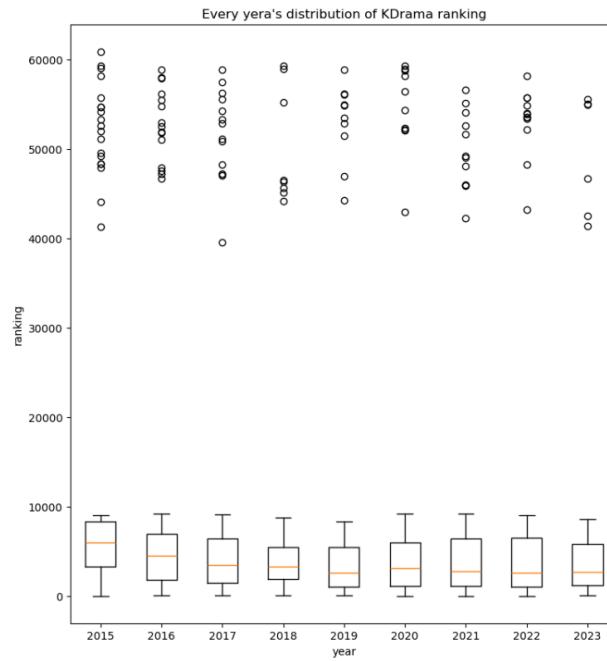




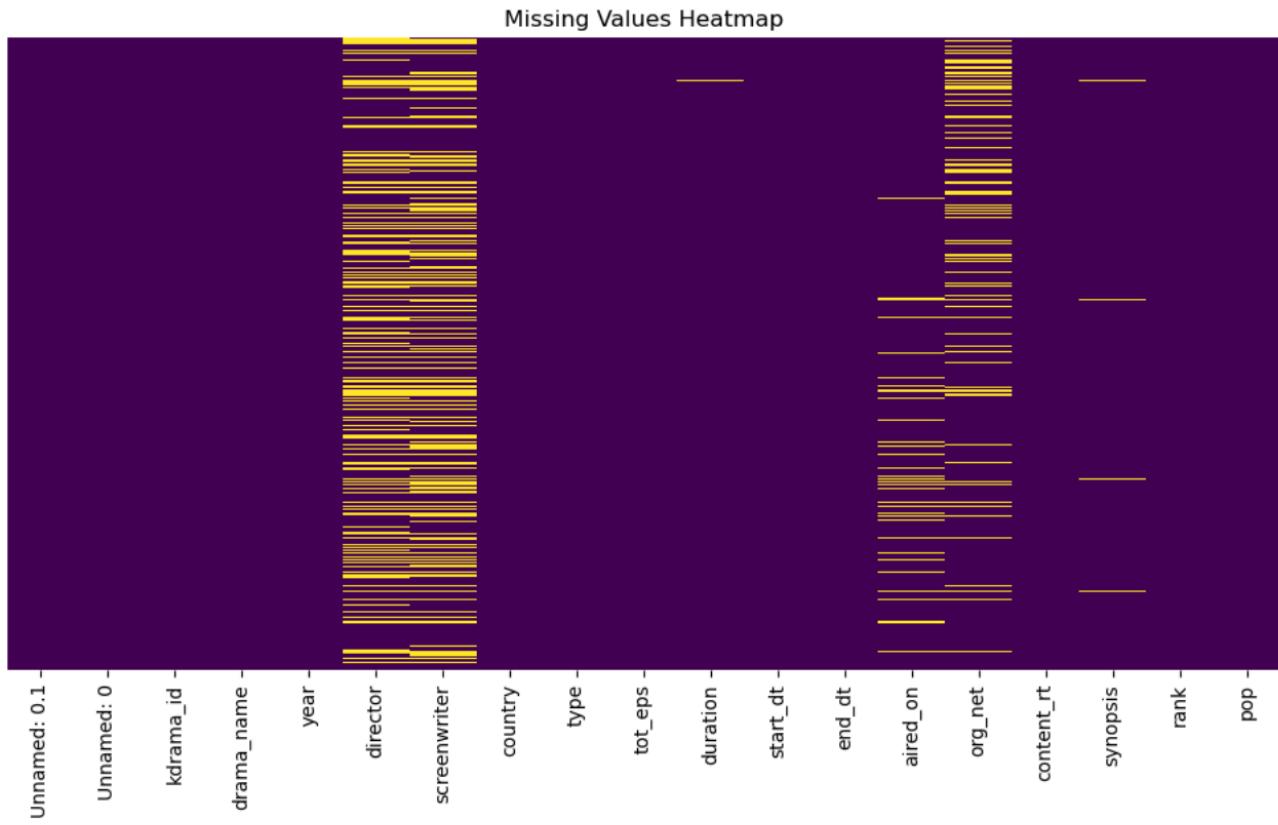
The top 10 and bottom 10 directors/screenwriters with respect to Korean Drama ranking and popularity are shown in the above charts. Besides, we also did some analysis based on time difference:

rank	
year	rank
2015	13557.000000
2016	10953.423077
2017	9029.570248
2018	6431.562500
2019	6738.812030
2020	7210.656250
2021	7429.543478
2022	7217.151316
2023	7251.461538

pop	
year	pop
2015	3986.105769
2016	3304.057692
2017	3019.719008
2018	2681.273438
2019	2635.781955
2020	2410.546875
2021	2207.514493
2022	2455.144737
2023	2518.200000



From the boxplots, we can see Korean Drama ranking increased from 2015 to 2018, and there was a slightly decreasing trend after 2019, which is also supported by the trend of mean ranking each year. There was an apparent increase in the overall popularity of Korean Drama (since lower number indicates higher popularity).



From the missing value heatmap, we can see most missing values are the name of directors and screenwriters, and Network that it aired on.

Feature Engineering

Variance and Box-Cox Transformation

Below are the feature engineering we did:

```

from sklearn.feature_selection import VarianceThreshold
# First, let's check the variance of the numerical features.
num_features = Data.select_dtypes(include=[np.number]).columns
variance = Data[num_features].var()
print("Variance of numerical features:")
print(variance)
# We choose a threshold. Here, we set a threshold of 0.01.
vt = VarianceThreshold(threshold=0.01)
X_num = Data[num_features].fillna(0) # temporarily fill missing values for variance calculation
vt.fit(X_num)
features_to_keep = X_num.columns[vt.get_support()]

print("\nNumerical features to keep (variance above threshold):")
print(list(features_to_keep))

Variance of numerical features:
tot_eps          7.146182e+02
duration         2.071882e+06
rank             2.103369e+08
pop              1.128234e+07
music_score      2.265932e+00
story_score      1.748072e+00
acting_cast_score 1.297782e+00
rewatch_value_score 3.432114e+00
overall_score    1.566452e+00
n_helpful        2.621988e+02
dtype: float64

Numerical features to keep (variance above threshold):
['tot_eps', 'duration', 'rank', 'pop', 'music_score', 'story_score', 'acting_cast_score', 'rewatch_value_score', 'overall_score', 'n_helpful']

```

First, we checked the variance of all numerical features and we set a threshold of 0.01. After calculation, no features has zero or near zero variance, thus our features are all good.

```

from scipy.stats import boxcox
from sklearn.preprocessing import StandardScaler, MinMaxScaler

# Numerical Data Transformations
num_cols = Data.select_dtypes(include=[np.number]).columns

# For Box-Cox transformation, the data must be positive.
for col in ['tot_eps', 'duration', 'rank', 'pop', 'music_score', 'story_score', 'acting_cast_score', 'rewatch_val']:
    if col in Data.columns:
        # Check for non-positive values and shift if necessary
        min_val = Data[col].min()
        if min_val <= 0:
            Data[col] = Data[col] + abs(min_val) + 1
        transformed, lam = boxcox(Data[col])
        Data[col + '_boxcox'] = transformed
        print(f"Applied Box-Cox transformation on {col} (lambda: {lam:.4f}).")
        # Optionally, drop or retain the original column

# Standardize numerical features
scaler = StandardScaler()
Data[num_cols] = scaler.fit_transform(Data[num_cols])

# Normalize (MinMax scaling)
minmax = MinMaxScaler()
Data[num_cols] = minmax.fit_transform(Data[num_cols])

# Categorical Data Transformations
cat_cols = Data.select_dtypes(include=["object"]).columns
print("\nCategorical columns before encoding:", list(cat_cols))

# One-hot encoding for categorical variables
df_encoded = pd.get_dummies(Data, columns=cat_cols, drop_first=True)

print("\nShape after one-hot encoding:", df_encoded.shape)

```

```

Applied Box-Cox transformation on tot_eps (lambda: -0.2501).
Applied Box-Cox transformation on duration (lambda: 0.7327).
Applied Box-Cox transformation on rank (lambda: 0.0761).
Applied Box-Cox transformation on pop (lambda: 0.2159).
Applied Box-Cox transformation on music_score (lambda: 2.5040).
Applied Box-Cox transformation on story_score (lambda: 2.2960).
Applied Box-Cox transformation on acting_cast_score (lambda: 3.7727).
Applied Box-Cox transformation on rewatch_value_score (lambda: 1.4414).
Applied Box-Cox transformation on overall_score (lambda: 2.5676).
Applied Box-Cox transformation on n_helpful (lambda: -0.0933).

Categorical columns before encoding: ['title', 'kdrama_id', 'director', 'screenwriter', 'start_dt', 'end_dt', 'aired_on', 'org_net', 'content_rt', 'synopsis']

Shape after one-hot encoding: (1248, 6997)

```

Next, we performed feature transformation. First, it is unnecessary to normalize the 5 score attributes since they are already with the range [0, 10]. It is sufficient to divide them by 10. Then, we performed box-cox transformation to the rest 5 numerical attributes to reduce their skewness. These 5 numerical attributes require rescaling, since their ranges are diverse. The Number of people-finding-this-comment-helpful (n_helpful) is applied normalization instead of standardization because it is skewed to 0. Since Rank and Popularity are both ranking-variables, they are applied the reverse function $f(x)=1-x$ after normalization. Total number of episodes (tot_eps) and Duration of single episode in second (duration) are applied normalization instead of standardization due to their limited value sets.

Not all categorical variables are encoded. For example, year is not encoded since it is useful

as the index for time series analysis; attributes like directors and screenwriters are impractical to encode directly, as this will produce hundreds of new variables. The only variables we applied one-hot encoding to are Content Rating (content_rt) and Aired-on Date (aired_on) because their types are limited. Content Rating was simply one-hot encoded. For Aired-on Date, since a show can be aired on several days of week, we created 7 new variables for each day of week, so each entry can have multiple True values among the 7.

The following is the refined version of feature engineering:

```
In [7]: from scipy.stats import boxcox
from sklearn.preprocessing import StandardScaler, MinMaxScaler

# Numerical Data Transformations
# Divide the scores by 10
Data[['music_score', 'story_score', 'acting_cast_score', 'rewatch_value_score', 'overall_score']] /= 10

# For Box-Cox transformation, the data must be positive.
num_cols = ['tot_eps', 'duration', 'rank', 'pop', 'n_helpful']
for col in num_cols:
    if col in Data.columns:
        # Check for non-positive values and shift if necessary
        min_val = Data[col].min()
        if min_val <= 0:
            Data[col] = Data[col] + abs(min_val) + 1
        transformed, lam = boxcox(Data[col])
        Data[col + '_boxcox'] = transformed
        print(f"Applied Box-Cox transformation on {col} (lambda: {lam:.4f}).")

# Normalize (MinMax scaling)
num_cols2 = ['tot_eps_boxcox', 'duration_boxcox', 'n_helpful_boxcox']
minmax = MinMaxScaler()
Data[num_cols2] = minmax.fit_transform(Data[num_cols2])

# Normalize with reversing
num_cols3 = ['rank_boxcox', 'pop_boxcox']
minmax = MinMaxScaler()
Data[num_cols3] = 1 - minmax.fit_transform(Data[num_cols3])

# Categorical Data Transformations
cat_cols = ['content_rt']
print("\nCategorical columns before encoding: ", list(cat_cols))

# One-hot encoding for categorical variables
df_encoded = pd.get_dummies(Data, columns=cat_cols, drop_first=True)

# One-hot encoding aired_on
import re
df_encoded['aired_on'] = df_encoded['aired_on'].apply(lambda x: re.split(r", [ \n]*", x))
days = ['Sunday', 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday']
one_hot = [day: [] for day in days]
for aired_days in df_encoded['aired_on']:
    for day in days:
        one_hot[day].append(day in aired_days)
one_hot_df = pd.DataFrame(one_hot)
df_encoded = pd.concat([df_encoded, one_hot_df], axis=1)
df_encoded.drop(['aired_on'], axis=1, inplace=True)

# Optionally, drop or retain the original column
df_encoded.drop(num_cols, axis=1, inplace=True)
print("\nShape after one-hot encoding: ", df_encoded.shape)

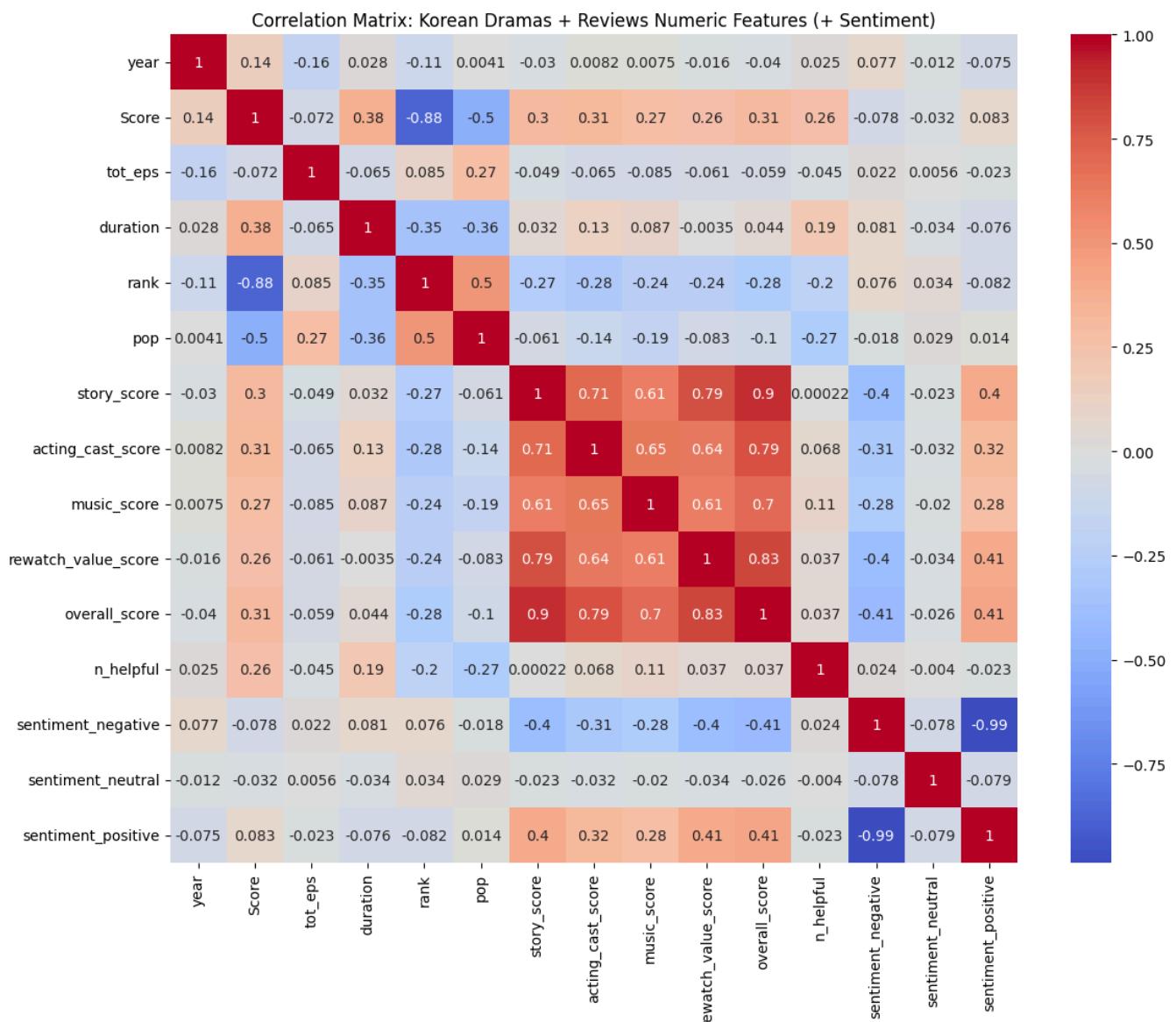
Applied Box-Cox transformation on tot_eps (lambda: -0.2501).
Applied Box-Cox transformation on duration (lambda: 0.7327).
Applied Box-Cox transformation on rank (lambda: 0.0761).
Applied Box-Cox transformation on pop (lambda: 0.2159).
Applied Box-Cox transformation on n_helpful (lambda: -0.0933).

Categorical columns before encoding: ['content_rt']

Shape after one-hot encoding: (1248, 30)
```

Below is the final cleaned data after feature engineering:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	U	W	Z	AA	AB	AC	AD	AE					
1	John	year	length	imdb_id	score	imdb_votes	adult	content	genre	music	score_story	score_acting	cast_rewatch	vt_overall	score_top250	length_top250	score_top250	content_top250	genre_top250	adult_top250												
2	1 The Family	2015-07-01	101	Doo Dong	Kim Shin	1/2/2015	2/5/2015	The prima	0.6	0.78	0.81	0.6	0.9	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77			
3	2 Love of Eve	2015-05-03	121	Kim Tae E/T Kim My	14/2/2015	6/25/2015	The series	0.7	0.77	0.85	0.9	0.8	0.9	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87			
4	3 Love on a	2015-04-27	11	Choi Ji Ye/ Choi My	14/6/2015	8/21/2015	Three won	0.1	0.3	0.3	0.1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3		
5	4 Kill Me, He	2015-06-09	107	Yoo In-na	14/6/2015	8/21/2015	A drama al	0.5	0.54	0.51	0.3	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51		
6	5 Detective	2015-07-01	101	Lee Joon-ki	14/6/2015	8/21/2015	Others	0.6	0.89	0.9	0.8	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87			
7	6 Love Detec	2015-05-02	126	Lee Hyo-jin	14/6/2015	8/21/2015	Others	0.35	0.35	0.37	0.1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3				
8	7 The Time V	2015-12-01	74	Choi Soo W/Jung Do	14/7/2015	9/16/2015	Choi Won	0.64	0.7	0.76	0.43	0.72	0.700701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631
9	8 The Dearie	2015-03-01	101	Lee Duk	14/7/2015	9/2/2015	An upbeat	0.3	0.42	0.5	0.1	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4		
10	9 The Entert	2015-01-11	117	Lee Yeo-ri	14/7/2015	9/2/2015	Oh In Yeol	0.78	0.73	0.78	0.62	0.79	0.6470715	0.76882	0.3115974	0.8115931	0.1081631	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
11	10 A Man's	2015-05-01	101	Kim Sun-kyun	14/7/2015	9/2/2015	Soohi has	0.56	0.66	0.74	0.58	0.71	0.6470715	0.76882	0.3115974	0.8115931	0.1081631	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
12	11 My Mother	2015-06-03	125	Kim Byung-ki	14/7/2015	9/2/2015	Re Hy	0.75	0.75	0.75	0.53	0.76	0.641024	0.7650671	0.5164237	0.8105671	0.1520630	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
13	12 Noble My	2015-05-03	55	Kim Heung/C Yang Others	14/8/2015	9/16/2015	Kang Hooy	0.64	0.67	0.75	0.65	0.74	0.7855	0.150502	0.2540648	0.5924009	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631			
14	13 My Mother	2015-05-03	55	Kim Heung/C Yang Others	14/8/2015	9/16/2015	Gyeong So	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
15	14 Love Cells	2015-03-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	Gyeng So	0.6	0.46	0.78	0.39	0.59	0.6470715	0.1303431	0.2851771	0.333656	0.2585343	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
16	15 Flower of	2015-04-01	101	Kim Sun-kyun	14/8/2015	9/2/2015	In this bitt	0.1	0.4	0.4	0.1	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4		
17	16 Flower of	2015-04-01	101	Kim Sun-kyun	14/8/2015	9/2/2015	Others	0.9	0.9	0.9	0.89	0.87	0.812732	0.728913	0.546242	0.4651854	0.5477949	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
18	17 Missing N	2015-04-01	101	Lee Seung/ Lee Yoo	14/8/2015	9/2/2015	Gil Su-hye	0.9	0.83	0.92	0.69	0.87	0.812732	0.728913	0.546242	0.4651854	0.5477949	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
19	18 Jumping G	2015-07-01	101	Lee Seung/ Lee Yoo	14/8/2015	9/2/2015	Others	0.65	0.64	0.76	0.68	0.68	0.6888485	0.1303431	0.2106808	0.1090805	0.251779	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
20	19 My Brother	2015-05-01	101	Kim Sang	14/8/2015	9/2/2015	Kim Do Hy	0.89	0.88	0.89	0.71	0.87	0.7000701	0.6441024	0.4417148	0.4527011	0.5182023	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
21	20 Shine or Gi	2015-04-01	101	Kim Sang	14/8/2015	9/2/2015	Wang So	0.8	0.73	0.86	0.51	0.78	0.6470715	0.698314	0.4430124	0.5481365	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631			
22	21 She Was P	2015-05-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	As a young	0.79	0.76	0.86	0.6	0.78	0.7000701	0.5441024	0.2746931	0.4827811	0.7592021	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
23	22 She Is 200	2015-05-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	200-year-old	0.6	0.61	0.61	0.42	0.64	0.6470715	0.283522	0.4530432	0.4531364	0.5920968	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
24	23 Girl of 04M	2015-05-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	Man and	0.82	0.76	0.89	0.71	0.85	0.7000701	0.3402404	0.5441024	0.5723028	0.5924998	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
25	24 Girl's Love	2015-04-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	Four hands	0.9	0.9	0.95	0.52	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55		
26	25 A Man's	2015-05-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	Four hands	0.95	0.98	0.95	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6		
27	26 A Man's	2015-05-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	Four hands	0.95	0.98	0.95	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6		
28	28 We Broke I	2015-05-01	101	Kim Sang	14/8/2015	9/16/2015	Yeon Yet	0.85	0.73	0.82	0.71	0.81	0.812732	0.1863208	0.3135989	0.4365854	0.3700537	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631		
29	30 Splendid Ju	2015-05-11	214	Princess Ju	14/8/2015	9/16/2015	Others	0.71	0.69	0.74	0.47	0.72	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
30	30 Spy	2015-05-01	101	Princess Ju	14/8/2015	9/16/2015	Others	0.67	0.63	0.82	0.44	0.69	0.7000701	0.5862971	0.3405152	0.4405923	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631	0.1661801	0.72000701	0.781105	0.363565	0.565763	0.9647949	0.9181631			
31	31 To Be Coni	2015-04-01	101	Kim Sang	14/8/2015	9/16/2015	This is a	0.88	0.86	0.87	0.76	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87		
32	32 Love at Night	2015-04-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	This is a	0.88	0.86	0.87	0.76	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87		
33	33 Sweet Savi	2015-05-01	101	Lee Hyo-jin	14/8/2015	9/16/2015	At home	0.52	0.69	0.75	0.45	0.64	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	
34	34 Sweet Tem	2015-05-01	101</																													



- Strong negative correlation (-0.99) between `sentiment_negative` and `sentiment_positive`, which is expected as they are opposing measures
- Positive sentiment shows moderate positive correlations with `story_score` (0.4), `rewatch_value_score` (0.41), and `overall_score` (0.41)—suggesting that better-rated dramas tend to generate more positive sentiment in reviews. This also demonstrates that the sentiment analysis is working as intended; the text of the reviews is being analyzed correctly.
- `sentiment_neutral` shows very weak correlations across the board, suggesting neutral reviews don't strongly relate to any other metrics

Key Findings

Our comprehensive analysis of **Korean drama data from 2015-2023** revealed several significant patterns across multiple dimensions of the industry. Through statistical analysis of features and audience reception, we identified clear trends in viewing preferences and content evolution.

Statistical Validity: The statistical analysis demonstrated **meaningful variance** across all examined features, validating their inclusion in our study. Numerical variables consistently showed asymmetrical distributions, necessitating careful statistical analysis through **Box-Cox transformations and standardization**.

Audience Reception Analysis: Our sentiment analysis pipeline, using the **DistilBERT model**, revealed fascinating insights into viewer behavior and preferences. Review scores demonstrated a pronounced left-skewed distribution, with the majority of dramas receiving **ratings between 8-10**. The analysis uncovered an almost perfect split between **positive (4,891)** and **negative (4,871)** reviews, with remarkably few neutral sentiments (61 reviews). This polarization suggests that K-dramas tend to elicit **strong emotional responses** from viewers, with **robust correlations between positive sentiment and high scores** across story quality (0.4), rewatch value (0.41), and overall rating (0.41).

Audience Reception Analysis (Reviews): Content format preferences emerged as a crucial factor in drama success. Our analysis revealed that dramas with **fewer than 20 episodes** consistently achieved higher popularity rankings, indicating a clear audience preference for more concise storytelling formats. This finding was reinforced by the **negative correlation (-0.38)** between episode duration and popularity ranking. The **bi-modal distribution of rewatch value scores** further suggested that dramas tend to fall into two distinct categories: highly rewatchable content (scoring 9-10) or single-viewing experiences (scoring 1-2). Similarly, since there is a negative relationship between overall score/rewatch value score/music score and popularity ranking and there is a positive relationship between story score/acting cast score and popularity ranking, it's likely that most of the audience enjoys a Korean drama because of its story and acting cast, but higher level of music or rewatch value means less popularity. In other words, most audience may not like a Korean drama due to its music or rewatch value.

Challenges and Future Recommendations

Our research encountered several significant challenges that inform our recommendations for future studies. **Data quality** presented persistent challenges, particularly regarding missing information for directors and screenwriters, inconsistent date formatting across sources, and varying levels of completeness in our multiple data sources. These issues required substantial preprocessing and standardization efforts.

Technical limitations also impacted our research scope. Web scraping restrictions from [MyDramaList](#)'s anti-bot protections constrained our ability to gather additional historical data. Resource constraints affected our processing of large-scale sentiment analysis, and memory management became crucial when handling high-dimensional feature spaces after categorical encoding.

Looking forward, we recommend several key areas for future research enhancement. First, implementing **advanced web scraping** techniques using rotating proxies and user agents would enable more comprehensive data collection. We also suggest developing automated data validation and standardization pipelines to improve data quality consistency. Establishing direct partnerships with data providers for **API access** would significantly enhance data reliability and completeness.

Additionally, incorporating production **budget data**, international **viewing statistics**, and **marketing spend** information would provide a more complete picture of drama success factors. These enhancements, combined with robust cross-validation methods and ensemble prediction approaches, would significantly advance our understanding of the Korean drama industry's evolution and success factors. Going even further, given the abundance of text data in our dataset, implementation of **topic modeling** on review text and synopses could uncover deeper thematic trends.

Contribution:

- **Mengyan Li (ml4779)**: Finding the data, Data Cleaning(fixing date format, remove outliers, merge datasets, drop/change to others for missing values),

EDA(summary statistics, linear regression for the merged data and review.csv, correlation heatmap, missing value heatmap, histogram for numerical variables), feature engineering(filter out zero/near-zero variance features, Box-Cox transformation, standardization/normalization, one-hot encoding), Report drafting

- **Zishun Shen (zs2695):** EDA(tentatively approach to visualize the dataset, top and bottom ranking of the directors and screenwriters through popularity rank and rank, ranking through year boxplot), feature engineering (one-hot encoding)
- **Zhisheng Yang (zy2675):** Data Cleaning(tentatively approach to drop missing values), EDA and feature engineering(linear regression visualization and standardize/normalize data, t test, feature importance bar chart, The actual vs. predicted popularity of K-dramas)
- **Shayan Chowdhury (sc4040):** report drafting and editing, web scraping script, combining popularity/ranking and revenue over the year, ML-based sentiment analysis of reviews using HuggingFace transformers and DistilBERT (feature engineering), merging of multiple kdramas datasets, merging of kdramas + revenue dataset