Report for Applied Data Science Project One

Report for Applied Data Science Project One

Group12: Mengyan Li (ml4779), Zishun Shen (zs2695), Zhisheng Yang (zy2675), Shayan Huda Chowdhury (sc4040)

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

Introduction and Data:

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 scrapping. 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 17 columns in korean_drama.csv including drama ID, drama name, Released year, Director name, Screenwriter name, country of origin, type, Total number of episodes, Each episode duration in second, First aired date, End date, Day of the week that it was aired on, Network that it aired on, Content Rating, Short synopsis, Ranking on the website, and Popularity Ranking on the website.

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.

Methodology:

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 scrapping 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_nam	year	director	screenwriter	country	type	tot_eps	duration	start_dt	end_dt	aired_on	org_net	content_rt	synopsis	rank	pop	
2	661d419391	Sing My Crus	2023	['So Joon Moon']		South Korea	Drama	8	1500	2-Aug-23	2-Aug-23	Wednesday		Not Yet Rate	Follow the st	1484	2238	
3	5ffcbeaa171	D.P. Season 2	2023		['Kim Bo Ton	South Korea	Drama	6	3000	28-Jul-23	28-Jul-23	Friday	Netflix	15+ - Teens	This unfoldin	164	1084	
4	65075cb9c1a	Shadow Dete	2023	['Han Dong H	['Song Jung \	South Korea	Drama	8	3300	7/5/23	7/26/23	Wednesday	Disney+ Hulu	15+ - Teens	Unfolds the	2443	6915	
5	df0f0ac4b3ff	To Be Hones	2023			South Korea	Drama	3	600	6/30/23	7/14/23	Friday		Not Yet Rate	Don't you ha	49895	99999	
6	04c1fe41948	Celebrity	2023	['Kim Chul G	['Kim Yi Your	South Korea	Drama	12	2700	30-Jun-23	30-Jun-23	Friday	Netflix	18+ Restricte	Fame. Mone	826	547	
7	2e06f3f0c28	Blue Temper	2023			South Korea	Drama	4	420	6/28/23	7/19/23	Wednesday	Naver TV Ca	Not Yet Rate	Set in the pic	47245	15405	
8	e632ba7628	Numbers	2023	['Kim Chil Bo	['Oh Hye Seo	South Korea	Drama	12	3600	6/23/23	7/29/23	Friday, Satur	MBC	15+ - Teens	Ho Woo is th	2685	1546	
9	4c99e91ea0	Revenant	2023	['Lee Jung Ri	['Kim Eun He	South Korea	Drama	12	4200	6/23/23	7/29/23	Friday, Satur	SBS	15+ - Teens	When the do	236	769	
10	4afb1765522	Adult Kids	2023			South Korea	Drama	8	600	6/19/23	8/7/23	Monday		Not Yet Rate	A hyper-reali	47813	10649	
11	565dbf05b12	Lies Hidden i	2023	['Jung Ji Hyui	['Ji Ah Nee']	South Korea	Drama	8	3600	6/19/23	7/11/23	Monday, Tue	ENA Genie T	15+ - Teens	Moon Joo Ra	3381	1843	
12	24ecf5e1ed6	King the Lan	2023	['Im Hyun W	['Choi Rom']	South Korea	Drama	16	4800	6/17/23	8/6/23	Saturday, Su	jTBC Netflix	15+ - Teens	Heir Goo Wo	605	237	
13	46b4058342	See You in M	2023	['Lee Na Jung	['Han Ah Reu	South Korea	Drama	12	4200	6/17/23	7/23/23	Saturday, Su	Netflix TVIN	15+ - Teens	Ban Ji Eum h	467	297	
14	6dca9e88afd	Bloodhounds	2023			South Korea	Drama	8	3600	9-Jun-23	9-Jun-23	Friday	Netflix	18+ Restricte	When reserv	261	444	
15	9213b6aa40	Love Tractor	2023	['Yang Kyung	Hee']	South Korea	Drama	8	1500	6/7/23	6/21/23	Wednesday		13+ - Teens	Seon Yul, a t	2342	1099	
16	e16e394bd6	Romance by	2023			South Korea	Drama	10	900	6/7/23	6/28/23	Wednesday		15+ - Teens	Having never	6232	7118	
17	b5b3c87821	The Villain o	2023			South Korea	Drama	10	2700	6/5/23	6/27/23	Monday, Tue	MBC Dramar	15+ - Teens	A coming-of-	56335	7766	
18	e40375435a	Sound Candy	2023	['Kang Hee Ju']		South Korea	Drama	10	900	6/3/23	7/1/23	Saturday	TVING	G - All Ages	A group of yo	54892	8575	
19	ede6d42db8	Bitch X Rich	2023			South Korea	Drama	10	2100	5/31/23	6/28/23	Wednesday	Netflix	15+ - Teens	Baek Je Na is	5659	1731	
20	a4b48d65e0	Battle for Ha	2023	['Kim Yoon C	['Joo Young I	South Korea	Drama	16	4200	5/31/23	7/20/23	Wednesday,	Amazon Prin	Not Yet Rate	A suspense d	2952	3767	
21	147ea55381	Delightfully I	2023	['Lee Soo Hyi	['Han Woo Jc	South Korea	Drama	16	4200	5/29/23	7/18/23	Monday, Tue	tvN	15+ - Teens	Clever con ar	3486	1302	
22	0c9607d2eca	One Day Off	2023	['Lee Jong Pi	['Son Mi']	South Korea	Drama	8	1500	5/24/23	5/31/23	Wednesday		Not Yet Rate	Set in the 19	1223	3710	
23	47b6544945	Star Struck	2023	['Park Sun Ja	['Jung Hyun \	South Korea	Drama	8	1020	5/18/23	6/8/23	Thursday		15+ - Teens	When love is	8567	1323	
24	213b74d7e6	Oh! Youngsir	2023	['Kim Eun Ky	['Jeon Seon \	South Korea	Drama	10	3000	5/15/23	6/13/23	Monday, Tue	ENA Genie T	Not Yet Rate	Oh Young Sir	8020	2632	
25	e27854d516	Black Knight	2023			South Korea	Drama	6	2520	12-May-23	12-May-23	Friday	Netflix	15+ - Teens	In 2071, toxi	1422	608	
26	131df87dd3	Race	2023	['Lee Dong Y	['Kim Roo Ri	South Korea	Drama	12	3480	5/10/23	6/14/23	Wednesday	Disney+ Hulu	Not Yet Rate	Office politic	6265	4178	
27	d623b6c105	It Was Spring	2023	['Park Joon Sik']		South Korea	Drama	8	1020	5/8/23	6/26/23	Monday	KBS1	Not Yet Rate	Set in a high	55545	13235	
28	2ace8af95b1	Tale of the N	2023	['Kang Shin H	['Han Woo R	South Korea	Drama	12	4200	5/6/23	6/11/23	Saturday, Su	tvN	15+ - Teens	An unexpecte	101	546	
29	1470331acb	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	756ab26b55	Love Mate	2023	['So Joon Moon']		South Korea	Drama	8	1200	5/4/23	5/25/23	Thursday		15+ - Teens	As a team le	5695	1135	
31	3974d947cd	My Perfect S	2023	['Kang Soo Y	['Baek So Ye	South Korea	Drama	16	4200	5/1/23	6/20/23	Monday, Tue	KBS2 ViuTV	15+ - Teens	Yoon Hae Joo	310	981	
32	69218fd653	Finland Papa	2023			South Korea	Drama	6	1680	4/29/23	5/14/23	Saturday, Sunday		Not Yet Rate	Lee Yu Ri is a	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 ENI	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 GOOI	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 MOI	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 MORI	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	01b015525e8	Sing My Crus	8.5	9	10	10	9	Second Winc	8 of 8 episod	37	
27	01b015525e8	Star Struck	8	8.5	7.5	6.5	7.5	More like a g	8 of 8 episod	24	
28	01b015525e8	Happy Merry	6	7	7.5	3.5	6	They got the	8 of 8 episod	25	
29	01b015525e8	The Eighth S	10	10	10	10	10	Blooming lik	10 of 10 epis	12	
30	01b015525e8	Individual Cir	9	9.5	8	8.5	8	Why did the	8 of 8 episod	4	
31	01b015525e8	The Director	6.5	7.5	7.5	5	6.5	Just . . . Bad.	10 of 10 epis	8	
32	01b015525e8	Happy Ending	8	9	9	8	8	An Empty Sp	8 of 8 episod	3	
33	01b015525e8	Summer Stri	7.5	8.5	8.5	6	7.5	Tumbling do	12 of 12 epis	17	
34	01b015525e8	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:

The first step is removing outliers. Because we focus on popularity, and we plan to use popularity and overall score as the dependent variables for two different linear regression, we removed the outliers in pop and overall_score. We built a boxplot with the interquartile range and removed any outliers above the upper limit and lower limit. In total, we removed 609 outliers.

The second step is fixing the date. Like said, some dates are month-date-year but some are date-month-year. Thus we used Python to correct the date. Some dates only has month and year but no dates, thus we dropped these data.

The third step is filling in missing values. After analyzing the data's missing value pattern, we found out that most missing values are the name of directors and screenwriters, and Network that it aired on. At first we want to fill in these missing value. However, the only way is through web scrapping. Based on the two week deadline, there is not enough time for us to

learn web scrapping and finish the project on time. Thus we fill in the missing value with "Others" and remove any missing values in the numerical variables.

The fourth step is mergeing dataset. The two datasets korean_drama.csv and review.csv have different numbers of rows. However, after analyzing datasets, we found out that the drama names in both datasets are very similar. Therefore, we used groupby function in pandas and grouped the review.csv by drama names. And we calculated the mean of the scores by the users who marked the same drama. We used the mean score as the score for this specific drama. We did this for Score for Story, Score for acting, Score for music, Score for rewatch value and Overall Score. Therefore, we merged the two datatables.

The following is the final cleaned data table:

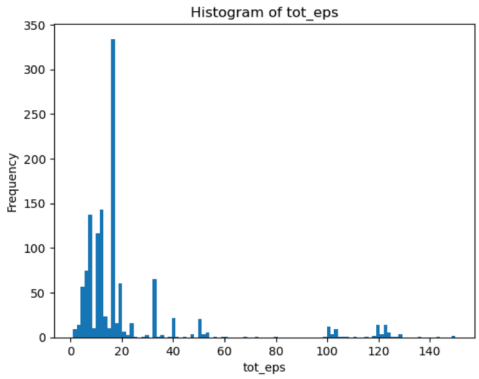
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	title	year	kdrama_id	director	screenwriter	tot_eps	duration	start_dt	end_dt	aired_on	org_net	content_rt	synopsis	rank	pop	music_score	story_score	acting_cast_	rewatch_val	overall_scor	n_helpful
2	The Family Is	2015	9170c31a02(	['Joo Dong M	['Kim Shin Hy	20	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	2dc531911b(	['Kim Tae Eu	['Kim Min Se	12	3000	4/2/15	6/25/15	Thursday	Mnet	18+ Restricte	The series ta	5486	527	7.7	8.5	9	8	8.7	15
4	Love of Eve	2015	832b2dcd1a(	['Lee Gye Joo	Others	120	2400	5/18/15	10/30/15	Monday, Tue	MBC	15+ - Teens	Three wome	54646	12387	1	3	3	1	3	1
5	Love on a Ro	2015	5eda227cf63	['Choi Ji Yeor	['Choi Min Ki	101	2100	4/6/15	8/21/15	Monday, Tue	KBS2	15+ - Teens	A drama abo	8341	7469	5	5.4	5.1	3	5.1	5
6	Kill Me, Heal	2015	e8c09f0f7d0	['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	59
7	Delicious Lov	2015	a38a585a6c(	Others	Others	3	1800	11/5/15	11/10/15	Monday, Tue	Naver TV Ca	Not Yet Rate	A teen roma	9163	4826	5.2	2.8	3.8	1	3.5	3
8	Love Detectiv	2015	fff500c6bfe7	Others	Others	10	600	11/11/15	11/20/15	Monday, Tue	Naver TV Ca	Not Yet Rate	Sherlock K is	8964	4942	3.5	3.5	7	1	4	2
9	The Time We	2015	12208c74d65	['Jo Soo Wor	['Jung Do Yo	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	12
10	The Dearest	2015	729e5a41ac(	['Choi Chang	['Seo Hyun Jo	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	1
11	The Eccentric	2015	cc11a722be1	['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	3
12	Awl	2015	b5446cda988	['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	7
13	Bubblegum	2015	f8ecb7e2038	['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	12
14	Noble, My Lo	2015	e2328f79d0(	['Kim Yang H	Others	20	900	8/23/15	9/16/15	Monday, Tue	Naver TV Ca	13+ - Teens	Kang Hoon is	5090	310	6.4	6.7	7.5	6.5	7.4	14
15	My Mother Is	2015	65e26535bd(	['Ko Heung S	['Lee Geun Y	136	2400	6/22/15	12/31/15	Monday, We	SBS	15+ - Teens	Gyeong Sook	44090	14364	10	10	10	10	10	0
16	Love Cells Se	2015	3bef90625c1	Others	Others	12	600	9/14/15	10/1/15	Monday, Tue	Naver TV Ca	13+ - Teens	In this bitter-	8646	3225	6	4.6	7.8	3.9	5.9	2
17	Flower of the	2015	62df2045415	['Kim Min Sh	['Park Hyun J	50	3900	3/14/15	8/30/15	Saturday, Su	MBC	15+ - Teens	South Korear	6954	4493	1	4	4.5	1	4	2
18	Six Flying Dra	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	27
19	Missing Noir	2015	64309e69d8(	['Lee Seung '	['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	12
20	Jumping Girl	2015	5ef07c61a4d	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	6.8	3
21	My Beautiful	2015	e265f560507	['Kim Chul Gi	['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	9
22	Shine or Go (	2015	45af025f19d	['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	11
23	Missing Kore	2015	0f989ec2490	['Min Doo Sh	Others	6	600	11/3/15	11/12/15	Tuesday, We	Naver TV Ca	Not Yet Rate	¸ÃûMissing K	8905	4337	3.9	6.4	7.1	3.4	6.4	1
24	She Was Pret	2015	1985f18b5a1	['Jeong Dae '	['Jo Sung Hee	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	36
25	She Is 200 Ye	2015	ffaec6267d9	Others	Others	5	600	10/27/15	10/27/15	Others	Naver TV Ca	Not Yet Rate	200-year-old	8782	2678	6	6.1	6.1	4.2	6.4	3
26	Warm and Cl	2015	81c8f1d83d7	['Park Hong k	['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	14
27	Girl of 0AM	2015	634519fd3e7	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	3
28	Girls' Love Sh	2015	8a5f7bb42f2	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	3
29	Assembly	2015	56f7a1b9c72	['Hwang In H	['Jung Hyun I	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	6
30	We Broke Up	2015	dce6fe763f2(	Others	['Jeon Seon Y	10	900	6/29/15	7/17/15	Monday, We	Naver TV Ca	Not Yet Rate	Ji Won Yeon	6979	1712	8.5	7.3	8.2	7.1	8.1	4
31	Splendid Poli	2015	75c1221d1d(	['Kim Sang H	['Kim Yi Youn	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	6
32	Spy	2015	14fb3da5d0(	['Park Hyun S	['Han Sang W	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	5
33	To Be Contin	2015	05be099411!	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	3
34	Late Night R	2015	1ede9e8879(	['Hwang In R	Others	20	1800	7/4/15	9/5/15	Saturday	SBS	Not Yet Rate	A late night (	3327	3914	8.9	9.4	9.4	9.3	9.4	6
35	Sweet, Savag	2015	53d882ba89(	['Kang Dae S	['Son Geun Ja	16	3780	11/18/15	1/14/16	Wednesday,	MBC	15+ - Teens	At home, Tae	7123	4165	5.2	6.8	7.5	4.5	6.8	4

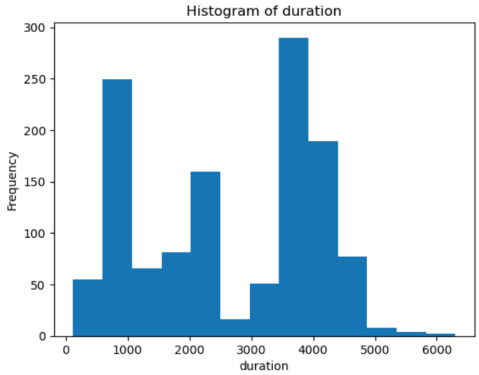
EDA:

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.

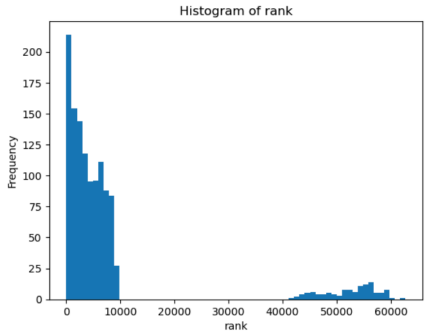
	Total number of episodes	Duration	Rank	Popularity Ranking	Music score	Story score	Acting cast score	Rewatch cast score	Overall score	Number of people that found commit helpful
count	1248	1248	1248	1248	1248	1248	1248	1248	1248	1248
mean	22	2633	8390	3028	7.3	7.4	8.3	6	7.6	12
std	26	1439	14502	3358	1.5	1.3	1.1	1	1.3	16
min	1	120	9	1	1	1	2	1	2.5	0
25%	10	1140	1592	571	6.7	6.8	7.8	5	7	3
50%	16	2700	3876	1713	7.6	7.6	8.5	6.2	7.8	6
75%	18	3900	6896	4299	8.3	8.4	9.1	7.3	8.5	15
max	150	6300	62723	15980	10	10	10	10	10	158

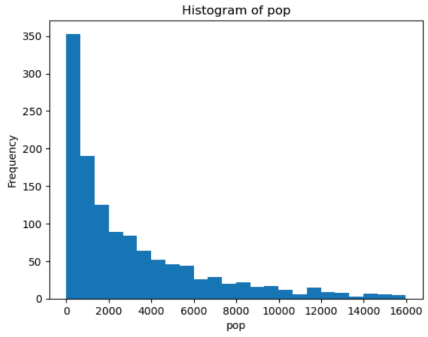
Below are the histograms of all numeric variables:



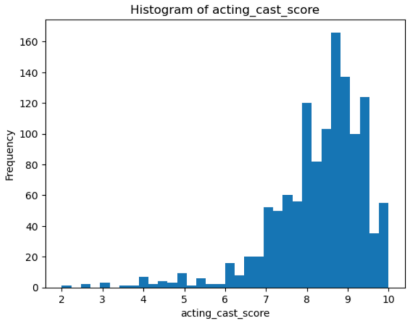


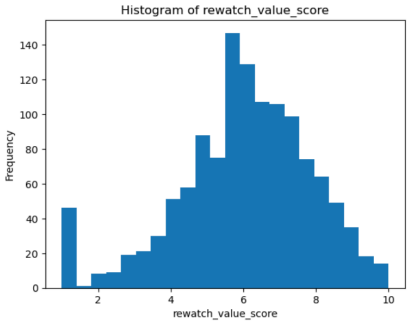
Looking at data for Total number of episodes, we can see that there is only one peak and the data is right skewed. Looking at data for Duration, we can see that there are two peaks, one at 4000 and the other at 1000, meaning there are mainly two types of drama, one is longer version for each episode and the other is shorter version for each episode.



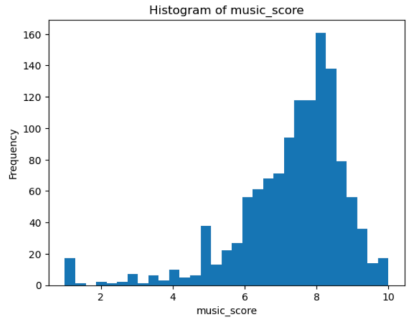


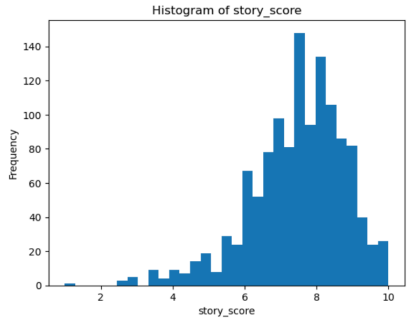
The data for rank and popularity ranking are also right skewed. Since higher number in ranking means lower popularity, most drama ranked high in the ranking and only a few have low popularity.



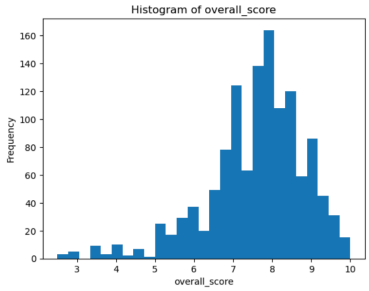


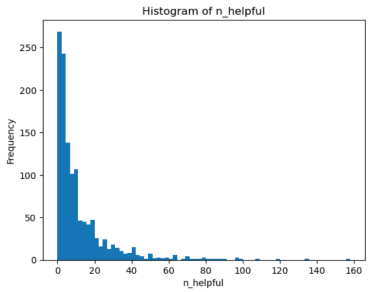
The data for acting cast score is left skewed, meaning many drama have high score on acting cast. The audience provide good reviews on the acting cast of Korean drama. The data for rewatch value score is also left skewed. However, there are more scores below the median meaning more audience do not want to rewatch the drama.





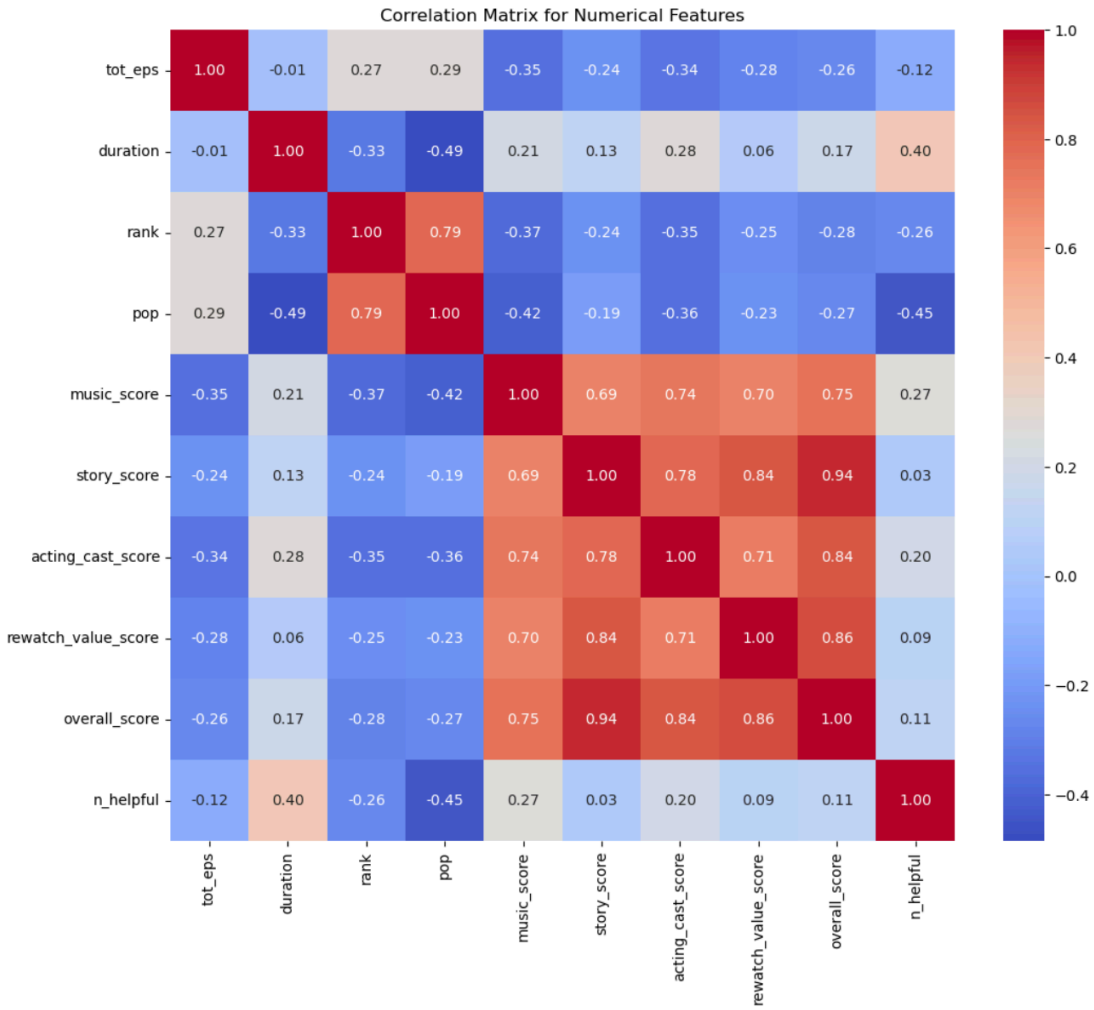
The data for music score and story score are both left skewed meaning most audience gave high scores to the music and the story and only a few audience gave low scores.



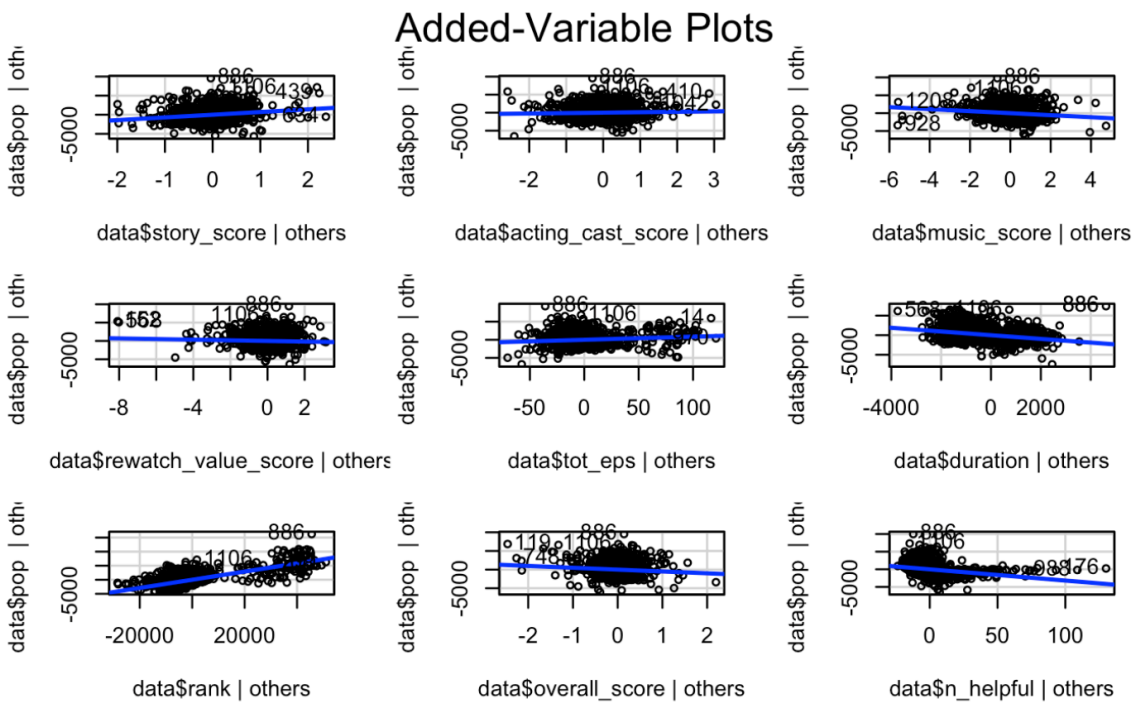


The data for the overall score is left skewed, meaning most audience gave high scores to the drama overall. The data for the Number of people that found commit helpful is right skewed, meaning the commit are very personalized so that not so many people hold same opinion.

Below is the correlation heatmap of all numerical values:

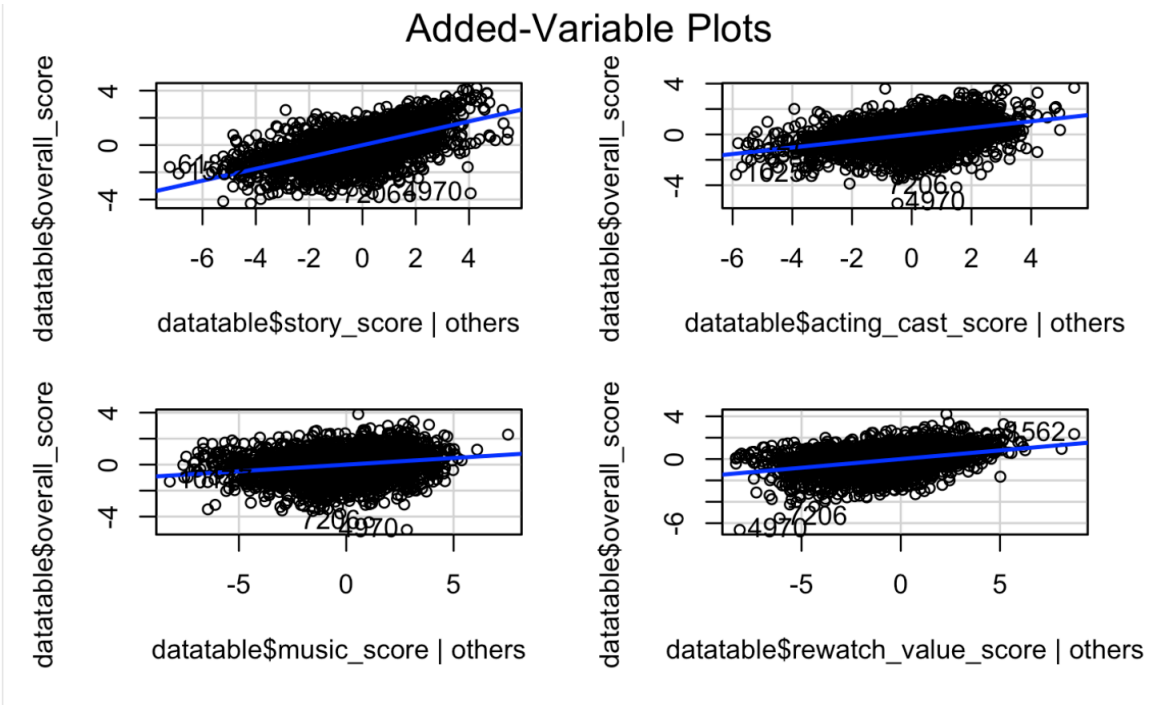


We can see that popularity ranking and ranking are highly correlated. And music_score, story_score, acting_cast_score, rewatch_value_score, and overall_score are highly correlated.



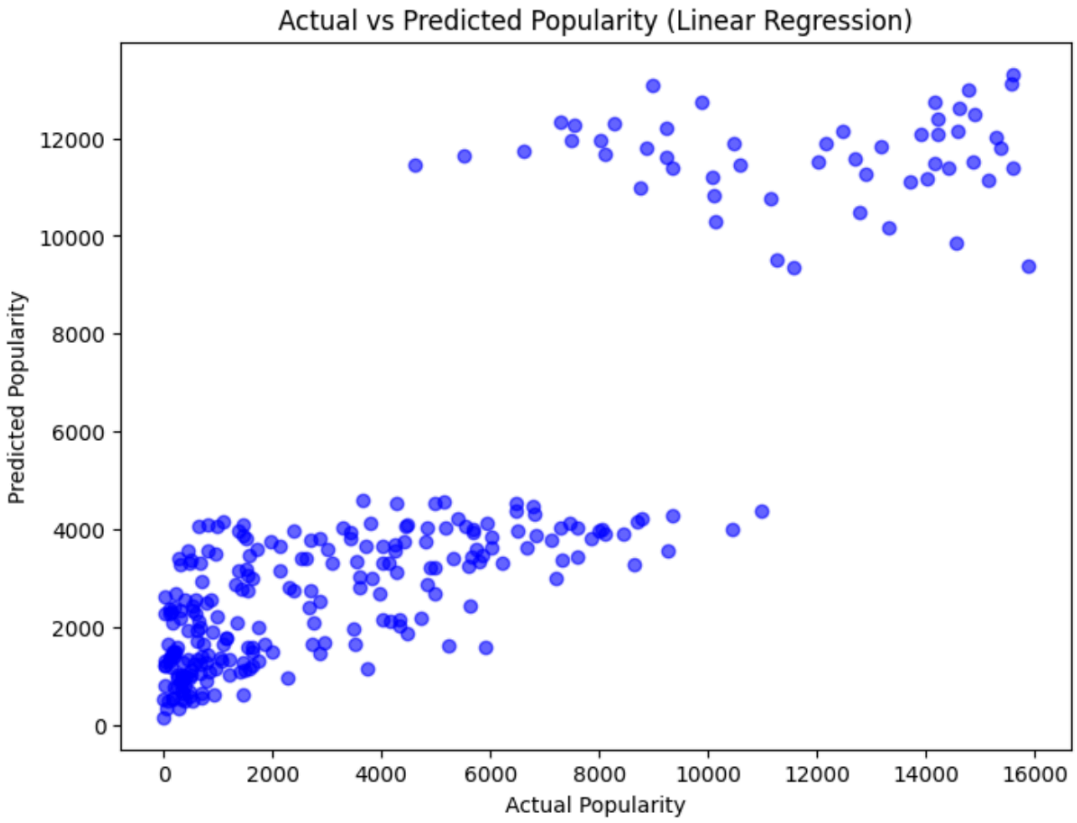
Same thing happened when we look at the linear regression plot. We can see that there is a positive relationship between popularity ranking and rank. We also found that there is a negative relationship between duration and popularity ranking, meaning longer drama has higher popularity. Interesting thing to see is that there is a negative relationship between number of people found the commit helpful and popularity ranking, 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:



This is the linear regression analysis between the overall score and story score/acting cast score/music music/rewatch value score. We can see that these are all positive relationship, meaning the story score/acting cast score/music music/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.

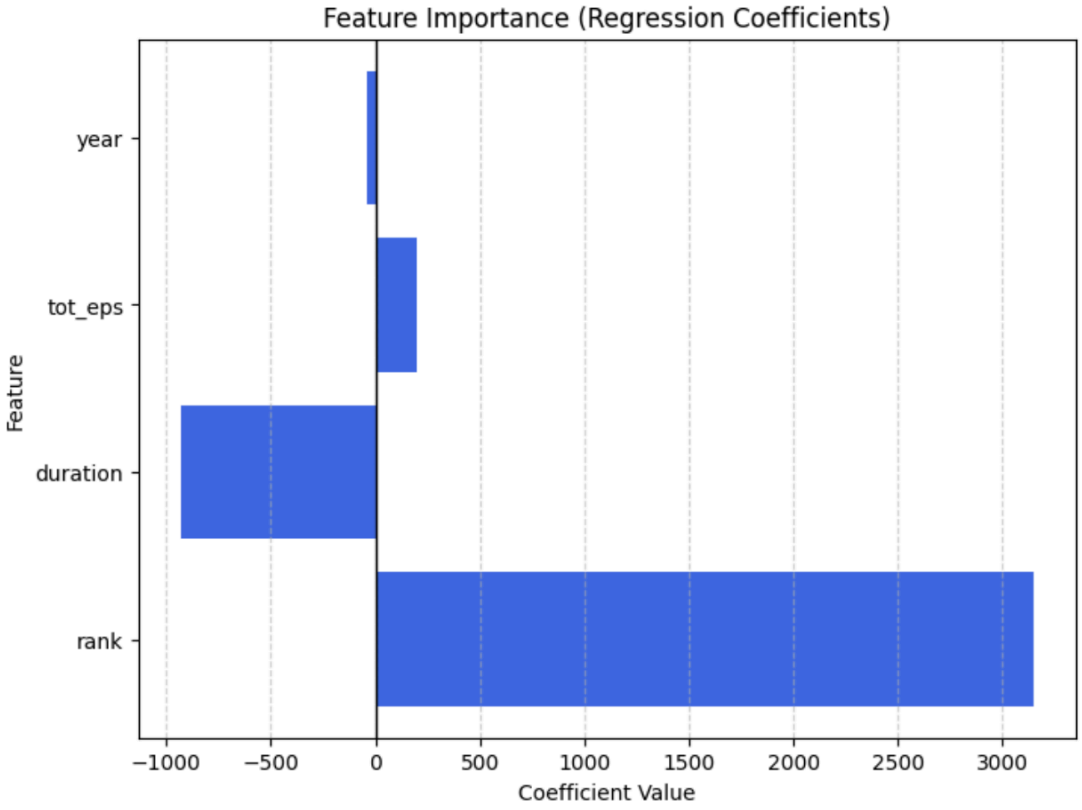
Comparison: Without standardization and normalization and with standardization and normalization:

Feature	Coefficient
year	-16
Total number of episodes	7
Duration of each episode	-0.6
rank	0.2

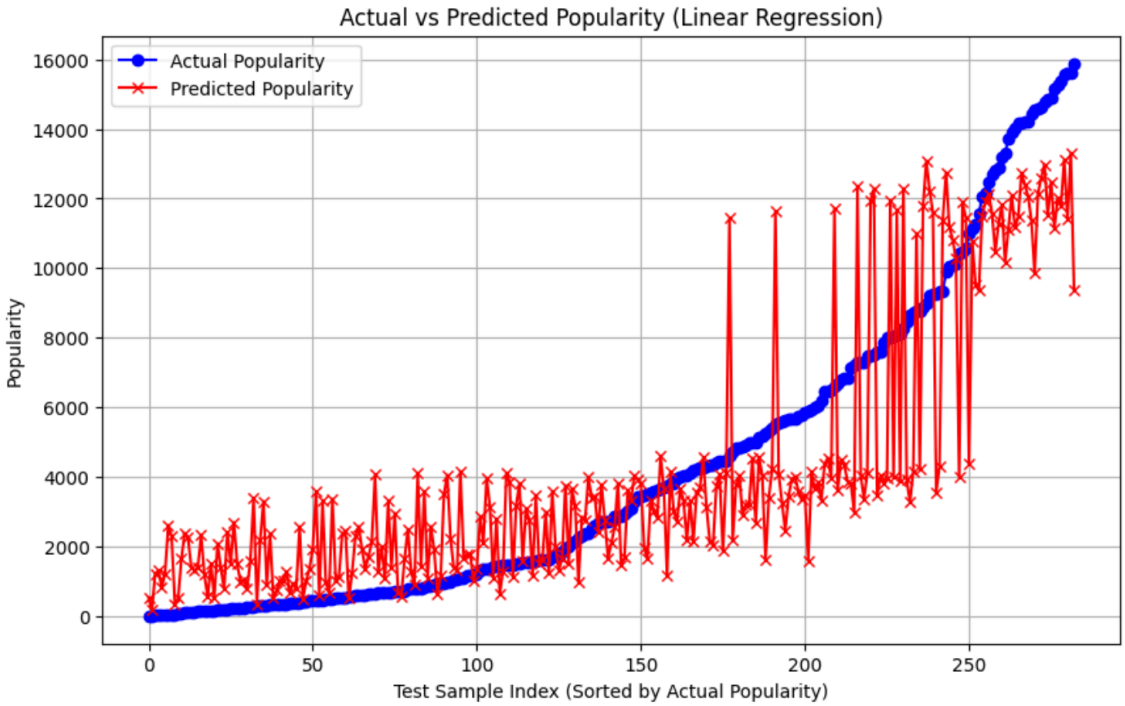
Feature	Coefficient
year	-38
Total number of episodes	195
Duration of each episode	-929
rank	3146

This is with standardization and normalization.

Rank was already strongly correlated with popularity, 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. The number of episodes 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.

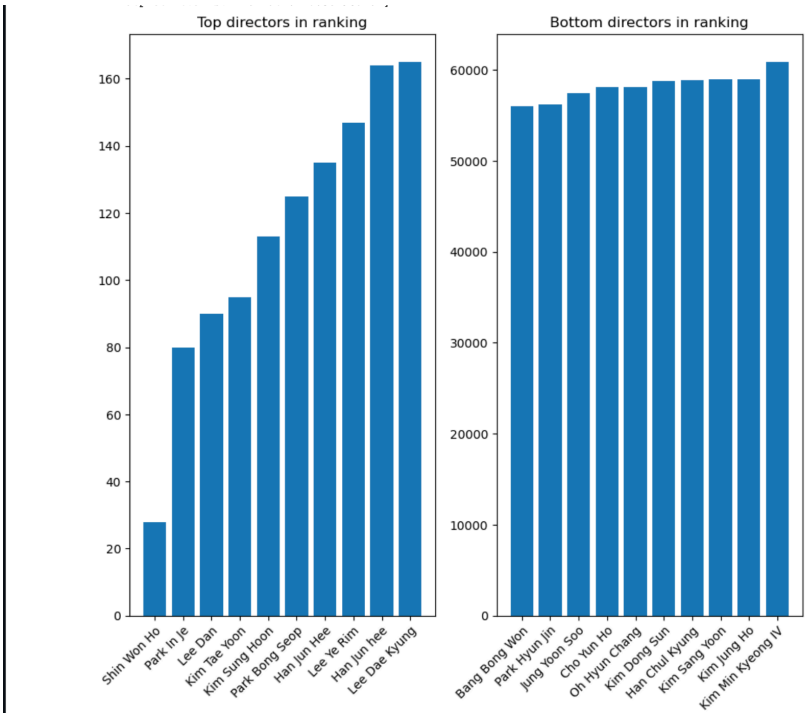
There is significant scatter in the predicted values, especially in the higher popularity range. This suggests that the model struggles to accurately predict very popular dramas. The linear regression model seems to perform better for less popular dramas (left side of the graph), where the red points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

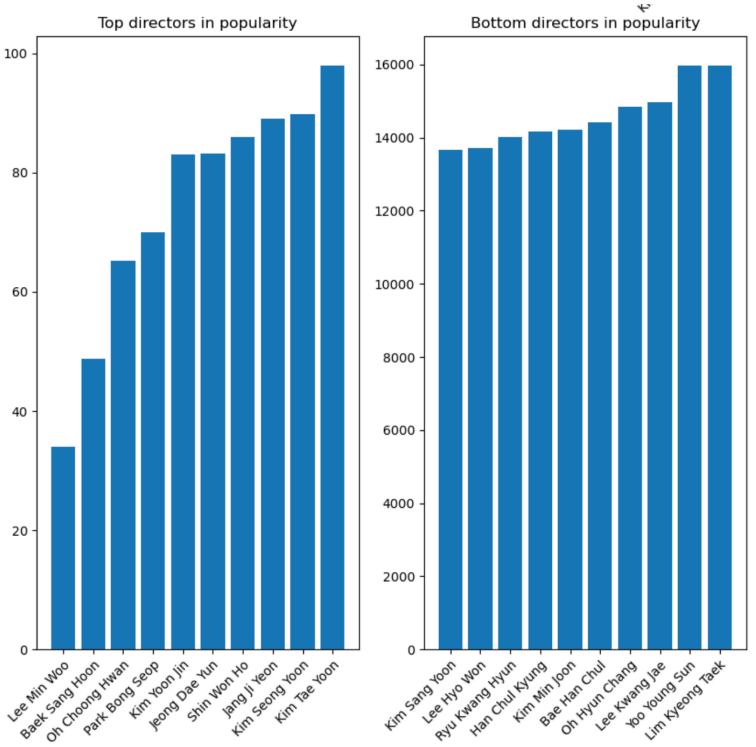
	T statistics	p value
Year	-0.03	0.98
Total number of episodes	0.30	0.76
Duration of each episode	2.65	0.01
rank	-0.79	0.43

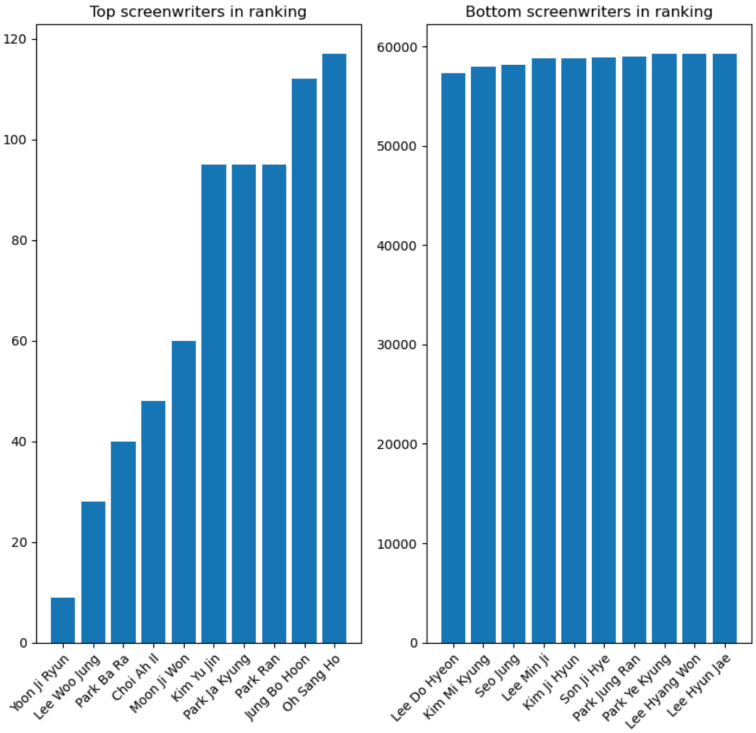
In this T-test on standardized variables, the null hypothesis (H) states that there is no significant difference in the mean values of each feature between the training and test sets. The

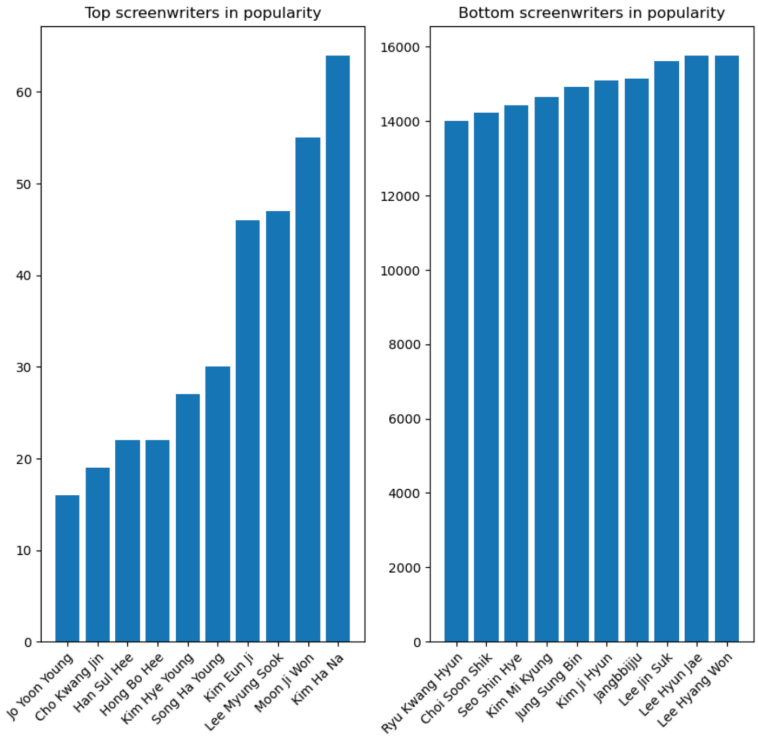
results indicate that most features, including year, total episodes (tot_eps), and rank, are well-balanced between the training and test sets, as their high p-values suggest no significant difference in their distributions. However, duration stands out with a statistically significant p-value (0.008), 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 screenwritters of Korean drama:




Top directors in popularity
Lee Min Woo
Baek Sang Hoon
Oh Choong Hwan
Park Bong Seop
Kim Yoon Jin
Jeong Dae Yun
Shin Won Ho
Jang Ji Yeon
Kim Seong Yoon
Kim Tae Yoon
Bottom directors in popularity
Kim Sang Yoon
Lee Hyo Won
Ryu Kwang Hyun
Han Chul Kyung
Kim Min Joon
Bae Han Chul
Oh Hyun Chang
Lee Kwang Jae
Yoo Young Sun
Lim Kyeong Taek

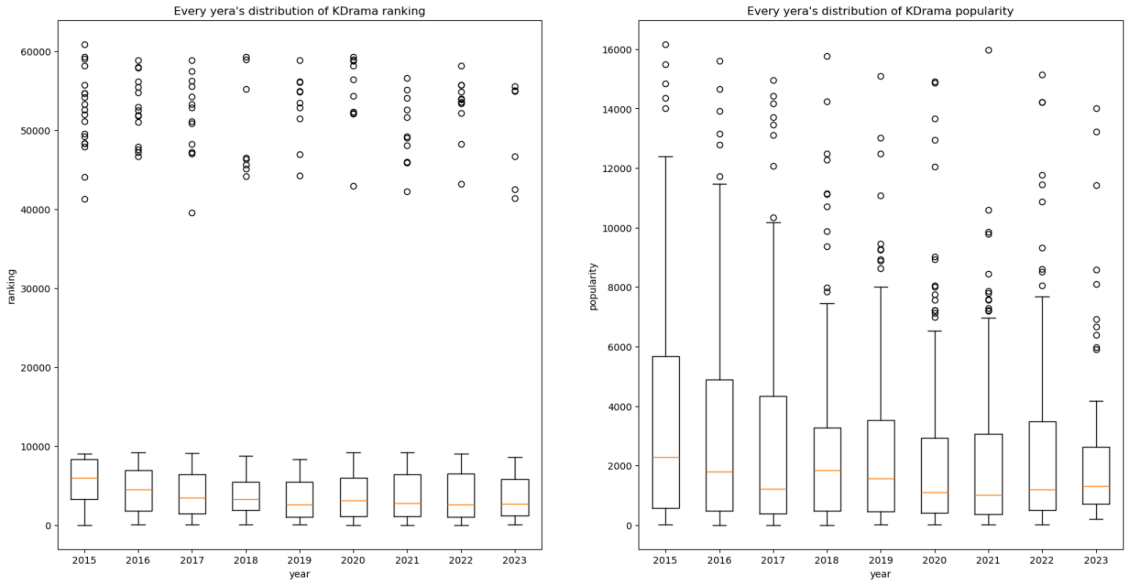

Top screenwriters in ranking
Yoon Ji Ryun
Lee Woo Jung
Park Ba Ra
Choi Ah Il
Moon Ji Won
Kim Yu Jin
Park Ja Kyung
Park Ran
Jung Bo Hoon
Oh Sang Ho
Bottom screenwriters in ranking
Lee Do Hyeon
Kim Mi Kyung
Seo Jung
Lee Min Ji
Kim Ji Hyun
Son Ji Hye
Park Jung Ran
Park Ye Kyung
Lee Hyang Won
Lee Hyun Jae



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
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
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:

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', 'aire
d_on', 'org_net', 'content_rt', 'synopsis']

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

Next, we did box-cox transformation to all numerical variables and did one-hot encoding to all categorical variables. We also tried standardization and normalization to all numerical variables. However, after analyzing the data itself, we think this is unnecessary bacause all the scores are in 10-point system. There is no necessity to normalize/standardize it. However, the number of people who finds the commit helpful and the popularity ranking should be normalized/standardized because we can not easily tell the popularity from the raw data. And the percentage of people who find the commit helpful is better than the raw data of number of people who find helpful. For other numerical variables—Total number of episodes, Each episode duration in second, and rank, we also think this is unnecessary to normalize/standardize them because they all have their unique meaning. And normalize/standardize them makes no sense.

We also think it unnecessary to one-hot encoding all categorical variables. For example, the

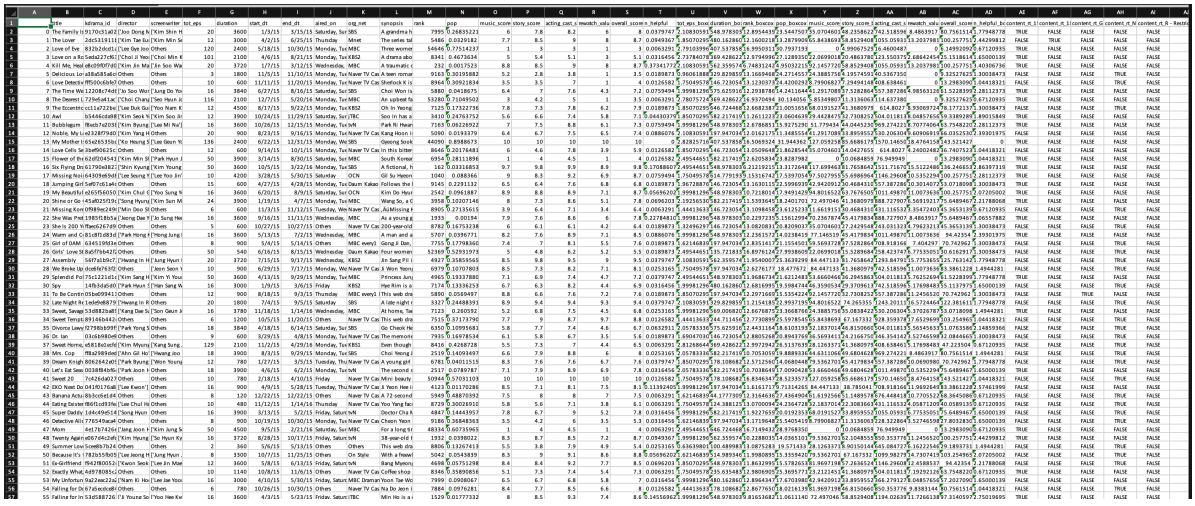
title of the drama does not need one-hot encoding because each drama name is special and one-hot encoding the title will produce 1248 different variables. However, only one variable needs one-hot encoding—Content Rating because it only has a few different types that need to be one-hot encoded. We converted categorical data into numerical values so that we could use machine learning algorithms to it. Thus, since Content Rating included different types of content rating, we one-hot encoded it and converted it to numerical variable so that machine learning algorithms can use it.

Thus, below is the refined version of feature engineering:

```
: 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_value_
      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
  num_cols2 = ['pop','n_helpful']
  # Standardize numerical features
  scaler = StandardScaler()
  Data[num_cols2] = scaler.fit_transform(Data[num_cols2])

  # Normalize (MinMax scaling)
  minmax = MinMaxScaler()
  Data[num_cols2] = minmax.fit_transform(Data[num_cols2])

  # 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)

  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: ['content_rt']

Shape after one-hot encoding: (1248, 34)
```

Below is the final cleaned data after feature engineering:



Key Findings:

After EDA and feature engineering, we found out the following things:

First, since no feature in a dataset has zero variance, it means that every feature contains some degree of variation or difference among its data points, indicating that each feature provides potentially useful information for analysis as no feature is completely constant with only one unique value across all samples.

Second, since the data for every numerical variable in a dataset is skewed, it indicates that the distribution of values for each numerical variable is not symmetrical, meaning there are more data points clustered towards one end of the scale with a "tail" extending towards the other.

Third, most audience gave these Korean drama high score while a few audience gave low score, meaning Korean drama are well made with good feedback from the audience.

Forth, most Korean drama has less than 20 episodes. Some have very long episodes and some have very short episodes. There is a negative relationship between total number of episodes and popularity ranking and a positive relationship between Total number of episodes and popularity ranking. Since higher number in ranking means lower popularity, we can tell that the audience likes longer duration for each episode but they prefer shorter numbers of total episodes.

Fifth, normally, there are less than 20 people who found any commit helpful, meaning most people have their own opinion on the drama. But, in general, if more people share same opinion, then the drama is more popular.

Sixth, there is a positive relationship between popularity ranking and rank, meaning popularity ranking and rank are very close measurement of the drama.

Seventh, 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. We can tell from these that most 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:

The major challenge is missing value and time management. We cannot use imputation for the missing value in director name/screenwriter name, thus the only way is to find other data to replace it. However, we searched IMDb and other drama sources, the only way to get data from IMDb is web scrapping. However, based on the two week deadline, there is not enough time for us to learn web scrapping and finish the project on time.

For the future, if we had enough time, we would learn web scrapping and filling in all missing values. In addition, our study mainly focus on the numerical feature and we did not include actors names or recommendations given to different drama which could be important variables to study. Therefore, if we had enough time, we would include more text data and set them as other variables.

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, histrogram 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 engneering(linear regression visualization and standardize/normalize data, t test, feature importance bar chart, The actual vs. predicted popularity of K-dramas)

Shayan Huda Chowdhury (sc4040): tentatively approach to report drafting, Working on web scrapping. Combining popularity/ranking and revenue over the year, Some feature engineering, Merge datasets, Wrote functions to web scrape missing info, Classified sentiment of reviews using Hugging Face, Combine data