K-Drama Recommendation System from Top 100 K-Dramas

```
In [1]:   #Modules for EDA
          import numpy as np
          import pandas as pd
          import seaborn as sns
          from matplotlib import pyplot as plt
          plt.style.use('fivethirtyeight')

          #Modules for ML(Recommendation)
          from sklearn.preprocessing import MinMaxScaler
          from sklearn.neighbors import NearestNeighbors
          from sklearn.metrics.pairwise import cosine_similarity

          %matplotlib inline
```

```
In [2]:   pwd
```

```
Out[2]:   'C:\\Users\\manis'
```

```
In [7]:   cd D:\documents\snigdha documents
```

D:\documents\snigdha documents

Importing dataset

```
In [8]:   kd = pd.read_csv('top100_kdrama.csv')
```

Understanding the data

```
In [9]: kd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Name              100 non-null    object
 1   Year of release   100 non-null    int64
 2   Aired Date        100 non-null    object
 3   Aired On          100 non-null    object
 4   Number of Episode 100 non-null    int64
 5   Network           100 non-null    object
 6   Duration          100 non-null    object
 7   Content Rating    100 non-null    object
 8   Synopsis          100 non-null    object
 9   Cast              100 non-null    object
 10  Genre             100 non-null    object
 11  Tags              100 non-null    object
 12  Rank              100 non-null    object
 13  Rating            100 non-null    float64
dtypes: float64(1), int64(2), object(11)
memory usage: 11.1+ KB
```

```
In [10]: kd.describe()
```

Out[10]:

	Year of release	Number of Episode	Rating
count	100.000000	100.000000	100.000000
mean	2017.970000	19.070000	8.723000
std	2.869044	12.378096	0.174573
min	2003.000000	6.000000	8.500000
25%	2017.000000	16.000000	8.600000
50%	2019.000000	16.000000	8.700000
75%	2020.000000	20.000000	8.800000
max	2021.000000	100.000000	9.200000

```
In [11]: kd.head()
```

Out[11]:

	Name	Year of release	Aired Date	Aired On	Number of Episode	Network	Duration	Content Rating	Synopsis	Cast	Genre	Tags	Rank	Rating
0	Move to Heaven	2021	May 14, 2021	Friday	10	Netflix	52 min.	18+ Restricted (violence & profanity)	Geu Roo is a young autistic man. He works for ...	Lee Je Hoon, Tang Jun Sang, Hong Seung Hee, Ju...	Life, Drama, Family	Autism, Uncle-Nephew Relationship, Death, Sava...	#1	9.2
1	Hospital Playlist	2020	Mar 12, 2020 - May 28, 2020	Thursday	12	Netflix, tvN	1 hr. 30 min.	15+ - Teens 15 or older	The stories of people going through their days...	Jo Jung Suk, Yoo Yeon Seok, Jung Kyung Ho, Kim...	Friendship, Romance, Life, Medical	Strong Friendship, Multiple Mains, Best Friend...	#2	9.1
2	Flower of Evil	2020	Jul 29, 2020 - Sep 23, 2020	Wednesday, Thursday	16	tvN	1 hr. 10 min.	15+ - Teens 15 or older	Although Baek Hee Sung is hiding a dark secret...	Lee Joon Gi, Moon Chae Won, Jang Hee Jin, Seo ...	Thriller, Romance, Crime, Melodrama	Married Couple, Deception, Suspense, Family Se...	#3	9.1
3	Hospital Playlist 2	2021	Jun 17, 2021 - Sep 16, 2021	Thursday	12	Netflix, tvN	1 hr. 40 min.	15+ - Teens 15 or older	Everyday is extraordinary for five doctors and...	Jo Jung Suk, Yoo Yeon Seok, Jung Kyung Ho, Kim...	Friendship, Romance, Life, Medical	Workplace, Strong Friendship, Best Friends, Mu...	#4	9.1
4	My Mister	2018	Mar 21, 2018 - May 17, 2018	Wednesday, Thursday	16	tvN	1 hr. 17 min.	15+ - Teens 15 or older	Park Dong Hoon is a middle-aged engineer who i...	Lee Sun Kyun, IU, Park Ho San, Song Sae Byuk, ...	Psychological, Life, Drama, Family	Age Gap, Nice Male Lead, Strong Female Lead, H...	#5	9.1

Name as seperate Data Frame

In [12]:
```python
kdrama_names = kd[['Name']]
kdrama_names.head()
```

Out[12]:

	Name
0	Move to Heaven
1	Hospital Playlist
2	Flower of Evil
3	Hospital Playlist 2
4	My Mister

Features using for recommendation

```
In [13]: cols_for_recommend = ['Year of release', 'Number of Episode', 'Network', 'Duration', 'Content Rating', 'Rating']
         kd = kd[cols_for_recommend]
         kd.head()
```

Out[13]:

	Year of release	Number of Episode	Network	Duration	Content Rating	Rating
0	2021	10	Netflix	52 min.	18+ Restricted (violence & profanity)	9.2
1	2020	12	Netflix, tvN	1 hr. 30 min.	15+ - Teens 15 or older	9.1
2	2020	16	tvN	1 hr. 10 min.	15+ - Teens 15 or older	9.1
3	2021	12	Netflix, tvN	1 hr. 40 min.	15+ - Teens 15 or older	9.1
4	2018	16	tvN	1 hr. 17 min.	15+ - Teens 15 or older	9.1

Feature Engineering

Removing duplicate values in Network column

```
In [14]: networks = []
         [networks.append(list(set(network.replace(' ','').split(',')))[0]) for network in kd['Network']]
         networks[:5]
```

Out[14]: ['Netflix', 'tvN', 'tvN', 'tvN', 'tvN']

```
In [16]: kd['Network'] = networks
         kd['Network'].unique()
```

Out[16]: array(['Netflix', 'tvN', 'jTBC', 'KBS2', 'OCN', 'SBS', 'MBC', 'Viki'],
               dtype=object)
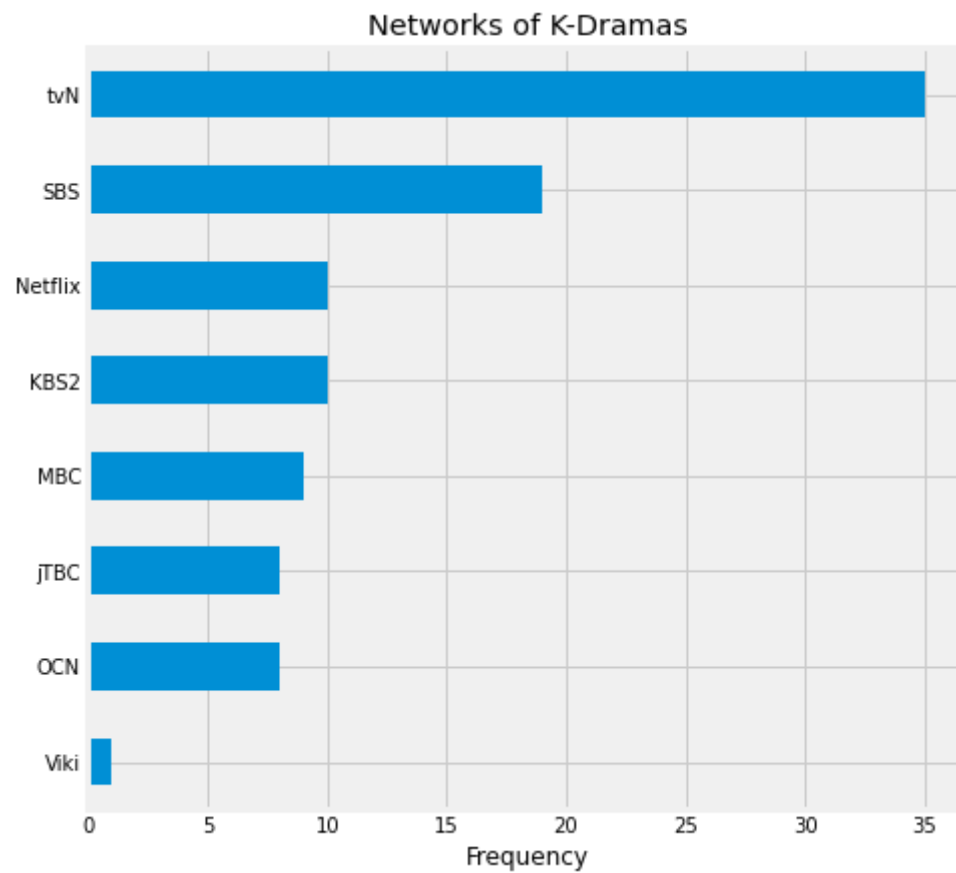
Network and Total K-Dramas

```
In [17]: plt.figure(figsize=(7,7))

         kd['Network'].value_counts().plot(kind='barh')

         plt.gca().invert_yaxis()
         plt.title("Networks of K-Dramas")
         plt.xlabel('Frequency')
         plt.show()

         kd['Network'].value_counts()
```


Networks of K-Dramas

Out[17]: tvN   35

```
SBS         19
Netflix     10
KBS2        10
MBC          9
jTBC         8
OCN          8
Viki         1
Name: Network, dtype: int64
```

In [19]:
```python
kd['Duration'] = kd['Duration'].str.replace('[A-Za-z]\D+','',regex=True)
kd['Duration'].head()
```
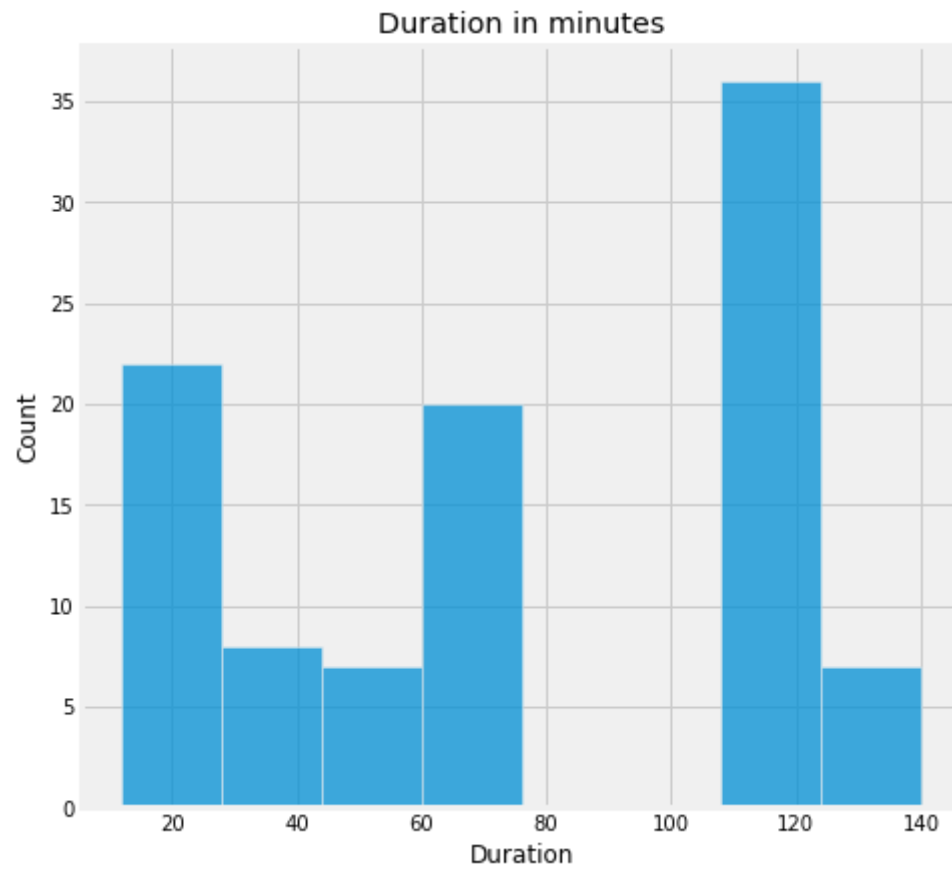
Out[19]:
```
0     52
1    130
2    110
3    140
4    117
Name: Duration, dtype: object
```

In [21]:
```python
kd['Duration'] = kd['Duration'].str.replace(' ','',regex=True)
kd['Duration'] = pd.to_numeric(kd['Duration'])
kd['Duration'].head()
```

Out[21]:
```
0     52
1    130
2    110
3    140
4    117
Name: Duration, dtype: int64
```
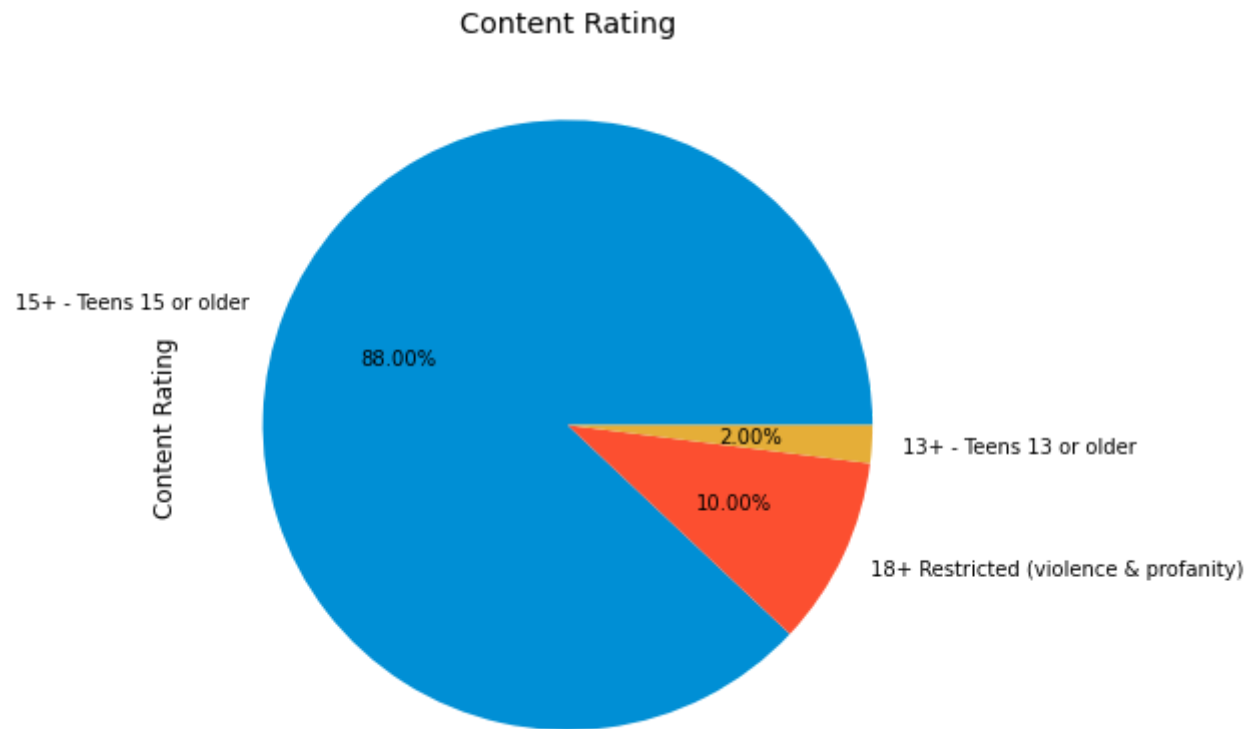
```
In [22]: plt.figure(figsize=(7,7))
         sns.histplot(data=kd['Duration'])
         plt.title('Duration in minutes')
         plt.show()
```


Duration in minutes

Content Rating

```python
plt.figure(figsize=(7,7))
kd['Content Rating'].value_counts().plot(kind='pie',autopct='%.2f%%')
plt.title("Content Rating")
plt.show()
```


Content Rating

```
In [24]: kd['Content Rating'].value_counts()
```
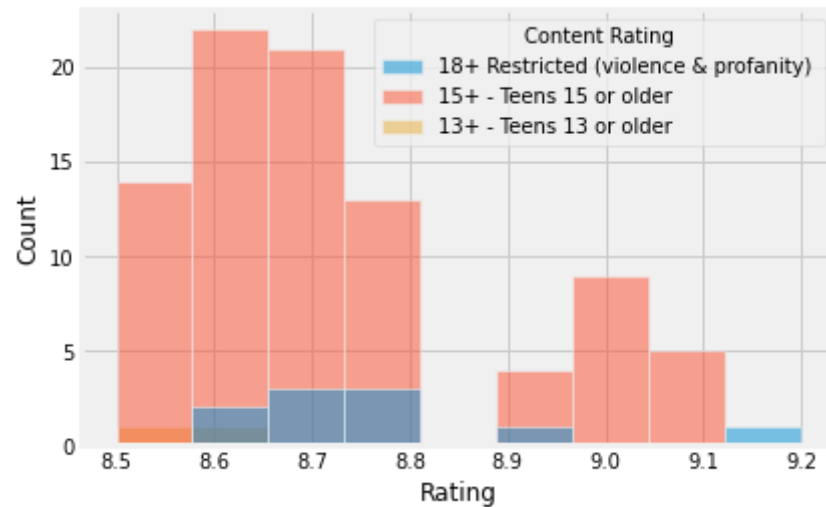
```
Out[24]: 15+ - Teens 15 or older                88
         18+ Restricted (violence & profanity)   10
         13+ - Teens 13 or older                  2
         Name: Content Rating, dtype: int64
```

Rating and Content Rating

```
In [25]: sns.histplot(data=kd[['Rating','Content Rating']],x='Rating',hue='Content Rating')
         plt.show()
```

`kd[['Rating']].describe()`

|	Rating
count	100.000000
mean	8.723000
std	0.174573
min	8.500000
25%	8.600000
50%	8.700000
75%	8.800000
max	9.200000

Number of K-Dramas released in a year

```
In [27]: plt.figure(figsize=(7,7))

         kd['Year of release'].value_counts().plot(kind='barh')

         plt.gca().invert_yaxis()
         plt.title("Number of K-Dramas released per Year")
         plt.xlabel('Frequency')
         plt.show()

         kd['Year of release'].value_counts()
```

Number of K-Dramas released per Year
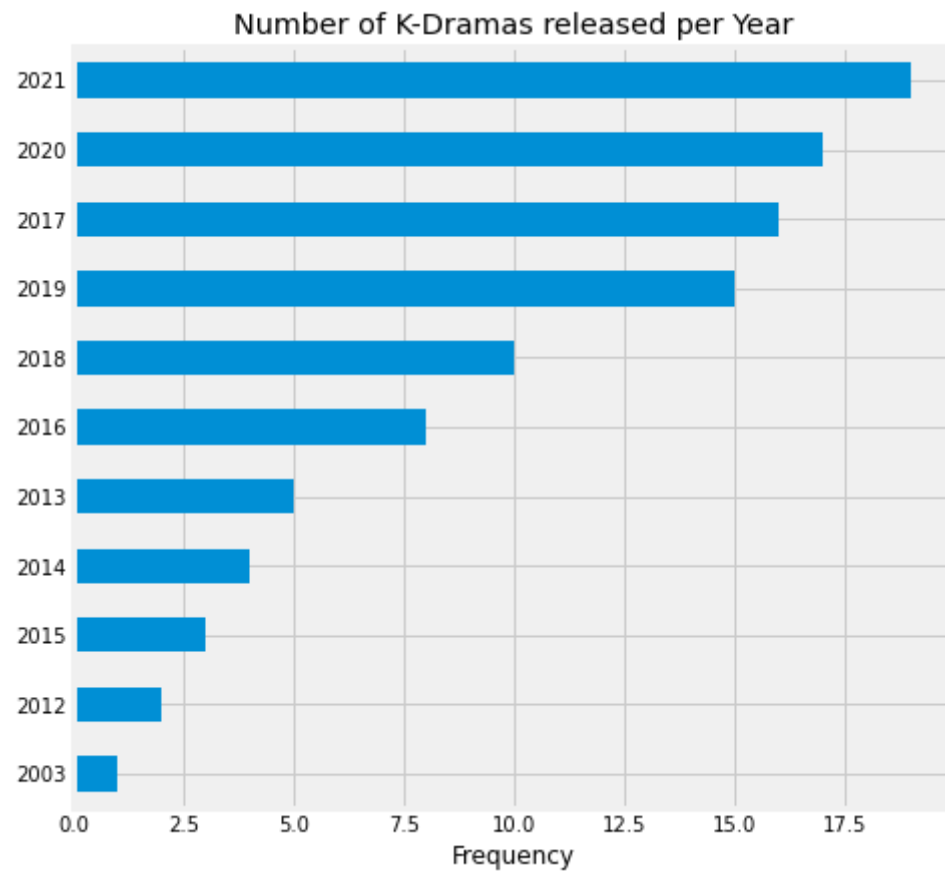
2021   19
        2020   17
        2017   16
        2019   15
        2018   10
        2016   8
        2013   5
        2014   4
        2015   3
        2012   2
        2003   1
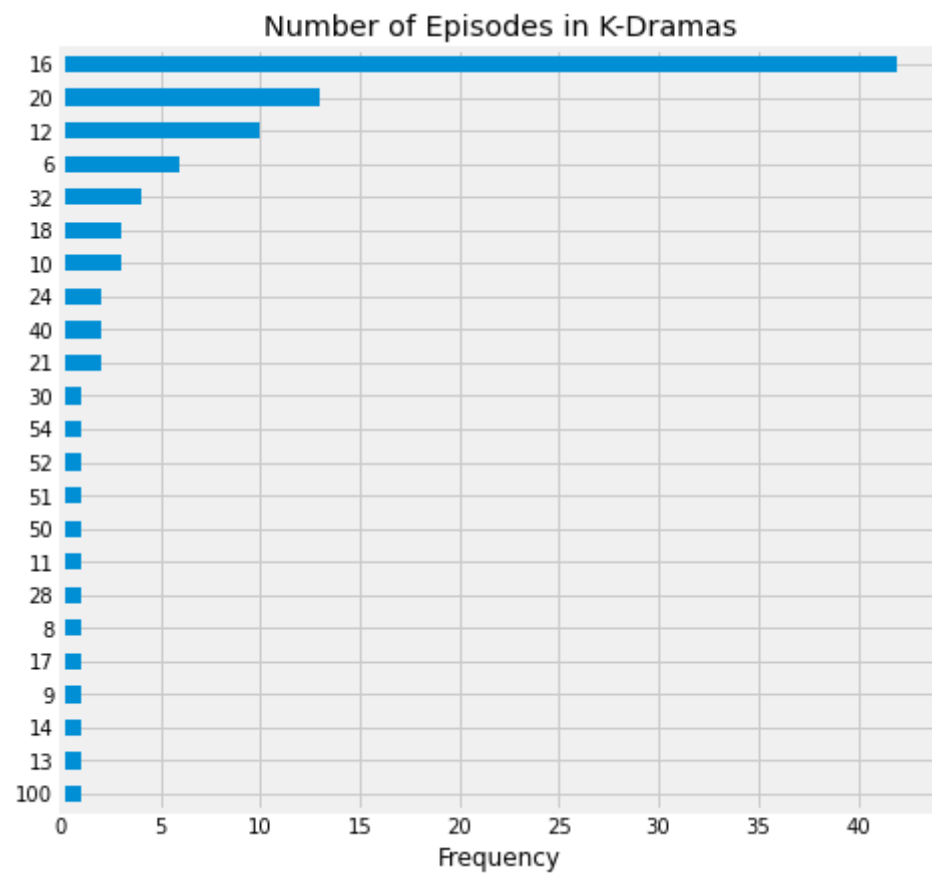   Name: Year of release, dtype: int64

Number of Episodes Distribution

```
In [30]: plt.figure(figsize=(7,7))

         kd['Number of Episode'].value_counts().plot(kind='barh')

         plt.gca().invert_yaxis()
         plt.title("Number of Episodes in K-Dramas")
         plt.xlabel('Frequency')
         plt.show()

         kd['Number of Episode'].value_counts()
```


Number of Episodes in K-Dramas

```
Out[30]:  16      42
          20      13
          12      10
          6        6
          32       4
          18       3
          10       3
          24       2
          40       2
          21       2
          30       1
          54       1
          52       1
          51       1
          50       1
          11       1
          28       1
          8        1
          17       1
          9        1
          14       1
          13       1
          100      1
          Name: Number of Episode, dtype: int64
```

One Hot Encoding

```
In [31]: kd.head()
```

Out[31]:

	Year of release	Number of Episode	Network	Duration	Content Rating	Rating
0	2021	10	Netflix	52	18+ Restricted (violence & profanity)	9.2
1	2020	12	tvN	130	15+ - Teens 15 or older	9.1
2	2020	16	tvN	110	15+ - Teens 15 or older	9.1
3	2021	12	tvN	140	15+ - Teens 15 or older	9.1
4	2018	16	tvN	117	15+ - Teens 15 or older	9.1

```
In [32]: cols_to_encode = ['Network','Content Rating']
         dummies = pd.get_dummies(kd[cols_to_encode],drop_first=True)
         dummies.head()
```

Out[32]:

	Network_MBC	Network_Netflix	Network_OCN	Network_SBS	Network_Viki	Network_jTBC	Network_tvN	Content Rating_15+ - Teens 15 or older	Content Rating_18+ Restricted (violence & profanity)
0	0	1	0	0	0	0	0	0	1
1	0	0	0	0	0	0	1	1	0
2	0	0	0	0	0	0	1	1	0
3	0	0	0	0	0	0	1	1	0
4	0	0	0	0	0	0	1	1	0

```
In [33]: kd.drop(cols_to_encode, axis=1,inplace=True)
         kd.head()
```

Out[33]:

	Year of release	Number of Episode	Duration	Rating
0	2021	10	52	9.2
1	2020	12	130	9.1
2	2020	16	110	9.1
3	2021	12	140	9.1
4	2018	16	117	9.1

Feature Scaling

```
In [34]: scale = MinMaxScaler()
         scalled = scale.fit_transform(kd)
```

```
In [36]: i=0
         for col in kd.columns:
             kd[col] = scalled[:,i]
             i += 1
```

```
In [37]: kd.head()
```

Out[37]:

	Year of release	Number of Episode	Duration	Rating
0	1.000000	0.042553	0.312500	1.000000
1	0.944444	0.063830	0.921875	0.857143
2	0.944444	0.106383	0.765625	0.857143
3	1.000000	0.063830	1.000000	0.857143
4	0.833333	0.106383	0.820312	0.857143

```
In [38]: new_kd = pd.concat([kd, dummies],axis=1)
         new_kd.head()
```

Out[38]:

	Year of release	Number of Episode	Duration	Rating	Network_MBC	Network_Netflix	Network_OCN	Network_SBS	Network_Viki	Network_jTBC	Network_tvN
0	1.000000	0.042553	0.312500	1.000000	0	1	0	0	0	0	0
1	0.944444	0.063830	0.921875	0.857143	0	0	0	0	0	0	1
2	0.944444	0.106383	0.765625	0.857143	0	0	0	0	0	0	1
3	1.000000	0.063830	1.000000	0.857143	0	0	0	0	0	0	1
4	0.833333	0.106383	0.820312	0.857143	0	0	0	0	0	0	1

```
In [41]: synopsis = pd.read_csv('top100_kdrama.csv',usecols=['Synopsis'])
         synopsis.head()
```

Out[41]:

	Synopsis
0	Geu Roo is a young autistic man. He works for ...
1	The stories of people going through their days...
2	Although Baek Hee Sung is hiding a dark secret...
3	Everyday is extraordinary for five doctors and...
4	Park Dong Hoon is a middle-aged engineer who i...

```
In [42]: kdrama_names['Name'].loc[23]='kingdom'
         new_kd.index = [drama for drama in kdrama_names['Name']]
         synopsis.index = [drama for drama in kdrama_names['Name']]
         new_kd.head()
```

Out[42]:

	Year of release	Number of Episode	Duration	Rating	Network_MBC	Network_Netflix	Network_OCN	Network_SBS	Network_Viki	Network_jTBC	Networl
Move to Heaven	1.000000	0.042553	0.312500	1.000000	0	1	0	0	0	0	
Hospital Playlist	0.944444	0.063830	0.921875	0.857143	0	0	0	0	0	0	
Flower of Evil	0.944444	0.106383	0.765625	0.857143	0	0	0	0	0	0	
Hospital Playlist 2	1.000000	0.063830	1.000000	0.857143	0	0	0	0	0	0	
My Mister	0.833333	0.106383	0.820312	0.857143	0	0	0	0	0	0	

Recommendation System

```
In [43]: def getRecommendation_dramas_for(drama_name,no_of_recommend=5,get_similarity_rate=False):

             kn = NearestNeighbors(n_neighbors=no_of_recommend+1,metric='manhattan')
             kn.fit(new_kd)

             distances, indices = kn.kneighbors(new_kd.loc[drama_name])

             print(f'Similar K-Dramas for "{drama_name[0]}":')
             nearest_dramas = [kdrama_names.loc[i][0] for i in indices.flatten()][1:]
             if not get_similarity_rate:
                 return nearest_dramas
             sim_rates = []
             synopsis_ = []
             for drama in nearest_dramas:
                 synopsis_.append(synopsis.loc[drama][0])
                 sim = cosine_similarity(new_kd.loc[drama_name],[new_kd.loc[drama]]).flatten()
                 sim_rates.append(sim[0])
             recommended_dramas = pd.DataFrame({'Recommended Drama':nearest_dramas,'Similarity':sim_rates,'Synopsis':synopsis_})
             recommended_dramas.sort_values(by='Similarity',ascending=True)
             return recommended_dramas
```

```
In [44]: def print_similiar_drama_Synopsis(recommended_kd):
             rkd = recommended_kd
             rkd_cols = rkd['Synopsis']
             dramas = rkd['Recommended Drama']
             for i in range(5):
                 print(dramas[i])
                 print(rkd_cols[i])
                 print('\n')
```

Predicting Drama Recommendation

```
In [45]: rd1 = kdrama_names.loc[0]
         rd1
```

```
Out[45]: Name     Move to Heaven
         Name: 0, dtype: object
```

```
In [46]: getRecommendation_dramas_for(rd1,no_of_recommend=5)
```

Similar K-Dramas for "Move to Heaven":

```
Out[46]: ['Kingdom', 'Kingdom', 'My Name', 'Sweet Home', 'Squid Game']
```

```
In [47]: rd2 = kdrama_names.loc[10]
         rd2
```

```
Out[47]: Name    Signal
         Name: 10, dtype: object
```

```
In [48]: getRecommendation_dramas_for(rd2,get_similarity_rate=True)
```

Similar K-Dramas for "Signal":

Out[48]:

	Recommended Drama	Similarity	Synopsis
0	It's Okay to Not Be Okay	0.994766	Moon Gang Tae is a community health worker at ...
1	Stranger	0.996784	Hwang Shi Mok underwent brain surgery as a chi...
2	Crash Landing on You	0.996966	After getting into a paragliding accident, Sou...
3	My Mister	0.997236	Park Dong Hoon is a middle-aged engineer who i...
4	Reply 1988	0.995079	Five childhood friends, who all live in the sa...

```
In [49]: rd3 = kdrama_names.loc[1]
         rd3
```

```
Out[49]: Name    Hospital Playlist
         Name: 1, dtype: object
```

```
In [54]: getRecommendation_dramas_for(rd3,no_of_recommend=10,get_similarity_rate=True)
```

Similar K-Dramas for "Hospital Playlist":

Out[54]:

	Recommended Drama	Similarity	Synopsis
0	Hospital Playlist 2	0.999395	Everyday is extraordinary for five doctors and...
1	Flower of Evil	0.997420	Although Baek Hee Sung is hiding a dark secret...
2	Prison Playbook	0.996988	Kim Je Hyuk, a famous baseball player, is arre...
3	My Mister	0.998064	Park Dong Hoon is a middle-aged engineer who i...
4	Crash Landing on You	0.997901	After getting into a paragliding accident, Sou...
5	It's Okay to Not Be Okay	0.996995	Moon Gang Tae is a community health worker at ...
6	Mr. Queen	0.996879	Jang Bong Hwan is a South Korean chef who has ...
7	Vincenzo	0.996514	At the age of eight, Park Joo Hyeong went to I...
8	Signal	0.994214	Fifteen years ago, a young girl was kidnapped ...
9	Reply 1988	0.988641	Five childhood friends, who all live in the sa...

```
In [51]: rd4 = kdrama_names.loc[8]
         rd4
```

Out[51]: Name   Mr. Queen
         Name: 8, dtype: object

```
In [52]: rdf4 = getRecommendation_dramas_for(rd4,no_of_recommend=10,get_similarity_rate=True)
         print_similiar_drama_Synopsis(rdf4)
```

Similar K-Dramas for "Mr. Queen":
It's Okay to Not Be Okay
Moon Gang Tae is a community health worker at a psychiatric ward who was blessed with everything including a great body, smarts, ability to sympathize with others, patience, ability to react quickly, stamina, and more. Meanwhile, Ko Moon Young is a popular writer of children's literature who, due to suffering from an antisocial personality disorder, seems extremely selfish, arrogant, and rude.

Vincenzo
At the age of eight, Park Joo Hyeong went to Italy after being adopted. Now an adult, he is known as Vincenzo Cassano to the Mafia, who employ him as a consigliere. Because mafia factions are at war with each other, he flees to South Korea, where he gets involved with Lawyer Hong Cha Young. She is the type of attorney who will do anything to win a case. Now back at his motherland, he gives an unrivaled conglomerate a taste of its own medicine with a side of justice.

Crash Landing on You
After getting into a paragliding accident, South Korean heiress Yoon Se Ri crash lands in North Korea. There, she meets North Korean army officer Ri Jung Hyuk, who agrees to help her return to South Korea. Despite the tension between their countries, the two of them start falling for one another.

Flower of Evil
Although Baek Hee Sung is hiding a dark secret surrounding his true identity, he has established a happy family life and a successful career. He is a loving husband and doting father to his young daughter. But his perfect façade begins to crumble when his wife, Cha Ji Won, a homicide detective, begins investigating a string of serial murders from 15 years ago. Ji Won notices changes in Hee Sung's behavior and begins to wonder if he could possibly be hiding something from her.

Mr. Sunshine
Mr. Sunshine centers on a young boy born into a house servant's family and travels to the United States during the 1871 Shinmiyangyo (U.S. expedition to Korea). He returns to his homeland later as a U.S. marine officer. He meets and falls in love with an aristocrat's daughter. At the same time, he discovers a plot by foreign forces to colonize Korea. Edit Translation English 한국어 中文(简体) Русский

```
In [61]: rd5 = kdrama_names.loc[99]
         rd5
```

Out[61]: Name   Fight For My Way
         Name: 99, dtype: object

```
In [62]: rdf5=getRecommendation_dramas_for(rd5,no_of_recommend=5,get_similarity_rate=True)
         print_similiar_drama_Synopsis(rdf5)
```

Similar K-Dramas for "Fight For My Way":
Good Manager
Can corporate politics turn a bad person into a good person? Kim Sung Ryong is a skilled accountant who works for ga
ngsters. He makes his way into a company called TQ Group as a middle manager in the accounting department with the i
ntention of embezzling money from the company. As he contends with hot-shot accountant Yoon Ha Kyung, financial prod
igy Seo Yul and peppy intern Hong Ga Eun, Sung Ryong finds himself embroiled in office politics and notices other sh
enanigans by those in power that threaten to bring down the company. Sung Ryong slowly comes to realize that he'd ra
ther fight against corporate corruption and fight for employees' rights than complete the mission he originally came
to do. Can he succeed at his new objective against all odds?

Descendants of the Sun
A love story that develops between a surgeon and a special forces officer.

Dali and the Cocky Prince
As a young boy, Moo Hak grew up in the market as a peddler. Although he is ignorant, he is strong-willed and knows h
ow to make money. He is now the managing director of Dondon F and B, a global restaurant chain that his family start
ed as a small gamjatang diner. However, he isn't dignified and only cares about money. He meets Da Li, the only daug
hter and child of an upper-class family who runs an art gallery. She has a profound deep knowledge of things. She ge
ts faced with bankruptcy due to her father's sudden death and begins a hard life. Moo Hak and Da Li start a relation
ship as a creditor and debtor over the art gallery. Will these two polar opposites understand each other and fall in
love?

Go Back Couple
38-years-old married couple, Choi Ban Do and Ma Jin Joo, were in love when they got married but now, hate each other
and regret marrying at such a young age. Choi Ban Do has been burdened with being the breadwinner, and Ma Jin Joo is
a housewife with low self-esteem. The couple travels through time and finds themselves as 20-year-old university stu
dents when they met for the first time.

Healer
Seo Jung Hoo is a special kind of night courier, known only as "Healer" by his clients. For the right price and with
the help of a genius hacker, he gets his clients whatever they want, as long as it doesn't involve murder. His lates
t job leads him to a second-rate tabloid writer, Chae Young Shin, and the successful reporter, Kim Moon Ho. He begin
s to uncover the mystery of his own shared past with the two reporters, thus putting them all in danger. Edit Transl
ation English Русский Ελληνικά Italiano