data_understanding_describe_data

August 11, 2025

1 Data Understanding

1.1 Describe Data

1.1.1 Anime Recommendations Database

[1]:
```
import pandas as pd
```

[3]:
```
cuAnime = pd.read_csv("E:\\applied data science␣
↪capstone\\data\\CooperUnion\\archive\\anime.csv")
```

Anime Listing

[4]:
```
cuAnime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12294 entries, 0 to 12293
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   anime_id  12294 non-null  int64
 1   name      12294 non-null  object
 2   genre     12232 non-null  object
 3   type      12269 non-null  object
 4   episodes  12294 non-null  object
 5   rating    12064 non-null  float64
 6   members   12294 non-null  int64
dtypes: float64(1), int64(2), object(4)
memory usage: 672.5+ KB
```

The size of the anime listing data is 915 KB
There are 12,294 records and 7 fields.
Episodes is the number of episodes an anime has and is a ratio
Rating is the overall rating of the anime and is ordinal.
Members is the number of users that have added it to their list of anime and is also a ratio.
Genre is the type of story and is categorical.
Type is the format of the production of the anime and is categorical.
Name is the title of the anime and is nominal
Anime_id is the unique identifier provided by myanimelist.net and is nominal

[5]:
```
cuAnime.describe()
```

[5]:

	anime_id	rating	members
count	12294.000000	12064.000000	1.229400e+04
mean	14058.221653	6.473902	1.807134e+04
std	11455.294701	1.026746	5.482068e+04
min	1.000000	1.670000	5.000000e+00
25%	3484.250000	5.880000	2.250000e+02
50%	10260.500000	6.570000	1.550000e+03
75%	24794.500000	7.180000	9.437000e+03
max	34527.000000	10.000000	1.013917e+06

[6]:
```
cuAnime.describe(include=object)
```

[6]:

	name	genre	type	episodes
count	12294	12232	12269	12294
unique	12292	3264	6	187
top	Saru Kani Gassen	Hentai	TV	1
freq	2	823	3787	5677

Anime ratings for each user

[7]:
```
cuRating = pd.read_csv("E:\\applied data science␣
↪capstone\\data\\CooperUnion\\archive\\rating.csv")
```

[8]:
```
cuRating.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7813737 entries, 0 to 7813736
Data columns (total 3 columns):
 #   Column    Dtype
---  ------    -----
 0   user_id   int64
 1   anime_id  int64
 2   rating    int64
dtypes: int64(3)
memory usage: 178.8 MB
```

[9]:
```
cuRating.describe()
```

[9]:

	user_id	anime_id	rating
count	7.813737e+06	7.813737e+06	7.813737e+06
mean	3.672796e+04	8.909072e+03	6.144030e+00
std	2.099795e+04	8.883950e+03	3.727800e+00
min	1.000000e+00	1.000000e+00	-1.000000e+00
25%	1.897400e+04	1.240000e+03	6.000000e+00
50%	3.679100e+04	6.213000e+03	7.000000e+00
75%	5.475700e+04	1.409300e+04	9.000000e+00
max	7.351600e+04	3.451900e+04	1.000000e+01

The size of the anime listing data is 108,794 KB
There are 7,813,737 records and 3 fields.
User_id is the unique identifier for the user that provided the rating and it is nominal.
Rating is the rating given to the anime by the user and it is ordinal.
Anime_id is the unique identifier for the anime and it is nominal.

Rating from the anime rating for each user and the rating from the anime listing are not the same. The rating for the entire anime is determined by combining the rating supplied by each user to obtain the overall rating.

1.1.2 Anime Recommendations Database 2020

Anime Listing

[10]:
```
hernanAnime = pd.read_csv("E:\\applied data science␣
↪capstone\\data\\hernan4444\\archive\\anime-hernan.csv")
```

[11]:
```
hernanAnime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17562 entries, 0 to 17561
Data columns (total 35 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   MAL_ID         17562 non-null  int64
 1   Name           17562 non-null  object
 2   Score          17562 non-null  object
 3   Genres         17562 non-null  object
 4   English name   17562 non-null  object
 5   Japanese name  17562 non-null  object
 6   Type           17562 non-null  object
 7   Episodes       17562 non-null  object
 8   Aired          17562 non-null  object
 9   Premiered      17562 non-null  object
 10  Producers      17562 non-null  object
 11  Licensors      17562 non-null  object
 12  Studios        17562 non-null  object
 13  Source         17562 non-null  object
 14  Duration       17562 non-null  object
 15  Rating         17562 non-null  object
 16  Ranked         17562 non-null  object
 17  Popularity     17562 non-null  int64
 18  Members        17562 non-null  int64
 19  Favorites      17562 non-null  int64
 20  Watching       17562 non-null  int64
 21  Completed      17562 non-null  int64
 22  On-Hold        17562 non-null  int64
 23  Dropped        17562 non-null  int64
 24  Plan to Watch  17562 non-null  int64
 25  Score-10       17562 non-null  object
```

```
 26  Score-9        17562 non-null  object
 27  Score-8        17562 non-null  object
 28  Score-7        17562 non-null  object
 29  Score-6        17562 non-null  object
 30  Score-5        17562 non-null  object
 31  Score-4        17562 non-null  object
 32  Score-3        17562 non-null  object
 33  Score-2        17562 non-null  object
 34  Score-1        17562 non-null  object
dtypes: int64(9), object(26)
memory usage: 4.7+ MB
```

The size of the anime listing data is 5,530 KB
There are 17,562 records and 35 fields.
MAL_ID is the unique identifier provided by myanimelist.net and is nominal
Name is the title of the anime and is nominal
Score is the overall rating of the anime and is ordinal. **Genres** is the type of story and is categorical.
English name is the title of the anime in english and is nominal
Japanese name is the title of the anime in japanese and is nominal
Type is the format of the production of the anime and is categorical.
Episodes is the number of episodes an anime has and is a ratio
Aired is the time period during which the anime was released and it is an interval
Premiered is the date the first episode was released it is an interval
Producers companies that finance the production of the anime and is categorical
Licensors companies that have the right to distribute the anime outside of Japan and is categorical
Studios companies that make the anime and is categorical
Source tells the source material that was used to create the anime and is categorical
Duration tells the duration of the anime and is categorical
Rating is the content suitability and is categorical
Ranked is the position of the anime on the rankings for the website. It is based on the average user rating and is ordinal
Popularity indicates how much an anime is interacted with by counting the number of users that have added it to their anime list and is ordinal
Members indicates the number of users that have added the anime to their anime list and is a ratio
Favorites indicates the number of users that have added the anime to their favourites list and is a ratio
Watching indicates the number of users that have indicated that they are currently watching the anime and is a ratio
Completed indicates the number of users that have indicated that they have watched all episodes for the anime and is a ratio
On-Hold indicates the number of users that have indicated that they started watching the anime but are not currently and is a ratio
Dropped indicates the number of users that have indicated that they started watching the anime but are no longer and is a ratio
Plan to Watch indicates the number of users that have indicated that they are not currently watching the anime, but intend to and is a ratio
Score-10 indicates the number of users that gave the anime a score of 10 and is a ratio

Score-9 indicates the number of users that gave the anime a score of 9 and is a ratio
Score-8 indicates the number of users that gave the anime a score of 8 and is a ratio
Score-7 indicates the number of users that gave the anime a score of 7 and is a ratio
Score-6 indicates the number of users that gave the anime a score of 6 and is a ratio
Score-5 indicates the number of users that gave the anime a score of 5 and is a ratio
Score-4 indicates the number of users that gave the anime a score of 4 and is a ratio
Score-3 indicates the number of users that gave the anime a score of 3 and is a ratio
Score-2 indicates the number of users that gave the anime a score of 2 and is a ratio
Score-1 indicates the number of users that gave the anime a score of 1 and is a ratio

Name, English name and *Japanese name* are all titles for the anime and are closely related.
Premiered can be determined from the *aired* field as *premiered* is simply the start date for the *aired* field.
There may be overlap with *producers, licensors* and *studios* as many companies serve multiple roles in the production and distribution process.
Ranked and *popularity* are both popularity measures derived from user interaction. *Ranked* is in fact directly linked to the *score* of the anime as anime with higher scores generally place higher on the rankings.
Members, favorites, watching, completed, on-hold, dropped and *plan to watch* are all metrics used to gauge user interaction with the anime.
Score-n for n = 1 to 10 are all combined to generate the *score* field for the anime.

[12]:
```
hernanAnime.describe()
```

[12]:

	MAL_ID	Popularity	Members	Favorites	Watching
count	17562.000000	17562.000000	1.756200e+04	17562.000000	17562.000000
mean	21477.192347	8763.452340	3.465854e+04	457.746270	2231.487758
std	14900.093170	5059.327278	1.252821e+05	4063.473313	14046.688133
min	1.000000	0.000000	1.000000e+00	0.000000	0.000000
25%	5953.500000	4383.500000	3.360000e+02	0.000000	13.000000
50%	22820.000000	8762.500000	2.065000e+03	3.000000	73.000000
75%	35624.750000	13145.000000	1.322325e+04	31.000000	522.000000
max	48492.000000	17565.000000	2.589552e+06	183914.000000	887333.000000

	Completed	On-Hold	Dropped	Plan to Watch
count	1.756200e+04	17562.000000	17562.000000	17562.000000
mean	2.209557e+04	955.049653	1176.599533	8199.831227
std	9.100919e+04	4275.675096	4740.348653	23777.691963
min	0.000000e+00	0.000000	0.000000	1.000000
25%	1.110000e+02	6.000000	37.000000	112.000000
50%	8.175000e+02	45.000000	77.000000	752.500000
75%	6.478000e+03	291.750000	271.000000	4135.500000
max	2.182587e+06	187919.000000	174710.000000	425531.000000

[13]:
```
hernanAnime[['Name', 'Score', 'Genres', 'English name', 'Japanese name',
       'Type','Episodes', 'Aired', 'Premiered', 'Producers', 'Licensors',
       'Studios', 'Source']].describe(include=object)
```

[13]:
```
                                                    Name    Score  Genres  \
count                                              17562    17562   17562
unique                                             17558      533    5034
top     Maou Gakuin no Futekigousha: Shijou Saikyou no…  Unknown  Hentai
freq                                                   3     5141     969

       English name Japanese name   Type Episodes    Aired Premiered  \
count         17562         17562  17562    17562    17562     17562
unique         6831         16679      7      201    11947       231
top         Unknown       Unknown     TV        1  Unknown   Unknown
freq          10565            48   4996     8381      309     12817

       Producers Licensors  Studios    Source
count      17562     17562    17562     17562
unique      3783       231     1090        16
top      Unknown   Unknown  Unknown  Original
freq        7794     13616     7079      5215
```

[14]:
```
hernanAnime[['Duration', 'Rating', 'Ranked', 'Popularity',
       'Members', 'Favorites', 'Watching', 'Completed', 'On-Hold', 'Dropped',
       'Plan to Watch', 'Score-10', 'Score-9', 'Score-8', 'Score-7', 'Score-6',
       'Score-5', 'Score-4', 'Score-3', 'Score-2', 'Score-1']].
 ↪describe(include=object)
```

[14]:
```
               Duration                     Rating   Ranked Score-10  Score-9  \
count             17562                      17562    17562    17562    17562
unique              313                          7    10490     3379     3645
top     24 min. per ep.  PG-13 - Teens 13 or older  Unknown      4.0  Unknown
freq               1723                       6132     1762      936     3167

        Score-8 Score-7  Score-6  Score-5  Score-4  Score-3  Score-2 Score-1
count     17562   17562    17562    17562    17562    17562    17562   17562
unique     4515    4933     4236     3288     2235     1506     1110    1084
top     Unknown     2.0  Unknown  Unknown  Unknown  Unknown  Unknown     4.0
freq       1371     610      511      584      977     1307     1597     955
```

Anime ratings for each user

[15]:
```
hernanRatings = pd.read_csv("E:\\applied data science␣
 ↪capstone\\data\\hernan4444\\archive\\rating_complete.csv")
```

[16]:
```
hernanRatings.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 57633278 entries, 0 to 57633277
Data columns (total 3 columns):
 #   Column    Dtype
---  ------    -----
 0   user_id   int64
```

```
 1   anime_id  int64
 2   rating    int64
dtypes: int64(3)
memory usage: 1.3 GB
```

The size of the anime listing data is 798,729 KB
There are 57,633,278 records and 3 fields.
User_id is the unique identifier for the user that provided the rating and it is nominal.
Rating is the rating given to the anime by the user and it is ordinal.
Anime_id is the unique identifier for the anime and it is nominal.

[17]:
```
hernanRatings.describe()
```

[17]:

	user_id	anime_id	rating
count	5.763328e+07	5.763328e+07	5.763328e+07
mean	1.768878e+05	1.583147e+04	7.510789e+00
std	1.020117e+05	1.326114e+04	1.697722e+00
min	0.000000e+00	1.000000e+00	1.000000e+00
25%	8.827800e+04	3.091000e+03	7.000000e+00
50%	1.772910e+05	1.188700e+04	8.000000e+00
75%	2.654190e+05	2.899900e+04	9.000000e+00
max	3.534040e+05	4.845600e+04	1.000000e+01

Once more *rating from the anime rating for each user and the rating from the anime listing are not the same. The rating for the entire anime is determined by combining the rating supplied by each user to obtain the overall rating.*

Synopsis Data

[18]:
```
synopsisDf = pd.read_csv("E:\\applied data science␣
 ↪capstone\\data\\hernan4444\\archive\\anime_with_synopsis.csv")
```

[19]:
```
synopsisDf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16214 entries, 0 to 16213
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   MAL_ID     16214 non-null  int64
 1   Name       16214 non-null  object
 2   Score      16214 non-null  object
 3   Genres     16214 non-null  object
 4   sypnopsis  16206 non-null  object
dtypes: int64(1), object(4)
memory usage: 633.5+ KB
```

The size of the anime listing data is 7,053 KB
There are 16,214 records and 5 fields.
MAL_ID is the unique identifier provided by myanimelist.net and is nominal
Name is the title of the anime and is nominal

Score is the overall rating of the anime and is ordinal.
Genres is the type of story and is categorical.
Synopsis is a brief description of the plot of the anime and is nominal.

[20]:
```
synopsisDf.describe()
```

[20]:
```
              MAL_ID
count   16214.000000
mean    22069.271555
std     14849.798248
min         1.000000
25%      6728.500000
50%     24164.000000
75%     35978.750000
max     48492.000000
```

[21]:
```
synopsisDf.describe(include=object)
```

[21]:
```
                                                    Name    Score Genres  \
count                                              16214    16214  16214
unique                                             16210      532   4857
top     Maou Gakuin no Futekigousha: Shijou Saikyou no…  Unknown  Music
freq                                                   3     5123    790

                                                 sypnopsis
count                                                16206
unique                                               15221
top     No synopsis information has been added to this…
freq                                                   709
```

1.1.3 MyAnimeList Comment Dataset V2

[22]:
```
natleeAnime = pd.read_csv("E:\\applied data science␣
 ↪capstone\\data\\natlee\\archive\\anime_list.csv")
```

Anime Listing

[23]:
```
natleeAnime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24594 entries, 0 to 24593
Data columns (total 28 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              24594 non-null  int64
 1   workId          24594 non-null  int64
 2   url             24594 non-null  object
 3   jpName          24504 non-null  object
 4   engName         10181 non-null  object
```

```
 5   synonymsName    12821 non-null  object
 6   workType        21988 non-null  object
 7   episodes        24594 non-null  object
 8   status          24594 non-null  object
 9   aired           24594 non-null  object
 10  premiered       5498 non-null   object
 11  producer        24594 non-null  object
 12  broadcast       7576 non-null   object
 13  licensors       24594 non-null  object
 14  studios         24594 non-null  object
 15  genres          19795 non-null  object
 16  themes          13612 non-null  object
 17  demographic     9392 non-null   object
 18  source          24594 non-null  object
 19  duration        24594 non-null  object
 20  rating          24013 non-null  object
 21  score           15647 non-null  float64
 22  allRank         20073 non-null  object
 23  popularityRank  24594 non-null  object
 24  members         24594 non-null  object
 25  favorites       24594 non-null  object
 26  scoredByUser    15647 non-null  float64
 27  lastUpdate      24594 non-null  object
dtypes: float64(2), int64(2), object(24)
memory usage: 5.3+ MB
```

id is just an identifier for the dataset
workId is the myanimelist.net id for the anime
jpName, *engName*, *synonymsName* all provide titles for the anime.
Status, aired and *premiered* are all related as *status* and *premiered* can be derived from the *aired* field.
Once more, *producers, licensors* and *studios* tend to have some overlap due to companies playing multiple roles in the anime production process.
Genres, themes and *demographic* can all be related as the *demographic* tends to dictate which *genre* and *themes* are apparent in an anime. *Genre* of course is a more broader category and *themes* tend to be a bit more specific.
Score, allRank, popularityRank, members, favorites, scoredByUser are all related as they are used to determine the popularity of the anime.

The size of the anime listing data is 8,296 KB
There are 24,594 records and 28 fields.
Id is the unique identifier provided by creator of the dataset and is nominal
WorkId is the unique identifier provided by myanimelist.net and is nominal
Url is the url that brings you to the description page for the anime on myanimelist.net
JpName is the title of the anime in japanese and is nominal
EngName is the title of the anime in english and is nominal
SynonymsName is another title for the anime and is nominal
WorkType is the format of the production of the anime and is categorical.
Episodes is the number of episodes an anime has and is a ratio

Status is the release state of the anime, whether ongoing or not and is categorical **Aired** is the time period during which the anime was released and it is an interval
Premiered is the date the first episode was released it is an interval
Producer are the companies that finance the production of the anime and is categorical
Broadcast is the scheduled airing time for the anime and is an interval
Licensors are the companies that have the right to distribute the anime outside of Japan and is categorical
Studios are the companies that make the anime and is categorical
Genres is the type of story and is categorical.
Themes are the more specific ideas or situations that appear in the anime and is categorical
Demographic refer to the type of audience the anime is geared towards and is categorical
Source tells the source material that was used to create the anime and is categorical
Duration tells the duration of the anime and is categorical
Rating is the content suitability and is categorical
Score is the overall rating of the anime and is ordinal. **AllRank** is the position of the anime on the rankings for the website. It is based on the average user rating and is ordinal
PopularityRank indicates how much an anime is interacted with by counting the number of users that have added it to their anime list and is ordinal
Members indicates the number of users that have added the anime to their anime list and is a ratio
Favorites indicates the number of users that have added the anime to their favourites list and is a ratio
ScoredByUser is the average score given to the anime by the users and is a ratio
LastUpdate is the date and time of the last update made to the anime and is interval

[24]:
```
natleeAnime.describe()
```

[24]:

	id	workId	score	scoredByUser
count	24594.000000	24594.000000	15647.000000	1.564700e+04
mean	12297.500000	29456.666057	6.382660	2.992858e+04
std	7099.820596	17860.097090	0.928238	1.167218e+05
min	1.000000	1.000000	1.840000	1.010000e+02
25%	6149.250000	10335.250000	5.730000	3.835000e+02
50%	12297.500000	34330.000000	6.390000	1.768000e+03
75%	18445.750000	44896.750000	7.060000	1.084700e+04
max	24594.000000	55566.000000	9.100000	2.654325e+06

[25]:
```
natleeAnime[['url', 'jpName', 'engName', 'synonymsName', 'workType',
       'episodes', 'status', 'aired', 'premiered', 'producer']].
 ↪describe(include=object)
```

[25]:

	url	jpName	\
count	24594	24504	
unique	24594	23530	
top	https://myanimelist.net/anime/47917/Bocchi_the…		
freq	1	8	

	engName	synonymsName	workType	episodes	status	\

```
count              10181         12821    21988    24594            24594
unique              9992         12128        5      252                3
top     Spirit Guardians  Minna no Uta       TV        1  Finished Airing
freq                   5           386     7576    11314            23810

                aired    premiered  producer
count           24594         5498     24594
unique          15095          243      4402
top     Not available  Spring 2017  add some
freq              887           88     13075
```

[26]:
```
natleeAnime[['broadcast',
       'licensors', 'studios', 'genres', 'themes', 'demographic', 'source',
       'duration', 'rating', 'allRank', 'popularityRank', 'members',
       'favorites', 'lastUpdate']].describe(include=object)
```

[26]:
```
       broadcast licensors   studios  genres themes demographic    source  \
count       7576     24594     24594   19795  13612        9392     24594
unique       591       265      1547    1000    831           5        17
top      Unknown  add some  add some  Comedy  Music        Kids  Original
freq        4197     19863     10279    2266   2674        5846      9446

               duration                     rating allRank popularityRank  \
count             24594                      24013   20073          24594
unique              328                          6   18499          18253
top     24 min. per ep.  PG-13 - Teens 13 or older  #10290         #18698
freq               1957                       8372       3              6

       members favorites           lastUpdate
count    24594     24594                24594
unique   10990      1787                24593
top         38         0  2023-06-02 11:06:59
freq       165     10546                    2
```

Anime Reviews The file was corrupted and could not be opened or loaded into a dataframe. Had to obtain new reviews data by polling the unofficial myanimelist.net api (https://api.jikan.moe/v4)