Monash University: Assessment Cover Sheet

Student name	Sivakumar	Harisriguhan	
School/Campus	CLAYTON	**Student's I.D. number**	31386709
Unit name	FIT5147 Data exploration and visualisation - S2 2021		
Lecturer's name	SARAH GOODWIN	**Tutor's name**	ANGEL DAS
Assignment name	Data Exploration Project Submission (33%)	**Group Assignment: No** **Note, each student must attach a coversheet**	
Lab/Tute Class: TUTORIAL 4	**Lab/Tute Time:** 12:00-14:00		**Word Count:** 1741
Due date: 09-09-2021	**Submit Date:** 08-09-2021		**Extension granted** ☐

If an extension of work is granted, specify date and provide the signature of the lecturer/tutor. Alternatively, attach an email printout or handwritten and signed notice from your lecturer/tutor verifying an extension has been granted.

Extension granted until (date): ...../...../..... Signature of lecturer/tutor: .....

Late submissions policy	**Days late**	**Penalty applied**
Penalties apply to late submissions and may vary between faculties. Please refer to your faculty's late assessment policy for details.		

Patient/client confidentiality: Where a patient/client case study is undertaken a signed Consent Form must be obtained.

Intentional plagiarism or collusion amounts to cheating under Part 7 of the Monash University (Council) Regulations

Plagiarism: Plagiarism means to take and use another person's ideas and or manner of expressing them and to pass these off as one's own by failing to give appropriate acknowledgement. This includes material from any source, staff, students or the Internet - published and unpublished works.

Collusion: Collusion means unauthorised collaboration on assessable written, oral or practical work with another person. Where there are reasonable grounds for believing that intentional plagiarism or collusion has occurred, this will be reported to the Associate Dean (Education) or nominee, who may disallow the work concerned by prohibiting assessment or refer the matter to the Faculty Discipline Panel for a hearing.

Student Statement:

- I have read the university's Student Academic Integrity Policy and Procedures
- I understand the consequences of engaging in plagiarism and collusion as described in Part 7 of the Monash University (Council) Regulations (academic misconduct).
- I have taken proper care to safeguard this work and made all reasonable efforts to ensure it could not be copied.
- No part of this assignment has been previously submitted as part of another unit/course.
- I acknowledge and agree that the assessor of this assignment may, for the purposes of assessment, reproduce the assignment and:

  i. provide it to another member of faculty and any external marker; and/or

  ii. submit to a text matching/originality checking software; and/or

  iii. submit it to a text matching/originality checking software which may then retain a copy of the assignment on its database for the purpose of future plagiarism checking.
- I certify that I have not plagiarised the work of others or participated in unauthorised collaboration or otherwise breached the academic integrity requirements in the Student Academic Integrity Policy.

Date: 08/09/2021 Signature: Hari *

Privacy Statement:

For information about how the University deals with your personal information go to http://privacy.monash.edu.au/guidelines/collection-personal-information.html#enrol

FIT5147 – Data Exploration and Visualization

Data Exploration Project – Report

Topic: The Influence of Anime

Done by

Harisriguhan Sivakumar

Student ID: 31386709

Course: Master of Data Science

Introduction

In the past few years, Anime has become a phenomenon across the world [1]. Anime (animation made in Japan) has started to influence people worldwide. Being influenced by it myself, I wanted to dive into the market share of Anime. By doing so, we will come to know how Anime has grown over the years to the point where it has become a global sensation. In addition, we will be discovering the factors that make an anime series successful.

Data Wrangling

We will start of our adventure by getting our datasets from Kaggle named “MyAnimeList Dataset” [2] which is a collection of metadata and information about various Anime as well as its viewers. For this project we will be taking the raw version. we will take 3 datasets [2] (“UserList.csv”,” UserAnimeList.csv” and “AnimeList.csv”) which contains the data about the Viewers, Anime seen by the Viewers and Anime series respectively. The total size of the 3 datasets combined is 4.69 GB! So it is necessary to perform data wrangling or else we may not be able to visualize our data due to the issues caused by its Volume (For Instance, slow performance).

We start of our wrangling with the elephant in the room “UserAnimeList.csv” which is a massive 4.6 GB file that has all the data about the anime that each user has seen. So it is no surprise that it is so large. For wrangling we will be using R over tableau for Quality of life purposes. When we open the dataset we can see that there are 80 million records with 11 attributes.

```
UAL <- read_csv("UserAnimeList.csv")
head(UAL)
```

R Console | tbl_df 6 x 11

A tibble: 6 x 11

username <chr>	anime_id <dbl>	my_watched_episodes <dbl>	my_start_date <date>	my_finish_date <date>	my_score <dbl>	my_status <dbl>
karthiga	21	586	<NA>	<NA>	9	1
karthiga	59	26	<NA>	<NA>	7	2
karthiga	74	26	<NA>	<NA>	7	2
karthiga	120	26	<NA>	<NA>	7	2
karthiga	178	26	<NA>	<NA>	7	2
karthiga	210	161	<NA>	<NA>	7	2

6 rows | 1-7 of 11 columns

Figure 1: A glimpse at our initial Data frame

We remove the unwanted columns as it removes a significant portion of the dataset. The only columns that are required by us is username, anime_id, my_watched_episodes, my_score, my_status.

```
```{r}
names(UAL)
```

[1] "username"   "anime_id"   "my_watched_episodes" "my_start_date"
[5] "my_finish_date"   "my_score"   "my_status"   "my_rewatching"
[9] "my_rewatching_ep"   "my_last_updated"   "my_tags"
```

Figure 2: The attributes/columns of UserAnimeList.csv

We then remove the other attributes from the data frame as a result we will only have the attributes we require.

```r
head(UAL)
```

A tibble: 6 x 5

username <chr>	anime_id <dbl>	my_watched_episodes <dbl>	my_score <dbl>	my_status <dbl>
karthiga	21	586	9	1
karthiga	59	26	7	2
karthiga	74	26	7	2
karthiga	120	26	7	2
karthiga	178	26	7	2
karthiga	210	161	7	2

6 rows

Figure 3: The Wrangled UserAnimeList.csv

Before proceeding, we must look into what the values in my_status mean. The image below gives us an idea of it.

my_status in animelists tables contains integer values. This is their semantics:

- 1: watching
- 2: completed
- 3: on hold
- 4: dropped
- 6: plan to watch

other values are not known.

Figure 4: Legend for my_status column in UserAnimeList.csv

The data that is relevant to our analysis are with status 1,2,3 or 4. We do not have to consider the records with status 6 as planning to see does not guarantee them seeing it. We will remove records that are not of the status 1,2,3 or 4.

```r
unique(UAL$my_status)
```

```
[1]  1  2  3  6  4  0  5 33 55
```

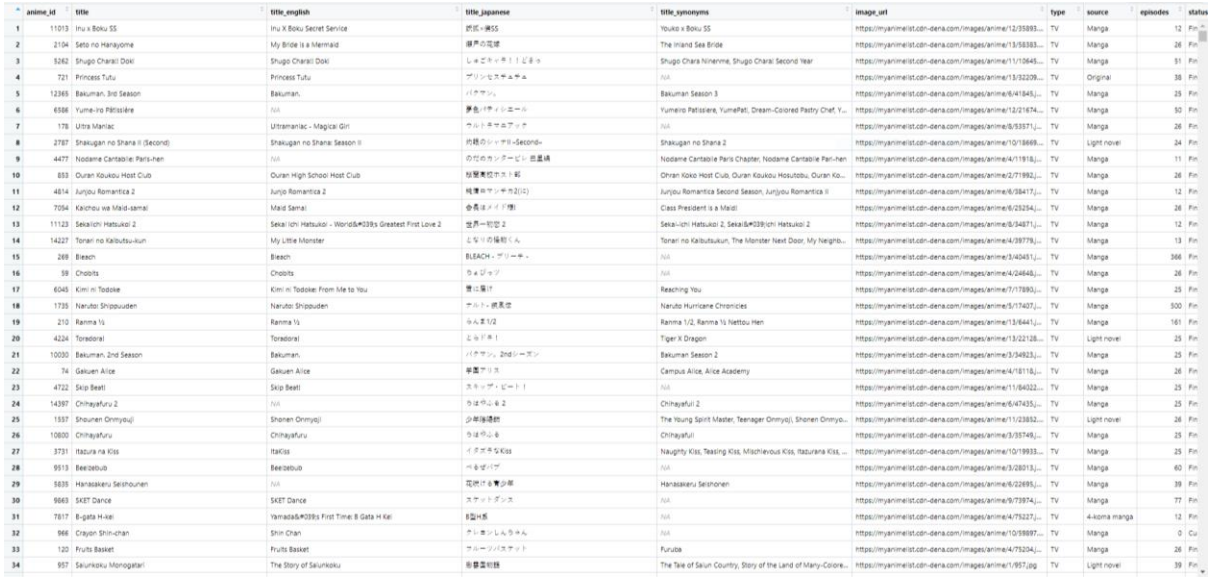
```r
UAL <- UAL[UAL$my_status %in% c(1,2,3,4),]
```

```r
unique(UAL$my_status)
```

```
[1] 1 2 3 4
```

Figure 5: Filtering based on my_status

Now we will move on to the next data set “AnimeList.csv”. Looking at the dataset there seem we only require a few attributes only. As we did with the previous data set we will remove them.



	anime_id	title	title_english	title_japanese	title_synonyms	image_url	type	source	episodes	status
1	11013	Inu x Boku SS	Inu X Boku Secret Service	妖狐×僕SS	Youko x Boku SS	https://myanimelist.cdn-dena.com/images/anime/12/35893...	TV	Manga	12	Fin
2	2104	Seto no Hanayome	My Bride is a Mermaid	瀬戸の花嫁	The Inland Sea Bride	https://myanimelist.cdn-dena.com/images/anime/13/58383...	TV	Manga	26	Fin
3	5262	Shugo Chara!! Doki	Shugo Chara!! Doki	しゅごキャラ!!どきっ	Shugo Chara Ninenme, Shugo Chara! Second Year	https://myanimelist.cdn-dena.com/images/anime/11/10645...	TV	Manga	51	Fin
4	721	Princess Tutu	Princess Tutu	プリンセスチュチュ	NA	https://myanimelist.cdn-dena.com/images/anime/13/32209...	TV	Original	38	Fin
5	12365	Bakuman. 3rd Season	Bakuman.	バクマン。	Bakuman Season 3	https://myanimelist.cdn-dena.com/images/anime/6/41845...	TV	Manga	25	Fin
6	6586	Yume-iro Pâtissière	NA	夢色パティシエール	Yumeiro Patissiere, YumePati, Dream-Colored Pastry Chef, Y...	https://myanimelist.cdn-dena.com/images/anime/12/21674...	TV	Manga	50	Fin
7	178	Ultra Maniac	Ultramaniac - Magical Girl	ウルトラマニアック	NA	https://myanimelist.cdn-dena.com/images/anime/8/53571...	TV	Manga	26	Fin
8	2787	Shakugan no Shana II (Second)	Shakugan no Shana: Season II	灼眼のシャナII-Second-	Shakugan no Shana 2	https://myanimelist.cdn-dena.com/images/anime/10/18669...	TV	Light novel	24	Fin
9	4477	Nodame Cantabile: Paris-hen	NA	のだめカンタービレ 巴里編	Nodame Cantabile Paris Chapter, Nodame Cantabile Pari-hen	https://myanimelist.cdn-dena.com/images/anime/4/1191&...	TV	Manga	11	Fin
10	853	Ouran Koukou Host Club	Ouran High School Host Club	桜蘭高校ホスト部	Ohran Koko Host Club, Ouran Koukou Hosutobu, Ouran Ko...	https://myanimelist.cdn-dena.com/images/anime/2/71992...	TV	Manga	26	Fin
11	4814	Junjou Romantica 2	Junjo Romantica 2	純情ロマンチカ2(仮)	Junjou Romantica Second Season, Junjyou Romantica II	https://myanimelist.cdn-dena.com/images/anime/6/38417...	TV	Manga	12	Fin
12	7054	Kaichou wa Maid-sama!	Maid Sama!	会長はメイド様!	Class President is a Maid!	https://myanimelist.cdn-dena.com/images/anime/6/25254...	TV	Manga	26	Fin
13	11123	Sekaiichi Hatsukoi 2	Sekai-ichi Hatsukoi - World's Greatest First Love 2	世界一初恋 2	Sekai-ichi Hatsukoi 2, Sekai'ichi Hatsukoi 2	https://myanimelist.cdn-dena.com/images/anime/8/34871...	TV	Manga	12	Fin
14	14227	Tonari no Kaibutsu-kun	My Little Monster	となりの怪物くん	Tonari no Kaibutsukun, The Monster Next Door, My Neighb...	https://myanimelist.cdn-dena.com/images/anime/4/39779...	TV	Manga	13	Fin
15	269	Bleach	Bleach	BLEACH - ブリーチ -	NA	https://myanimelist.cdn-dena.com/images/anime/3/40451...	TV	Manga	366	Fin
16	59	Chobits	Chobits	ちょびっツ	NA	https://myanimelist.cdn-dena.com/images/anime/4/24846...	TV	Manga	26	Fin
17	6045	Kimi ni Todoke	Kimi ni Todoke: From Me to You	君に届け	Reaching You	https://myanimelist.cdn-dena.com/images/anime/7/17880...	TV	Manga	25	Fin
18	1735	Naruto: Shippuuden	Naruto: Shippuden	ナルト- 疾風伝	Naruto Hurricane Chronicles	https://myanimelist.cdn-dena.com/images/anime/5/17407...	TV	Manga	500	Fin
19	210	Ranma ½	Ranma ½	らんま1/2	Ranma 1/2, Ranma ½ Nettou Hen	https://myanimelist.cdn-dena.com/images/anime/13/6441...	TV	Manga	161	Fin
20	4224	Toradora!	Toradora!	とらドラ!	Tiger X Dragon	https://myanimelist.cdn-dena.com/images/anime/13/22128...	TV	Light novel	25	Fin
21	10030	Bakuman. 2nd Season	Bakuman.	バクマン。2nd シーズン	Bakuman Season 2	https://myanimelist.cdn-dena.com/images/anime/3/34923...	TV	Manga	25	Fin
22	74	Gakuen Alice	Gakuen Alice	学園アリス	Campus Alice, Alice Academy	https://myanimelist.cdn-dena.com/images/anime/4/18116...	TV	Manga	26	Fin
23	4722	Skip Beat!	Skip Beat!	スキップ・ビート!	NA	https://myanimelist.cdn-dena.com/images/anime/11/64022...	TV	Manga	25	Fin
24	14397	Chihayafuru 2	NA	ちはやふる 2	Chihayafuri 2	https://myanimelist.cdn-dena.com/images/anime/6/47435...	TV	Manga	25	Fin
25	1557	Shounen Onmyouji	Shonen Onmyoji	少年陰陽師	The Young Spirit Master, Teenager Onmyoji, Shonen Onmyo...	https://myanimelist.cdn-dena.com/images/anime/11/23852...	TV	Light novel	26	Fin
26	10800	Chihayafuru	Chihayafuru	ちはやふる	Chihayafuri	https://myanimelist.cdn-dena.com/images/anime/3/35749...	TV	Manga	25	Fin
27	3731	Itazura na Kiss	ItaKiss	イタズラなKiss	Naughty Kiss, Teasing Kiss, Mischievous Kiss, Itazurana Kiss, ...	https://myanimelist.cdn-dena.com/images/anime/10/19933...	TV	Manga	25	Fin
28	9513	Beelzebub	Beelzebub	べるぜバブ	NA	https://myanimelist.cdn-dena.com/images/anime/3/28013...	TV	Manga	60	Fin
29	5835	Hanasakeru Seishounen	NA	花咲ける青少年	Hanasakeru Seishonen	https://myanimelist.cdn-dena.com/images/anime/6/22695...	TV	Manga	39	Fin
30	9663	SKET Dance	SKET Dance	スケットダンス	NA	https://myanimelist.cdn-dena.com/images/anime/9/73974...	TV	Manga	77	Fin
31	7817	B-gata H-kei	Yamada's First Time: B Gata H Kei	B型H系	NA	https://myanimelist.cdn-dena.com/images/anime/4/75227...	TV	4-koma manga	12	Fin
32	966	Crayon Shin-chan	Shin Chan	クレヨンしんちゃん	NA	https://myanimelist.cdn-dena.com/images/anime/10/59897...	TV	Manga	0	Cu
33	120	Fruits Basket	Fruits Basket	フルーツバスケット	Furuba	https://myanimelist.cdn-dena.com/images/anime/4/75204...	TV	Manga	26	Fin
34	957	Saiunkoku Monogatari	The Story of Saiunkoku	彩雲国物語	The Tale of Saiun Country, Story of the Land of Many-Colore...	https://myanimelist.cdn-dena.com/images/anime/1/957.jpg	TV	Light novel	39	Fin

Figure 6: Overview of AnimeList.csv

```
{r}
names(anime)
```

```
 [1] "anime_id"      "title"         "title_english"  "title_japanese"  "title_synonyms"  "image_url"
 [7] "type"          "source"        "episodes"       "status"          "airing"          "aired_string"
[13] "aired"         "duration"      "rating"         "score"           "scored_by"       "rank"
[19] "popularity"    "members"       "favorites"      "background"      "premiered"       "broadcast"
[25] "related"       "producer"      "licensor"       "studio"          "genre"           "opening_theme"
[31] "ending_theme"
```

Figure 7: Columns in AnimeList.csv before wrangling

```
{r}
colnames(anime)
```

```
[1] "anime_id"      "title"         "title_english" "type"          "source"        "episodes"
[7] "status"        "duration"      "rating"        "score"         "studio"        "genre"
```

Figure 8: Columns in AnimeList.csv after Wrangling

We will now go to the last dataset “UserList.csv”. As usual, we will we will see the attributes present in it.

username	user_id	user_watching	user_completed	user_onhold	user_dropped	user_plantowatch	user_days_spent_watching	gender	location	birth_date	access_rank	join_date	last_online
karthiga	2255153	3	49	1	0	0	55.31	Female	Chennai, India	1990-04-29	NA	2013-03-03	2014-02-04 01:32:00
RedvelvetDaisuki	1897606	61	396	39	0	206	118.07	Female	Manila	1995-01-01	NA	2012-12-13	1900-05-13 02:47:00
Damonashu	37326	45	195	27	25	59	83.70	Male	Detroit,Michigan	1991-08-01	NA	2008-02-13	1900-03-24 12:48:00
bskai	228342	25	414	2	5	11	167.16	Male	Nayarit, Mexico	1990-12-14	NA	2009-08-31	2014-05-12 16:35:00
shuzzable	2347781	36	72	16	2	25	35.48	NA	NA	<NA>	NA	2013-03-25	2015-09-09 21:54:00
terune_uzumaki	327311	5	5	0	0	0	15.20	Female	Malaysia, Kuantan	1998-08-24	NA	2010-05-10	2012-10-18 19:08:00
Bas_C	5015094	35	114	6	20	175	30.81	Male	Nijmegen, Nederland	1999-10-24	NA	2015-11-26	1900-05-10 13:53:00
punane	39250	1	0	0	0	0	2.97	NA	NA	<NA>	NA	2008-02-24	2011-02-13 02:42:00
-otaku-sama-	4583478	0	4	0	0	1	6.42	NA	NA	<NA>	NA	2015-04-27	2015-05-04 19:05:00
sprite1989	102436	2583	102	0	0	22	30.86	Male	NA	1989-07-29	NA	2008-10-06	2017-10-30 05:26:00
thetreedude	28328	11	703	27	29	79	206.68	Male	California	<NA>	NA	2007-12-29	2017-09-10 19:04:00
Heihouka	12489	8	69	0	8	0	112.73	Male	NA	<NA>	NA	2007-09-06	2015-08-30 04:36:00
HimeAria	3129315	2	87	2	0	28	20.12	Female	Poland	1996-09-26	NA	2013-09-08	1900-04-27 10:52:00
Skallington	326733	0	3	0	0	0	4.32	Male	Norberg Sweden	1991-03-15	NA	2010-05-09	2010-05-12 06:29:00
Slimak	61677	79	224	0	3	84	126.17	Male	Poland	1988-02-21	NA	2008-05-18	1900-05-01 05:04:00
Elysiun	5927342	7	4	0	0	4	2.09	Male	Nebraska	1984-11-16	NA	2016-12-25	2017-03-11 21:37:00
jimothy000	47167	6	0	0	0	2	7.71	Male	England,Manchester	1995-04-25	NA	2008-04-01	2008-04-27 16:00:00
MistButterfly	2485327	66	3923	115	0	368	614.96	Female	~ Hungary ~	1992-01-16	NA	2013-04-25	2018-05-17 13:31:00
BadAngel	15865	3	20	0	0	5	21.34	Male	NA	1991-06-14	NA	2007-10-03	2017-05-31 15:52:00
Not_Rijad	4980716	1	99	5	2	14	52.78	NA	Where even am I	<NA>	NA	2015-11-07	1900-05-10 17:59:00
ProperBritish	253613	6	233	7	9	62	44.63	Male	United Kingdom	1992-02-20	NA	2009-11-01	1900-04-19 03:47:00
Cranberry_cakes	32590	0	4	1	0	1	4.77	NA	NA	<NA>	NA	2008-01-20	2008-01-24 22:23:00
yen_otaku	211766	15	90	13	0	37	73.65	Female	manila	<NA>	NA	2009-07-22	2010-05-12 02:00:00
kioniel	144049	11	433	5	2	0	166.24	Male	home	1988-01-16	NA	2009-01-29	2017-05-31 05:57:00
iLLMaTiCc	2282031	47	80	10	14	151	61.82	Female	NA	<NA>	NA	2013-03-09	1900-01-07 20:41:00
Ximil	1	4	230	8	89	60	133.32	Male	California	1985-03-04	NA	2004-11-05	2018-05-14 18:02:55
xchatechasexx	118593	7	31	18	0	29	23.25	Female	socal =)	<NA>	NA	2008-11-28	2011-09-06 12:53:00
helenply	2529849	9	184	6	8	39	41.15	Female	Mexico	1999-08-30	NA	2013-05-06	1900-05-13 15:25:00
cfoordddd	61291	5	392	31	0	288	151.61	NA	Sydney Australia N.S.W	<NA>	NA	2008-05-17	1900-04-18 00:58:00
hood_rad	4538087	1	314	0	1	160	80.22	Male	Hamilton Ontario	1997-10-31	NA	2015-04-04	1900-04-21 19:44:00
sweetangie	452735	3	432	7	28	10	134.48	Female	Nevada USA	<NA>	NA	2011-02-23	1900-05-13 02:35:00
detestedlove311	5309	3	29	5	2	20	24.58	Female	Chicago, IL	1989-09-14	NA	2007-06-06	2009-11-16 16:48:00
ihasabucket	18867	6	146	15	25	15	63.02	Male	San Jose, CA	1984-09-08	NA	2007-10-26	2015-06-28 01:17:00
YunYun_Pachina	270169	27	270	16	78	240	66.97	Female	NA	<NA>	NA	2009-12-15	1900-05-11 18:17:00
morbid3500	189574	4	37	0	5	4	11.82	Female	NA	<NA>	NA	2009-06-02	2014-11-10 16:00:00
xTheFallenx	340873	22	655	2	3	212	131.89	Male	somewhere really hot	1993-05-29	NA	2010-06-13	1900-05-16 03:20:00
L-LaxlieIDN	14658	5	0	0	1	0	3.65	Male	Cuess	1995-05-29	NA	2007-09-23	2008-04-24 16:01:00
zZKinhoZz	6951406	10	65	3	8	22	73.30	Male	Camaragibe, Pernambuco	2000-04-04	NA	2018-02-03	2018-05-17 10:41:00
xXxToxic94xXx	98148	0	0	0	0	0	0.00	NA	NA	<NA>	NA	2008-09-24	2008-09-24 16:29:00
SaberWolfKaiNi	5844149	21	183	46	18	190	44.45	NA	NA	<NA>	NA	2016-11-17	2018-05-16 19:20:20
glittermilk	502599	15	35	2	6	70	17.20	Female	austria<U+3002>	1994-04-13	NA	2011-06-23	2012-09-16 06:47:00
asksam	60375	2	386	1	14	12	156.31	Male	NA	<NA>	NA	2008-05-14	2017-08-03 06:14:00
misumimi	844589	11	116	16	1	101	21.77	NA	NA	<NA>	NA	2011-10-30	2017-08-20 06:54:00
fluffylonewolf	1537661	6	121	11	12	175	27.63	Female	Canada	1989-01-01	NA	2012-09-24	2016-06-22 13:18:00

Figure 9:An overview of UserList.csv

```r
colnames(users)
```

```
 [1] "username"            "user_id"             "user_watching"          "user_completed"
 [5] "user_onhold"         "user_dropped"        "user_plantowatch"       "user_days_spent_watching"
 [9] "gender"              "location"            "birth_date"             "access_rank"
[13] "join_date"           "last_online"         "stats_mean_score"       "stats_rewatched"
[17] "stats_episodes"
```

Figure 10: The attributes of UserList.csv before wrangling

Out all the columns we need only need 3 columns. We will remove the other attributes as well as records with no location (location = NA) from the dataset.

```r
users <- users[is.na(users$location) == F,]
users
```

A tibble: 156,796 x 3

username <chr>	gender <chr>	location <chr>
karthiga	Female	Chennai, India
RedvelvetDaisuki	Female	Manila
Damonashu	Male	Detroit,Michigan
bskai	Male	Nayarit, Mexico
terune_uzumaki	Female	Malaysia, Kuantan
Bas_G	Male	Nijmegen, Nederland
thetreedude	Male	California
HimeAria	Female	Poland
Skallington	Male	Norberg Sweden

Figure 11: UserList.csv after filtering

After we perform our Data Checking which will be explained in the next section we will merge all the tables which will form our final dataset.

Data Checking

All the data sets are skimmed through in order to spot and rectify any anomalies. Let us start with UserAnimeList.csv. We can see a few records where the value of my_watched_episodes is 0. This does not make any sense. So we will remove them and the my_watched_episodes column as it has served its purpose.

```r
UAL[UAL$my_watched_episodes > 0,c("username","anime_id","my_score")]
```

A tibble: 57,934,010 x 3

username <chr>	anime_id <dbl>	my_score <dbl>
karthiga	21	9
karthiga	59	7
karthiga	74	7
karthiga	120	7
karthiga	178	7
karthiga	210	7
karthiga	232	6

Figure 12: The clean version of UserAnimeList.csv

We save the clean version of this file as a new csv called UserAnimeListFiltered.csv. Now we will check the next dataset AnimeList.csv here we can see 2 main errors. The episode duration is in the form “XX mins per episode” which isn’t suitable for visualisation. This will cause issues when we need to visualize based on the episode duration. So we create a user defined function called extract_mins that will return only the integer value. We will also change the column name to Duration_in_Minutes.

```
extract_mins <- function(duration){
  duration = strsplit(duration," ")[[1]][1]
  return (as.integer(duration))
}
```
```
```{r,warning=FALSE}
anime["Duration_in_Minutes"] <- sapply(anime$Duration_in_Minutes, extract_mins)
anime
```
```

A tibble: 14,478 x 11

status <chr>	Duration_in_Minutes <int>	rating <chr>
Finished Airing	24	PG-13 - Teens 13 or older
Finished Airing	24	PG-13 - Teens 13 or older
Finished Airing	24	PG - Children
Finished Airing	16	PG-13 - Teens 13 or older
Finished Airing	24	PG-13 - Teens 13 or older
Finished Airing	24	G - All Ages
Finished Airing	24	G - All Ages
Finished Airing	24	PG-13 - Teens 13 or older
Finished Airing	23	PG-13 - Teens 13 or older
Finished Airing	23	PG-13 - Teens 13 or older

Figure 13: Type conversion of Episode Duration

Next is the genre which is multivalued will be really hard for us to visualize based on genre. We need to turn all the genres to boolean attributes assign the appropriate TRUE/FALSE value to it.

```
genre_tweak <- function(row,animecolnames){
  genres = c(strsplit(row[match("genre",animecolnames)],", ")[[1]])
  for (i in genres){
    row[match(i,animecolnames)] <- T
  }
  return(row)
}
```
```
```{r}
anime <- as.data.frame(t(apply(anime,1,genre_tweak,animecolnames = colnames(anime))))
anime
```
```

Description: df [14,478 x 55]

Comedy <chr>	Supernatural <chr>	Romance <chr>	Shounen <chr>	Parody <chr>	School <chr>	Magic <chr>	Shoujo <chr>
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE

Figure 14: Assigning appropriate Boolean values

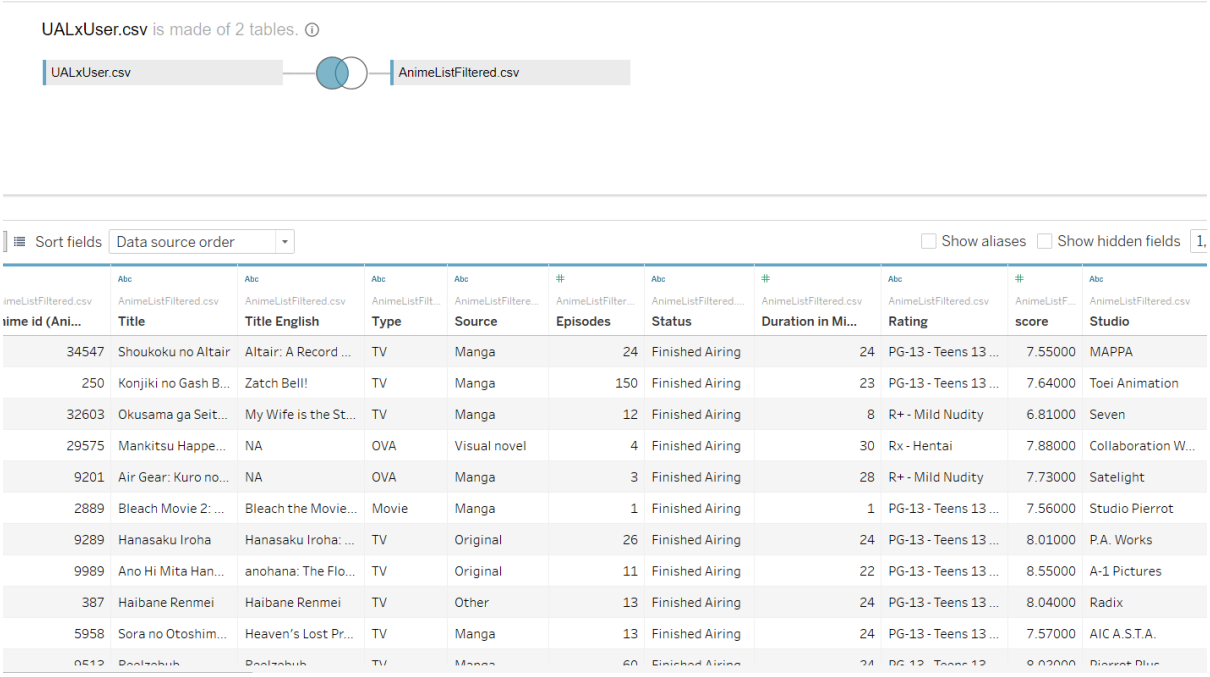
After removing the genre column, we save the cleaned dataset as a new csv called AnimeListFiltered.csv.

Now to the UsersList.csv dataset. The main issue is that the location is not in a tableau friendly format. In order for us to use the spatial data, the library countrycode [3] was used to convert the location. Once converted, we will remove the records where conversion wasn't possible. We then save it as a new csv called UsersFiltered.csv

username <chr>	gender <chr>	location <chr>	Location_Name <chr>
karthiga	Female	IND	India
bskai	Male	MEX	Mexico
terune_uzumaki	Female	MYS	Malaysia
HimeAria	Female	POL	Poland
Skallington	Male	SWE	Sweden
Slimak	Male	POL	Poland
MistButterfly	Female	HUN	Hungary
ProperBritish	Male	GBR	United Kingdom
helenply	Female	MEX	Mexico
cfoordddd	*NA*	AUS	Australia

Figure 15: The Cleaned UserList.csv

Now we need to an inner join between the datasets UsersFiltered.csv and UserAnimeListFiltered.csv. We will join the resulting data frame with the AnimeListFiltered.csv data set in tableau itself as performing an inner join increases the size of the data set to 7 GB.



ime id (Ani...	Title	Title English	Type	Source	Episodes	Status	Duration in Mi...	Rating	score	Studio
34547	Shoukoku no Altair	Altair: A Record ...	TV	Manga	24	Finished Airing	24	PG-13 - Teens 13 ...	7.55000	MAPPA
250	Konjiki no Gash B...	Zatch Bell!	TV	Manga	150	Finished Airing	23	PG-13 - Teens 13 ...	7.64000	Toei Animation
32603	Okusama ga Seit...	My Wife is the St...	TV	Manga	12	Finished Airing	8	R+ - Mild Nudity	6.81000	Seven
29575	Mankitsu Happe...	NA	OVA	Visual novel	4	Finished Airing	30	Rx - Hentai	7.88000	Collaboration W...
9201	Air Gear: Kuro no...	NA	OVA	Manga	3	Finished Airing	28	R+ - Mild Nudity	7.73000	Satelight
2889	Bleach Movie 2: ...	Bleach the Movie...	Movie	Manga	1	Finished Airing	1	PG-13 - Teens 13 ...	7.56000	Studio Pierrot
9289	Hanasaku Iroha	Hanasaku Iroha: ...	TV	Original	26	Finished Airing	24	PG-13 - Teens 13 ...	8.01000	P.A. Works
9989	Ano Hi Mita Han...	anohana: The Flo...	TV	Original	11	Finished Airing	22	PG-13 - Teens 13 ...	8.55000	A-1 Pictures
387	Haibane Renmei	Haibane Renmei	TV	Other	13	Finished Airing	24	PG-13 - Teens 13 ...	8.04000	Radix
5958	Sora no Otoshim...	Heaven's Lost Pr...	TV	Manga	13	Finished Airing	24	PG-13 - Teens 13 ...	7.57000	AIC A.S.T.A.

Figure 16: The final Dataset.

Data Exploration

Now we will proceed with the visualisations in Tableau. We will start by analysing the market/viewership share of Anime. As anime is a visual medium, we can find its market by checking it viewership or TRP. We will be analysing it in the following 6 ways.

- **User Distribution**: the number of anime fans country wise.

- **Anime Distribution**: Tells us about the no of anime series that have caught the attention of the users. This will be done genre wise as well as overall.
- **Anime Viewership**: Gives us information about the overall TRP (total views overall) of anime in a country.
- **Most Viewed anime**: Gives us information about the most viewed anime in a particular country.
- **Highest rated anime**: Gives us the highest rated anime for each country. This is calculated by taking the average scores of each anime seen in the country and choosing the anime with the maximum average.
- **Anime's key of success**: we will be plot the scores of various anime series against factors like studio, episode length, Rating, Type and Genre.

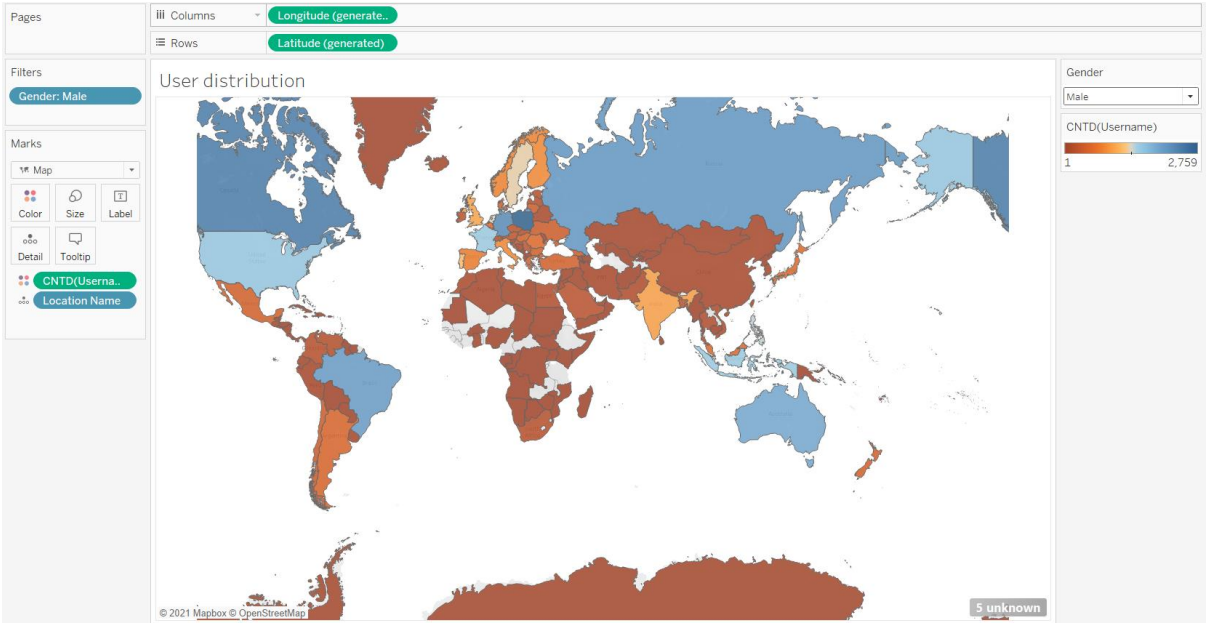


Figure 17: The number of Anime fans around the world

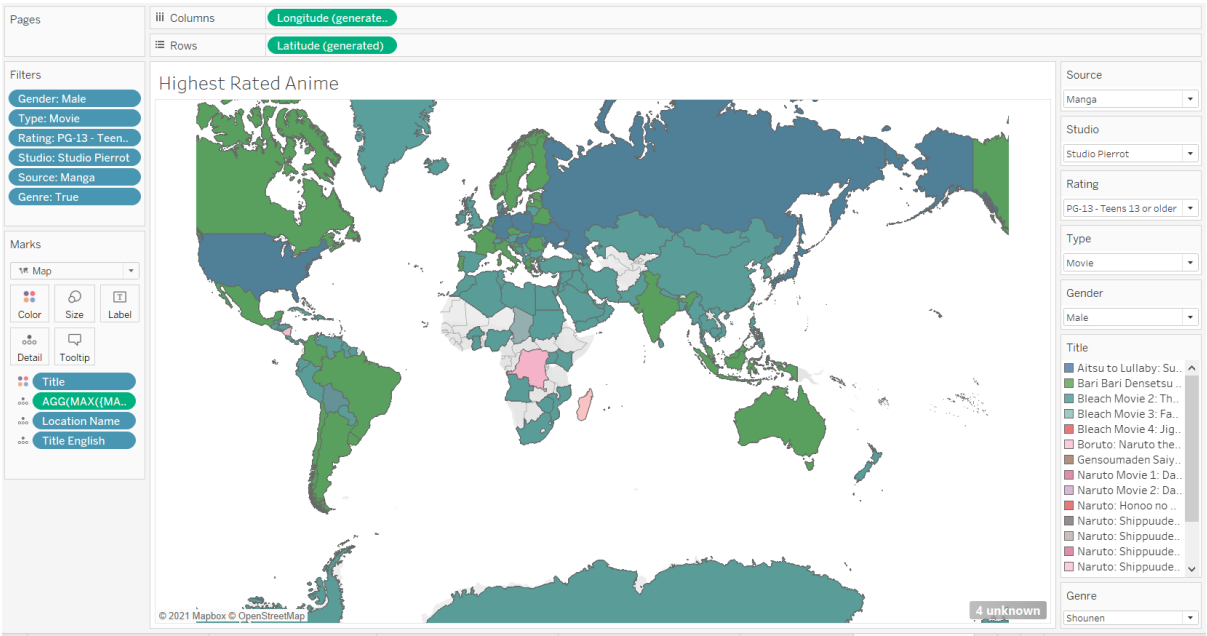


Figure 18: The visualisation of the Highest rated anime series

All the visualisations except the anime's keys of success will use spatial data. In every visualisation there will be the following filters

- **Gender**: We will have an option to filter our visualisation by choosing a gender. There are 4 unique types of genders Male, Female, Non-Binary(Transgender) and NA.
- **Type**: Type is the form in which it was released. The major types are Movie, TV, OVA (Original Video Animation), ONA (Original Net Animation), Special and Unknown.
- **Studio**: The Studio that made the anime. A few instances are Studio Bones, MAPPA, etc.
- **Rating**: Rating gives us the age restrictions given to an anime series. For Example, G, PG, etc.
- **Source**: It tells us the kind of source material from the anime and its characters are based on. Manga and Light Novels are examples of it.

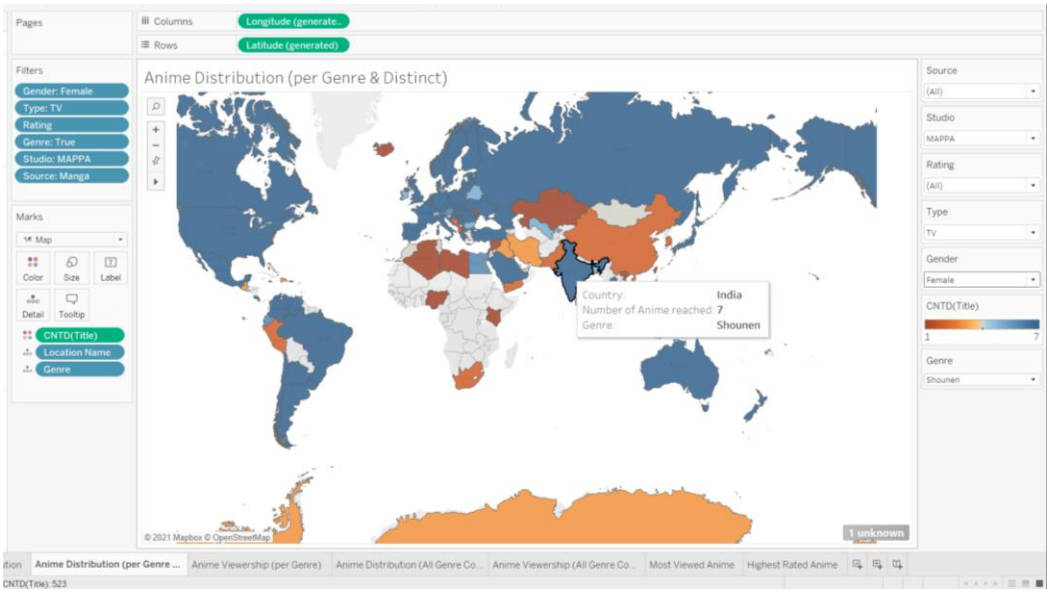


Figure 19: The anime Reach based on many factors

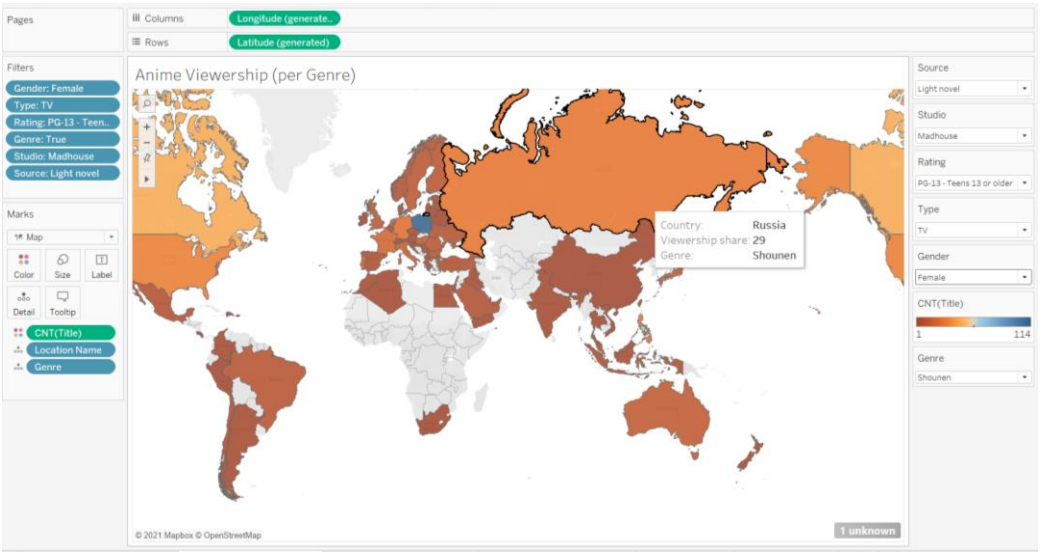


Figure 20: The Viewership Share of Anime based on many factors

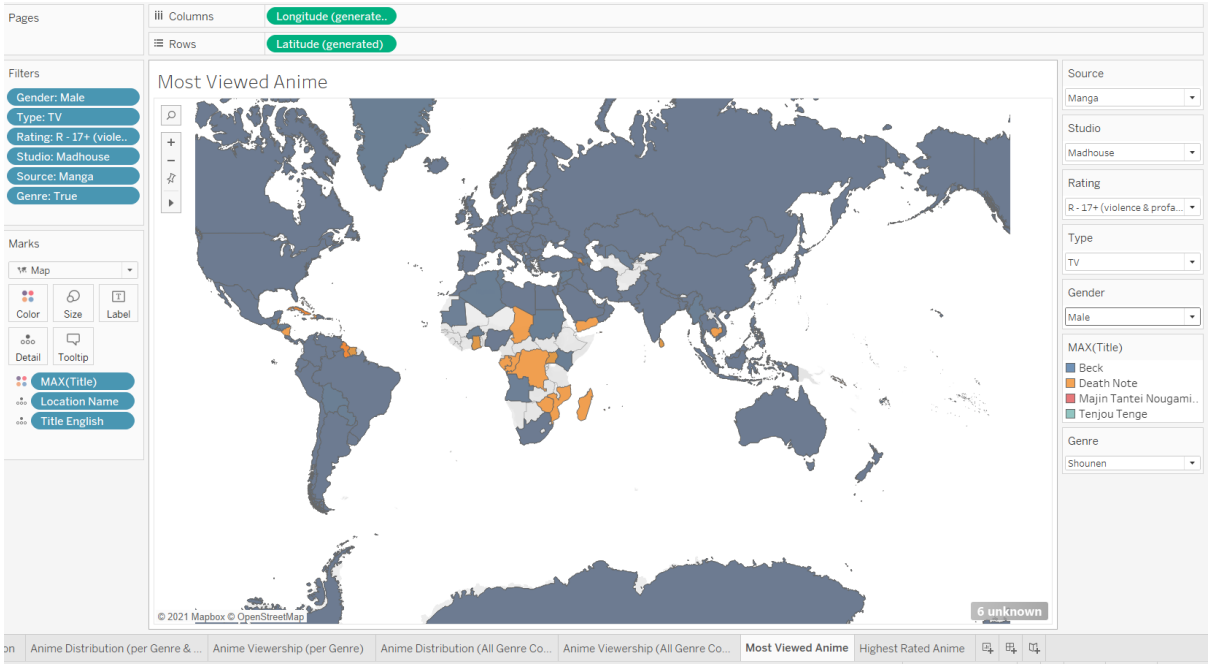


Figure 21: The most Viewed Country based on the filters

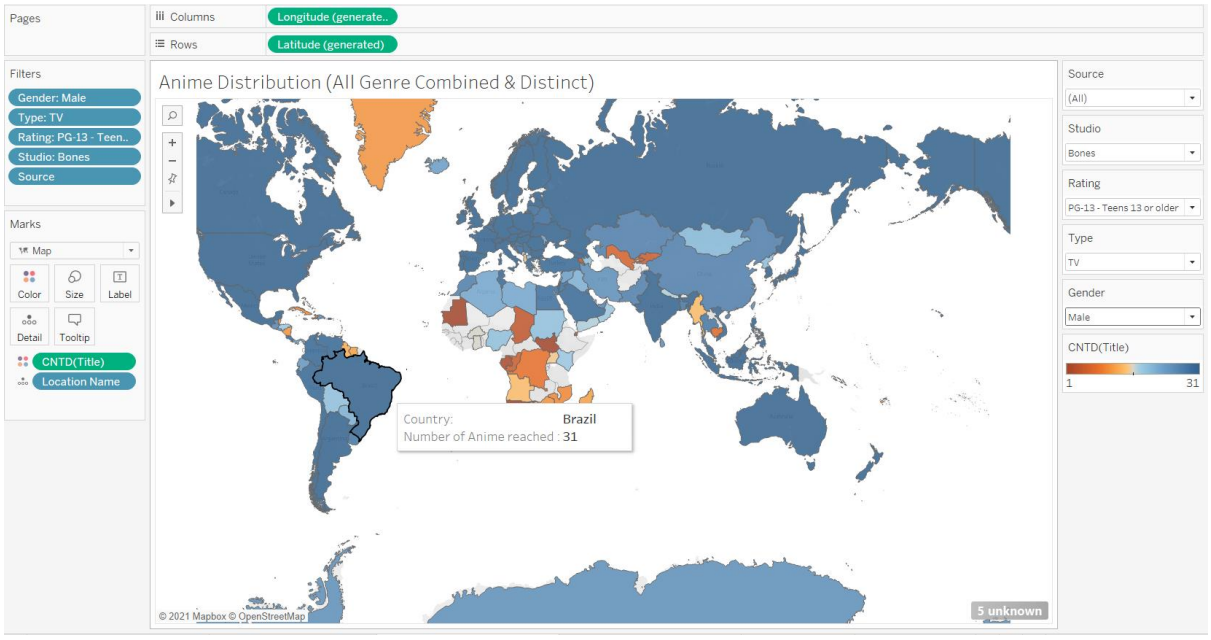


Figure 22: Overall Anime Distribution

By visualising we are able to find out the how different types of anime are received around the world. This will help us understand the global reach of the genre as well as the sub genres that are preferred by the community. Now we understand the tastes of the community and will give animation studios an idea before they produce another anime. We can also see the factors that will lead to an anime's success and how they affect them.

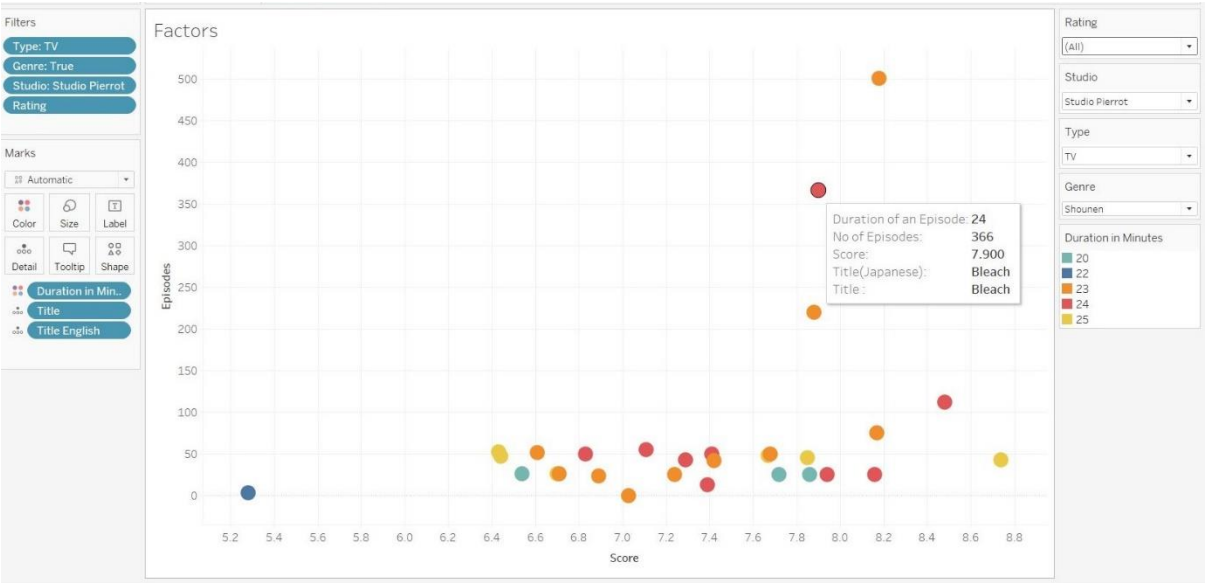


Figure 23: A Scatter plot displaying the factors influencing an Anime's score

Conclusion

Anime has definitely grown to a point where its influence is significant. Although there may be variation in preference regarding the genre we can all agree this is a testament to its range.

There are a few factors like the number of episodes, studio, rating and type have some sort of influence on its success. We found that the episode duration is not a factor as the duration is the same in most cases.

I can say for sure that it did answer to the questions I had. It has also made think about diving deeper into this in the future. When I get hands on better datasets.

Reflection

During my exploration I managed to learn many things that will definitely help me in the future as I work in this field. They are as follows.

- The importance of wrangling / cleaning and how difficult it is to master. I had a first hand experience on how dirty real world data is and the challenges that may arise while dealing with it.
- How spatial data can make a visualization a lot more alluring even though it is difficult as well as have a level of complexity.

In hindsight, I would have loved to convert the state names to its country name while I was wrangling but unfortunately due to the lack of resources I wasn't able to pull it off with the choroplethAdmin1 package in R. It would have definitely given me more data to analyse. Also I wanted to display the image of the anime in the tooltip. Unluckily, the poster link that was in the dataset itself was invalid. So it wasn't possible for me to do it.

Bibliography

[1] Henerson, Evan,"The Influence of Anime",Keyframe Magazine. https://keyframemagazine.org/2020/08/14/the-influence-of-anime/ (Accessed 2020).

[2] Račinský, Matěj ,"MyAnimeList Dataset." Kaggle, 2018, doi: 10.34740/KAGGLE/DSV/45582. URL: https://www.kaggle.com/azathoth42/myanimelist .

[3] B. Arel, Vincent, E. Nils,C.J. Yetman, "countrycode: An R package to convert country names and country codes", Journal of Open Source Software, vol. 3, no. 28 ,pp. 848, 2018.

Appendix

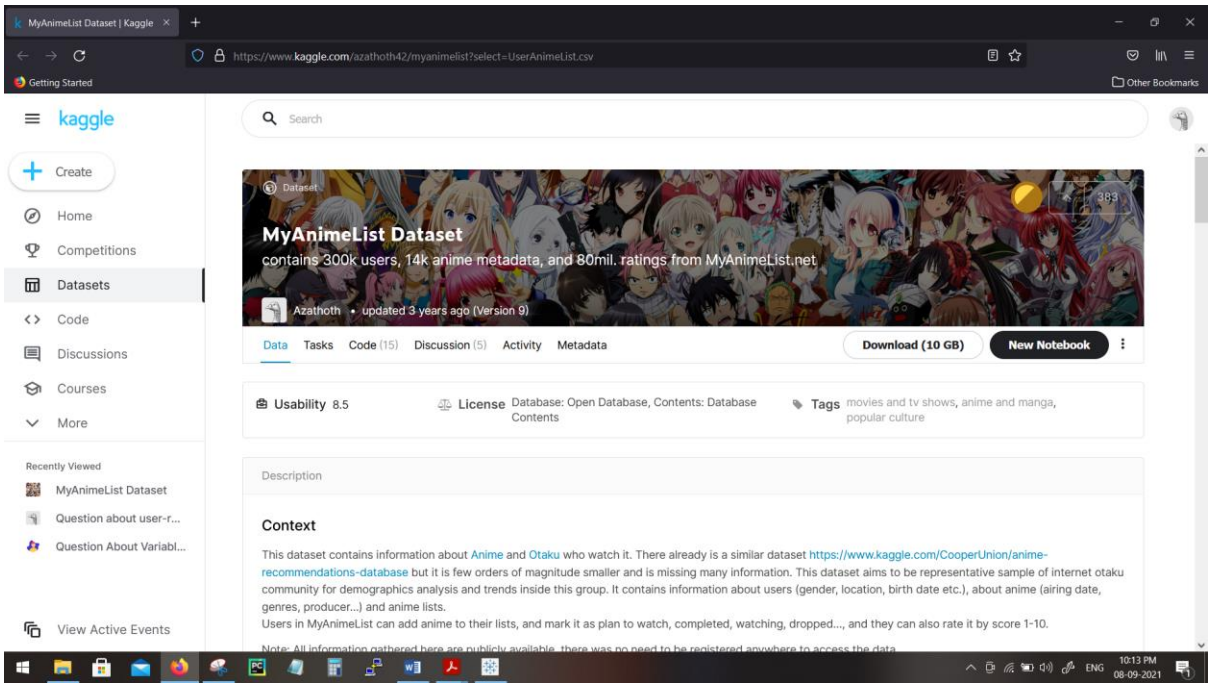


Figure 24: The Kaggle Datasets website screenshot

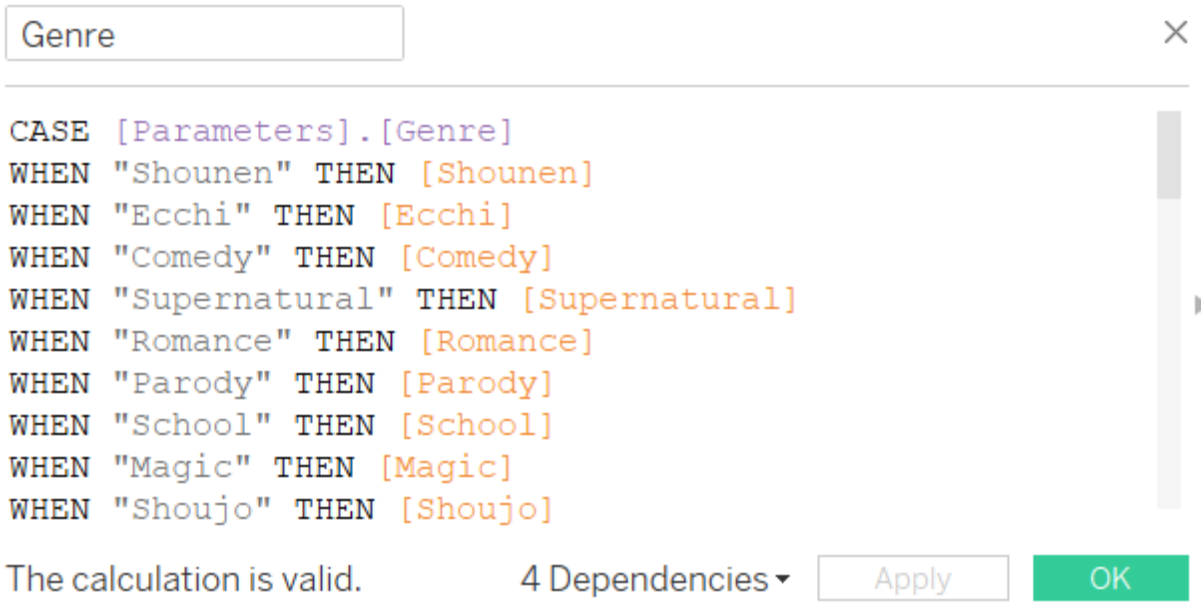


Figure 25: A screenshot of the tableau coding I used for the filters

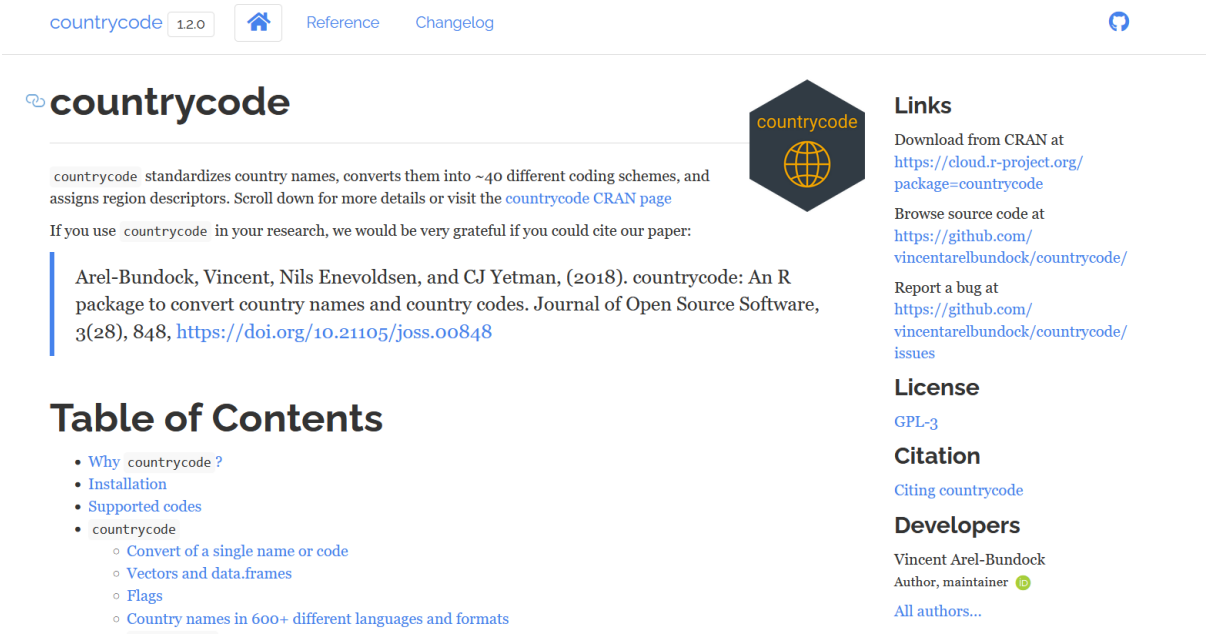


Figure 26:A look at countrycode's official website

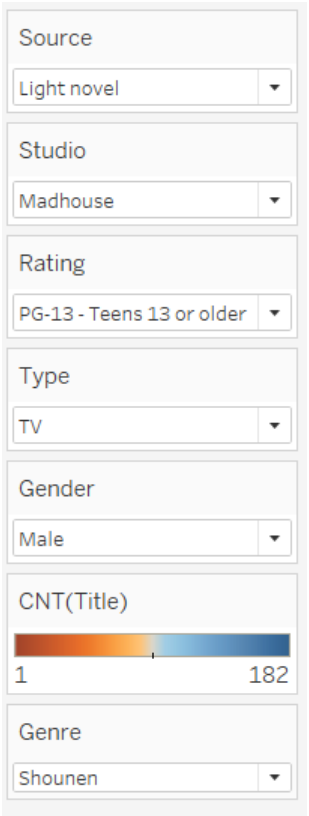


Figure 27: A closer look at the filters