



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Data Mining Report - The Movies Dataset

Contents

1. Introduction	1
2. The Data Set.....	1
2.1 Dataset Attributes.....	4
3. Data Pre-Processing	5
3.1 Removal of unsuitable attributes	5
3.2 Attributes with missing values.....	6
3.3 Illegal Characters	8
3.3 Checking for nonsensical data and incomplete records	9
3.4 Splitting an attribute into separate groups	9
3.5 Restructuring date values.....	11
3.6 Creating the ARFF file for Weka.....	13
4. Data Visualisation in Weka	14
5. Data Analysis.....	15
5.1 Question 1 - Is it possible to predict the popularity score for a movie based on the existing records in the dataset?	15
5.2 Question 2 - Are there any patterns contained in the genres? Are there any genres that feature frequently in the dataset?.....	23
5.3 Question 3 - Is there any relation between the budget for a movie and the revenue it makes?	31
5.4 Question 4 – Is there a relationship between Runtime and Revenue?.....	33
6. Reflection	36
References	39

Table of Figures

Figure 1: The MovieLens website [2]	2
Figure 2: The 'movies_metadata' CSV file.....	3
Figure 3: The reorganised dataset in Excel	6
Figure 4: Missing Runtime attribute	7
Figure 5: Using the Find and Replace tool in Excel	8
Figure 6: Incomplete or nonsensical records which were removed.....	9
Figure 7: The Genre attribute before splitting.....	9
Figure 8: The keyword columns added to the dataset file	10
Figure 9: Using a function to find a keyword in a string	10
Figure 10: Changing the date structure for Weka	12
Figure 11: Initial Weka visualisation	14
Figure 12: Popularity attribute values	15
Figure 13: Top 10 movies with highest Popularity score	16
Figure 14: The dataset loaded into Weka.....	17
Figure 15: Setting the Popularity attribute as a class	18
Figure 16: The NumericToNominal filter being applied.....	18
Figure 17: Results of 75% split classification	20
Figure 18: Updated dataset groupings	21
Figure 19: Results of the 66% percentage split classification.....	22
Figure 20: Confusion matrix for 66% split / 4 buckets.....	23
Figure 21: The dataset loaded into Weka	24
Figure 22: Results of the Apriori algorithm.....	25
Figure 23: Updated results of the Apriori algorithm	26
Figure 24: Adjusting the Apriori algorithm	27
Figure 25: Results of the Apriori algorithm.....	28
Figure 26: Results of the Apriori algorithm with 50% confidence	29
Figure 27: Results of the Apriori algorithm 3rd run.....	30
Figure 28: Top 10 grossing movies in the dataset – sorted by Revenue	31
Figure 29: Top 10 grossing movies in the dataset – sorted by Budget.....	31
Figure 30: Scatter plot of Budget against Revenue	32
Figure 31: Count of assigned Genres in the full dataset.....	33
Figure 32: Scatter plot of Runtime vs Revenue	34
Figure 33: Scatter plot of Runtime vs Revenue (Animation only)	35

Figure 34: Top 10 Animation movies with highest revenue	36
---	----

1. Introduction

This aim of this report is to analyse a database extract and perform data mining processes on it to identify if any significant patterns or trends can be found. The dataset will go through many pre-processing operations before being imported into industry standard data mining and visualisation software including Weka and Tableau. Once successfully imported, various methods of processing which could include classification, clustering and association analysis will be used to create models, visualisations, and to identify patterns. There will be pre-defined questions that will attempt to be answered using the different models. Any significant findings will then be discussed and reported and finally a conclusion will be drawn on the overall process.

2. The Data Set

The dataset selected for analysis and investigation in this report is based on data collected about movies. The dataset has been retrieved from the Kaggle website. Kaggle are specialists in providing data for data science projects [1]. The origins of the dataset come from the MovieLens [2] website, which is an online movie recommender service, and in turn the dataset has been collected and made available through GroupLens Research [3]. Although the data has been collected between 1995 and 2017, it does not imply that a movie must have been released between those years to be featured in the dataset. This is simply the date range in which data was collected about each individual movie.

Some elements of the dataset are blended from other sources, including selected movie details, credits and keywords, which have been retrieved from The Movie Database (TMDb) website. TMDb is a user editable database built for movies and TV shows and provides extensive metadata, with over 200,000 developers and organisations using the platform [4].

As well as data collected about movies, part of the dataset includes data collected from users of the website such as their ratings for movies that have been viewed. These users were selected randomly for inclusion and must have previously rated at least twenty movies in order to be included. Each user in the dataset is represented by an ID number and there is no demographic information included that could be used to identify individuals [5].

The MovieLens website works as a movie recommendation system, where users create an account and rate movies that they like. The recommendation system is then able to gather data from the users, build a profile on their specific likes and dislikes and use association methods to recommend other similar movies which may also appeal to the user. An example of the MovieLens website can be found

in Figure 1. Here you can see the website has made a prediction about the rating that a user is likely to give based on the data that has been gathered about that user to date.

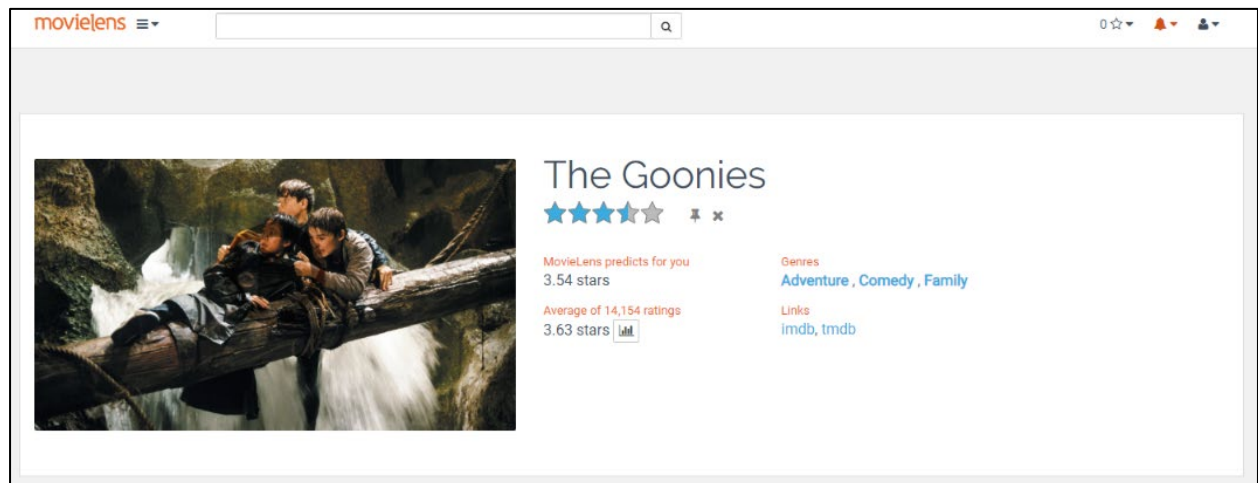


Figure 1: The MovieLens website [2]

The main purpose for this dataset according to Kaggle is to build a recommendation engine for users. Whilst this goes beyond the scope of this report, the dataset has been chosen because it is still substantial enough for analysis and it is hoped that there could be hidden patterns or trends which can be explored further.

The full dataset consists of seven files in CSV format, however on closer inspection it became clear that not all of them would be relevant to this report and some could be disregarded. The primary dataset file for analysis is called 'movies_metadata.csv'. Within this file there is data about 45,467 movies.

In total, there are 24 attributes of varying data types and sizes. On first examination of the dataset, it was found that not all of the attributes were suitable to be included for analysis and some were therefore disregarded. It was also discovered that the data is not yet suitable to be loaded into any data mining software as it requires a considerable amount of pre-processing and cleaning. This is due to many factors including null values, illegal characters, as well as missing data. The pre-processing methods used to clean the data and prepare it for analysis will be explained later in this report.

An extract of the raw dataset prior to any pre-processing can be seen in Figure 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	adult	belongs_to_collection	budget	genres	homepage	id	imdb_id	original_lang	original_title	overview	popularity	poster_path	production_compa	production_coun	release_date	revenue	runtime	spoken_la_status	tagline	title	video	vote_aver	vote_count	
2	FALSE	{'id': 10194, 'name': 'Toy	300000000	{'id': 16, 'name': 'http://toy		862	tt0114709	en	Toy Story	Led by Woo	21.946943	/rhlRbceoe9lI	{'name': 'Pixar Ani	{'iso_3166_1': 'US	30/10/1995	373554033	81	{'iso_639_1': 'Released		Toy Story	FALSE	7.7	5415	
3	FALSE		650000000	{'id': 12, 'name': 'Adve		8844	tt0113497	en	Jumanji	When siblin	17.015539	/vzmL6f7PaPi	{'name': 'TriStar Pi	{'iso_3166_1': 'US	15/12/1995	262797249	104	{'iso_639_1': 'Released		Roll the d	Jumanji	FALSE	6.9	2413
4	FALSE	{'id': 119050, 'name': 'Gr	0	{'id': 10749, 'name': 'Com		15602	tt0113228	en	Grumpier Old A family we		11.7129	/6ksm1sJKMF	{'name': 'Warner B	{'iso_3166_1': 'US	22/12/1995	0	101	{'iso_639_1': 'Released		Still Yellin	Grumpier	FALSE	6.5	92
5	FALSE		160000000	{'id': 35, 'name': 'name': 'Rc		31357	tt0114885	en	Waiting to Exit	Cheated on	3.859495	/16XOMPcEaL	{'name': 'Twentiet	{'iso_3166_1': 'US	22/12/1995	81452156	127	{'iso_639_1': 'Released		Friends ar	Waiting to	FALSE	6.1	34
6	FALSE	{'id': 96871, 'name': 'Fat	0	{'id': 35, 'name': 'name': 'Com		11862	tt0113041	en	Father of the	Just when G	8.387519	/e64sO48hQ	{'name': 'Sandollar	{'iso_3166_1': 'US	10/02/1995	76578911	106	{'iso_639_1': 'Released		Just Wher	Father of f	FALSE	5.7	173
7	FALSE		600000000	{'id': 28, 'name': 'Action		949	tt0113277	en	Heat	Obsessive n	17.924927	/zMyfPUelun	{'name': 'Regency	{'iso_3166_1': 'US	15/12/1995	187436818	170	{'iso_639_1': 'Released		A Los Angl	Heat	FALSE	7.7	1886
8	FALSE		580000000	{'id': 35, 'name': 'Com		11860	tt0114319	en	Sabrina	An ugly duc	6.677277	/jQh15y5YB7k	{'name': 'Paramou	{'iso_3166_1': 'DE	15/12/1995	0	127	{'iso_639_1': 'Released		You are cc	Sabrina	FALSE	6.2	141
9	FALSE		0	{'id': 28, 'name': 'Action		45325	tt0112302	en	Tom and Huck A mischiev		2.561161	/sGO5Qa55p7	{'name': 'Walt Disr	{'iso_3166_1': 'US	22/12/1995	0	97	{'iso_639_1': 'Released		The Origir	Tom and H	FALSE	5.4	45
10	FALSE		350000000	{'id': 28, 'name': 'Action		9091	tt0114576	en	Sudden Deat	Internation	5.23158	/eoWvKD60lT	{'name': 'Universal	{'iso_3166_1': 'US	22/12/1995	64350171	106	{'iso_639_1': 'Released		Terror goe	Sudden Di	FALSE	5.5	174
11	FALSE	{'id': 645, 'name': 'Jame	580000000	{'id': 12, 'name': 'http://www		710	tt0113189	en	GoldenEye	James Bond	14.686036	/5c0ovjT41Kn	{'name': 'United Ai	{'iso_3166_1': 'GE	16/11/1995	352194034	130	{'iso_639_1': 'Released		No limits.	GoldenEyi	FALSE	6.6	1194
12	FALSE		620000000	{'id': 35, 'name': 'Com		9087	tt0112346	en	The American	Widowed U	6.318445	/lymPnGLZgF	{'name': 'Columbia	{'iso_3166_1': 'US	17/11/1995	107879496	106	{'iso_639_1': 'Released		Why can't	The Amer	FALSE	6.5	199
13	FALSE		0	{'id': 35, 'name': 'Com		12110	tt0112896	en	Dracula: Dead	When a law	5.430331	/xve4cgFYItnc	{'name': 'Columbia	{'iso_3166_1': 'FR	22/12/1995	0	88	{'iso_639_1': 'Released		Dracula: D		FALSE	5.7	210
14	FALSE	{'id': 117693, 'name': 'Be	0	{'id': 10751, 'name': 'Fant		21032	tt0112453	en	Balto	An outcast f	12.140733	/gV5PCAVCVP	{'name': 'Universal	{'iso_3166_1': 'US	22/12/1995	11348324	78	{'iso_639_1': 'Released		Part Dog.	Balto	FALSE	7.1	423
15	FALSE		440000000	{'id': 36, 'name': 'name': 'Histo		10858	tt0113987	en	Nixon	An all-star c	5.092	/clCkmCEIXRf	{'name': 'Hollywoc	{'iso_3166_1': 'US	22/12/1995	13681765	192	{'iso_639_1': 'Released		Triumphai	Nixon	FALSE	7.1	72
16	FALSE		980000000	{'id': 28, 'name': 'Action		1408	tt0112760	en	Cutthroat Isla	Morgan Ada	7.284477	/odM9973kliv	{'name': 'Le Studio	{'iso_3166_1': 'FR	22/12/1995	10017322	119	{'iso_639_1': 'Released		The Cours	Cutthroat	FALSE	5.7	137
17	FALSE		520000000	{'id': 18, 'name': 'Dram		524	tt0112641	en	Casino	The life of t	10.137389	/xo517ibXBd	{'name': 'Universal	{'iso_3166_1': 'FR	22/11/1995	116112375	178	{'iso_639_1': 'Released		No one st	Casino	FALSE	7.8	1343
18	FALSE		165000000	{'id': 18, 'name': 'Dram		4584	tt0114388	en	Sense and Sei	Rich Ric. Ma	10.673167	/IA9HTy84Bb	{'name': 'Columbia	{'iso_3166_1': 'GE	13/12/1995	135000000	136	{'iso_639_1': 'Released		Lose your	Sense and	FALSE	7.2	364
19	FALSE		400000000	{'id': 80, 'name': 'Crim		5	tt0113101	en	Four Rooms	It's Ted the	9.026586	/eQs5hh9xrkk	{'name': 'Miramax	{'iso_3166_1': 'US	09/12/1995	43000000	98	{'iso_639_1': 'Released		Twelve ou	Four Room	FALSE	6.5	539
20	FALSE	{'id': 3167, 'name': 'Ace	300000000	{'id': 80, 'name': 'Crim		9273	tt0112281	en	Ace Ventura:	Summoned	8.205448	/wRlGnJhEzC	{'name': 'O Enterta	{'iso_3166_1': 'US	10/11/1995	212385533	90	{'iso_639_1': 'Released		New anim	Ace Ventu	FALSE	6.1	1128
21	FALSE		600000000	{'id': 28, 'name': 'Action		11517	tt0113845	en	Money Train	A vengeful	7.337906	/jSozzzVOR2k	{'name': 'Columbia	{'iso_3166_1': 'US	21/11/1995	35431113	103	{'iso_639_1': 'Released		Get on, or	Money Tri	FALSE	5.4	224
22	FALSE	{'id': 91698, 'name': 'Chi	302500000	{'id': 35, 'name': 'Com		8012	tt0113161	en	Get Shorty	Chilli Palmei	12.669608	/vWtDUUgQA	{'name': 'Jersey Fil	{'iso_3166_1': 'US	20/10/1995	115101622	105	{'iso_639_1': 'Released		The mob i	Get Shorty	FALSE	6.4	305
23	FALSE		0	{'id': 18, 'name': 'Dram		1710	tt0112722	en	Copycat	An agoraphi	10.701801	/80czeJGSoiK	{'name': 'Regency	{'iso_3166_1': 'US	27/10/1995	0	124	{'iso_639_1': 'Released		One man	Copycat	FALSE	6.5	199
24	FALSE		500000000	{'id': 28, 'name': 'Action		9691	tt0112401	en	Assassins	Assassin Ro	11.065939	/xAX5MP7Dg	{'name': 'Silver Pic	{'iso_3166_1': 'FR	06/10/1995	30303072	132	{'iso_639_1': 'Released		In the sha	Assassins	FALSE	6	394
25	FALSE		0	{'id': 18, 'name': 'Dram		12665	tt0114168	en	Powder	Harassed by	12.133094	/1uRKsXOCtg	{'name': 'Caravan f	{'iso_3166_1': 'US	27/10/1995	0	111	{'iso_639_1': 'Released		An extrao	Powder	FALSE	6.3	143
26	FALSE		360000000	{'id': 18, 'name': 'http://www		451	tt0113627	en	Leaving Las V	Ben Sander	10.332025	/37qHRJxnSh	{'name': 'United Ai	{'iso_3166_1': 'US	27/10/1995	498000000	112	{'iso_639_1': 'Released		I Love You	Leaving La	FALSE	7.1	365
27	FALSE		0	{'id': 18, 'name': 'Dram		16420	tt0114057	en	Othello	The evil lag	1.845899	/gM0BXEQjm	{'name': 'Columbia	{'iso_3166_1': 'IT	15/12/1995	0	123	{'iso_639_1': 'Released		Envy, gree	Othello	FALSE	7	33
28	FALSE		120000000	{'id': 35, 'name': 'Com		9263	tt0114011	en	Now and Ther	Waxing nos	8.681325	/wD6rldD2lX	{'name': 'New Line	{'iso_3166_1': 'US	20/10/1995	274000000	100	{'iso_639_1': 'Released		In every w	Now and	FALSE	6.6	91
29	FALSE		0	{'id': 18, 'name': 'Dram		17015	tt0114117	en	Persuasion	This film ad	2.228434	/si8911IezMv	{'name': 'BBC Film	{'iso_3166_1': 'GE	27/09/1995	0	104	{'iso_639_1': 'Released		Persuasio		FALSE	7.4	36
30	FALSE		180000000	{'id': 14, 'name': 'Fanta		902	tt0112682	fr	La Cité@ des	A scientiat	9.822423	/eV06gewq4a	{'name': 'Procirop	{'iso_3166_1': 'FR	16/05/1995	1738611	108	{'iso_639_1': 'Released		Where ha	The City o	FALSE	7.6	308
31	FALSE		0	{'id': 18, 'name': 'Dram		37557	tt0115012	zh	æ'tʃʌ•sæ'tʃɪk	A provincial	1.100915	/qcoOCoN7vi	{'name': 'MinistA	{'iso_3166_1': 'CN	30/04/1995	0	108	{'iso_639_1': 'Released		In 1930's	S Shanghai	FALSE	6.5	17
32	FALSE		0	{'id': 18, 'name': 'Dram		9909	tt0112792	en	Dangerous Mi	Former Mar	9.481338	/yJee3QmYC	{'name': 'Hollywoc	{'iso_3166_1': 'US	11/08/1995	180000000	99	{'iso_639_1': 'Released		She broke	Dangerou	FALSE	6.4	249
33	FALSE		295000000	{'id': 878, 'name': 'Scie		63	tt0114746	en	Twelve Monk	In the year	12.297305	/6Sj9wDu3Yu	{'name': 'Universal	{'iso_3166_1': 'US	29/12/1995	168840000	129	{'iso_639_1': 'Released		The future	Twelve M	FALSE	7.4	2470
34	FALSE		0	{'id': 10749, 'name': 'Rom		78802	tt0114952	fr	Guillaumet, les	alles du cc	0.745542	/k6ODtR38dK	{'name': 'Uwerks Er	{'iso_3166_1': 'FR	18/09/1996	0	50	{'iso_639_1': 'Released		Wings of		FALSE	6.8	4
35	FALSE	{'id': 9435, 'name': 'Bab	300000000	{'id': 14, 'name': 'Fant		9598	tt0112431	en	Babe	Babe is a lit	14.404764	/gN6X3fwPya	{'name': 'Universal	{'iso_3166_1': 'AI	18/07/1995	254134910	89	{'iso_639_1': 'Released		A little pig	Babe	FALSE	6	756
36	FALSE		0	{'id': 36, 'name': 'name': 'Histo		47018	tt0112637	en	Carrington	The story of	1.493361	/a7w6rPdTBg	{'name': 'StudioCa	{'iso_3166_1': 'FR	08/11/1995	0	121	{'iso_639_1': 'Released		A Love so	Carrington	FALSE	6.4	16
37	FALSE		110000000	{'id': 18, 'name': 'Dram		687	tt0112818	en	Dead Man Wa	A justice dri	6.891317	/y19uRKAHXC	{'name': 'Havoc', 'i	{'iso_3166_1': 'US	29/12/1995	39363635	122	{'iso_639_1': 'Released		Dead Man		FALSE	7.3	350
38	FALSE		0	{'id': 12, 'name': 'Adve		139405	tt0112286	en	Across the Se	A young Rus	0.114469	/IKOAYy098y	{'name': 'Sony Nev	{'iso_3166_1': 'US	20/10/1995	0	51	{'iso_639_1': 'Released		Across the		FALSE	3.5	2

Figure 2: The 'movies_metadata' CSV file

Microsoft Excel was used to load the dataset to begin the process of exploring what data was included and to identify any attributes that may not be relevant. It also enabled a plan to be drawn up to outline the pre-processing operations that would be performed to make the data suitable for analysis in data mining software.

2.1 Dataset Attributes

Attribute	Data Type	Description
adult	Boolean	Indicates whether a movie has been rated as having adult content
belongs_to_collection	String	Gives information on whether a movie belongs to a series or collection e.g. Star Wars
budget	Numeric	The budget of the movie given in dollars
genres	String Array	A list of all genres that can be associated with a movie
homepage	String	The location of the official homepage URL of a movie
id	Numeric	The official ID number allocated to the movie by the MovieLens website
imdb_id	String	The unique ID allocated to the movie by the IMDB website. There is no association with the ID given by the MovieLens website
original_language	String	The spoken language that the movie was originally filmed in
original_title	String	The original title given to the movie when it began production. On rare occasions, movies can change titles during production and this is represented by the <i>title</i> attribute instead
overview	String	Gives potential viewers a brief outline of the plot and storyline to engage their interest
popularity	Real	This is the popularity score assigned to a movie by the TMDB website. The scores allocated range from 0 to 547, with a mean value of 2.9
poster_path	String	A URL given to locate the official movie poster image
production_companies	String	A list of all production companies who were involved in the making of the movie
production_countries	String	A list of all the countries that the movie was filmed in
release_date	Date	The theatrical release date of the movie
revenue	Numeric	The total revenue of the movie in dollars. The revenue ranges from 0 to £2,787,965,087 with a mean of £11,209,348
runtime	Numeric	The runtime of the movie given in minutes
spoken_language	String	The spoken languages that are featured in the movie. This can be different from the language that the movie was filmed in
status	String	The status of the movie at the time of data collection. Values can include Released, Announced, Post-Production, etc
tagline	String	The tagline or caption for the movie
title	String	The official final title of the movie. This can sometimes differ from the original title
video	Boolean	This indicates if there is a video of the movie available on the TMDB website
vote_average	Real	The average rating given to the movie by users of the website. The score is based on a range from 1 to 10. The mean value of votes is 5.6
vote_count	Numeric	The number of votes given by users to each particular movie. The count of votes ranges from 0 to 14075, with a mean value of 109.8

3. Data Pre-Processing

3.1 Removal of unsuitable attributes

The first step in the pre-processing stage was to remove the attributes that were deemed to be unnecessary or unsuitable for further analysis. The following attributes were removed:

Attribute	Example Data	Reason
adult	FALSE	There were only 13 movies in the dataset categorised with adult content. It was therefore deemed unsuitable for further analysis
belongs_to_collection	{'id': 10, 'name': 'Star Wars Collection'}	Less than 10% of the total movies in the dataset belonged to a collection. As this report is focused on individual movies it was decided to remove this attribute
homepage	http://www.starwars.com/films/star-wars-episode-v-the-empire-strikes-back	There is no further analysis that can be performed on this attribute, so it was removed
original_language	en	The vast majority of movies in the dataset were filmed in English. This attribute may be useful in future exploration but on this occasion, it was removed
original_title	Return of the Jedi	There were few instances where the original title was different to the final title attribute, so it was felt that this attribute was not necessary
overview	The epic saga continues as Luke Skywalker tries to defeat the Galactic Empire	There is no further analysis that can be performed on this attribute, so it was removed
poster_path	/n8V09dBX6hN8X.jpg	There is no further analysis that can be performed on this attribute, so it was removed
production_companies	{'name': 'Lucasfilm', 'id': 1}	This attribute may be useful in future exploration of the dataset to help predict the success of movies, but on this occasion, it was removed from the dataset
production_countries	{'iso_3166_1': 'US'}	This attribute may be useful in future exploration of the dataset to help predict the success of movies, but on this occasion, it was removed from the dataset
spoken_language	{'iso_639_1': 'en'}	Similar to the original_language attribute, this may be useful in future exploration but on this occasion, it was removed
status	Released	There is no further analysis that can be performed on this attribute, so it was removed
tagline	A long time ago in a galaxy far, far away...	Similar to overview, there is no further analysis that can be performed on this attribute, so it was removed
video	FALSE	Only 100 movies had the full video available to view on the TMDb website. It was therefore deemed unsuitable for further analysis

Following the removal of these attributes, the dataset was then left with 11 attributes that were either considered to be suitable for further analysis and exploration or deemed necessary to provide context to the data. In Excel, some formatting was done to tidy up the appearance of the dataset and reorganise the columns into a logical order, which can be seen in Figure 3.

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Title	Genres	Budget	Revenue	Release_Date	Runtime	Popularity	Vote_Average	Vote_Count	IMDB_ID
2	11	Star Wars	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	11000000	775398007	25/05/1977	121	42.149697	8.1	6778	tt0076759
3	1891	The Empire Strikes Back	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	18000000	538400000	17/05/1980	124	19.470959	8.2	5998	tt0080684
4	1892	Return of the Jedi	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	32350000	572700000	23/05/1983	135	14.586087	7.9	4763	tt0086190
5	1893	Star Wars: Episode I - The Phantom Menace	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	115000000	924317558	19/05/1999	136	15.649091	6.4	4526	tt0120915
6	1894	Star Wars: Episode II - Attack of the Clones	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	120000000	649398328	15/05/2002	142	14.072511	6.4	4074	tt0121765
7	1895	Star Wars: Episode III - Revenge of the Sith	{{'id': 878, 'name': 'Science Fiction'}, {'id': 12, 'name': 'Adventure'}}	113000000	850000000	17/05/2005	140	13.165421	7.1	4200	tt0121766
8	140607	Star Wars: The Force Awakens	{{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}}	245000000	2068223624	15/12/2015	136	31.626013	7.5	7993	tt2488496
9	330459	Rogue One: A Star Wars Story	{{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}}	200000000	1056057273	14/12/2016	133	36.567575	7.4	5111	tt3748528
10	4518	Elizabeth	{{'id': 18, 'name': 'Drama'}, {'id': 34, 'name': 'Biography'}}	30000000	82150642	13/09/1998	124	7.566321	7.2	346	tt0127536
11	4517	Elizabeth: The Golden Age	{{'id': 18, 'name': 'Drama'}, {'id': 34, 'name': 'Biography'}}	55000000	74237563	09/09/2007	114	7.608243	6.6	310	tt0414055
12	14962	Thumb Wars: The Phantom	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	0	0	02/11/1999	29	1.12808	6.8	16	tt0206341
13	26743	Thumbtanic	{{'id': 35, 'name': 'Comedy'}}	0	0	07/07/2000	26	0.871762	6.4	8	tt0234889
14	816	Austin Powers: International Man of Mystery	{{'id': 878, 'name': 'Science Fiction'}, {'id': 12, 'name': 'Adventure'}}	16500000	67683989	02/05/1997	94	11.749159	6.5	1033	tt0118655
15	817	Austin Powers: The International Man of Mystery	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	33000000	310940086	08/06/1999	95	13.984182	6.1	1034	tt0145660
16	818	Austin Powers in Gold	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	63000000	296655431	26/07/2002	94	8.786653	5.9	997	tt0295178
17	12230	One Hundred and One Dalmatians	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	4000000	215880014	25/01/1961	79	15.727519	6.8	1643	tt0055254
18	13654	101 Dalmatians II: Patch's London Adventure	{{'id': 16, 'name': 'Animation'}, {'id': 12, 'name': 'Adventure'}}	0	0	21/01/2003	70	10.398905	5.7	231	tt0324941
19	10865	Atlantis: The Lost Empire	{{'id': 16, 'name': 'Animation'}, {'id': 12, 'name': 'Adventure'}}	120000000	186053725	02/06/2001	95	10.901602	6.7	1257	tt0230011
20	8965	Atlantis: Milo's Return	{{'id': 14, 'name': 'Fantasy'}, {'id': 12, 'name': 'Adventure'}}	0	0	25/02/2003	70	7.751536	5.4	166	tt0344864
21	10948	The Fox and the Hound	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	12000000	29800000	10/07/1981	82	10.197106	7	884	tt0082406
22	9948	The Fox and the Hound	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	24000000	0	09/11/2006	69	7.432461	5.6	115	tt0465997
23	18992	My Friend Flicka	{{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}}	0	0	26/05/1943	89	0.70993	3.8	2	tt0036182
24	52827	Thunderhead - Son of Thunder	{{'id': 18, 'name': 'Drama'}, {'id': 14, 'name': 'Fantasy'}}	0	0	15/03/1945	78	0.37161	0	1	tt0038172
25	5889	Azumi	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	0	0	10/05/2003	128	6.415721	6.5	91	tt0384819
26	5890	Azumi 2: Death or Love	{{'id': 28, 'name': 'Action'}}	0	0	12/03/2005	108	6.754744	6.4	45	tt0431641
27	9958	Trailer Park Boys: The Movie	{{'id': 35, 'name': 'Comedy'}}	5000000	3833507	06/10/2006	95	3.172765	6.8	53	tt0425601
28	27561	Trailer Park Boys: Countdown to Liquidation	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	3500000	0	25/09/2009	101	3.093829	7	42	tt1337032
29	38674	Trailer Park Boys	{{'id': 35, 'name': 'Comedy'}}	0	0	15/07/1999	67	2.158798	7.4	21	tt0383678
30	257451	Trailer Park Boys: Don't Starve	{{'id': 35, 'name': 'Comedy'}}	0	0	18/04/2014	95	1.335959	6.6	28	tt3244992
31	303742	Trailer Park Boys: Live at the Bar	[[]]	0	0	15/11/2014	88	1.43617	7.6	8	tt3596200
32	26461	Trailer Park Boys: Say Goodnight to the World	{{'id': 35, 'name': 'Comedy'}}	0	0	07/12/2008	60	0.658229	7.6	18	tt1324027
33	24929	Return of the Living Dead	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	6200000	9205924	15/01/1988	89	6.606325	5.8	95	tt0095990
34	10925	The Return of the Living Dead	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	4000000	14237000	15/05/1985	91	6.258984	7.3	270	tt0089907
35	28260	Return of the Living Dead	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	2000000	54207	01/10/1993	97	9.796755	5.9	88	tt0107953
36	24935	Return of the Living Dead	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	0	0	15/10/2005	88	8.195958	3.7	27	tt0411805
37	24939	Return of the Living Dead	{{'id': 35, 'name': 'Comedy'}, {'id': 12, 'name': 'Adventure'}}	0	0	15/10/2005	86	8.186979	4.3	21	tt0411806
38	37988	Vares: Private Eye	{{'id': 28, 'name': 'Action'}, {'id': 34, 'name': 'Biography'}}	0	0	22/07/2004	95	0.944547	6.2	16	tt0372594

Figure 3: The reorganised dataset in Excel

3.2 Attributes with missing values

Being able to use Microsoft Excel to inspect the dataset prior to loading into the data mining and visualisation software was very helpful. By applying filters to the various attribute columns and changing the properties, it was discovered that there were numerous missing values throughout the dataset. Through experience of working with Weka, it is necessary for all values to be present when importing the dataset into the software. Weka is able to process null values but if a value is missing or blank it will cause issues during the data import operation whereby the software thinks there is an error in the data file.

The first task was to deal with the missing values and replace them with a question mark character '?' so that Weka would interpret and understand them as being intentionally missing, rather than assuming there was an error in the dataset.

Figure 4 shows an issue that was identified with the Runtime attribute in column G. There were approximately 264 records with this attribute missing. To rectify this, a question mark character was added to all of the missing values.

A cursory glance through the movie titles in column B with missing runtimes showed that the vast majority of them tended not to be movies with English as the predominant language. There is no known explanation for this, however it could possibly be due to differences in how movies that originated in different countries were originally recorded in the database.

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Title	Genres	Budget	Revenue	Release Date	Runtime	Popularity	Vote Average	Vote Count	IMDB Rating
933	55254	Uuno Turhapuro - pisinä	[[{"id": 35, "name": "Comedy"}]]	0	0	28/08/1998		0.107109	6.7	3	tt0161052
1335	62725	Delitto a Porta Romana	[[{"id": 35, "name": "Comedy"}]]	0	0	01/01/1980		0.027777	6.6	6	tt0080606
2350	290043	Mörderische Erpressung	[[{"id": 80, "name": "Crime"}]]	0	0	03/11/2006		0.003565	0	0	tt0884138
2791	367596	Täter-Ärztin - Die Klinik	[[{"id": 18, "name": "Drama"}, {"id": 35, "name": "Comedy"}]]	0	0	25/06/2015		0.291696	4.5	4	tt5133288
3146	63179	Totò & Sciecò	[[{"id": 35, "name": "Comedy"}]]	0	0	30/11/1950		0.566432	6.5	8	tt0043059
3512	278475	Äzvekat Yerimdar	[[{"id": 10749, "name": "Romance"}, {"id": 35, "name": "Comedy"}]]	0	0	25/10/2013		0.098004	7.5	2	tt3283664
5037	287305	Peanuts & Die Bank	[[{"id": 35, "name": "Comedy"}]]	0	0	21/03/1998		0.066123	4	1	tt0117312
5038	339428	Happy Weekend	[[{"id": 35, "name": "Comedy"}]]	0	65335	14/03/1998		0.002229	0	0	tt0116485
5045	278978	Und keiner weint mir	[[{"id": 18, "name": "Drama"}]]	0	0	29/02/1998		0.439989	0	0	tt0118026
5188	282919	Diebinnen	[[{"id": 18, "name": "Drama"}]]	0	0	20/06/1998		0.106345	4	1	tt0112865
5242	253632	Baton Rouge	[[{"id": 53, "name": "Thriller"}]]	0	0	08/10/1988		0.437895	0	0	tt0094822
6291	77314	The Cabinet of Dr. Rami	[[{"id": 35, "name": "Comedy"}]]	0	0	04/12/1993		0.0375	0	0	tt0101530
16224	55748	Ystäväni Henry	[[{"id": 35, "name": "Comedy"}]]	0	0	20/08/2004		0.002524	0	0	tt0387701
16237	61070	Pekka ja Pätkä lumin	[[{"id": 35, "name": "Comedy"}]]	0	0	30/07/1954		0.107042	5.7	3	tt0047338
16308	210487	Człowiek z M-3	[[{"id": 35, "name": "Comedy"}]]	0	0	20/02/1963		0.162404	7	1	tt0064203
16394	200796	Show	[[{"id": 35, "name": "Comedy"}, {"id": 18, "name": "Drama"}]]	0	0	19/03/2003		0.149818	6.3	2	
17327	62775	Havukka-Ahon Ajattelij	[[{"id": 35, "name": "Comedy"}, {"id": 18, "name": "Drama"}]]	2225000	0	15/01/2010		0.164919	6.3	3	tt1404661
17436	174748	Lars i porten	[[{"id": 35, "name": "Comedy"}]]	0	0	30/12/1984		0	0	0	tt0087587
17748	173865	Esa ja Vesa & auringor	[[{"id": 35, "name": "Comedy"}]]	3120375	0	10/11/1994		0.205786	6.5	6	tt0109744
17875	46341	El príncimo Oriente	[[{"id": 35, "name": "Comedy"}]]	0	0	14/08/2008		0.131963	5.2	3	tt0484064
18201	55403	Kulkuri ja joutsen	[[{"id": 35, "name": "Comedy"}]]	0	0	19/02/1995		0.00125	0	0	tt0154750
19281	54354	Eila	[[{"id": 35, "name": "Comedy"}]]	0	0	14/03/2003		0.077438	7.5	2	tt0343624
19811	230680	Nachbarinnen	[[{"id": 18, "name": "Drama"}]]	0	0	10/11/2004		0.147153	6	1	tt0366808
20243	434973	Sokkotanssi	[[{"id": 35, "name": "Comedy"}]]	0	0	12/03/1995		0.001194	0	0	tt0189102
20322	344741	Pajäczki	[[{"id": 35, "name": "Comedy"}]]	0	0			0	0	0	tt0105087
20815	404471	Pälynimurikauppiat	[[{"id": 99, "name": "Documentary"}]]	0	0			0	0	0	tt0107903
20908	399229	Kovat Miehet	[[{"id": 35, "name": "Comedy"}]]	0	0	01/01/1995		0.312543	7	1	tt0236398
21133	61361	L'outremangeur	[[{"id": 80, "name": "Crime"}, {"id": 18, "name": "Drama"}]]	0	0	10/09/2002		0.550466	8	2	tt0339482
21179	362151	Nuoruuteni savotat	[[{"id": 18, "name": "Drama"}]]	0	0	18/11/1988		0.049706	6	1	tt0095764
21476	458428	Operaciön Cannabis	[[{"id": 28, "name": "Action"}, {"id": 11, "name": "Mystery"}]]	0	0	01/11/2005		0.005935	0	0	tt1542928
21530	79034	La Californie	[[{"id": 18, "name": "Drama"}]]	0	0	26/10/2008		0.004195	0	0	tt0465160
21851	82663		[[{"id": 28, "name": "Action"}, {"id": 53, "name": "Thriller"}]]	0	0			0	0	0	tt0113002
21983	328712	Tarda estate	[[{"id": 18, "name": "Drama"}]]	0	0	04/09/2010		0.005935	0	0	tt2406990
22243	139909	Aurinkotuuli	[[{"id": 35, "name": "Comedy"}]]	0	0			0.000127	0	0	tt0140826
22431	334317	Dä@sirä	[[{"id": 878, "name": "Science Fiction"}]]	0	0	05/04/1993		0.581963	7.5	2	tt0242424
22671	62130	El Big Bang de Tunguska	[[{"id": 99, "name": "Documentary"}]]	0	0	01/06/2008		0.043159	7	1	tt1459858
22736	54356	Emmauksen tiellä	[[{"id": 35, "name": "Comedy"}]]	0	0	26/09/2003		0.07607	7	2	tt0298317

Figure 4: Missing Runtime attribute

The Runtime attribute had the most missing values; however, it was not the only one. A similar check was also performed on the other attributes and there were a few instances where missing values had to be replaced in the Release_Date and Revenue attributes, although not to such a great extent.

The filters in Excel proved to be very effective at rectifying these issues.

3.3 Illegal Characters

The next step in cleaning the data was to ensure that no illegal characters were present in the dataset as Weka would identify these and reject the data. Illegal characters can exist in many different types of software and operating systems and can vary between them meaning that some illegal characters may be unintentionally used.

It was decided to use the find and replace tools in Excel as they would be able to quickly and efficiently perform updates to the whole dataset. The vast majority of the data is in numeric format and it was identified that the primary attribute that would need to be checked for illegal characters was Title.

In particular, the characters/issues that needed to be checked and removed included:

- Spaces between words (replaced with an underscore character)
- Apostrophes - ‘
- Commas - ,
- Percentage symbols - %
- Quotation marks - “
- Question marks - ?

With the exception of spaces between words, which were replaced with an underscore, all other illegal characters were simply removed. This process was not expected to cause any issues with understanding the data.

It was a relatively simple process to use the find and replace tool to address these issues. One problem that became apparent was with the question mark symbol. When using the find and replace tool, Excel identifies a question mark as a wildcard and therefore removes every character from the specified attribute during the replace option. In a dataset of this size, it took some time for Excel to complete this process which then had to be rectified, taking a similar amount of time.

To resolve this issue, a tilde ~ symbol was used in front of the question mark character so that Excel would interpret the query correctly and only replace question mark characters within the text rather than replaces all characters as a wildcard. An example of this can be seen in Figure 5:

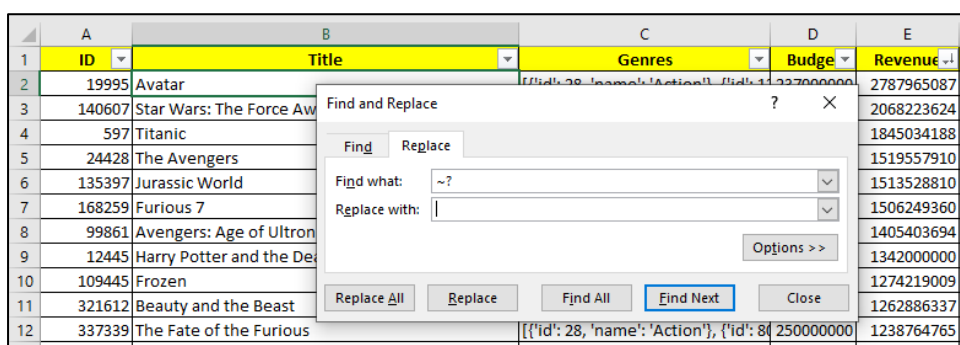


Figure 5: Using the Find and Replace tool in Excel

3.3 Checking for nonsensical data and incomplete records

By using the custom sort tool within Excel, it was possible to sort by different attributes to identify records that did not make sense or were incomplete. These records would not have imported correctly into Weka and need to be dealt with. In total, there were 8 records found out of the 45,467 in the dataset that needed to be removed for these reasons. This figure represented only 0.02% of the total records which was deemed an acceptable figure.

Figure 6 shows how the incomplete records were originally found in the dataset.

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Title	Genres	Budget	Revenue	Release_Dat	Runtime	Popularity	Vote_Average	Vote_Count	IMDB_Rating
45460	11193	TRUE	[{"id": 18, "name": "Drama"}]	0	0	06/02/2004	10	1.680105	6	13	tt0397693
45461	72478	#NAME?	[{"id": 16, "name": "Animation"}, {"id": 10751, "name": "Family"}]	0	0	04/03/2011	110	3.742632	7.7	113	tt1808454
45462	122662		[{"id": 16, "name": "Animation"}, {"id": 10751, "name": "Family"}]	0							tt2423504
45463	82663		[{"id": 28, "name": "Action"}, {"id": 35, "name": "Comedy"}]	0							tt0113002
45464	249260		[{"id": 10770, "name": "TV Movie"}, {"id": 10751, "name": "Family"}]	0							tt2622826
45465		20/08/1997		0	1	/ff9qCepilowshEtG2GYWwzt2bs4.jpg					
45466		29/09/2012		0	12	/zV8bHuSL6WxoD6FWogP9j4x80bL.jpg					
45467		01/01/2014		0	22	/zaSf5OG7V8X8gqFvly88zDdR	Beware Of Frost Bites				

Figure 6: Incomplete or nonsensical records which were removed

3.4 Splitting an attribute into separate groups

In order to perform association analysis algorithms on the Genre attribute in Weka, the current structure of the attribute will need to be amended. Presently, the data is all grouped together in a similar format which looks as if it has been previously exported from another software. Column C in Figure 7 shows this:

	A	B	C
1	ID	Title	Genres
21	10193	Toy Story 3	[{"id": 16, "name": "Animation"}, {"id": 10751, "name": "Family"}, {"id": 35, "name": "Comedy"}]

Figure 7: The Genre attribute before splitting

Here you can see that the movie Toy Story 3 has been assigned the following genres:

- Animation
- Family
- Comedy

In order to split the genres and assign them individually to each movie, it was decided to use the built-in functions in Excel.

The first step in the process was to identify all of the different genres that existed in the dataset. This resulted in 20 genres being found:

- Action
- Adventure
- Animation
- Comedy
- Crime
- Documentary
- Drama
- Family
- Fantasy
- Foreign
- History
- Horror
- Music
- Mystery
- Romance
- Science Fiction
- Thriller
- TV Movie
- War
- Western

Next, it was necessary to add columns to the end of the dataset file using the different genres as column headings, as seen in Figure 8:

	A	B	D	E	H	I	J	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	ID	Title	Budget	Revenue	Popularity	Vote_Average	Vote_Count	Action	Adventure	Animation	Comedy	Crime	Documentary	Drama	Family	Fantasy	Foreign	History	Horror	Music	Mystery	Romance	Science Fiction	Thriller	TV Movie	War	Western
21	10193	Toy Story 3	200000000	1066969703	16.96647	7.6	4710																				
22	38	Pirates of the Caribbean: The Curse of the Black Pearl	200000000	106659812	29.455314	7	5380																				
23	330459	Rogue One: A Star Wars Story	200000000	1056057273	36.567575	7.4	5111																				
24	1865	Pirates of the Caribbean: On Stranger Tides	380000000	1045713802	27.88772	6.4	5068																				
25	127386	Finding Dory	200000000	1028570899	14.477877	6.8	4393																				
26	12155	Alice in Wonderland	200000000	1025491110	17.285093	6.4	4773																				
27	269149	Zootopia	150000000	1023784195	26.024868	7.7	4961																				
28	49051	The Hobbit: An Unexpected Journey	250000000	1021103568	23.253089	7	8427																				
29	324852	Despicable Me 3	80000000	1020683384	36.813129	6.2	2002																				
30	155	The Dark Knight	185000000	100458444	123.167299	8.3	12269																				
31	671	Harry Potter and the Chamber of Secrets	125000000	976475550	38.187238	7.5	7188																				
32	93456	Despicable Me 2	76000000	970761885	24.82355	7	4729																				
33	278927	The Jungle Book	175000000	966550600	17.432758	6.7	2977																				
34	285	Pirates of the Caribbean: The Curse of the Black Pearl	200000000	961000000	31.363664	6.9	4627																				
35	57158	The Hobbit: The Desolation of Smaug	250000000	958400000	20.644776	7.6	4633																				
36	122917	The Hobbit: The Battle of the Five Armies	250000000	956019788	31.718977	7.1	4884																				

Figure 8: The keyword columns added to the dataset file

Once the columns had been added it was then possible to use Excel functions to identify if the genre keyword was present in the string of the existing genre attribute. This was accomplished by using the following function:

=ISNUMBER(SEARCH("Action",C1))

An example of this can be seen in the formula bar in Figure 9:

L21												=ISNUMBER(SEARCH("Action",C1))	
	A	B	C	D	E	H	I	J	L	M	N		
1	ID	Title	Genres	Budget	Revenue	Popularity	Vote_Average	Vote_Count	Action	Adventure	Animation		
21	10193	Toy Story 3	{{'id': 16, 'name': 'Animation'}, {'id': 16, 'name': 'Animation'}}	200000000	1066969703	16.96647	7.6	4710	FALSE				
22	38	Pirates of the Caribbean: The Curse of the Black Pearl	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	200000000	1065659812	29.455314	7	5380					
23	330459	Rogue One: A Star Wars Story	{{'id': 28, 'name': 'Action'}, {'id': 12, 'name': 'Adventure'}}	200000000	1056057273	36.567575	7.4	5111					
24	1865	Pirates of the Caribbean: On Stranger Tides	{{'id': 12, 'name': 'Adventure'}, {'id': 12, 'name': 'Adventure'}}	380000000	1045713802	27.88772	6.4	5068					

Figure 9: Using a function to find a keyword in a string

This function would search through the string within the cell and return True if the genre was found and False if it was not found. Once the results of this function had been checked to see if it was working correctly, it could then be replicated down to the following rows and across columns. Upon successful completion, the original Genre attribute could then be removed from the dataset as it is no longer needed.

One issue that became evident at this stage was the need to ensure the genre name had been spelt correctly within the function. If for any reason a genre name was typed incorrectly the function would have been unable to find the text within the string and this could easily have resulted in the results being distorted during the data mining stage.

Completing this process and adding the extra columns has meant that the number of attributes in the dataset has now grown to 30, although this does mean that the size of the dataset file has increased to 7Mb.

However, despite a marginal increase in the file size it is still relatively small. Having all of the attributes stored together in one dataset file allows flexibility when analysing the data in Weka. The pre-processing tools in Weka will enable attributes to be temporarily removed during different operations and will also allow them to be restored at a later time. These temporary changes will not affect the master dataset.

3.5 Restructuring date values

The next step in the pre-processing stage was to rearrange the date values in the dataset as they were not in the correct format to be imported into Weka.

The original date values were all structured as “DD-MM-YYYY”, but for working with Weka they need to be rearranged to “YYYY-MM-DD”. These amendments can easily be made using the Format Cells options in Excel and choosing a different method of displaying the date, as seen in Figure 10:

The date attribute in its original format:

The screenshot shows an Excel spreadsheet with columns A through L. Column A contains 'ID', B contains 'Title', C contains 'Budget', D contains 'Revenue', and E contains 'Release Date'. The 'Release Date' column is highlighted with a red box. A 'Format Cells' dialog box is open, showing the 'Date' category and the '2012-03-14' format selected. The dialog box also shows a 'Sample' of the format: '2009-12-10'.

ID	Title	Budget	Revenue	Release Date
19995	Avatar	237000000	2787965087	10/12/2009
140607	Star Wars: The Force A	245000000	2068223624	15/12/2015
597	Titanic	200000000	1845034188	18/11/1997
24428	The Avengers	220000000	1519557910	25/04/2012
135397	Jurassic World	150000000	1513528810	09/06/2015
168259	Furious 7	190000000	1506249360	01/04/2015
99861	Avengers: Age of Ultr	280000000	1405403694	22/04/2015
12445	Harry Potter and the D	125000000	1342000000	07/07/2011
109445	Frozen	150000000	1274219005	27/11/2013
321612	Beauty and the Beast	160000000	1262886337	16/03/2017
337339	The Fate of the Furiou	250000000	1238764765	12/04/2017
68721	Iron Man 3	200000000	1215439994	18/04/2013
211672	Minions	74000000	1156730962	17/06/2015
271110	Captain America: Civil	250000000	1153304495	27/04/2016
38356	Transformers: Dark of	195000000	1123746996	28/06/2011
122	The Lord of the Rings:	94000000	1118888979	01/12/2003
37724	Skyfall	200000000	1108561013	25/10/2012
91314	Transformers: Age of t	210000000	1091405097	25/06/2014
49026	The Dark Knight Rises	250000000	1084939099	16/07/2012
10193	Toy Story 3	200000000	1066969703	16/06/2010
58	Pirates of the Caribbe	200000000	1065659812	20/06/2006
330459	Rogue One: A Star Wa	200000000	1056057273	14/12/2016
1865	Pirates of the Caribbe	380000000	1045713802	14/05/2011

The date attribute in its amended format:

ID	Title	Budget	Revenue	Release Date	Runtime	Popularity	Vote Ave	Vote Cou	IMDB ID	Action	Adventure	Animation
19995	Avatar	237000000	2787965087	2009-12-10	162	185.0709	7.2	12114	tt0499549	TRUE	TRUE	FALSE
140607	Star Wars: The Force A	245000000	2068223624	2015-12-15	136	31.62601	7.5	7993	tt2488496	TRUE	TRUE	FALSE
597	Titanic	200000000	1845034188	1997-11-18	194	26.88907	7.5	7770	tt0120338	FALSE	FALSE	FALSE
24428	The Aven	220000000	1519557910	2012-04-25	143	89.88765	7.4	12000	tt0848228	TRUE	TRUE	FALSE
135397	Jurassic W	150000000	1513528810	2015-06-09	124	32.79048	6.5	8842	tt0369610	TRUE	TRUE	FALSE
168259	Furious 7	190000000	1506249360	2015-04-01	137	27.27569	7.3	4253	tt2820852	TRUE	FALSE	FALSE
99861	Avengers:	280000000	1405403694	2015-04-22	141	37.37942	7.3	6908	tt2395427	TRUE	TRUE	FALSE
12445	Harry Pott	125000000	1342000000	2011-07-07	130	24.99074	7.9	6141	tt1201607	FALSE	TRUE	FALSE
109445	Frozen	150000000	1274219005	2013-11-27	102	24.24824	7.3	5440	tt2294629	FALSE	TRUE	TRUE
321612	Beauty an	160000000	1262886337	2017-03-16	129	287.2537	6.8	5530	tt2771200	FALSE	FALSE	FALSE
337339	The Fate d	250000000	1238764765	2017-04-12	136	48.57329	6.8	3803	tt4630562	TRUE	FALSE	FALSE
68721	Iron Man	200000000	1215439994	2013-04-18	130	23.72124	6.8	8951	tt1300854	TRUE	TRUE	FALSE
211672	Minions	74000000	1156730962	2015-06-17	91	547.4883	6.4	4729	tt2293640	FALSE	TRUE	TRUE
271110	Captain A	250000000	1153304495	2016-04-27	147	145.8821	7.1	7462	tt3498820	TRUE	TRUE	FALSE
38356	Transform	195000000	1123746996	2011-06-28	154	4.503505	6.1	3351	tt1399103	TRUE	TRUE	FALSE
122	The Lord o	94000000	1118888979	2003-12-01	201	29.32436	8.1	8226	tt0167260	TRUE	TRUE	FALSE
37724	Skyfall	200000000	1108561013	2012-10-25	143	20.30959	6.9	7718	tt1074638	TRUE	TRUE	FALSE
91314	Transform	210000000	1091405097	2014-06-25	165	18.4956	5.8	3136	tt2109248	TRUE	TRUE	FALSE
49026	The Dark	250000000	1084939099	2012-07-16	165	20.58258	7.6	9263	tt1345836	TRUE	FALSE	FALSE
10193	Toy Story	200000000	1066969703	2010-06-16	103	16.96647	7.6	4710	tt0435761	FALSE	FALSE	TRUE
58	Pirates of	200000000	1065659812	2006-06-20	151	29.45531	7	5380	tt0383574	TRUE	TRUE	FALSE
330459	Rogue On	200000000	1056057273	2016-12-14	133	36.56758	7.4	5111	tt3748528	TRUE	TRUE	FALSE

Figure 10: Changing the date structure for Weka

3.6 Creating the ARFF file for Weka

Once the dataset has been cleaned and made ready for Weka it needs to be structured correctly in ARFF format so that it will be recognised by the software. This was achieved by opening the updated CSV dataset file in Notepad++ and adding the following header section for the file:

```
@RELATION genres
@ATTRIBUTE ID Numeric
@ATTRIBUTE Budget Numeric
@ATTRIBUTE Revenue Numeric
@ATTRIBUTE Release_Date Date "YYYY-mm-DD"
@ATTRIBUTE Runtime Numeric
@ATTRIBUTE Popularity Numeric
@ATTRIBUTE Vote_Average Numeric
@ATTRIBUTE Vote_Count Numeric
@ATTRIBUTE Action {FALSE,TRUE}
@ATTRIBUTE Adventure {FALSE,TRUE}
@ATTRIBUTE Animation {FALSE,TRUE}
@ATTRIBUTE Comedy {FALSE,TRUE}
@ATTRIBUTE Crime {FALSE,TRUE}
@ATTRIBUTE Documentary {FALSE,TRUE}
@ATTRIBUTE Drama {FALSE,TRUE}
@ATTRIBUTE Family {FALSE,TRUE}
@ATTRIBUTE Fantasy {FALSE,TRUE}
@ATTRIBUTE Foreign {FALSE,TRUE}
@ATTRIBUTE History {FALSE,TRUE}
@ATTRIBUTE Horror {FALSE,TRUE}
@ATTRIBUTE Music {FALSE,TRUE}
@ATTRIBUTE Mystery {FALSE,TRUE}
@ATTRIBUTE Romance {FALSE,TRUE}
@ATTRIBUTE Science_Fiction {FALSE,TRUE}
@ATTRIBUTE Thriller {FALSE,TRUE}
@ATTRIBUTE TV_Movie {FALSE,TRUE}
@ATTRIBUTE War {FALSE,TRUE}
@ATTRIBUTE Western {FALSE,TRUE}
@DATA
```

It is important to ensure that the data types alongside each attribute are correct otherwise Weka will report errors during the import stage. Once complete, the file then needed to be saved with a '.arff' extension so that it would be recognised and opened by Weka.

4. Data Visualisation in Weka

The data was successfully imported into Weka . The first step was then to create a visualization to see if any interesting patterns could be found prior to the data mining processes being run. The ID and release date of each movie as well as all 20 of the genres were removed from the visualisation as they were not deemed to be suitable for this stage.

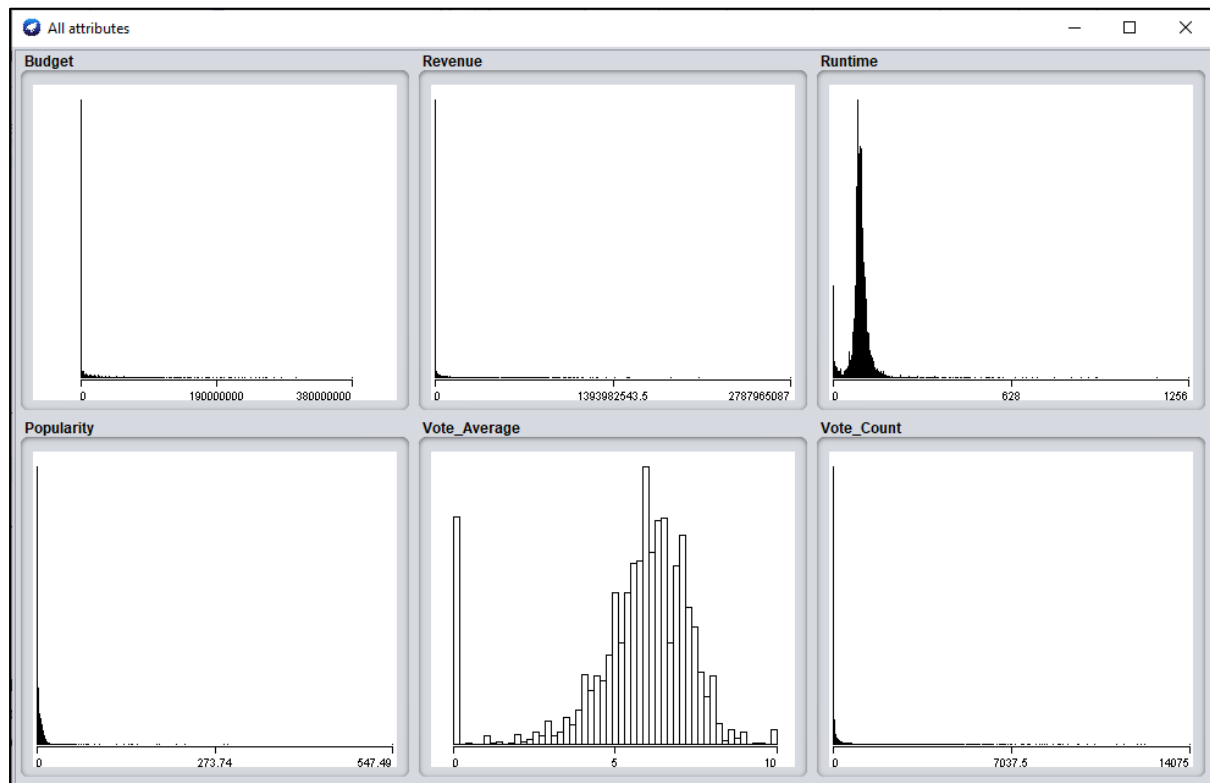


Figure 11: Initial Weka visualisation

From this visualisation, it would appear that the `Vote_Average` attribute seems to peak around a value of 6 with a fairly even distribution either side. There also appears to be a spike in the `Runtime` attribute at the lower end of the graph, suggesting that the majority of the movies in the dataset were of a shorter runtime.

However, it is clear to see that not much else can be concluded from this visualisation, therefore it will be necessary to use Tableau in order to explore the dataset in greater detail using the sophisticated tools and techniques available.

5. Data Analysis

There are 4 questions that will attempt to be answered by investigating the dataset:

Question	Technique Used
1. Is it possible to predict the popularity score for a movie based on the existing records in the dataset?	Weka Classification
2. Are there any patterns contained in the genres? Are there any genres that feature frequently in the dataset?	Weka Association Analysis
3. Is there any relation between the budget for a movie and the revenue it makes?	Tableau Visualisation
4. Is there a relationship between Runtime and Revenue?	Tableau Visualisation

5.1 Question 1 - Is it possible to predict the popularity score for a movie based on the existing records in the dataset?

This question requires classification techniques to be applied to the dataset. Figure 12 shows that the Popularity attribute values range from 0 to 547.488, with a mean value of 2.921 and a standard deviation of 6.005. This would suggest that the Popularity values are somewhat skewed and the values at the higher end of the scale are not commonly found in the dataset.

Selected attribute	
Name: Popularity	Type: Numeric
Missing: 0 (0%)	Distinct: 43757
	Unique: 42977 (95%)
Statistic	Value
Minimum	0
Maximum	547.488
Mean	2.921
StdDev	6.005

Figure 12: Popularity attribute values

The top 10 movies with the highest popularity score are shown in Figure 13:

	A	B	C	D	E	F	G
1	ID	Title	Budget	Revenue	Release_Date	Runtime	Popularity
2	211672	Minions	74000000	1156730962	17/06/2015	91	547.4883
3	297762	Wonder Woman	149000000	820580447	30/05/2017	141	294.33704
4	321612	Beauty and the Beast	160000000	1262886337	16/03/2017	129	287.25365
5	339403	Baby Driver	34000000	224511319	28/06/2017	113	228.03274
6	177572	Big Hero 6	165000000	652105443	24/10/2014	102	213.84991
7	293660	Deadpool	58000000	783112979	09/02/2016	108	187.86049
8	283995	Guardians of the Galaxy	200000000	863416141	19/04/2017	137	185.33099
9	19995	Avatar	237000000	2787965087	10/12/2009	162	185.07089
10	245891	John Wick	20000000	88761661	22/10/2014	101	183.87037
11	210577	Gone Girl	61000000	369330363	01/10/2014	145	154.80101

Figure 13: Top 10 movies with highest Popularity score

There are some very well-known and blockbuster movies in this list. However, what is not clear is how the Popularity score is assigned by TMDb. Viewers of a movie are able to provide their own rating, but this is stored in a separate attribute and is on a scale of 1 to 10. There is nothing on the TMDb website to indicate how this score is calculated. One possible suggestion could be to do with the date that a movie was released. All of the movies in this list were released within the last decade. As the growth and popularity of the Internet and social media have continued to grow, maybe this has had an effect on the Popularity score of newer movies resulting in them being higher than older movies in the dataset.

For example, the TMDb website sometimes provides trailers for potential viewers to see if they would be interested in watching a film. Maybe the use of social media and internet enabled devices in the last decade has driven more users to the website and had an effect on the Popularity score. However, this is just a theory and cannot be proven.

As the Popularity scores varied so much, it was anticipated that this could cause inaccurate results to be reported by Weka. To try and combat this, it was decided to manually group movies together into 8 evenly spread groups or buckets numbered 1 to 8, where 1 would be movies that were least popular and 8 would be movies that were most popular.

This meant going back to Excel, adding an additional column to hold the bucket number, and manually assigning a bucket number to each movie. There were approximately 5,682 movies in each bucket. The data was then reloaded back into Weka.

Next, it was necessary to remove any attributes that were not suitable to answer this question. The attributes removed were ID, Title, Budget, Revenue, Release_Date, Runtime, Vote_Average, Vote_Count and IMDB_ID.

This left 21 attributes available as seen in Figure 14:

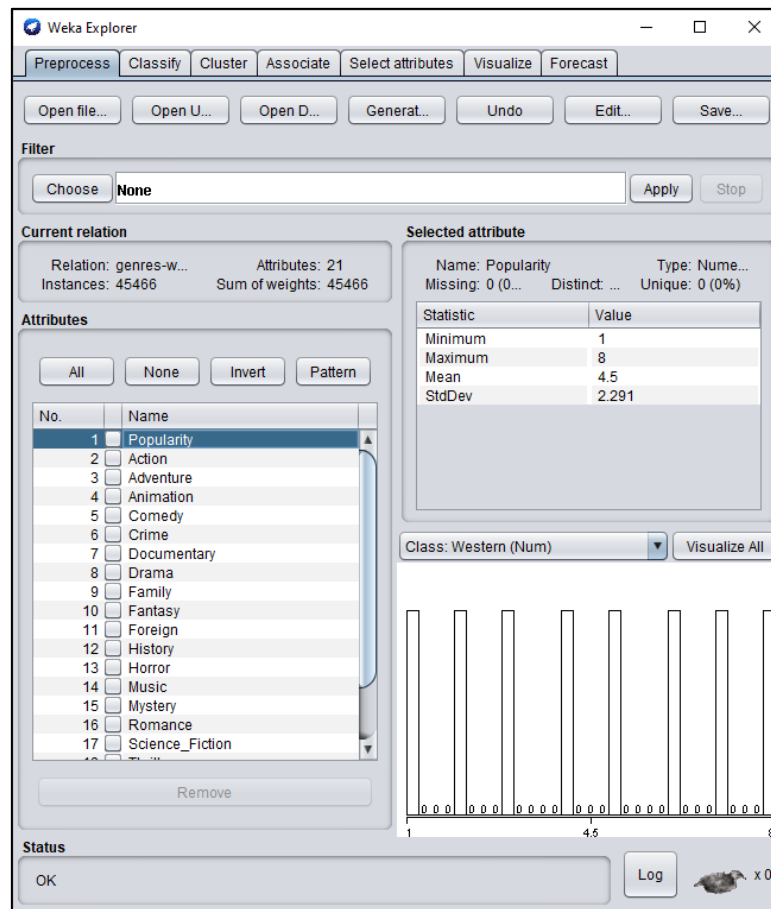


Figure 14: The dataset loaded into Weka

The Popularity attribute is the class in this investigation, so it was necessary to use the Edit tool to make this change, as seen in Figure 15.

By default, when a class is defined it then moves to the bottom of the list of attributes.

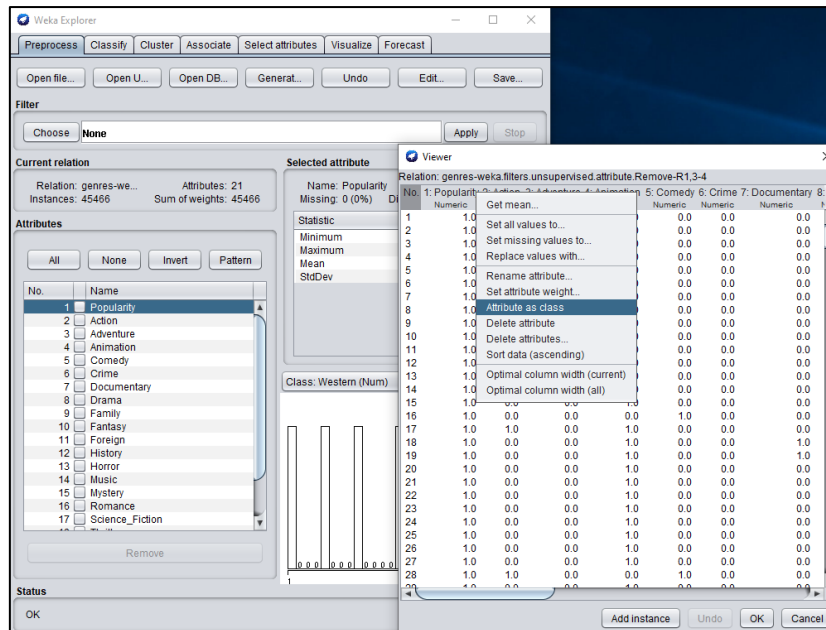


Figure 15: Setting the Popularity attribute as a class

Classification algorithms cannot be run on a dataset if the class attribute is numeric. This caused an issue with the Popularity data which was all in numeric format. Fortunately, Weka has built-in filters which can be used to convert data into nominal format to make it usable.

Figure 16 shows the NumericToNominal filter has been applied and has successfully changed the type of data in Popularity to nominal. You can also see the 8 evenly spread buckets in the preview window.

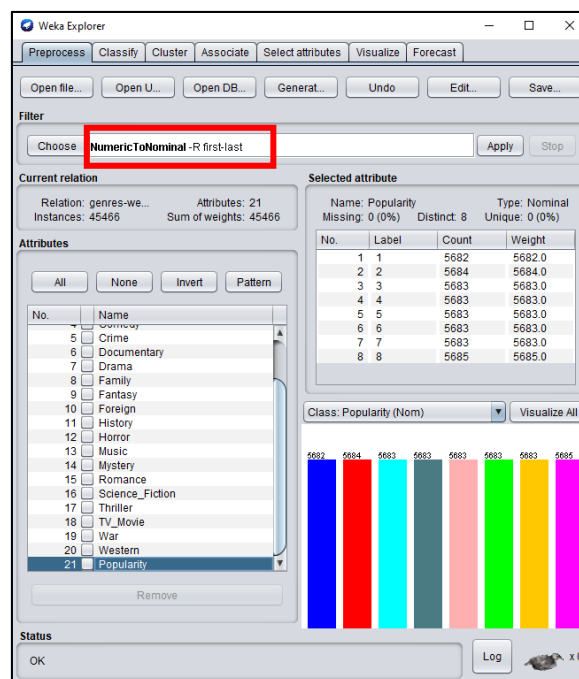


Figure 16: The NumericToNominal filter being applied

Now the dataset is ready for classification. The algorithm that will be used to try and extract rulesets from the data is J48. According to Witten, Eibe and Hall [6] J48 is a decision tree algorithm in Weka which is a derivative of the C4.5 algorithm. It uses simple divide and conquer methods in order to determine if rules can be found.

The main objective for this question is to find out if the algorithm will be able to identify how to classify the popularity of future movies based on existing data.

As the dataset is 45,000+ records, it was decided to use a percentage split. This will allow a subset of the dataset to be explored by the algorithm and the results can then be tested on the remaining data to check for accuracy. The algorithm was run 3 times using a different percentage split for each occurrence.

The results for this method can be found in the table below:

Percentage split	Correct classification	Incorrect classification
75%	21.13%	78.86%
66%	21.11%	78.88%
55%	20.75%	79.24%

There appeared to be very little difference in the results when running the different percentage splits. The highest number of correct classifications (21.13%) was achieved using a 75% split, the results of which can be seen in Figure 17.

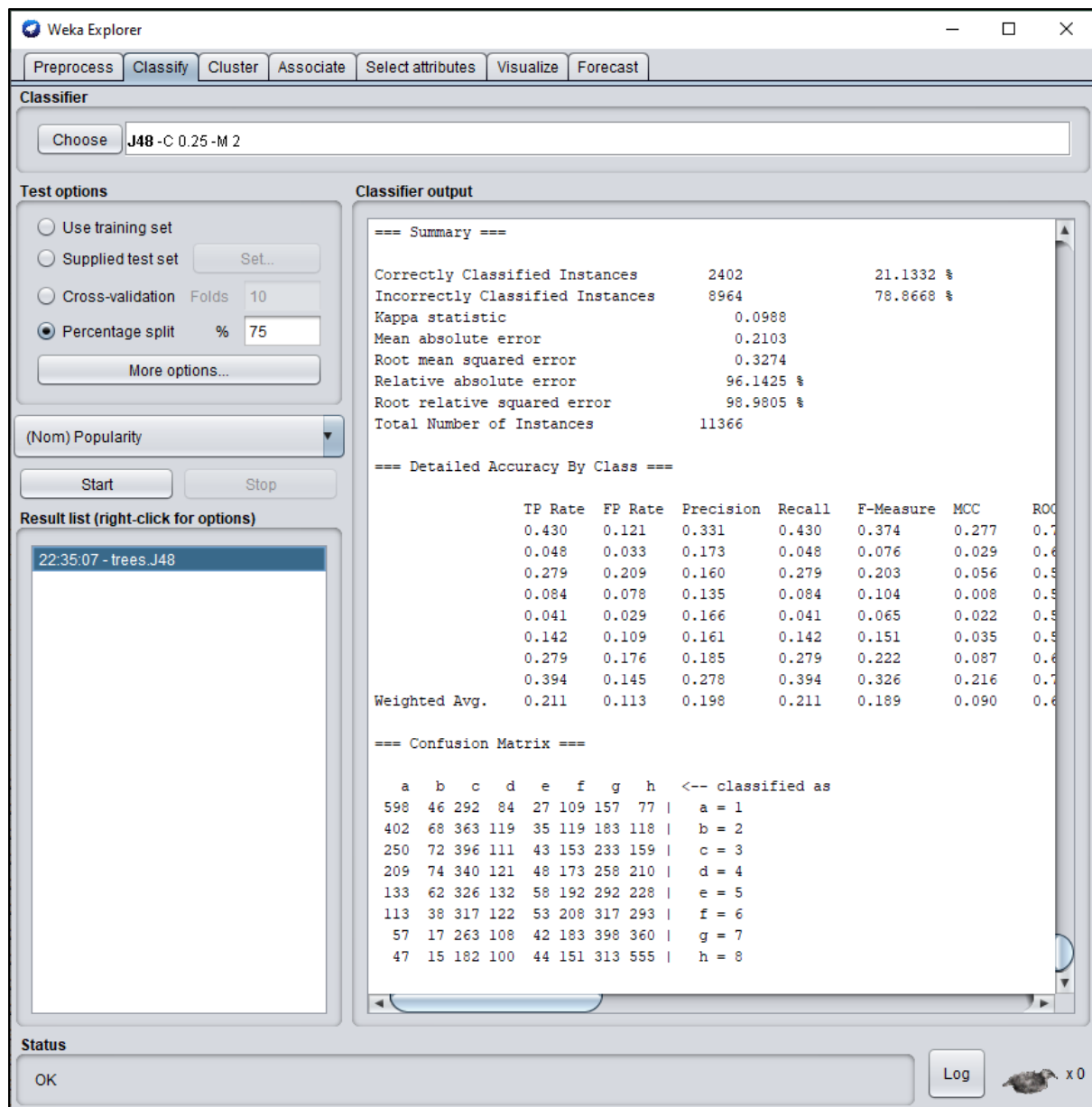


Figure 17: Results of 75% split classification

It was initially hoped that the algorithm would return a higher number of correctly classified instances. One theory is that having the 8 different buckets has resulted in too many different leaves of the decision tree and therefore produced more inaccurate classifications. Indeed looking at the confusion matrix in Figure 17, it would seem that there is a fairly even spread of numbers between the different groups.

In an attempt to try and improve the accuracy of the model, it was decided to modify the dataset by reducing the number of groups/buckets to 4 instead of 8. It was hoped that a reduction in groups could lead to a slimmed down decision tree and more movies being correctly classified.

The first step was to modify the dataset in Excel. The 8 assigned bucket numbers were removed and replaced with numbers 1 to 4 instead, with 1 being lower popularity and 4 being higher popularity. Again, the data was spread evenly between the 4 buckets which increased the number of movies per group/bucket to approximately 11,365.

Figure 18 shows the reimported dataset in Weka and the updated groupings for Popularity:

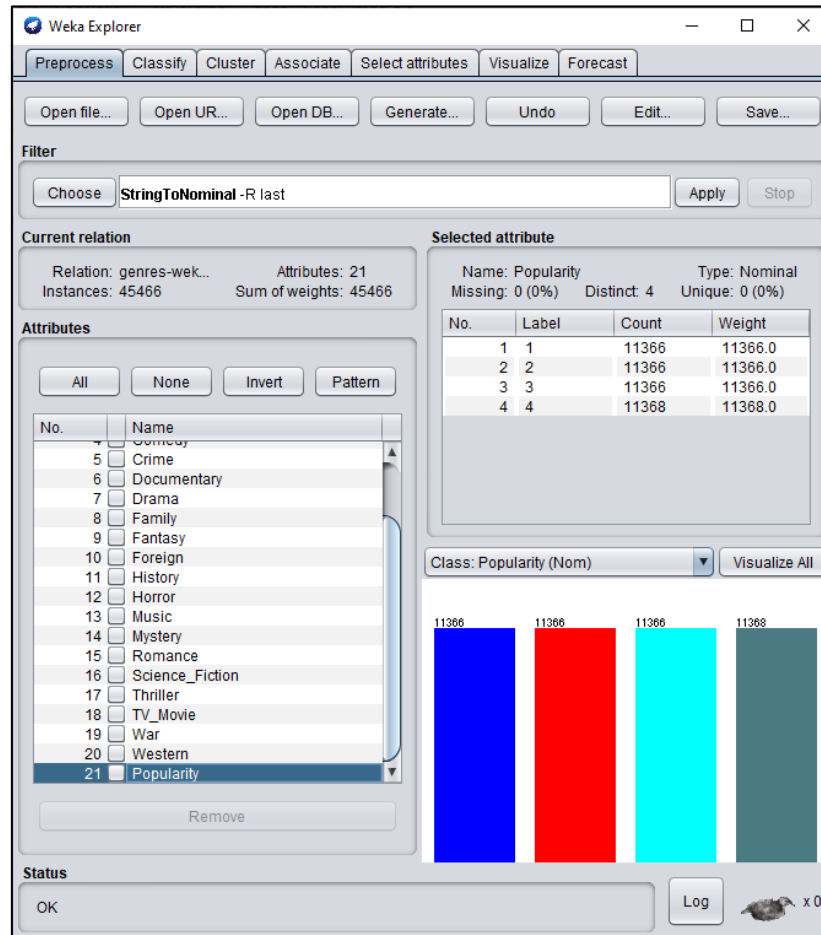


Figure 18: Updated dataset groupings

To ensure a fair comparison, the same percentage splits were used with the updated dataset. The results can be found in the table below:

Percentage split	Correct classification	Incorrect classification
75%	37.70%	62.30%
66%	38.10%	61.90%
55%	37.68%	62.31%

Again, there appeared to be very little difference in the results when running the different percentage splits. The highest number of correct classifications (38.10%) was achieved using a 66% split, as can be seen in Figure 19.

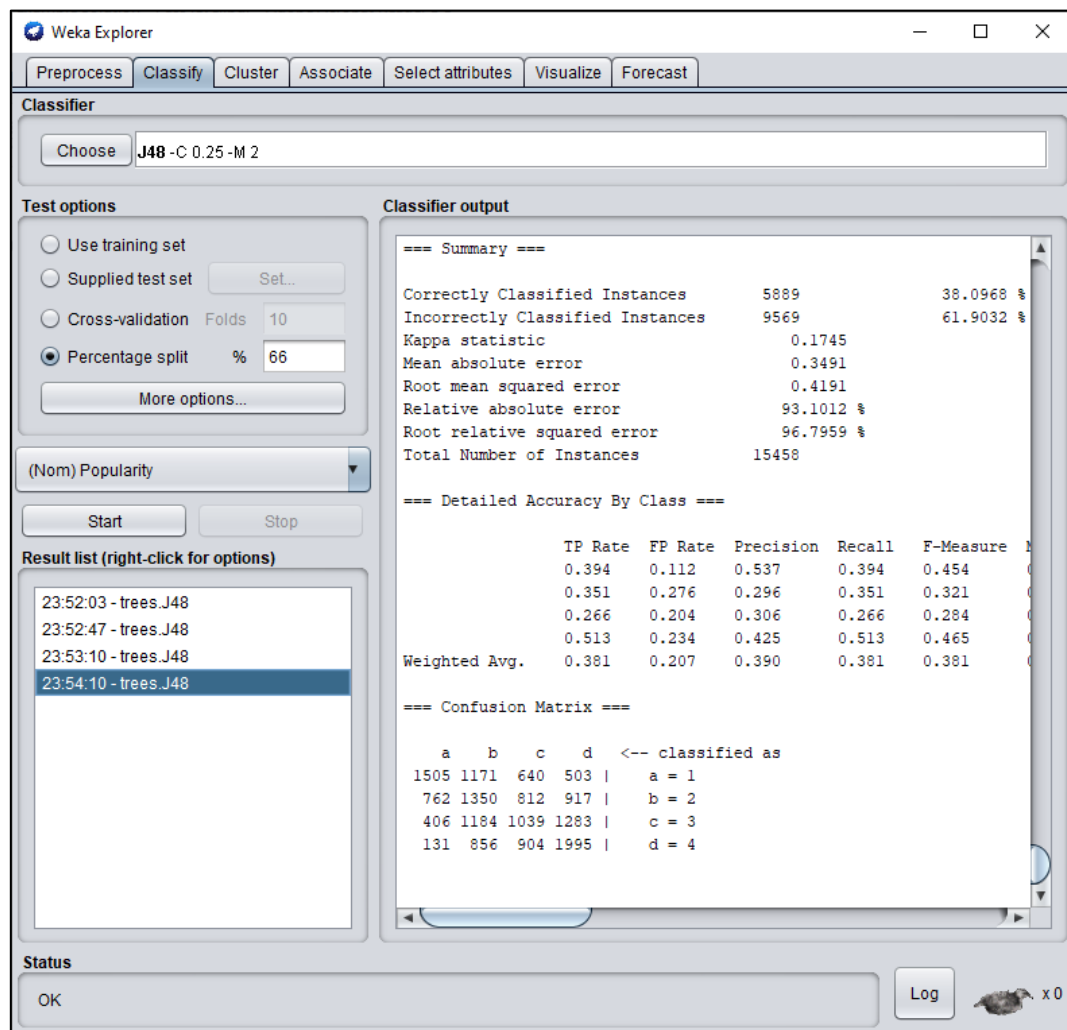


Figure 19: Results of the 66% percentage split classification

It is now evident that reducing the number of buckets from 8 to 4 has led to an improvement in the classification of movies. Similarly, if the number of buckets was reduced again it is very likely that the correctly classified number would increase.

However, this method of thinking presents problems. By reducing the number of buckets from 8 to 4, correct classifications have indeed gone up. This would be because the range of each bucket has become broader allowing more movies to be grouped together in the same bucket. However, there could be very big extremes in the Popularity score within each bucket. Depending on how these results were planning to be used, it might not be realistic to assume that there could only be 4 possible

buckets that a movie could be assigned to. This is where domain knowledge would come in useful to be able to help understand what would make this process more effective.

It is also worth looking at the new confusion matrix to see if any conclusions can be drawn:

=== Confusion Matrix ===					
a	b	c	d		<-- classified as
1505	1171	640	503		a = 1
762	1350	812	917		b = 2
406	1184	1039	1283		c = 3
131	856	904	1995		d = 4

Figure 20: Confusion matrix for 66% split / 4 buckets

It would appear that the highest number of correct classifications are achieved in group D (1995) for the most popular movies, followed by group A (1505) with the least popular movies. This would suggest that the model found it easier to classify those movies that were at the extremes of the Popularity score range, rather than those that were in the middle.

Bearing in mind that the Popularity attribute values originally ranged from 0 to 547.488, with a mean value of 2.921 and a standard deviation of 6.005, it was hoped that by allocating movies into a series of groups/buckets the model would be able to better classify them. This does not seem to have been the case for those movies that are not at the higher and lower ends of the Popularity score range. In conclusion, 1 in 3 movies can be correctly classified. Depending on how this model might be used, this may not be accurate enough to be trusted.

5.2 Question 2 - Are there any patterns contained in the genres? Are there any genres that feature frequently in the dataset?

This question requires association analysis techniques to be applied to the dataset. This will focus on applying an association algorithm to the individual Genre attributes that were added in the pre-processing stage.

It was necessary to remove any attributes that were not suitable to answer this question. The attributes removed were ID, Title, Budget, Revenue, Release_Date, Runtime, Popularity, Vote_Average, Vote_Count and IMDB_ID.

This left 20 attributes available as seen in Figure 21:

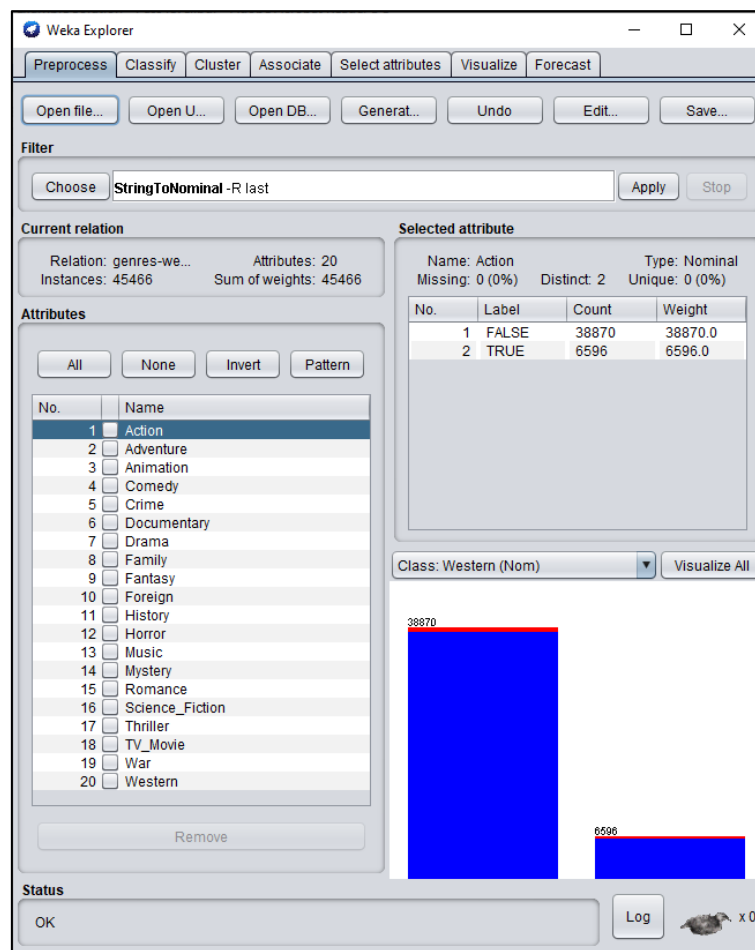


Figure 21: The dataset loaded into Weka

The algorithm chosen for association analysis is Apriori. Witten, Eibe and Hall [6] state that Apriori is effective for finding frequent item sets and looking for patterns between them to generate association rules. It is a popular data mining algorithm often used in market basket analysis for supermarkets. Witten, Eibe and Hall [6] go on to explain that although effective, the generation process of the algorithm can be computationally costly particularly in very large datasets.

The dataset file was again imported into Weka. There were no changes made to the default settings for the algorithm in the first instance. The algorithm should produce 10 rule sets with a minimum confidence level of 90% - which indicates how frequently a rule is found to be true.

The results from the first application of the algorithm can be seen in Figure 22:

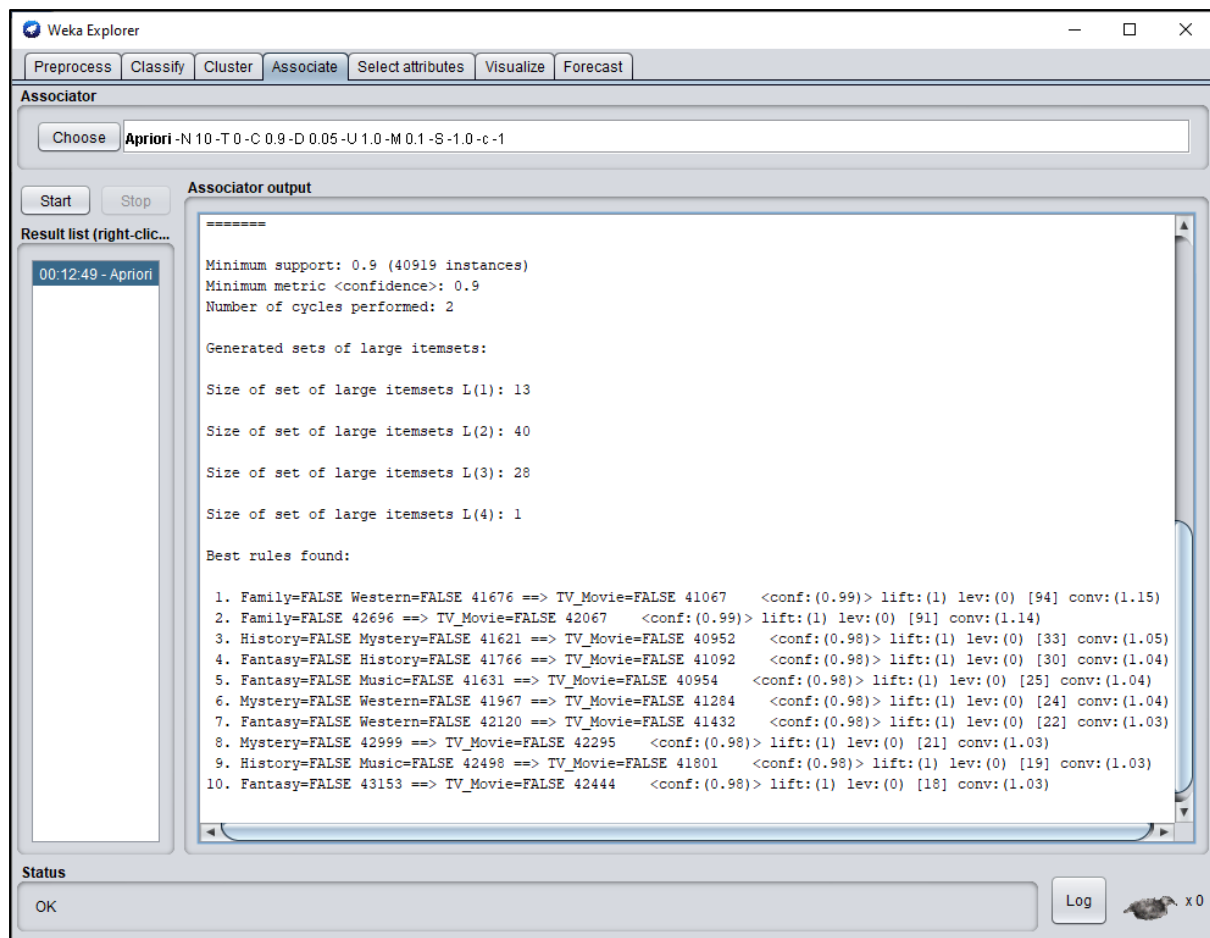


Figure 22: Results of the Apriori algorithm

The algorithm has successfully identified 4 large item sets and 10 association rules. The most confident rule found suggests that if a movie is categorised as Family and Western it is therefore not likely to be a TV Movie. As this is the most confident rule that the algorithm could find, it does raise some questions about the accuracy of the results. Whilst it is possible that a family movie could indeed also be a western, it is not a combination that would normally be associated together.

This prompted further investigation of the rules. It would appear that TV Movie seems to be used in all of the 10 rules found. As an experiment and as part of secondary pre-processing, TV Movie was removed from the attributes list in Weka and the algorithm was run again. The results of this can be seen in Figure 23:

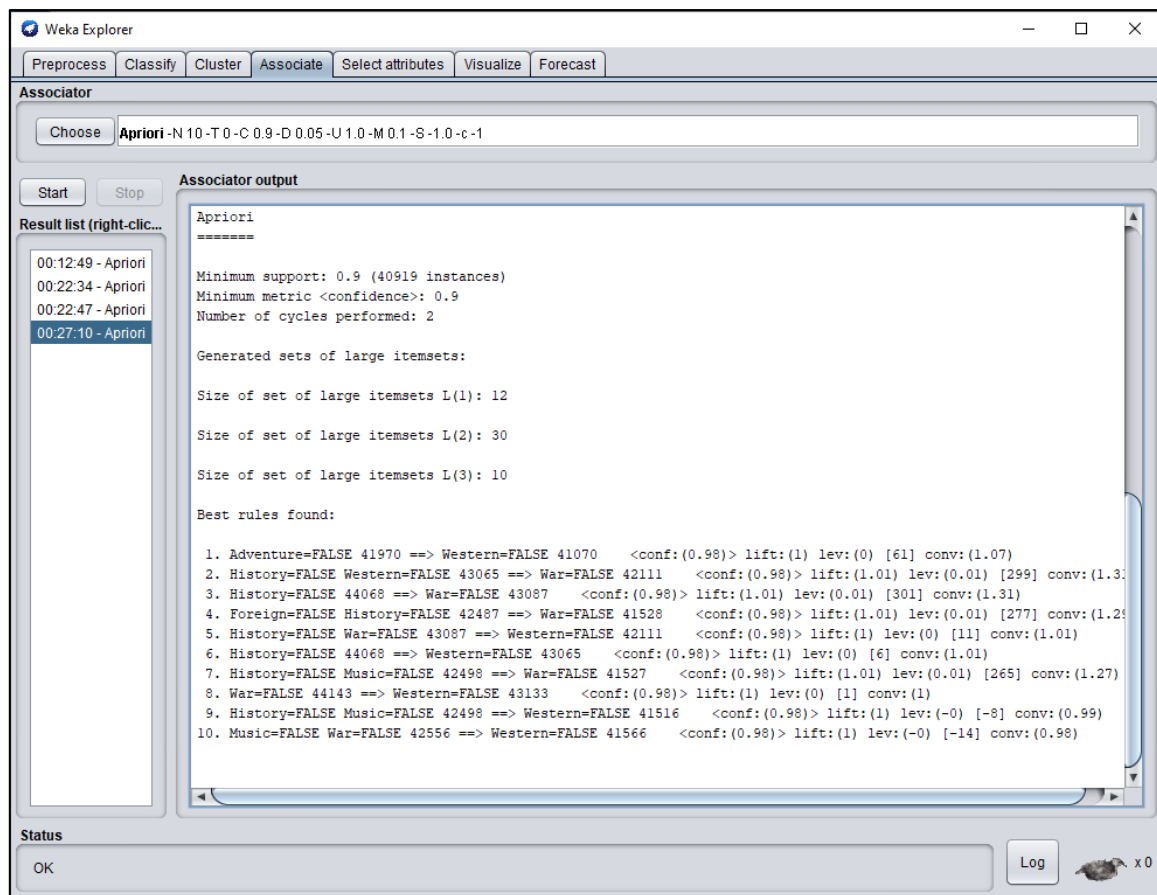


Figure 23: Updated results of the Apriori algorithm

This time TV Movie has been removed from the rules found. The most confident rule on this occasion suggests that a movie that is not categorised as Adventure would therefore not be a Western. The second rule found suggests that a movie that is not categorised as History or Western can therefore not be categorised as War.

There is not a lot of knowledge that can be gained from these rules as there does not seem to be any real patterns or trends.

In addition, upon closer examination of both implementations of the algorithm, it appears that all of the values in the rules have been returned as False. This is to be expected as in total there are 909,180 possible values that can be collected for each genre and each movie (45,459 records x 20 attributes).

91,094 of those values have been recorded as True, which leaves 818,086 False entries in the dataset (True – 10% vs False – 90%). This means that a significant proportion of the dataset will be False and this could therefore have impacted the result of the algorithm.

In order to test this theory, an adjustment was made to the properties of the Apriori algorithm. In order to compensate for the substantial amount of False entries in the dataset the 'treatZeroAsMissing' option has been changed to True, as seen in Figure 24.

This option will analyse the data as before but will treat all False values as being missing, rather than being False. The logic is that the balance should shift in favour of True values and the algorithm may produce different outcomes as a result.

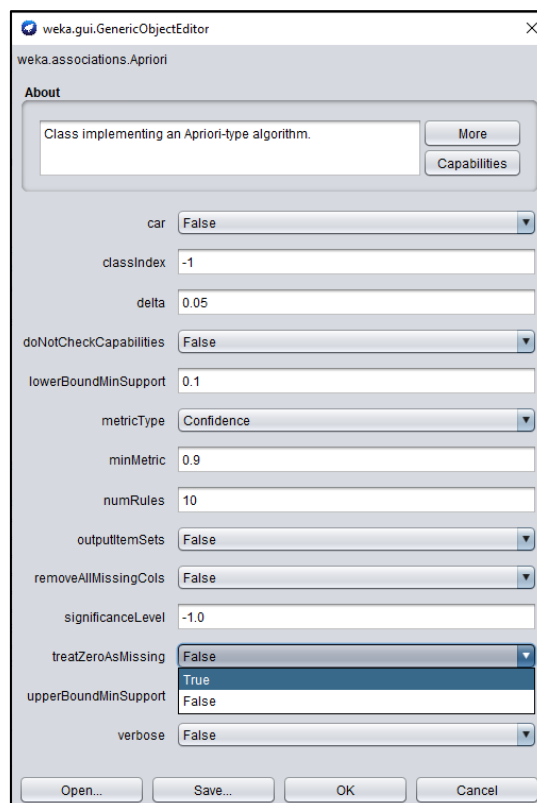


Figure 24: Adjusting the Apriori algorithm

The algorithm was run again and the results can be seen in Figure 25:

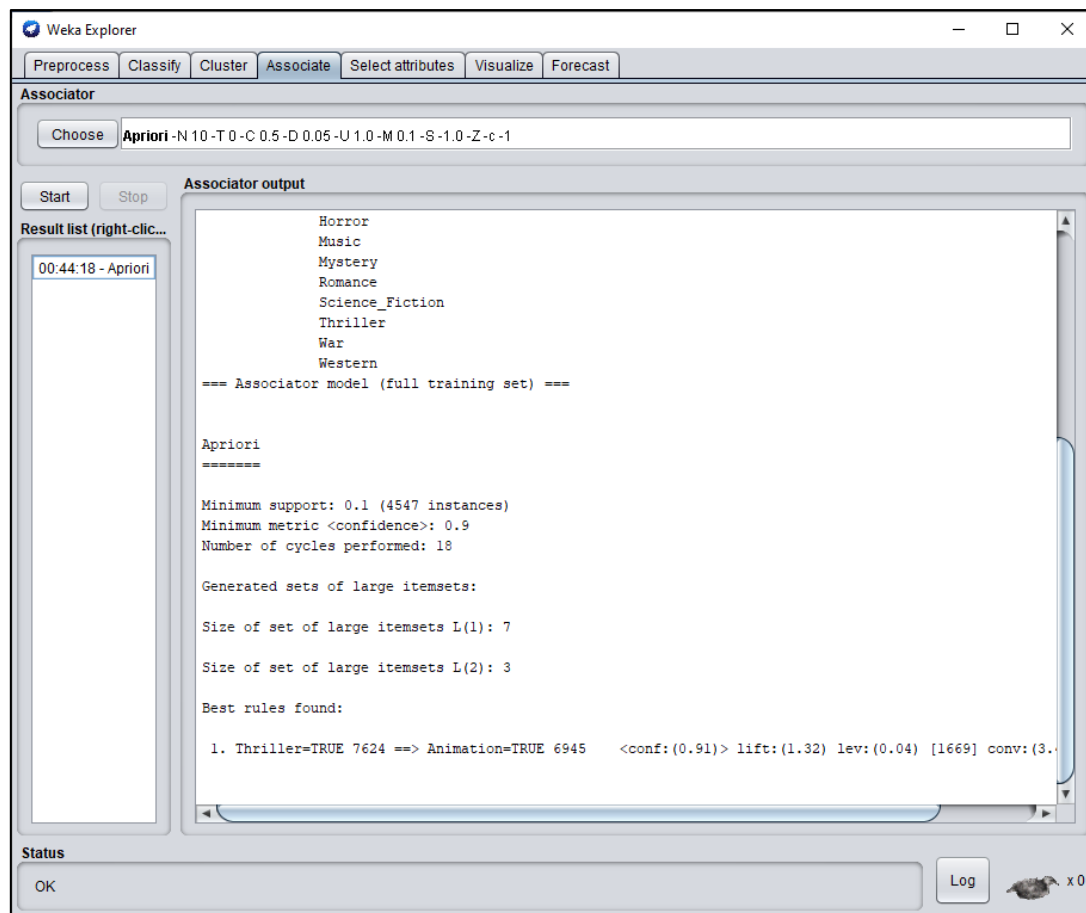


Figure 25: Results of the Apriori algorithm

This time, the results are very different. Despite the algorithm being instructed to find 10 rules it was only able to find 1. If a movie is categorised as a Thriller it is therefore also likely to be Animation as well. This does not seem to hold much logic and there is very little to associate the two genres together.

As no further results were found, it was decided to reduce the confidence level of the algorithm from 90% to 50%. Whilst this may increase the number of rules found, it does also mean that the reliability and association rate of those rules is reduced.

The results of this reduction can be seen in Figure 26:

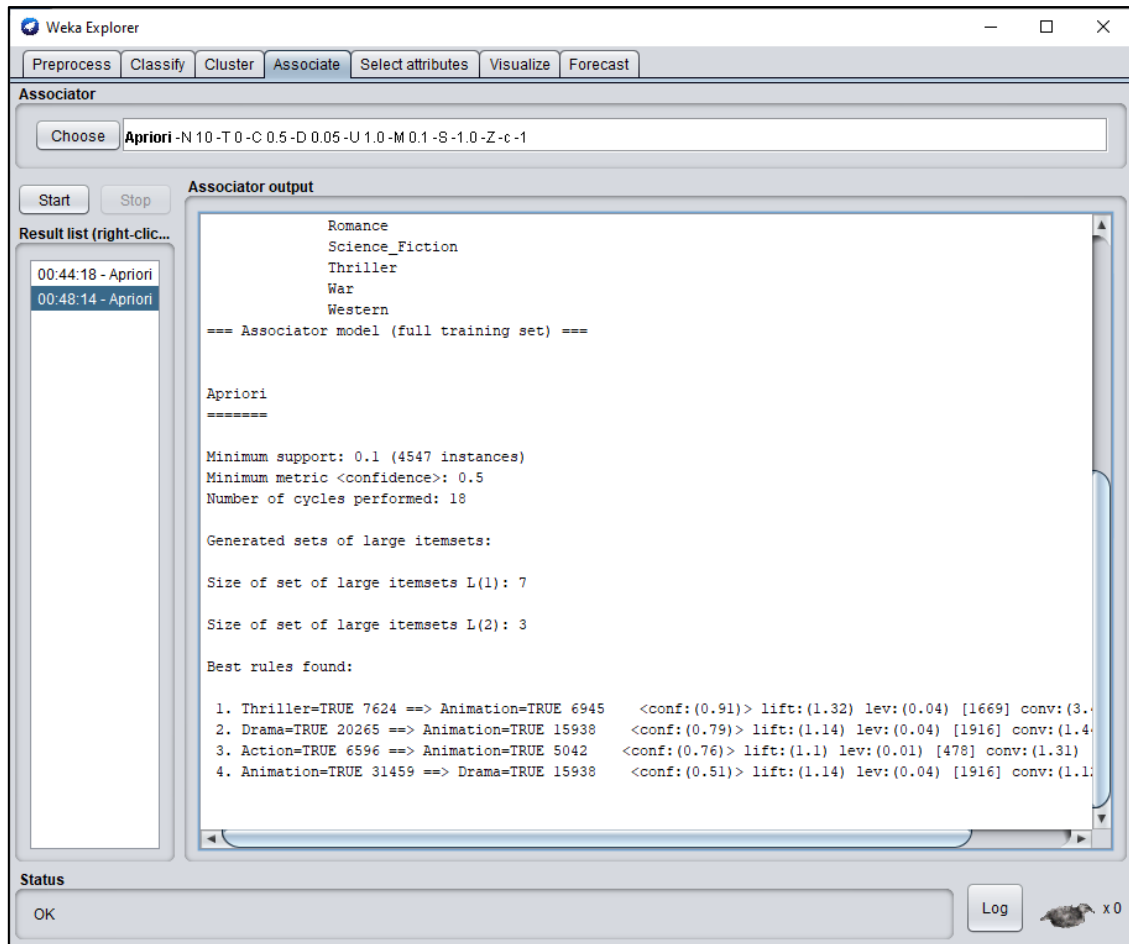


Figure 26: Results of the Apriori algorithm with 50% confidence

Even with confidence at 50%, the algorithm was only able to find 4 rules. Animation seems to feature heavily in the results but there still seems to be no real logic between the association of the genres.

As an experiment, the Animation genre was removed from the available attributes and the algorithm run for a third time. The results of this can be seen in Figure 27:

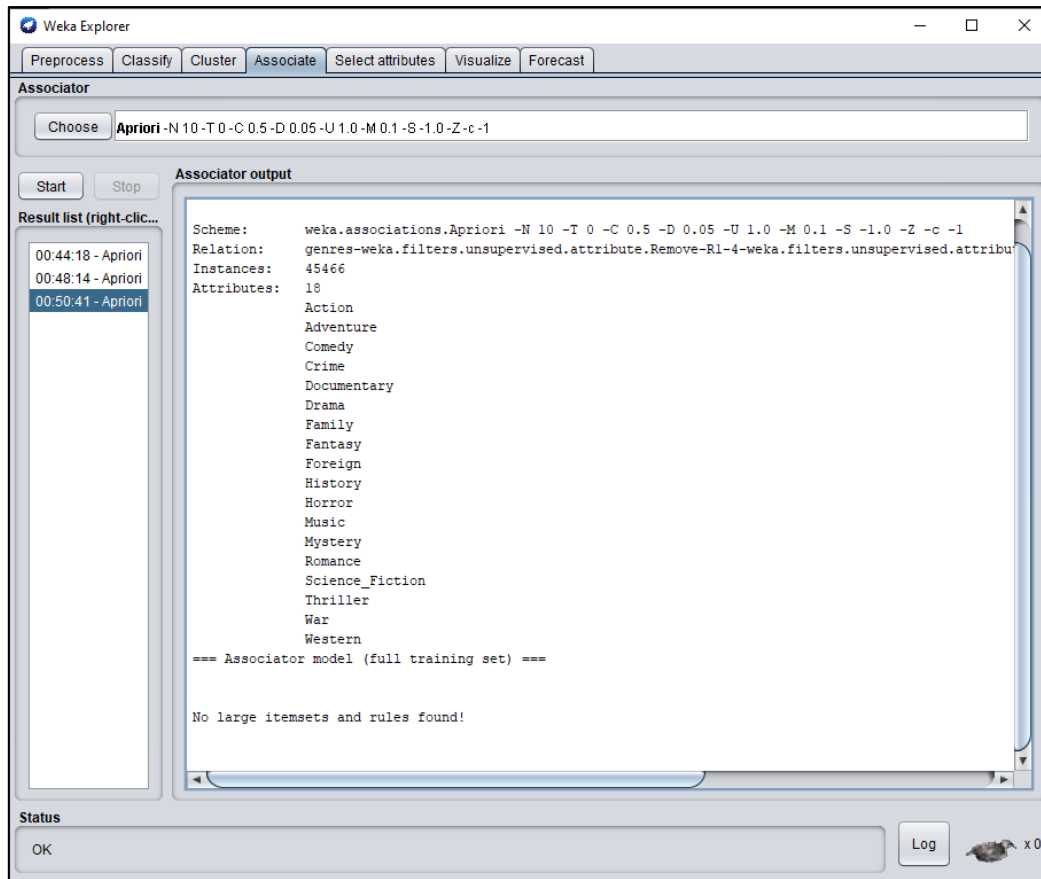


Figure 27: Results of the Apriori algorithm 3rd run

With the Animation genre removed from the attributes and even at 50% confidence, the algorithm is unable to find any rules at all. Despite changing a number of the properties available, no results could be found.

Unfortunately, it would appear that the attempt to associate movies together based on their allocated genres has been unsuccessful. There could be many reasons for this. For example, it is not known how the genres were assigned originally by the TMDb website or indeed how comprehensive the process of assignment was. It could be possible that if the dataset was compiled over a period of time, and possibly by different people and in different countries, different interpretations could have been made as to how to assign genres to movies.

There may be other software, tools or algorithms that may be able to determine if there is any association present but, on this occasion, it has not been possible to find anything of any significance.

5.3 Question 3 - Is there any relation between the budget for a movie and the revenue it makes?

It is sometimes thought that movies that earn higher revenues would have had a higher budget for production to start with. This is not always the case, but it is worth considering if there could be any correlation between the two.

Figure 28 shows the top 10 movies with the highest revenues:

	A	B	C	D	E	F	G	H	I
1	ID	Title	Budget	Revenue	Release_Date	Runtime	Popularity	Vote_Average	Vote_Count
2	19995	Avatar	£ 237,000,000	£ 2,787,965,087	2009-12-10	162	185.07089	7.2	12114
3	140607	Star Wars: The Force Awakens	£ 245,000,000	£ 2,068,223,624	2015-12-15	136	31.626013	7.5	7993
4	597	Titanic	£ 200,000,000	£ 1,845,034,188	1997-11-18	194	26.88907	7.5	7770
5	24428	The Avengers	£ 220,000,000	£ 1,519,557,910	2012-04-25	143	89.887648	7.4	12000
6	135397	Jurassic World	£ 150,000,000	£ 1,513,528,810	2015-06-09	124	32.790475	6.5	8842
7	168259	Furious 7	£ 190,000,000	£ 1,506,249,360	2015-04-01	137	27.275687	7.3	4253
8	99861	Avengers: Age of Ultron	£ 280,000,000	£ 1,405,403,694	2015-04-22	141	37.37942	7.3	6908
9	12445	Harry Potter and the Deathly H	£ 125,000,000	£ 1,342,000,000	2011-07-07	130	24.990737	7.9	6141
10	109445	Frozen	£ 150,000,000	£ 1,274,219,009	2013-11-27	102	24.248243	7.3	5440
11	321612	Beauty and the Beast	£ 160,000,000	£ 1,262,886,337	2017-03-16	129	287.25365	6.8	5530

Figure 28: Top 10 grossing movies in the dataset – sorted by Revenue

The budgets allocated to these movies do indeed seem to be quite substantial stretching to hundreds of millions. However, the budgets represented here are not the highest to be found in the dataset. A simple sort in Excel on the Budget attribute column has changed the top 10 list, as seen in Figure 29:

	A	B	C	D	E	F	G	H	I
1	ID	Title	Budget	Revenue	Release_Date	Runtime	Popularity	Vote_Average	Vote_Count
2	1865	Pirates of the Caribbean: On Stranger Tides	£ 380,000,000	£ 1,045,713,802	2011-05-14	136	27.88772	6.4	5068
3	285	Pirates of the Caribbean: At World's End	£ 300,000,000	£ 961,000,000	2007-05-19	169	31.363664	6.9	4627
4	99861	Avengers: Age of Ultron	£ 280,000,000	£ 1,405,403,694	2015-04-22	141	37.37942	7.3	6908
5	1452	Superman Returns	£ 270,000,000	£ 391,081,192	28/06/2006	154	13.284712	5.4	1429
6	335988	Transformers: The Last Knight	£ 260,000,000	£ 604,942,143	21/06/2017	149	39.186819	6.2	1440
7	38757	Tangled	£ 260,000,000	£ 591,794,936	24/11/2010	100	14.684761	7.4	3419
8	49529	John Carter	£ 260,000,000	£ 284,139,100	07/03/2012	132	14.670353	6.1	2170
9	559	Spider-Man 3	£ 258,000,000	£ 890,871,626	01/05/2007	139	27.614265	5.9	3687
10	57201	The Lone Ranger	£ 255,000,000	£ 89,289,910	03/07/2013	149	12.729104	5.9	2361
11	337339	The Fate of the Furious	£ 250,000,000	£ 1,238,764,765	2017-04-12	136	48.573287	6.8	3803

Figure 29: Top 10 grossing movies in the dataset – sorted by Budget

This time there is a different list of movies presented. Only 1 of the movies, Avengers: Age of Ultron, appears in both lists. This would suggest that there may not be a close relationship between budget and revenue. However, this is only a very superficial analysis and needs to be explored further.

To enable this exploration, it was decided that Weka would not be a suitable tool to use. Tableau is much better at visualising data and comes with a wide selection of tools and techniques that can make the process more effective.

The dataset file was loaded into Tableau. Unlike Weka, it does not need to be in a specific structure to be imported. Tableau is able to recognise CSV and spreadsheet files and interpret them correctly.

The most appropriate method of displaying this data was to use a scatter plot of Budget against Revenue. The results can be seen in Figure 30:

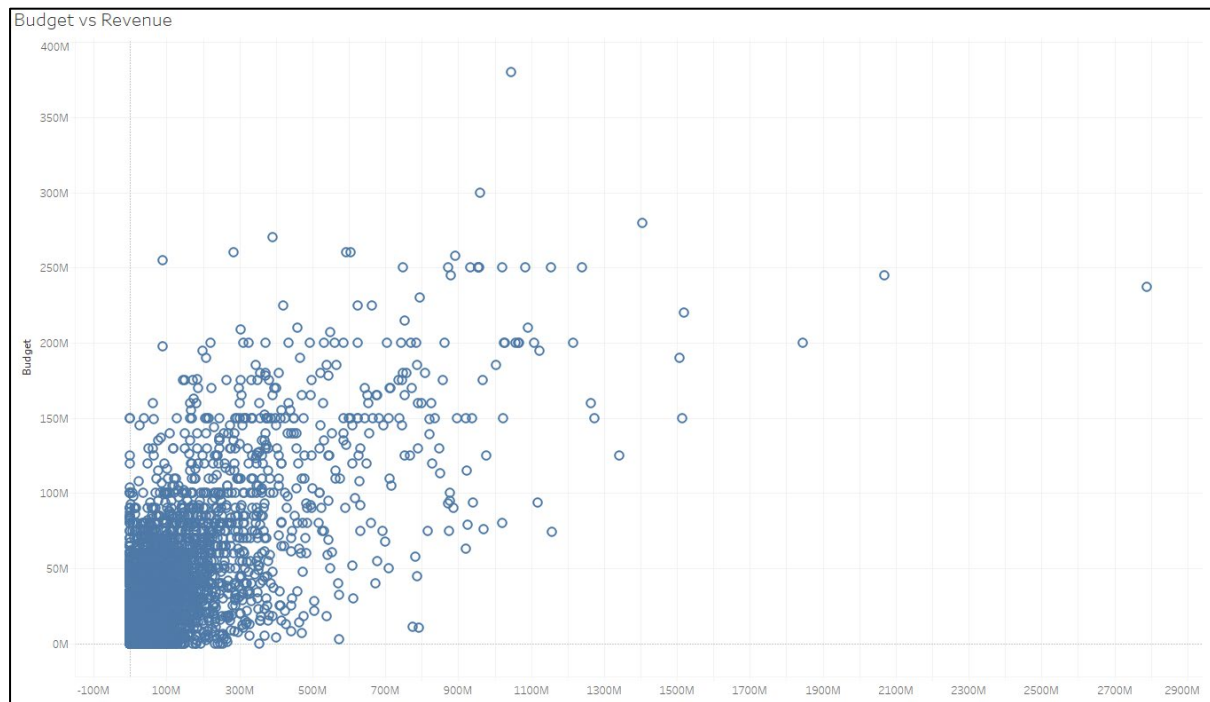


Figure 30: Scatter plot of Budget against Revenue

It would be difficult to draw a line of best fit on this scatter plot, and even if one was drawn there would be a considerable number of movies with some distance from it. This would suggest that there is no real strong correlation here between Budget and Revenue.

However, there is some correlation present but it is unlikely that the relationship between the two attributes would be strong enough to provide an accurate forecast for future movies.

5.4 Question 4 – Is there a relationship between Runtime and Revenue?

Upon analysing the count of genres associated with the movies in the dataset, it was found that the Animation genre features far higher than any other. There are 120,618 genres assigned to the full movie dataset in total, of which 31,459 or 26.08% are Animation. This can be seen in Figure 31:

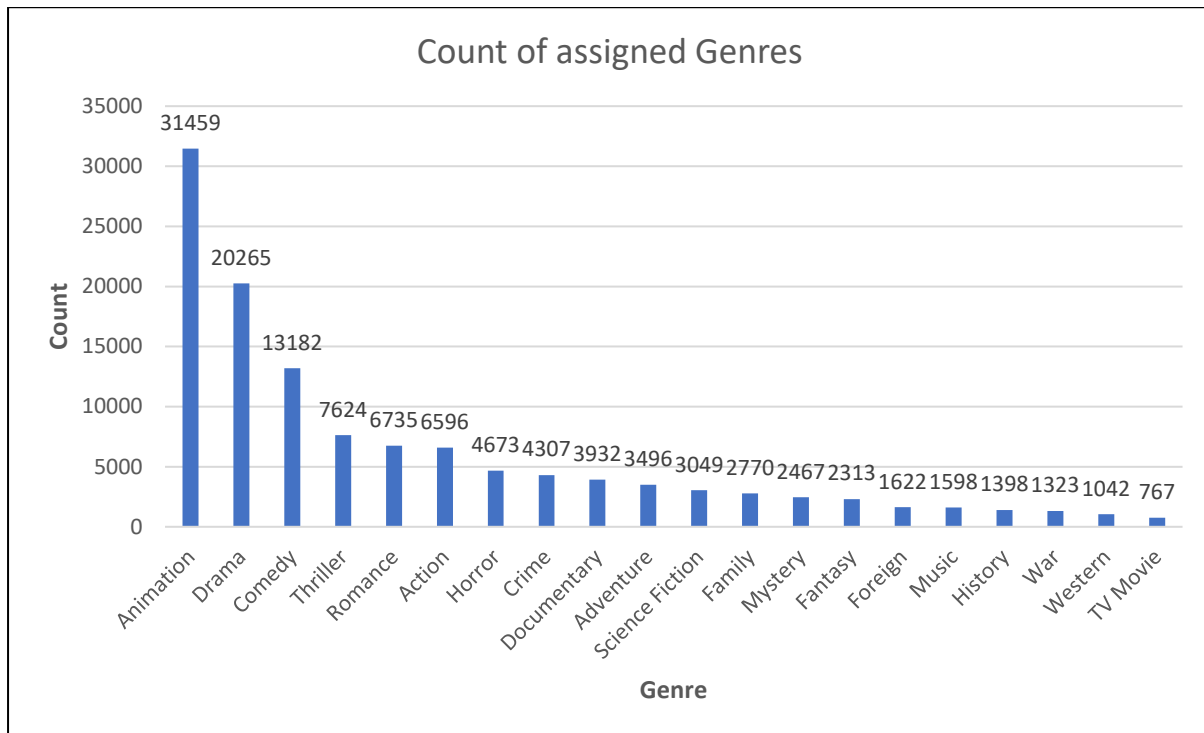


Figure 31: Count of assigned Genres in the full dataset

Traditionally, animated films have higher production costs and are normally assigned higher budgets as a result. However, it is not clear from the dataset what type of movies actually count as being Animation. For example, a live action movie can blend animation with real actors, crew and locations. This can be a costly process. Whereas a fully animated movie is much more time consuming and will need a higher number of technical staff to produce, which can also be costly.

The Runtime attribute contains a wide range of data. The longest length of a movie in the dataset is 1256 minutes long (20.9 hours), whilst the shortest length of a movie is 1 minute. The mean value for movie length is 94.12 minutes (1 hour 34 minutes).

Owing to the large costs involved, it is possible that animated movies would therefore have a shorter Runtime. A visualisation was created in Tableau to test if this theory is true, as seen in Figure 32:

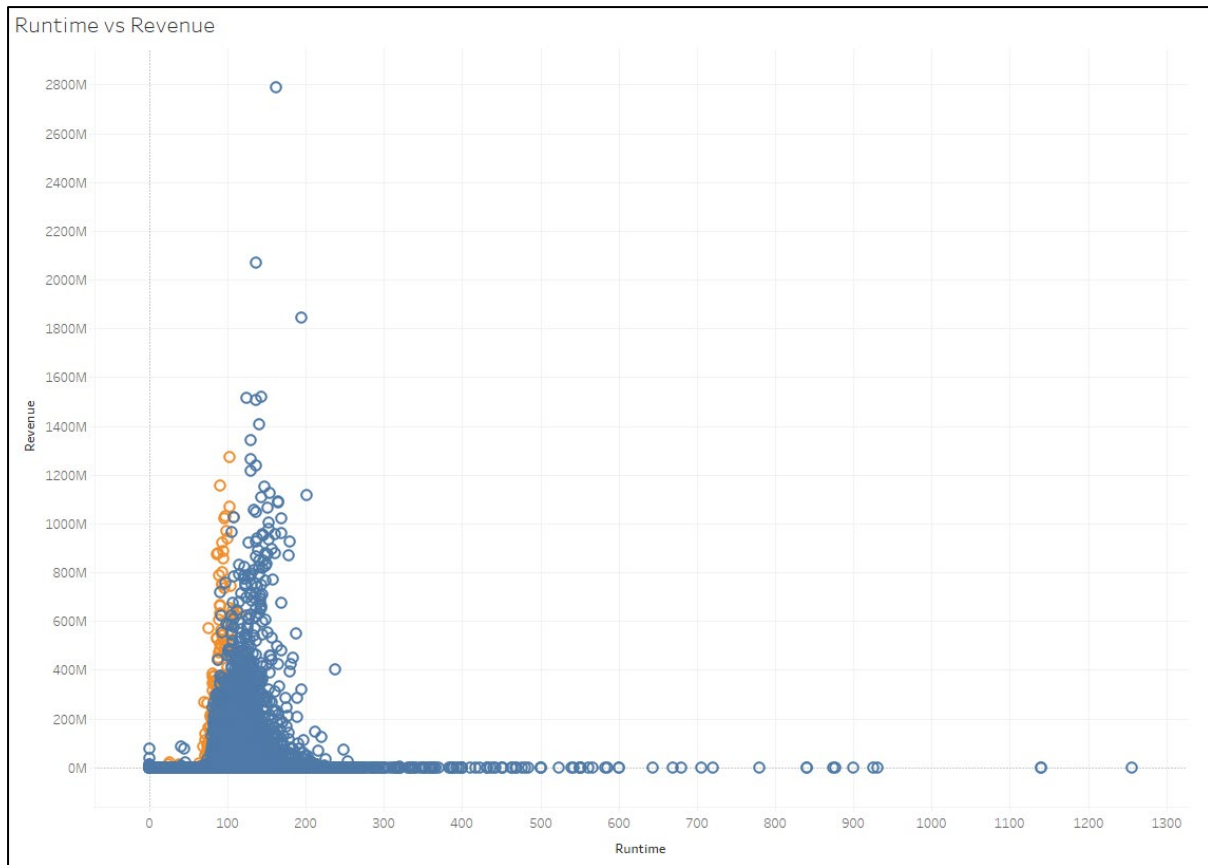


Figure 32: Scatter plot of Runtime vs Revenue

The scatter plot is showing all movies in the dataset. The blue instances are all movies that are not categorised as Animation and the orange instances are movies that are categorised as Animation. It can be seen, although not very clearly, from this view that the Runtime for Animation movies seems to be towards the lower end of Runtime on the X axis. There are no orange instances found at the higher end of the X axis.

Using the filter tools in Tableau, it is possible to add a filter to the data to remove those movies which are not categorised as Animation. The results of this can be seen in Figure 33:

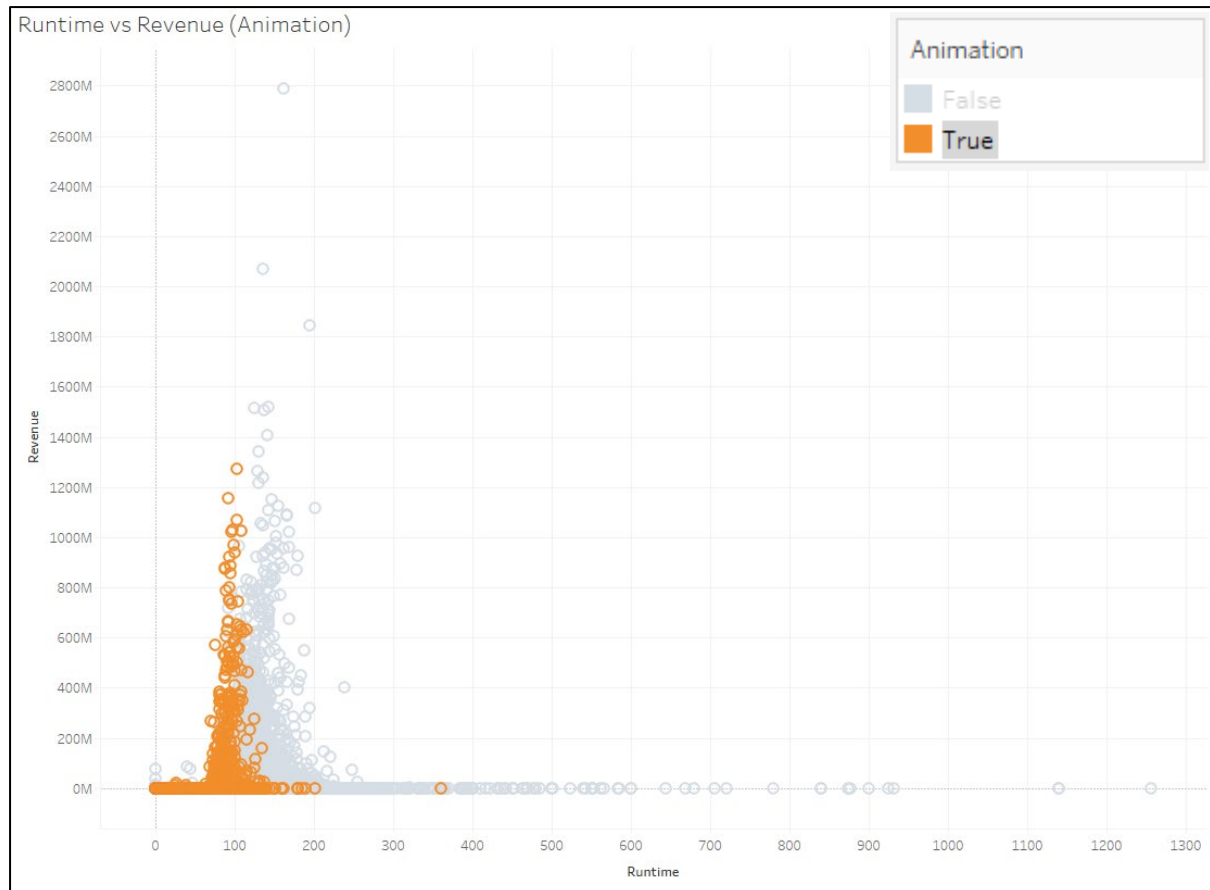


Figure 33: Scatter plot of Runtime vs Revenue (Animation only)

It is now much clearer from this scatter plot to see that Animation movies, again represented in orange, are definitely towards the lower end of the Runtime values. The average Runtime for each movie here seems to follow the mean value of 94 minutes (1 hour 30 minutes). The movies which are not classified as Animation are represented in grey.

There is 1 instance of an Animation movie between 300 and 400 minutes long. On closer examination, this belongs to a movie called *Guyver: The Bioboosted Armor*, which is actually an animated series of episodes each of which is approximately 55 minutes in length [7]. It would appear that multiple episodes have been combined together in the dataset to produce a higher than average calculated runtime, and as such this outlier can be disregarded.

The conclusion from looking at this visualisation would suggest that Animation movies are therefore in fact shorter in length, possibly due to financial and time restraints. However, despite being shorter in length there are a considerable number of Animation movies that have some of the highest recorded revenue values.

The top 10 Animation movies with the highest revenue are shown in Figure 34:

A	B	C	D	E	F	G	H	I	J	K	L	M
ID	Title	Budget	Revenue	Release_Dat	Runtime	Popular	Vote_Average	Vote_Count	IMDB	Action	Adventure	Animation
19995	Avatar	£ 237,000,000	£ 2,787,965,087	2009-12-10	162	185.07089	7.2	12114	tt0499549	TRUE	TRUE	TRUE
109445	Frozen	£ 150,000,000	£ 1,274,219,009	2013-11-27	102	24.248243	7.3	5440	tt2294629	FALSE	TRUE	TRUE
211672	Minions	£ 74,000,000	£ 1,156,730,962	2015-06-17	91	547.4883	6.4	4729	tt2293640	FALSE	TRUE	TRUE
10193	Toy Story 3	£ 200,000,000	£ 1,066,969,703	2010-06-16	103	16.96647	7.6	4710	tt0435761	FALSE	FALSE	TRUE
127380	Finding Dory	£ 200,000,000	£ 1,028,570,889	2016-06-16	97	14.477677	6.8	4333	tt2277860	FALSE	TRUE	TRUE
269149	Zootopia	£ 150,000,000	£ 1,023,784,195	2016-02-11	108	26.024868	7.7	4961	tt2948356	FALSE	TRUE	TRUE
324852	Despicable Me 3	£ 80,000,000	£ 1,020,063,384	2017-06-15	96	36.631519	6.2	2002	tt3469046	TRUE	TRUE	TRUE
93456	Despicable Me 2	£ 76,000,000	£ 970,761,885	2013-06-25	98	24.82355	7	4729	tt1690953	FALSE	FALSE	TRUE
12	Finding Nemo	£ 94,000,000	£ 940,335,536	2003-05-30	100	25.497794	7.6	6292	tt0266543	FALSE	FALSE	TRUE
809	Shrek 2	£ 150,000,000	£ 919,838,758	19/05/2004	93	16.22986	6.7	3082	tt0298148	FALSE	TRUE	TRUE

Figure 34: Top 10 Animation movies with highest revenue

There are certainly some very popular and recognisable movies in this list. Looking at the Runtime attribute in column F it can be seen that, with the exception of Avatar at 162 minutes, the rest of the top 10 have a runtime that is not far from the overall mean value of 94.10 minutes. This would suggest that the Runtime for Animation movies does not necessarily need to be longer, and therefore cost more to produce, in order to secure a good return.

6. Reflection

The process of understanding and applying data mining techniques to this dataset has resulted in a number of conclusions being drawn.

Firstly, prior to completing this investigation it was essential to spend time using the software that was chosen for analysis to establish the capabilities and functions that would be available. In addition to being familiar with the software, it is also essential to ensure that there is an awareness of what the software would be used for. For example, in Weka there are numerous different algorithms available for classification, clustering and association analysis. Each of these algorithms have their own strengths and can yield different results from the data. It is important to be able to understand what the differences are and how the results could potentially be affected.

In addition, each algorithm also has its own set of properties that can be amended to suit the nature of the dataset. For example, in this investigation the 'treatZeroAsMissing' option was enabled in the Apriori algorithm to treat all False values as being missing rather than actually treating them as False to see the effect on the results. All of this is beneficial to anybody wanting to apply data mining techniques, however there should still be a background understanding of how each method works. This can be achieved by reading around the different concepts and experimenting in the software. Both Weka and Tableau are industry standard software and are widely used meaning that there are numerous tutorials and videos available which can help with this process.

Before the data is imported into any software, it is essential that time is spent familiarising yourself with and understanding the dataset that you have chosen to work with. It is beneficial to take some time to observe the data that has been presented in its raw format and then consider what techniques could possibly be used in order to extract meaningful information from it.

Domain knowledge is critical here, as it will enable more effective questions to be asked and there will be a greater understanding of any analysis that takes place and interpretation of the results. It may also be the case in smaller datasets that some patterns can be identified even before data mining software is utilised and then confirmed once analysis has taken place.

It was also determined that the pre-processing of the data can take a considerable amount of time and should not be underestimated. There are multiple tools, techniques and varieties of software that can be used in order to clean the data and make it ready to be analysed. The simplest of these tools used in this investigation was spreadsheet software. If you are responsible for pre-processing data, it is beneficial if you are aware of simple formulas and functions which can help to restructure the data and prepare it for analysis in data mining or visualisation software. Weka, in particular, is quite strict on any data stored in the .ARFF file which is imported and it can take multiple attempts and iterations of the cleaning process to complete it successfully. Larger datasets that require substantial cleaning can make this a laborious task.

A valuable lesson learned in the pre-processing stage was the need to keep files and folders organised. It was important to keep copies of any files prior to any changes being made in case there was a need to revert back to them later on for any reason. This resulted in a lot of files being generated and required the use of a logical folder structure and file naming regime to maintain order.

The quality of the data collected is essential to ensuring that meaningful and valid conclusions can be drawn from any analysis that takes place. It was disappointing on this occasion that the dataset chosen for analysis did not produce many meaningful patterns or trends.

It was hoped that the association analysis applied to the Genre attribute would have produced more significant results. However, this was not the case. Despite many adjustments being made to the algorithm properties, the rules created did not lead to any meaningful conclusion. For association rules to be valid they must have significant strength to be considered, which was not the case in this investigation.

The classification techniques used in this report were moderately successful, however it was hoped that the accuracy rate would be more conclusive. Secondary pre-processing of the data did increase the accuracy of the model, however it only reached 38.10% by reducing the available groups of the

Popularity class attribute. This would have resulted in a broadening of each group and, whilst the accuracy rate in Weka was increased, in reality the groups were too broad to mean anything. For a classification model to be trusted it needs to be accurate enough.

Clustering techniques were considered for investigation in this report however on detailed analysis of the dataset it was felt that there was not likely to be any presence of clusters and therefore no valid use of clustering that would have generated any meaningful conclusion.

It is acknowledged that the dataset could be subjected to more advanced data mining processes and software in the future and these could potentially draw a wider scope of conclusions. In addition, as mentioned earlier in this report, the dataset from Kaggle was complete with seven CSV files containing other data regarding cast and crew amongst others. It was felt on this occasion that these files would not be necessary, however it does allow scope for further questions to be asked from this dataset in the future.

References

- [1] R. Banik, "The Movies Dataset," 2017. [Online]. Available: <https://www.kaggle.com/rounakbanik/the-movies-dataset>. [Accessed 1 November 2019].
- [2] MovieLens, "Recommendations," [Online]. Available: <https://movielens.org/>. [Accessed 2 November 2019].
- [3] GroupLens Research, "Datasets," [Online]. Available: <https://grouplens.org/datasets/movielens/>. [Accessed 2 November 2019].
- [4] The Movie Database (TMDb), [Online]. Available: <https://www.themoviedb.org/>. [Accessed 2 November 2019].
- [5] GroupLens Research, "Summary," [Online]. Available: <http://files.grouplens.org/datasets/movielens/ml-25m-README.html>. [Accessed 2 November 2019].
- [6] I. H. Witten, F. Eibe and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Amsterdam, 2011.
- [7] IMDb, "Guyver: The Bioboosted Armor," [Online]. Available: <https://www.imdb.com/title/tt0143353/>. [Accessed 6 January 2020].