

Contents

1	Abstract	6
2	Introduction	6
3	Data description	7
4	Data transformations	10
5	WEKA examination	17
5.1	ARFF header	17
5.2	Initial data visualisation and analysis	18
6	Data Exploration	20
6.1	Exploration 1: Is there a relation between Artifact and Enchantment?	20
6.2	Exploration 2: Is there a relation between Creature, Sorcery and Instant?	29
6.3	Exploration 3: Exponential smoothing/time series analysis to predict future wins and losses	35
7	Probelms encountered	44
7.1	Raw data extraction	44
7.2	Custom web crawler	47
7.3	WEKA 3.7.12	47
7.4	Pre-processing of data	48
8	Reflection	50
9	Research	51

List of Figures

1	Collected data in raw format	9
2	Initial raw data from web crawler	11
3	Encapsulation of player name in single quotes	12
4	Removal of non relevant data and encapsulation in single quotes of event number	12
5	Removal of hh:mm:ss component event date	13
6	Replacement of null values with Planeswalker (0), removal of excess text from left and right hand from Planeswalker	13

7	Replacement of null values with Creature (0), removal of excess text from left and right hand from Creature	13
8	Replacement of null values with Sorcery (0), removal of excess text from left and right hand from Sorcery	14
9	Replacement of null values with Instant (0), removal of excess text from left and right hand of Instant	14
10	Replacement of null values with Artifact (0), removal of excess text from left and right hand of Artifact	14
11	Replacement of null values with Enchantment (0), removal of excess text from left and right hand of Enchantment	15
12	Replacement of null values with Land (0), removal of excess text from left and right hand of Land	15
13	Replacement of null values with Other (0), removal of excess text from left and right hand of Other	15
14	Completion of data cleansing	16
15	ARFF header code	17
16	Initial WEKA visualisation of .ARFF	18
17	.ARFF header file for exploration 1	20
18	Exploration 1 setting of 25 percentage cluster split	21
19	25 percent split result: 8 cluster division	21
20	25 percent split cluster visualisation	21
21	Secondary setting of 30 percentage cluster split	22
22	30 percent split result: 3 cluster division	22
23	30 percent split cluster visualisation	22
24	Applying an EM cluster filter as an attribute	23
25	Training set applied to .ARFF with EM clustering	23
26	Result of EM training set with cluster filter	24
27	EM training set and cluster filter visualisation	24
28	EM clustering with 40 percentage split	24
29	EM clustering with 40 percent cluster split result	25
30	EM clustering with 40 percentage split visualisation	25
31	EM clustering with 50 percent cluster split	25
32	EM clustering with 50 percent split result	26
33	EM clustering with 50 percent split visualisation	26
34	Re-application of cluster filter for SimpleKMeans	27
35	SimpleKMeans clustering with 40 percent cluster split	27
36	SimpleKMeans clustering with 40 percent split result	27
37	SimpleKMeans clustering with 40 percent split visualisation	28
38	.ARFF header file for exploration 2	29
39	Exploration 2 setting of 25 percentage cluster split	29
40	25 percent split result: 2 cluster division	29

41	25 percent split cluster visualisation	30
42	Secondary setting of 30 percentage cluster split	30
43	30 percent split result: 2 cluster division	30
44	30 percent split cluster visualisation	30
45	Tertiary setting of 40 percentage cluster split	31
46	40 percent split result: 2 cluster division	31
47	40 percent split cluster visualisation	31
48	Applying an SimpleKMeans cluster filter as an attribute	32
49	Training set applied to .ARFF with SimpleKMeans clustering	32
50	Result of SimpleKMeans training set with cluster filter	32
51	SimpleKMeans training set and cluster filter visualisation	33
52	Applying an EM cluster filter as an attribute	33
53	Training set applied to .ARFF with EM clustering	33
54	Result of EM training set with cluster filter	34
55	EM training set and cluster filter visualisation	34
56	.ARFF header file for exploration 2	35
57	Initial setting of Holt Winters time series 10 fold cross validation	35
58	Result of 10 fold cross validation	36
59	Visualisation result of 10 fold cross validation (Wins)	36
60	Visualisation result of 10 fold cross validation (Loss)	37
61	Holt Winters time series with 66 percent data split	37
62	Result of time series with 66 percent data split	37
63	Visualisation of time series with 66 percent data split (Wins)	38
64	Visualisation of time series with 66 percent data split (Loss)	38
65	Application of TimeSeriesTranslate filter attribute	39
66	Holt Winters time series with 60 percent split	39
67	Result of time series with 60 percent split	39
68	Visualisation of time series with 60 percent split (Wins)	40
69	Visualisation of time series with 60 percent split (Loss)	40
70	Holt Winters time series with 50 percent split (E-Date)	41
71	Result of time series with 50 percent split	41
72	Visualisation of time series with 50 percent split (Wins)	41
73	Visualisation of time series with 50 percent split (Loss)	42
74	Holt Winters time series with 50 percent split (Rank)	42
75	Result of time series with 50 percent split	42
76	Visualisation of time series with 50 percent split (Wins)	43
77	Web page to search daily events	45
78	Raw web data	46
79	Raw web data (cont.)	46
80	Event Number pre-processing issue	49
81	Removal of Win/Loss duplicates	49

82	Record of abnormality identified	49
83	Secondary location of irregularity	49

1 Abstract

Data mining and analysis of large data is a process commonplace in the modern business world. To identify trends, patterns and spending habits can assist organisations in not only what products are popular but also at what time they are common, how many of that product were used and the times products were purchased or used.

Many data footprints can be traced through this method and by using data mining grants statistical scientists the ability to track these patterns. Using a collected and custom demographic of information, this report aims to conduct a series of pre processing techniques to the data and answer specified analysis questions using the custom data set.

2 Introduction

This report has been produced to provide a statistical analysis of data representative across a 6 month period for daily events played online for a popular card game. The dataset has been constructed to provide analysis into possible patterns and associations, which could give the business an overview of trends for players of the game and their thought processes when constructing card decks.

The first section provides an overview of the data including a description of each attribute. Data transformations are also documented explaining the process required to ensure the information was translatable for the data mining application. Upon achieving this an initial analysis was required to determine research questions.

In the second section, investigation questions are derived from the observations and a detailed analysis for every problem synthesized. Each process has been documented with screenshots outlining changes to the dataset for each specific inquiry, the process of data mining applied and a final conclusion produced from these findings.

Finally the closing section contains a detailed review of the problems encountered including solutions found as a work around to these issues. A final reflection of the process is included expressing the assets and flaws of the analysis conducted and a comparison of the findings.

3 Data description

The data that has been identified for data analysis is an online compendium of card duels from Magic: The Gathering Online [1]. Players are dealt a specified hand of 7 cards from a pre shuffled deck totalling 60 cards and the games objective is to wipe out the opponents life point total of 20. Selected for analysis are daily events played by players in a ranking system over a 6 month period. This will provide a fine granular level of detail for mining and represents data at a snapshot in time.

As shown in Figure 1 the data that has been extracted using a software based web crawler [2]. While it has extrapolated all required information, the data is not yet fit for mining. Cleansing the data is a paramount activity that will be required before any analysis can take place. The complete size of the data set is approximately 11055 records consisting of 15 separate attributes. Each record details how each player placed in that daily event based on the number of wins and losses; the event name, the event date and ratio of cards that player used within that specific event.

The data is catagorised under the following titles:

Rank: How the player placed during a specific event.

Player: Unique text based identification of the player.

Wins: How many games a player won (per event)

Losses: How many games a player lost (per event).

Event Number: Unique numeric identifier for each event scheduled.

Event Name: Textual name of event scheduled.

Event Date: The specified date of the event.

Planeswalker: Number of planeswalker cards (per player deck).

Creature: Number of creature cards (per player deck).

Sorcery: Number of sorcery cards (per player deck).

Instant: Number of instant cards (per player deck).

Artifact: Number of artifact cards (per player deck).

Enchantment: Number of enchantment cards (per player deck).

Land: Number of land cards (per player deck).

Other: Number of other cards (per player deck). Typically cards that do not fall into any of the above listed categories.

Rank	Player	Wins	Losses	Event Number	Event Name	Event Date	Planeswalker	Creature	Sorcery	Instant	Artifact	Enchantment	Land	Other
1	krom321	4	0	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (4)	Creature (20)		Instant (6)		Enchantment (3)	Land (25)	Other (2)
2	Chaindog313	4	0	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (5)	Creature (18)		Instant (9)		Enchantment (4)	Land (24)	
3	Seizurema	4	0	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (20)		Instant (5)		Enchantment (15)	Land (20)	
4	killabeeze	4	0	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (24)	Sorcery (7)			Enchantment (7)	Land (22)	
5	theeguy	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (6)	Creature (14)	Sorcery (4)	Instant (10)			Land (25)	Other (1)
6	Dremin	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (26)	Sorcery (4)	Instant (6)	Artifact (1)		Land (23)	
7	ability170146	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (10)	Creature (19)		Instant (6)			Land (25)	
8	blastoids	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (7)	Creature (18)		Instant (8)		Enchantment (3)	Land (24)	
9	TheKemplar	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (28)	Sorcery (4)	Instant (6)			Land (22)	
10	zibb	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (28)	Sorcery (4)	Instant (6)			Land (22)	
11	dRost	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (24)	Sorcery (4)	Instant (7)	Artifact (1)		Land (24)	
12	VonPapen	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (8)	Creature (18)		Instant (6)		Enchantment (4)	Land (24)	
13	Based	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (2)	Creature (21)		Instant (5)		Enchantment (7)	Land (25)	
14	shamanicblue	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (9)	Creature (20)				Enchantment (4)	Land (25)	Other (2)
15	Jilessoftrees	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (4)	Creature (18)	Sorcery (4)	Instant (8)		Enchantment (1)	Land (25)	
16	petomartinez	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (9)	Creature (18)		Instant (7)		Enchantment (2)	Land (24)	
17	TptBahamut	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (19)	Sorcery (4)	Instant (8)	Artifact (3)		Land (26)	
18	mikeman29	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00		Creature (22)	Sorcery (2)	Instant (4)	Artifact (1)	Enchantment (8)	Land (23)	
19	chivas73	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (3)	Creature (23)		Instant (11)			Land (23)	
20	Kaitai	3	1	#7143393	THS Block Constructed Daily	02/06/2014 00:00:00	Planeswalker (7)	Creature (13)	Sorcery (5)	Instant (9)			Land (25)	Other (1)
1	silviawataru	4	0	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00		Creature (13)	Sorcery (10)	Instant (18)			Land (19)	
2	_Nukesaku_	4	0	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00		Creature (14)	Sorcery (7)	Instant (21)			Land (18)	
3	sakimmd	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (6)	Creature (14)	Sorcery (5)	Instant (12)			Land (23)	
4	Ghesizzi	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (3)	Creature (2)	Sorcery (8)	Instant (17)	Artifact (5)	Enchantment (3)	Land (22)	
5	MatsOle	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00			Sorcery (8)	Instant (7)	Artifact (4)	Enchantment (6)	Land (35)	
6	gotsjo	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00		Creature (8)	Sorcery (2)	Instant (16)	Artifact (21)		Land (13)	
7	echechetmat35	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (3)	Creature (17)	Sorcery (4)	Instant (14)			Land (22)	
8	RaV_	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (2)	Creature (23)	Sorcery (4)	Instant (8)	Artifact (1)	Enchantment (2)	Land (20)	
9	VariableChange88	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (2)	Creature (14)	Sorcery (7)	Instant (17)			Land (20)	
10	kozel	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (3)	Creature (4)	Sorcery (7)	Instant (16)	Artifact (4)	Enchantment (3)	Land (23)	
11	Lejay	3	1	#7143420 on 06/02/20	Legacy Daily	02/06/2014 00:00:00	Planeswalker (5)	Creature (13)	Sorcery (6)	Instant (12)		Enchantment (1)	Land (23)	
1	giuliop91	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (15)	Sorcery (6)	Instant (12)		Enchantment (4)	Land (23)	
2	UglyCat	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (31)		Instant (2)	Artifact (4)		Land (23)	
3	Rolo_Horse	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (12)	Sorcery (8)	Instant (18)	Artifact (2)		Land (20)	
4	TheHamburglar	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (14)	Sorcery (12)	Instant (16)			Land (18)	
5	under_a	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (12)		Instant (22)		Enchantment (4)	Land (22)	
6	Freezerr	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (12)		Instant (22)		Enchantment (4)	Land (22)	
7	tyobe0440	4	0	#7143419 on 06/02/20	Modern Daily	02/06/2014 00:00:00		Creature (16)	Sorcery (4)	Instant (13)		Enchantment (4)	Land (23)	

Figure 1: Collected data in raw format

4 Data transformations

In order to get the data into WEKA [3], all data transformations were conducted in Excel, which was then output to a .CSV file before .ARFF conversion. The following steps were undertaken during this process:

- Concatenate all string based values into single quotes using the formulae =CONCATENATE("'",cell,"'"). The best way to do this is to write the formulae to a blank cell then copy and paste the new value into the original cell.
- Event Date fields have excess information relating to time. This field relies on a date format only. The formulae =LEFT(cell, LEN(cell) - 9 will remove the time characters that are not required.
- WEKA dislikes null values and as such each field will need to have a value assigned to it. These will be assigned a value based on surrounding records (e.g null PW cell will contain Planeswalker(0)).
- Remove excess characters from a string within given cells. This will be used primarily for the alteration to numerical fields. The =LEFT(cell, LEN(cell) - number of characters) formulae will be required. This will have multiple iterations and adaptations in order to get all of the records within the columns to a required level.

1	keltch	4	0	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (11)	Creature (12)	Sorcery (8)	Instant (4)				Land (24)	
2	_Shatun_	4	0	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (5)	Creature (21)	Sorcery (8)	Instant (3)				Land (23)	
3	Ifoldforfood	4	0	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (6)	Creature (31)						Land (23)	
4	Aegis HK	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00		Creature (15)	Sorcery (6)	Instant (9)		Enchantment (4)		Land (26)	
5	_Soku_	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (9)	Creature (14)	Sorcery (8)	Instant (5)				Land (24)	
6	shochoo	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00		Creature (8)		Instant (27)		Enchantment (1)		Land (24)	
7	_Marian_	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (9)	Creature (14)	Sorcery (8)	Instant (5)				Land (24)	
8	Mcrells	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00		Creature (23)		Instant (16)				Land (21)	
9	peter780107	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (4)		Sorcery (10)	Instant (18)	Artifact (1)			Land (26)	Other (1)
10	AleLawyer1	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (1)	Creature (14)	Sorcery (4)	Instant (10)		Enchantment (6)		Land (25)	
11	Diem4x	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00	Planeswalker (5)	Creature (21)	Sorcery (7)	Instant (4)				Land (23)	
12	machiko	3	1	#7500399 on 09/29/2014	Standard Daily	29/09/2014 00:00:00		Creature (10)		Instant (27)				Land (23)	
1	06nash	4	0	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (29)		Instant (2)	Artifact (4)	Enchantment (4)		Land (21)	
2	TheLostCouz3	4	0	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (18)	Sorcery (9)	Instant (9)		Enchantment (2)		Land (22)	
3	JDunkin00	4	0	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (17)	Sorcery (9)	Instant (11)				Land (23)	
4	kiyomori	4	0	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (23)	Sorcery (7)	Instant (8)				Land (22)	
5	simmsey	4	0	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (33)	Sorcery (4)	Instant (4)				Land (19)	
6	nickdebe	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (21)	Sorcery (8)	Instant (13)	Artifact (1)			Land (17)	
7	Hocking	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (34)	Sorcery (4)	Instant (5)				Land (17)	
8	TSAI_PO	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (22)	Sorcery (7)	Instant (14)	Artifact (1)			Land (16)	
9	Moi	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (18)	Sorcery (11)	Instant (13)				Land (18)	
10	joannesp	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (15)				Enchantment (28)		Land (17)	
11	cariocax	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (17)	Sorcery (8)	Instant (9)		Enchantment (2)		Land (24)	
12	TheBimdeezy	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (16)	Sorcery (12)	Instant (6)		Enchantment (3)		Land (23)	
13	takeonman	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (11)	Sorcery (8)	Instant (18)	Artifact (1)			Land (22)	
14	Zakurero22	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (21)	Sorcery (9)	Instant (8)				Land (22)	
15	DocPepper	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (13)	Sorcery (7)	Instant (4)	Artifact (16)			Land (20)	
16	JherjamesB	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (21)	Sorcery (4)	Instant (16)	Artifact (2)			Land (17)	
17	Gui_BR	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (13)	Sorcery (6)	Instant (8)	Artifact (16)	Enchantment (5)		Land (12)	
18	Torreth	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (29)		Instant (2)	Artifact (4)	Enchantment (4)		Land (21)	
19	DaveFlanagan	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00		Creature (36)	Sorcery (8)		Artifact (4)			Land (12)	
20	Abaris	3	1	#7500381 on 09/29/20	Pauper Daily	29/09/2014 00:00:00			Sorcery (16)	Instant (22)		Enchantment (4)		Land (18)	
1	Wuzlking	4	0	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (29)		Instant (4)	Artifact (4)			Land (23)	
2	Suczi	4	0	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (15)	Sorcery (8)	Instant (18)				Land (19)	
3	McWinSauce	4	0	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (5)	Sorcery (5)	Instant (17)	Artifact (7)	Enchantment (1)		Land (25)	
4	startoss	3	1	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (16)	Sorcery (1)	Instant (18)				Land (25)	
5	DC_Casanuda	3	1	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (10)	Sorcery (2)	Instant (16)	Artifact (9)			Land (23)	
6	Kawatori	3	1	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (14)	Sorcery (9)	Instant (16)	Artifact (1)			Land (20)	
7	ikko	3	1	#7500404 on 09/29/20	Modern Daily	29/09/2014 00:00:00		Creature (25)		Instant (7)		Enchantment (4)		Land (20)	

Figure 2: Initial raw data from web crawler

Player
'keltch '
'_Shatun_ '
'Ifoldforfood '
'Aegis HK '
'_Soku_ '
'shochoo '
'_Marian_ '
'Mcrells '
'peter780107 '
'AleLawyer1 '
'Diem4x '
'machiko '
'06nash '
'TheLostCouz3 '
'JDunkin00 '
'kiyomori '
'simmsey '
'nickdebe '
'Hocking '
'TSAI_PO '
'Moi '
'joannesp '

Figure 3: Encapsulation of player name in single quotes

E-Number	E-Number
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500399	' #7500399'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'
#7500381	' #7500381'

Figure 4: Removal of non relevant data and encapsulation in single quotes of event number

E-Date
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014
29/09/2014

Figure 5: Removal of hh:mm:ss component event date

PW	PW	PW
Planeswalker (11)	11)	11
Planeswalker (5)	5)	5
Planeswalker (6)	6)	6
Planeswalker (0)	0)	0
Planeswalker (9)	9)	9
Planeswalker (0)	0)	0
Planeswalker (9)	9)	9
Planeswalker (0)	0)	0
Planeswalker (9)	9)	9
Planeswalker (4)	4)	4
Planeswalker (1)	1)	1
Planeswalker (5)	5)	5

Figure 6: Replacement of null values with Planeswalker (0), removal of excess text from left and right hand from Planeswalker

C	C	C
Creature (12)	12)	12
Creature (21)	21)	21
Creature (31)	31)	31
Creature (15)	15)	15
Creature (14)	14)	14
Creature (8)	8)	8
Creature (14)	14)	14
Creature (23)	23)	23
Creature (0)	0)	0
Creature (14)	14)	14

Figure 7: Replacement of null values with Creature (0), removal of excess text from left and right hand from Creature

S	S	S
Sorcery (8)	8)	8
Sorcery (8)	8)	8
Sorcery (0)	0)	0
Sorcery (6)	6)	6
Sorcery (8)	8)	8
Sorcery (8)	0)	0
Sorcery (0)	8)	8
Sorcery (8)	0)	0
Sorcery (0)	10)	10
Sorcery (0)	4)	4

Figure 8: Replacement of null values with Sorcery (0), removal of excess text from left and right hand from Sorcery

I	I	I
Instant (4)	4)	4
Instant (3)	3)	3
Instant (0)	0)	0
Instant (9)	9)	9
Instant (5)	5)	5
Instant (27)	27)	27
Instant (5)	5)	5
Instant (16)	16)	16
Instant (18)	18)	18
Instant (10)	10)	10

Figure 9: Replacement of null values with Instant (0), removal of excess text from left and right hand of Instant

A	A	A
Artifact (0)	0)	0
Artifact (0)	0)	0
Artifact (0)	0)	0
Artifact (0)	0)	0
Artifact (0)	0)	0
Artifact (0)	0)	0
Artifact (0)	0)	0
Artifact (1)	1)	1
Artifact (0)	0)	0

Figure 10: Replacement of null values with Artifact (0), removal of excess text from left and right hand of Artifact

E	E	E
Enchantment (0)	0)	0
Enchantment (0)	0)	0
Enchantment (0)	0)	0
Enchantment (4)	4)	4
Enchantment (0)	0)	0
Enchantment (1)	1)	1
Enchantment (0)	0)	0
Enchantment (0)	0)	0
Enchantment (0)	0)	0
Enchantment (6)	6)	6

Figure 11: Replacement of null values with Enchantment (0), removal of excess text from left and right hand of Enchantment

L	L	L
Land (24)	24)	24
Land (23)	23)	23
Land (23)	23)	23
Land (26)	26)	26
Land (24)	24)	24
Land (24)	24)	24
Land (24)	24)	24
Land (21)	21)	21
Land (26)	26)	26

Figure 12: Replacement of null values with Land (0), removal of excess text from left and right hand of Land

O	O	O
Other (0)	0)	0
Other (0)	0)	0
Other (0)	0)	0
Other (0)	0)	0
Other (0)	0)	0
Other (0)	0)	0
Other (0)	0)	0
Other (0)	0)	0
Other (0)	1)	1
Other (1)	0)	0

Figure 13: Replacement of null values with Other (0), removal of excess text from left and right hand of Other

1	'keltch '	4	0	'#7500399'	Standard Daily	29/09/2014	11	12	8	4	0	0	24	0
2	'_Shatun_ '	4	0	'#7500399'	Standard Daily	29/09/2014	5	21	8	3	0	0	23	0
3	'Ifoldforfood '	4	0	'#7500399'	Standard Daily	29/09/2014	6	31	0	0	0	0	23	0
4	'Aegis HK '	3	1	'#7500399'	Standard Daily	29/09/2014	0	15	6	9	0	4	26	0
5	'_Soku_ '	3	1	'#7500399'	Standard Daily	29/09/2014	9	14	8	5	0	0	24	0
6	'shochoo '	3	1	'#7500399'	Standard Daily	29/09/2014	0	8	0	27	0	1	24	0
7	'_Marian_ '	3	1	'#7500399'	Standard Daily	29/09/2014	9	14	8	5	0	0	24	0
8	'Mcrells '	3	1	'#7500399'	Standard Daily	29/09/2014	0	23	0	16	0	0	21	0
9	'peter780107 '	3	1	'#7500399'	Standard Daily	29/09/2014	4	0	10	18	1	0	26	1
10	'AleLawyer1 '	3	1	'#7500399'	Standard Daily	29/09/2014	1	14	4	10	0	6	25	0
11	'Diem4x '	3	1	'#7500399'	Standard Daily	29/09/2014	5	21	7	4	0	0	23	0
12	'machiko '	3	1	'#7500399'	Standard Daily	29/09/2014	0	10	0	27	0	0	23	0
1	'06nash '	4	0	'#7500381'	Pauper Daily	29/09/2014	0	29	0	2	4	4	21	0
2	'TheLostCouz3 '	4	0	'#7500381'	Pauper Daily	29/09/2014	0	18	9	9	0	2	22	0
3	'JDunkin00 '	4	0	'#7500381'	Pauper Daily	29/09/2014	0	17	9	11	0	0	23	0
4	'kiyomori '	4	0	'#7500381'	Pauper Daily	29/09/2014	0	23	7	8	0	0	22	0
5	'simmsey '	4	0	'#7500381'	Pauper Daily	29/09/2014	0	33	4	4	0	0	19	0
6	'nickdebe '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	21	8	13	1	0	17	0
7	'Hocking '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	34	4	5	0	0	17	0
8	'TSAI_PO '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	22	7	14	1	0	16	0
9	'Moi '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	18	11	13	0	0	18	0
10	'joannesp '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	15	0	0	0	28	17	0
11	'cariocax '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	17	8	9	0	2	24	0
12	'TheBimdeezzy '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	16	12	6	0	3	23	0
13	'takeonman '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	11	8	18	1	0	22	0
14	'Zakurero22 '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	21	9	8	0	0	22	0
15	'DocPepper '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	13	7	4	16	0	20	0
16	'JherjamesB '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	21	4	16	2	0	17	0
17	'Gui_BR '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	13	6	8	16	5	12	0
18	'Torreth '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	29	0	2	4	4	21	0
19	'DaveFlanagan '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	36	8	0	4	0	12	0
20	'Abaris '	3	1	'#7500381'	Pauper Daily	29/09/2014	0	0	16	22	0	4	18	0
1	'Wuzlking '	4	0	'#7500404'	Modern Daily	29/09/2014	0	29	0	4	4	0	23	0
2	'Suczi '	4	0	'#7500404'	Modern Daily	29/09/2014	0	15	8	18	0	0	19	0
3	'McWinSauce '	4	0	'#7500404'	Modern Daily	29/09/2014	0	5	5	17	7	1	25	0
4	'startoss '	3	1	'#7500404'	Modern Daily	29/09/2014	0	16	1	18	0	0	25	0
5	'DC_Casanuda '	3	1	'#7500404'	Modern Daily	29/09/2014	0	10	2	16	9	0	23	0
6	'Kawatori '	3	1	'#7500404'	Modern Daily	29/09/2014	0	14	9	16	1	0	20	0
7	'ikko '	3	1	'#7500404'	Modern Daily	29/09/2014	0	25	0	7	0	4	20	0

Figure 14: Completion of data cleansing

5 WEKA examination

5.1 ARFF header

In figure 15, the heading code of the .arff file has been defined. To get the data into WEKA for meaningful data mining procedures the specific columns have been separated and identified accordingly. Fields with a number based value have been marked as numeric while records with text have been given a string allocation. These field divisions give WEKA a descriptive key word and determine how the data should be treated in the mining process.

It was concluded that some fields required classification above other variable allocations as the repetitive nature of the data within these fields is predictable, suggesting patterns of potential interest. Classification allows for similar data elements to be grouped together and should aid in the overall analysis of the data.

```
@relation 'MTGO Data'

@attribute Rank numeric
@attribute Player string
@attribute Win {3, 4}
@attribute Loss {0, 1}
@attribute E-Number string
@attribute E-Name string
@attribute E-Date DATE "dd/mm/yy"
@attribute PW numeric
@attribute C numeric
@attribute S numeric
@attribute I numeric
@attribute A numeric
@attribute E numeric
@attribute L numeric
@attribute O numeric
|
@data
```

Figure 15: ARFF header code

5.2 Initial data visualisation and analysis

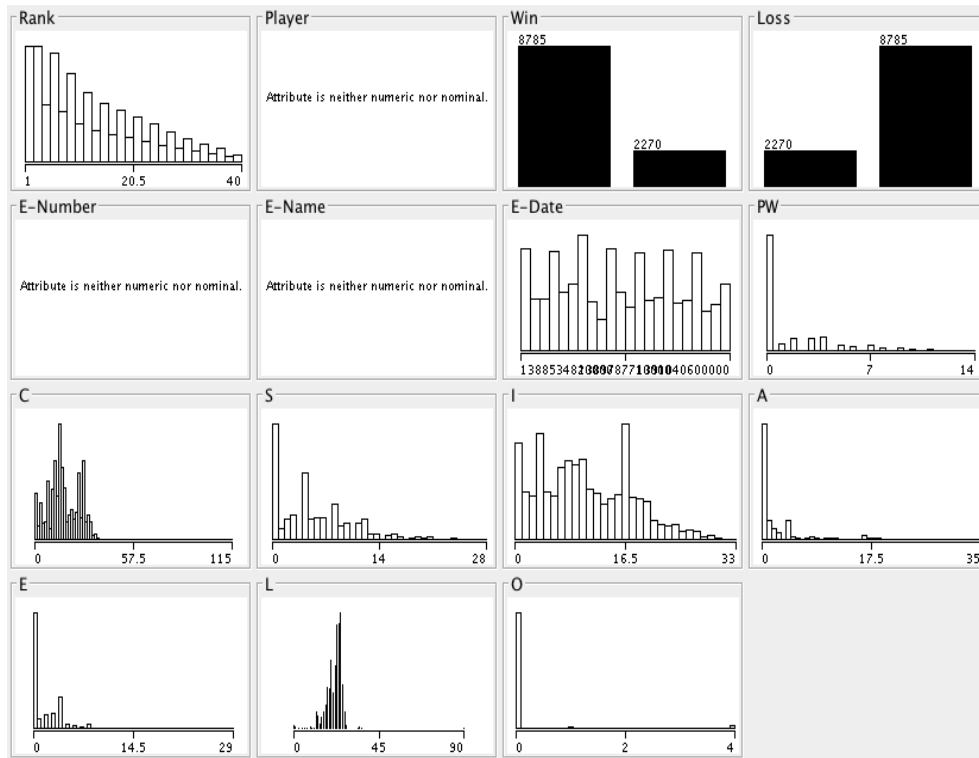


Figure 16: Initial WEKA visualisation of .ARFF

With the data cleansed, the information has been visualised as shown in figure 16. This representation of the data marks an analysis milestone and allows for questions to be asked from the data.

Rank - A reasonably standard depreciative shape. The field is representative of a players position within daily events, many participants have reached rank 1 and fewer have been placed at rank 40, the question is why?

Player - Text based field and provides no interesting information other than who played. Research questions will not be reliant on what player has participated in an event, but rather when they participated.

Win - Left column depicts 3 wins and the right represents 4 wins overall. This seems about right as 4 wins is less frequent whereas 3 appear more often.

Loss - Left column shows 0 losses and the right portrays 1 loss. This is a complete mirrored pattern of wins. Again this is expected as 0 losses are less

common than 1 loss per game.

E-Number - Text based field stating the events unique number. Not really needed as the data holds nothing of worth.

E-Name - Specified name of sectioned daily events. Again not really needed as the data is tailored for daily events only, so the names of events are not required.

E-Date - Very interesting pattern and shows a clear need for time based analysis. This field is vital to determine when an event was played and looks as though it will provide key information.

PW - Large spike of low planes walkers within decks overall and slight alterations across the board. Interesting demographic as typically a game changing card.

C - Characteristics of the data here seem worth investigation. Clearly a double hump in the data presented and could provide interesting analysis.

S - A standard curved shape. Range consists of high spike in lower numbers but also shows an interesting pattern of half size spikes of the initial spike every 4 columns. Worth potential investigation.

I - More records consist of lower numbers for instant based cards with 3 spikes ranging from low numbers to middle. An interesting pattern and feels like it does not follow convention.

A - Expected spike of low numbers with a very low prominence in the mid region. This does not come across as significant but still relevant to deck composition.

E - Similar to the previous but does contain a minor spike a few columns inward. Could be interpreted as a double hump, however this does not seem likely.

L - A very standard shape delimiting a clear region for typical card numbers of this type. While relevant it however does not present much that requires investigation.

O - This feels like a complete outlier and could skew the data in unfavourable ways. This field holds numbers that rarely deviate from 0 and 4, therefore the data within is of no interest.

6 Data Exploration

With the completion of pre-processing and cleansing the data can now be imported into WEKA for analysis. The initial data visualisation observation allowed for preliminary cogitation upon the data and from this, the identification of questions to answer using mining techniques.

For all exploration problems outlined the initial arff header must have all redundant data removed to ensure that only the relevant fields are kept for the analysis of described phenomena; each research question has been generated from observations of the data conducted.

6.1 Exploration 1: Is there a relation between Artifact and Enchantment?

Before starting this exploration, figure 17 shows alterations made to the .arff file. The attributes selected are Artifact (A) and Enchantment (E).

```
@relation 'MTGO exp 1'

@attribute A numeric
@attribute E numeric

@data
0,3
0,4
0,15
0,7
```

Figure 17: .ARFF header file for exploration 1

Predicted outcome: Both Artifact and Enchantment will correlate with one another. Association rules outlined below will be fulfilled.

Estimated rulesets:

$[A] \Rightarrow [E]$

$[E] \Rightarrow [A]$

Initial Exploration

Initial testing and exploration of the data does not yield promising results for proving the association rule sets. Figures 18, 19 and 20 display a non correlating pattern between the two attributes and suggests that the data cannot be associated in any way.

Figures 21, 22 and 23 yielded some rather promising results lowering the clusters from eight to three. While this is a good result and could lead to potentially

proving the association rules, this feels too good to be true. Producing a formal model should produce a better calibre of results.

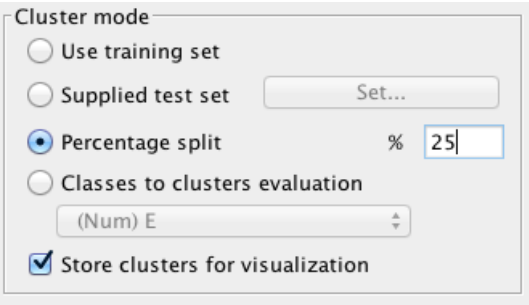


Figure 18: Exploration 1 setting of 25 percentage cluster split

Time taken to build model (percentage split) : 8.33 seconds

Clustered Instances

1	88	(1%)
2	199	(2%)
3	465	(6%)
4	530	(6%)
5	649	(8%)
6	2334	(28%)
7	3635	(44%)
8	392	(5%)

Log likelihood: 1.16942

Figure 19: 25 percent split result: 8 cluster division



Figure 20: 25 percent split cluster visualisation

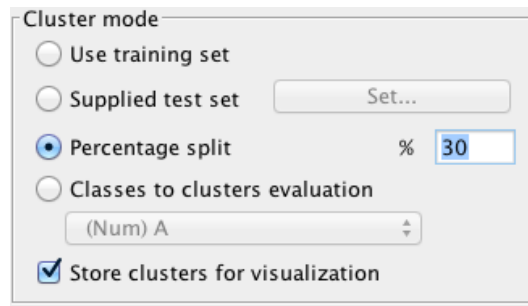


Figure 21: Secondary setting of 30 percentage cluster split

Time taken to build model (percentage split) : 2.29 seconds

Clustered Instances

0	2915 (38%)
1	3871 (50%)
2	953 (12%)

Log likelihood: -2.40087

Figure 22: 30 percent split result: 3 cluster division

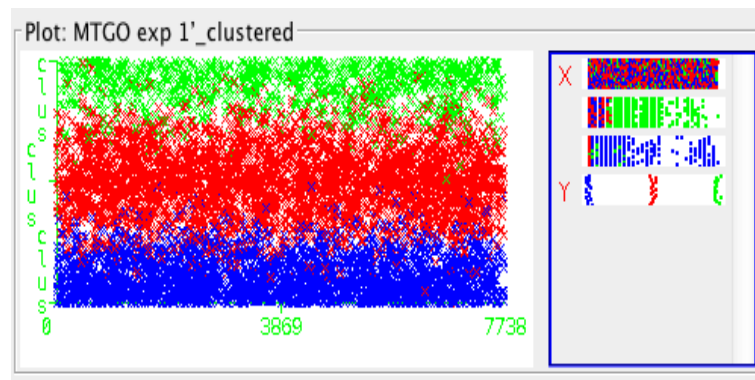


Figure 23: 30 percent split cluster visualisation

Formal Model

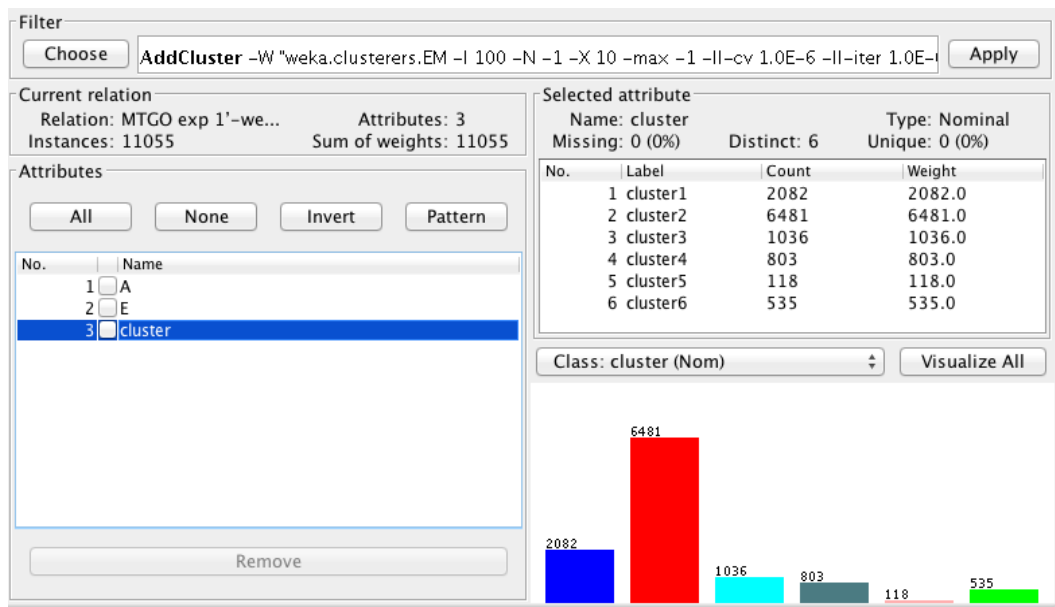


Figure 24: Applying an EM cluster filter as an attribute

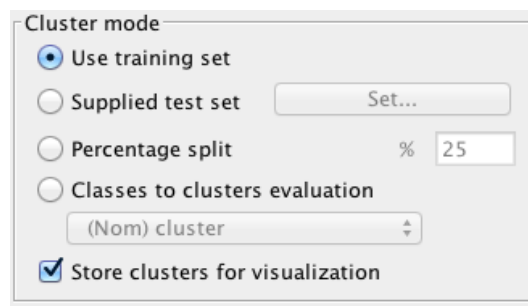


Figure 25: Training set applied to .ARFF with EM clustering

Time taken to build model (full training data) : 38.49 sec

=== Model and evaluation on training set ===

Clustered Instances

0	50 (0%)
1	779 (7%)
2	1301 (12%)
3	1036 (9%)
4	803 (7%)
5	118 (1%)
6	6315 (57%)
7	535 (5%)
8	118 (1%)

Log likelihood: 2.69566

Figure 26: Result of EM training set with cluster filter

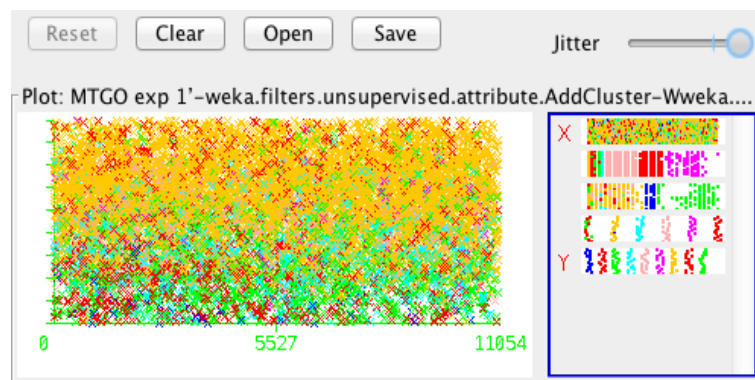


Figure 27: EM training set and cluster filter visualisation

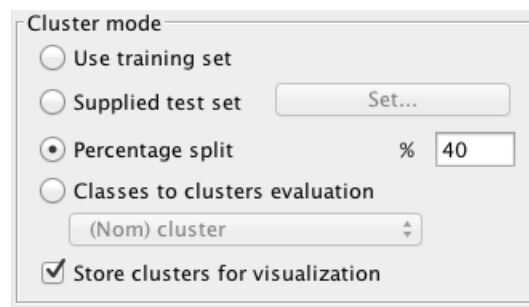


Figure 28: EM clustering with 40 percentage split

Time taken to build model (percentage split) : 3.61 seconds

Clustered Instances

0	3899 (59%)
1	1256 (19%)
2	618 (9%)
3	860 (13%)

Figure 29: EM clustering with 40 percent cluster split result

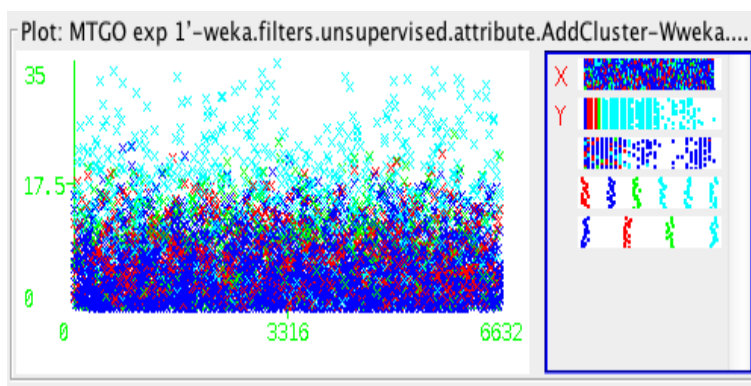


Figure 30: EM clustering with 40 percentage split visualisation

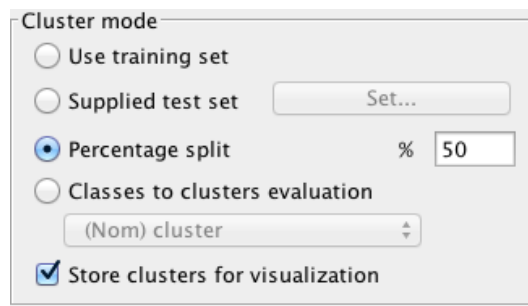


Figure 31: EM clustering with 50 percent cluster split

Time taken to build model (percentage split) : 13.79 sec

Clustered Instances

0	60	(1%)
1	1285	(23%)
2	1957	(35%)
3	673	(12%)
4	390	(7%)
5	512	(9%)
6	266	(5%)
7	385	(7%)

Log likelihood: 1.04076

Figure 32: EM clustering with 50 percent split result

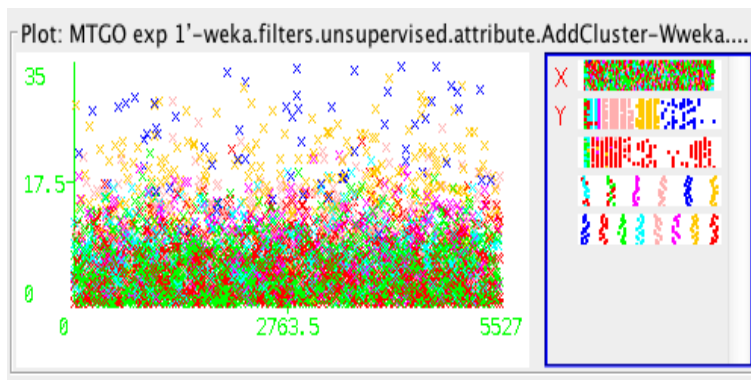


Figure 33: EM clustering with 50 percent split visualisation

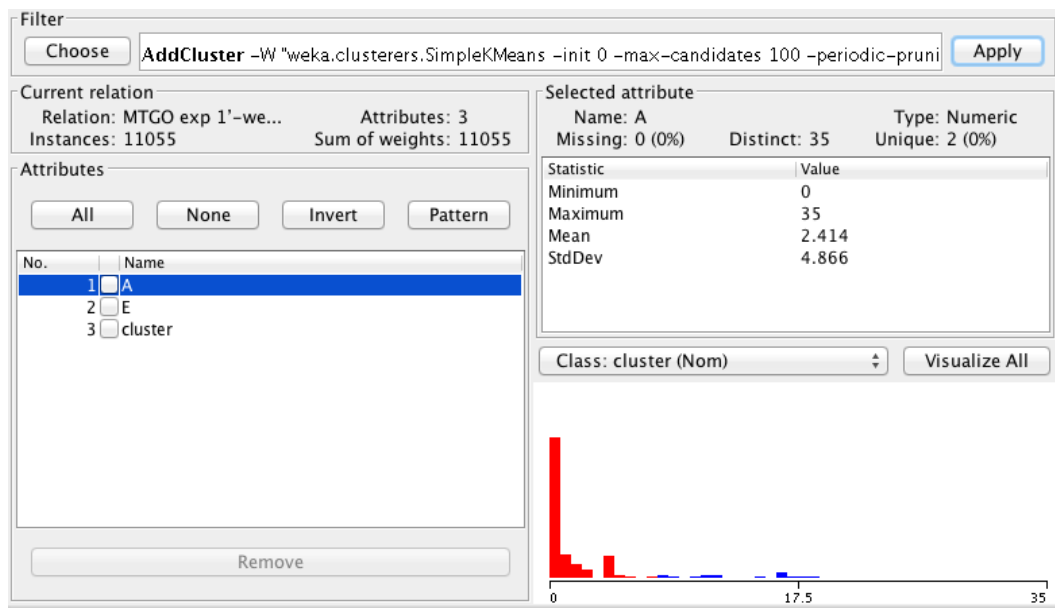


Figure 34: Re-application of cluster filter for SimpleKMeans

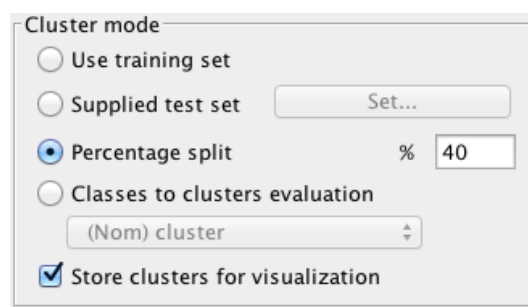


Figure 35: SimpleKMeans clustering with 40 percent cluster split

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0	701 (11%)
1	5932 (89%)

Figure 36: SimpleKMeans clustering with 40 percent split result

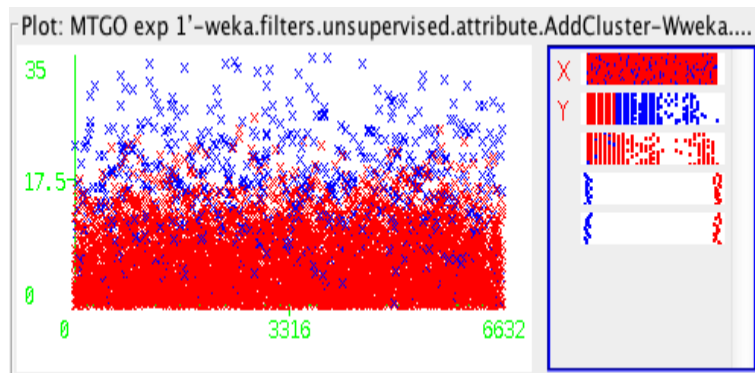


Figure 37: SimpleKMeans clustering with 40 percent split visualisation

Research Question: Is there a relation between Artifact and Enchantment?

Answer: No

Exploration 1 Conclusion

From the exploration conducted, it can be concluded that there is no association between Artifact and Event cards in deck construction. While it was initially conceived that there could be a link between the two, after careful analysis and investigation it is shown from figures 24 through to 37 that the predicted rule set does not apply to the data.

6.2 Exploration 2: Is there a relation between Creature, Sorcery and Instant?

Figure 38 shows alterations made to the .arff file. The attributes selected are Creature (C), Sorcery (S) and Instant (I).

```
@relation 'MTGO Data exp 2'

@attribute C numeric
@attribute S numeric
@attribute I numeric

@data
```

Figure 38: .ARFF header file for exploration 2

Predicted outcome: Creature, Sorcery and Instant will correlate with one another. Association rules outlined below will be fulfilled.

Estimated rulesets:

$[C, S] \Rightarrow [I]$

$[S, I] \Rightarrow [C]$

$[C, I] \Rightarrow [S]$

Initial Exploration

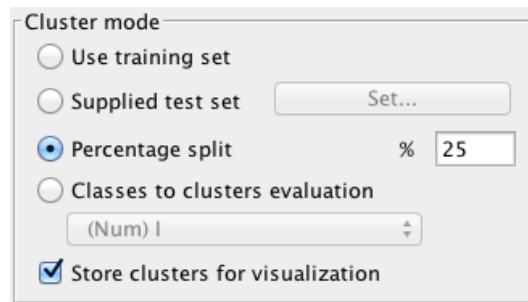


Figure 39: Exploration 2 setting of 25 percentage cluster split

Time taken to build model (percentage split) : 0.03 seconds

Clustered Instances

0	4958 (60%)
1	3334 (40%)

Figure 40: 25 percent split result: 2 cluster division

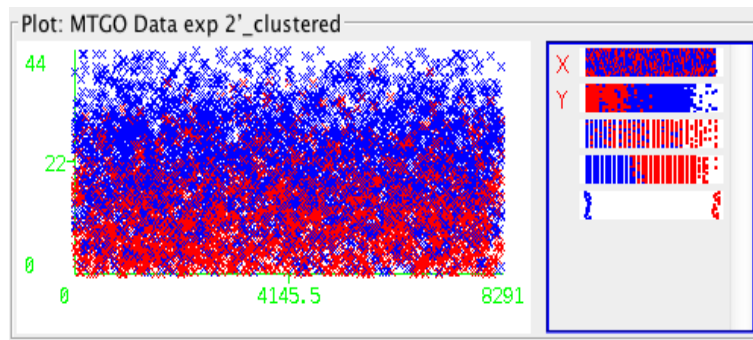


Figure 41: 25 percent split cluster visualisation

Cluster mode

☐ Use training set

☐ Supplied test set

☒ Percentage split %

☐ Classes to clusters evaluation

☒ Store clusters for visualization

Figure 42: Secondary setting of 30 percentage cluster split

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0	4572 (59%)
1	3167 (41%)

Figure 43: 30 percent split result: 2 cluster division

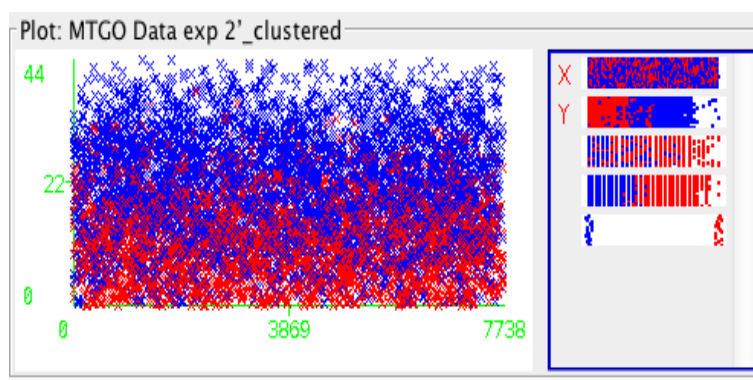


Figure 44: 30 percent split cluster visualisation

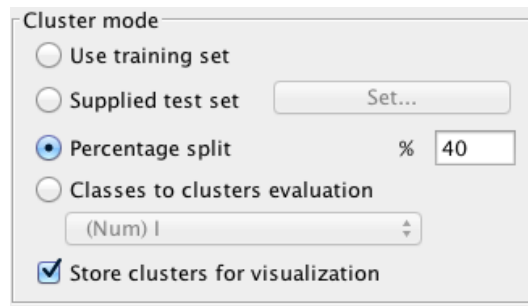


Figure 45: Tertiary setting of 40 percentage cluster split

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0	3835 (58%)
1	2798 (42%)

Figure 46: 40 percent split result: 2 cluster division

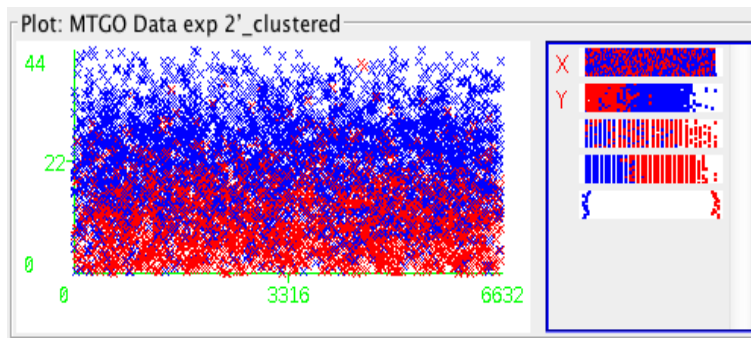


Figure 47: 40 percent split cluster visualisation

Formal Model

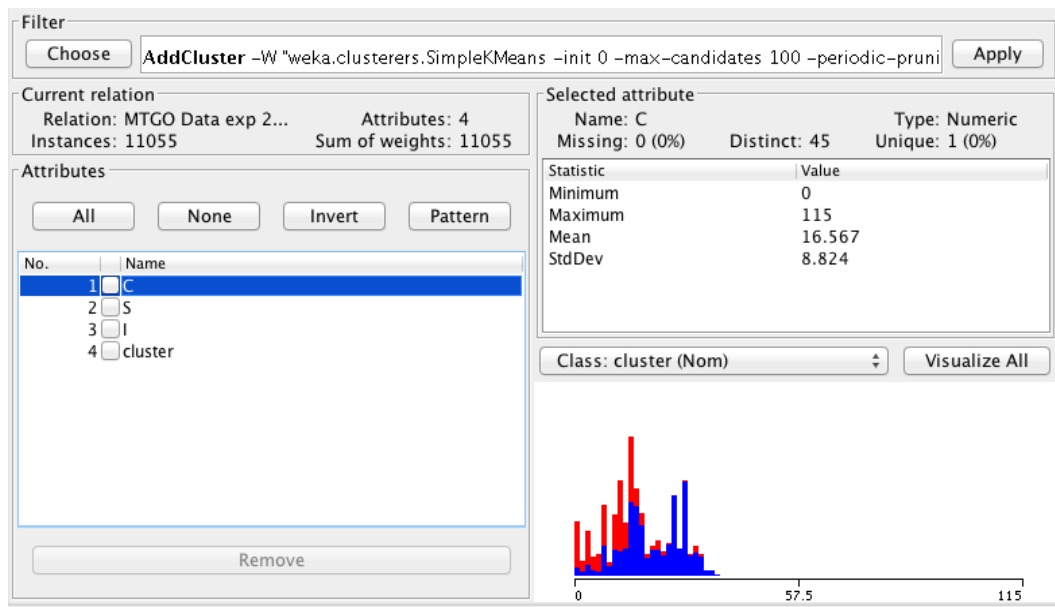


Figure 48: Applying an SimpleKMeans cluster filter as an attribute

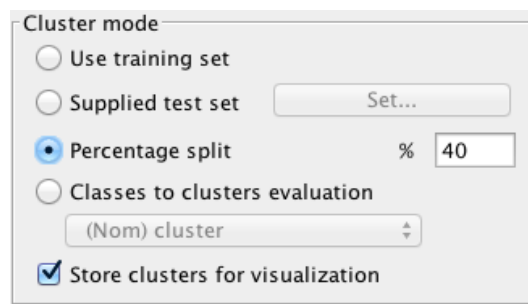


Figure 49: Training set applied to .ARFF with SimpleKMeans clustering

Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0	4220 (64%)
1	2413 (36%)

Figure 50: Result of SimpleKMeans training set with cluster filter

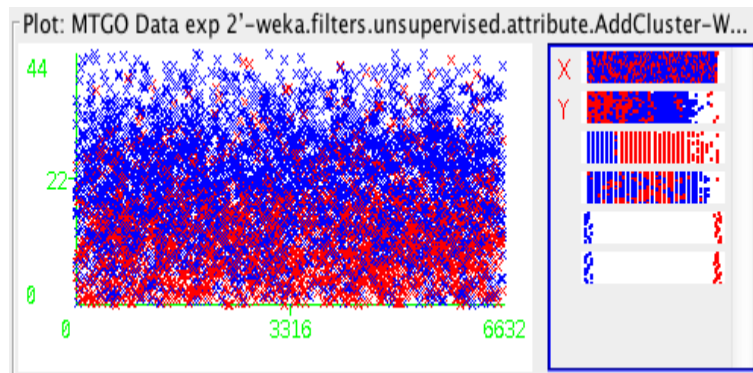


Figure 51: SimpleKMeans training set and cluster filter visualisation

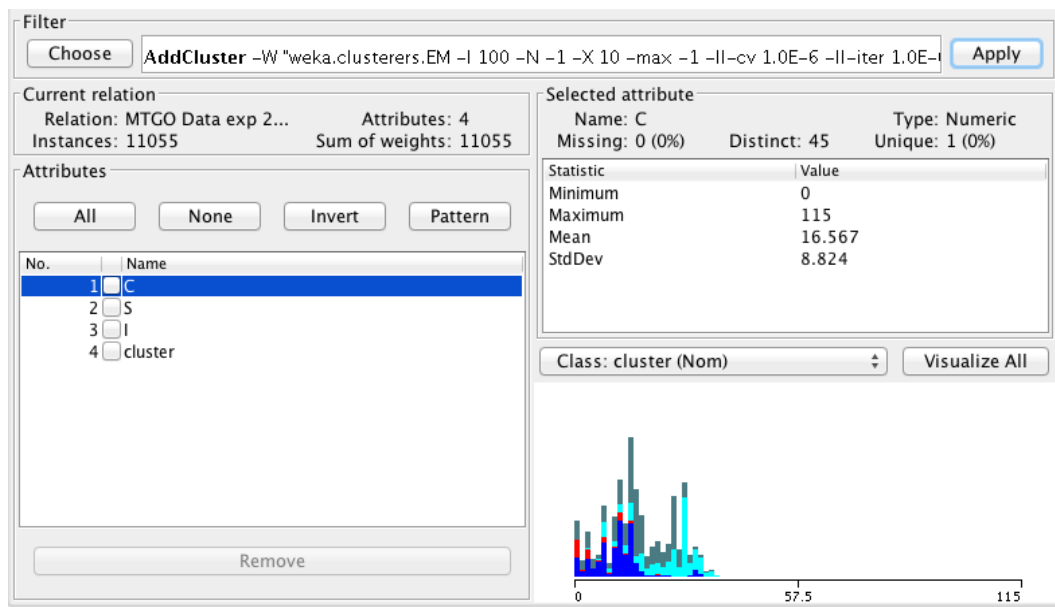


Figure 52: Applying an EM cluster filter as an attribute

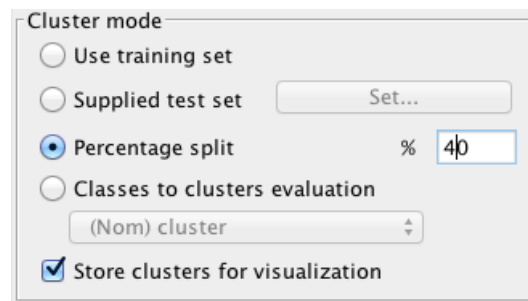


Figure 53: Training set applied to .ARFF with EM clustering

Time taken to build model (percentage split) : 3.83 seconds

Clustered Instances

0	1764 (27%)
1	2877 (43%)
2	1992 (30%)

Figure 54: Result of EM training set with cluster filter

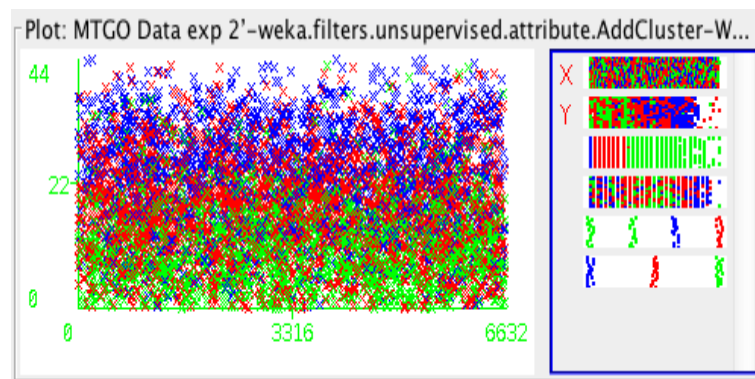


Figure 55: EM training set and cluster filter visualisation

Research Question: Is there a relation between Creature, Sorcery and Instant?

Answer: Yes

Exploration 2 Conclusion

From the examination, it can be concluded that there is a strong indication of association between Creature, Sorcery and Instant cards. Exploration 1 did not lead to promising results for this analysis, however figures 39 to 44 displayed intriguing outcomes.

The outcome from analysis led to a stronger identification of the association rules fulfilment, therefore the formal model was created shown in figures 48 to 55; results from the formal model if developed further could lead to prove outright the association rules application to the data.

6.3 Exploration 3: Exponential smoothing/time series analysis to predict future wins and losses

Figure 56 shows alterations made to the .arff file. The attributes selected are the Rank (R), number of player wins (Win), number of player losses (Loss) and the date of the event (E-Date).

```
@relation 'MTGO Data exp 3'

@attribute Rank numeric
@attribute Win {3,4}
@attribute Loss {0,1}
@attribute E-Date date dd/mm/yy

@data
1,4,0,02/06/14
2,4,0,02/06/14
```

Figure 56: .ARFF header file for exploration 2

Predicted outcome: Rough predictions of potential wins and losses for future events based on the temporal data presented.

Initial Exploration

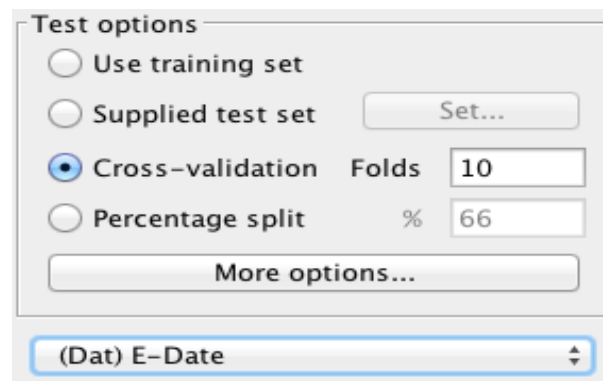


Figure 57: Initial setting of Holt Winters time series 10 fold cross validation

Time taken to build model: 0.01 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient	0.0037
Mean absolute error	700810614.2425
Root mean squared error	841181337.6254
Relative absolute error	111.0033 %
Root relative squared error	115.6129 %
Total Number of Instances	11055

Figure 58: Result of 10 fold cross validation

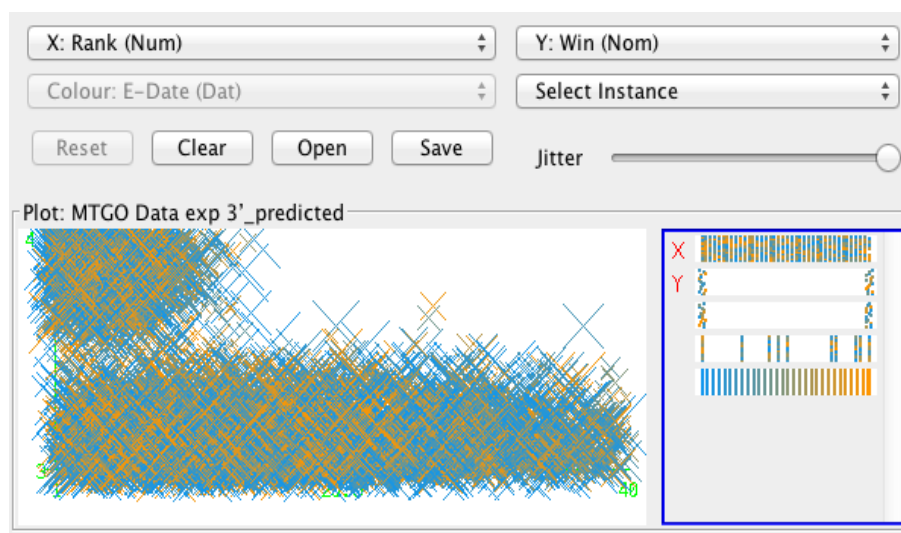


Figure 59: Visualisation result of 10 fold cross validation (Wins)

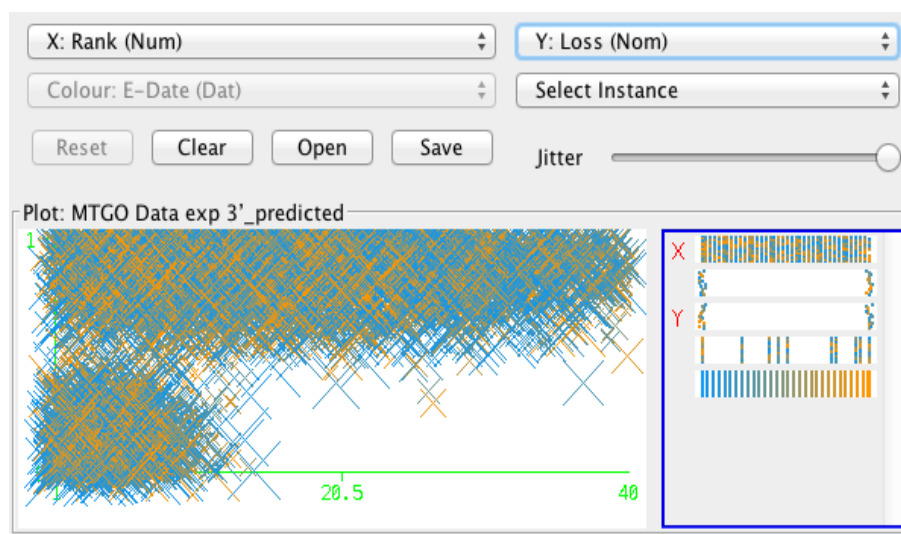


Figure 60: Visualisation result of 10 fold cross validation (Loss)

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds

☒ Percentage split %

More options...

(Dat) E-Date

Figure 61: Holt Winters time series with 66 percent data split

Time taken to test model on training split: 0 seconds

=== Summary ===

Correlation coefficient	0
Mean absolute error	645257601.9443
Root mean squared error	744481987.0281
Relative absolute error	102.7957 %
Root relative squared error	103.0326 %
Total Number of Instances	3759

Figure 62: Result of time series with 66 percent data split

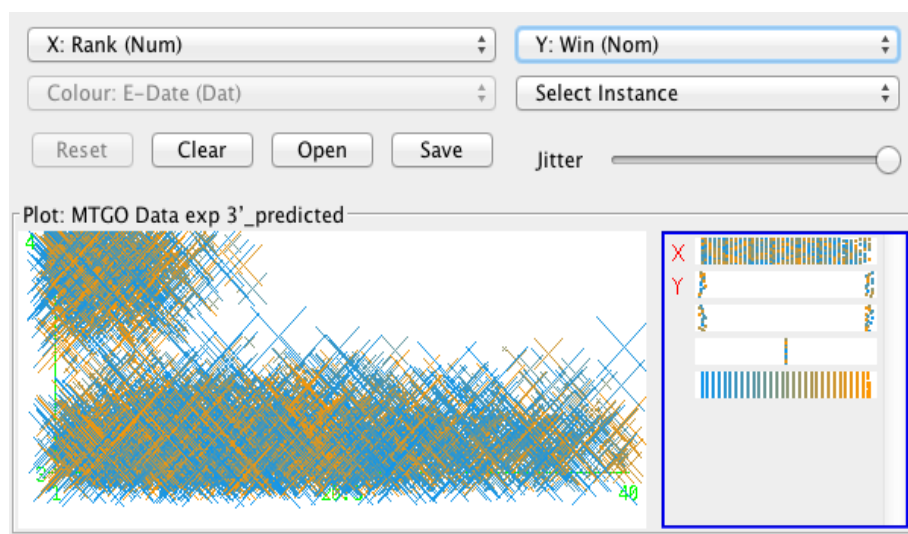


Figure 63: Visualisation of time series with 66 percent data split (Wins)

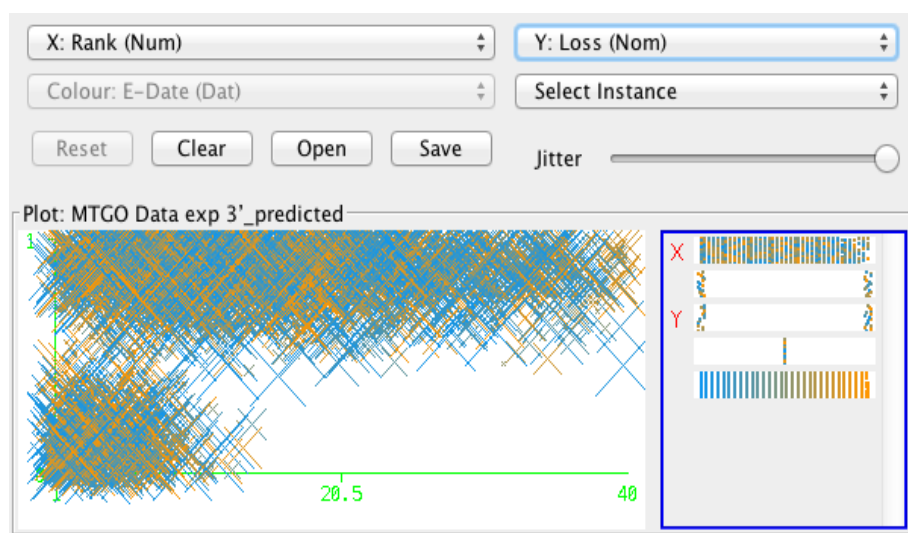


Figure 64: Visualisation of time series with 66 percent data split (Loss)

Formal Model

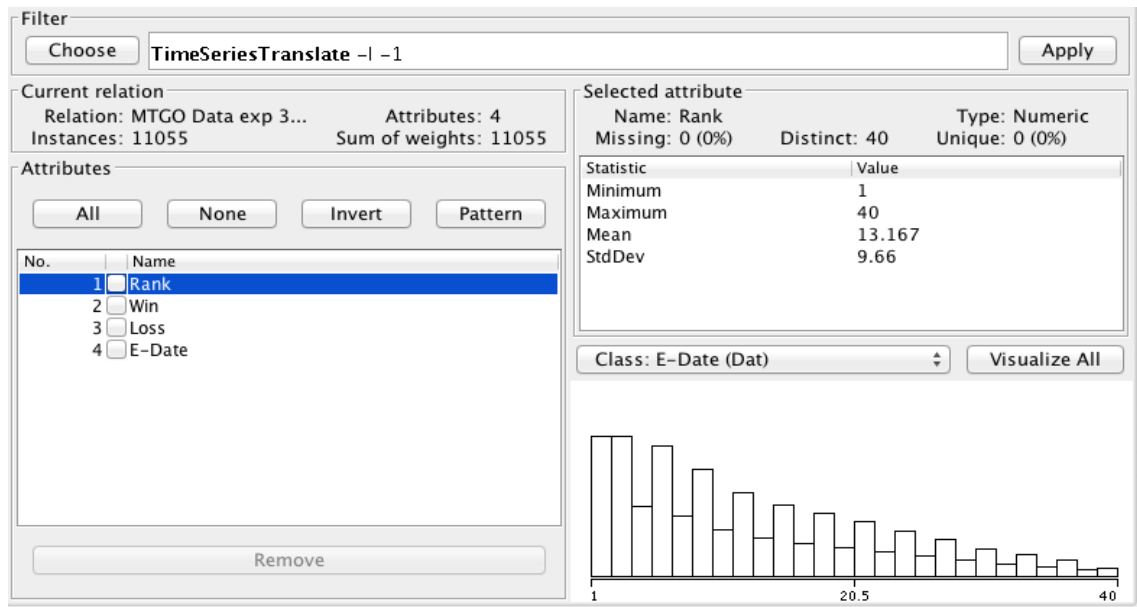


Figure 65: Application of TimeSeriesTranslate filter attribute

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 60

More options...

(Dat) E-Date

Figure 66: Holt Winters time series with 60 percent split

Time taken to test model on training split: 0.01 seconds

=== Summary ===

Correlation coefficient	0
Mean absolute error	637220924.4748
Root mean squared error	743712527.6533
Relative absolute error	101.6191 %
Root relative squared error	102.8871 %
Total Number of Instances	4422

Figure 67: Result of time series with 60 percent split

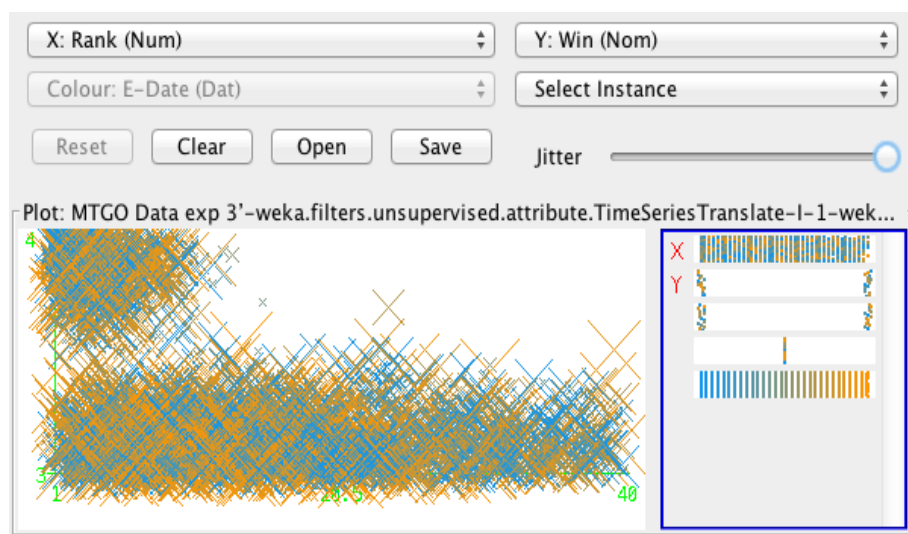


Figure 68: Visualisation of time series with 60 percent split (Wins)

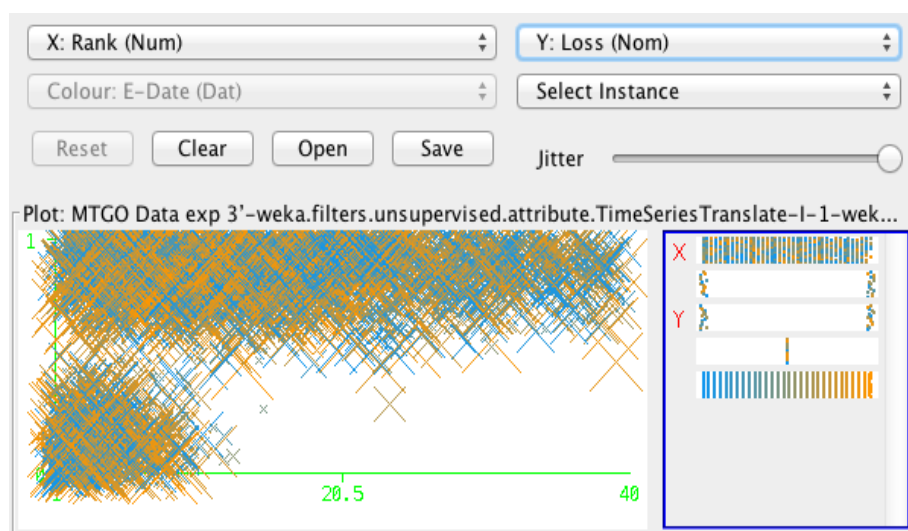


Figure 69: Visualisation of time series with 60 percent split (Loss)

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds

☒ Percentage split %

More options...

(Dat) E-Date

Figure 70: Holt Winters time series with 50 percent split (E-Date)

Time taken to test model on training split: 0.01 seconds

=== Summary ===

Correlation coefficient	0
Mean absolute error	967636208.9848
Root mean squared error	1181657066.6591
Relative absolute error	153.6046 %
Root relative squared error	162.8337 %
Total Number of Instances	5527

Figure 71: Result of time series with 50 percent split

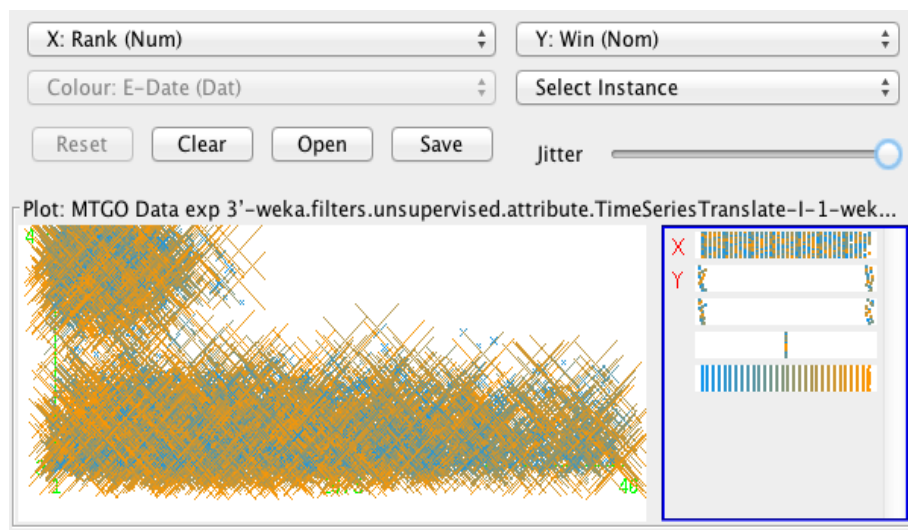


Figure 72: Visualisation of time series with 50 percent split (Wins)

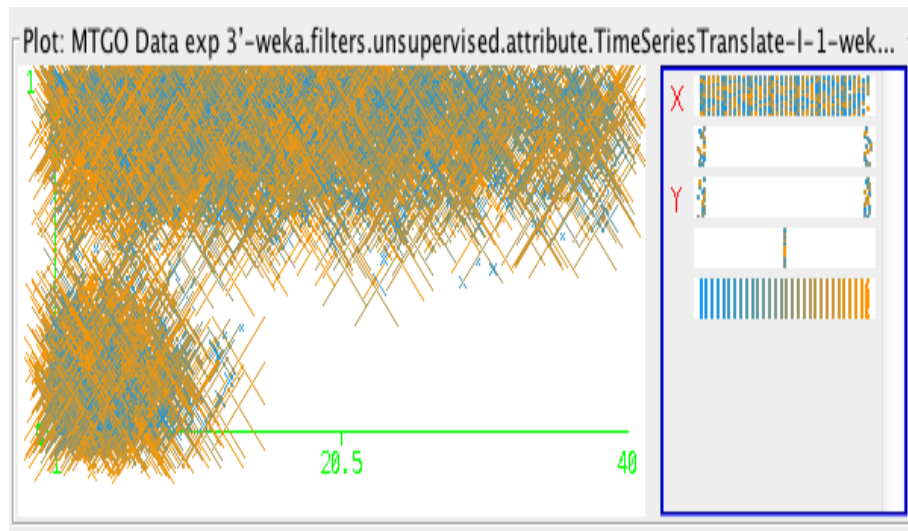


Figure 73: Visualisation of time series with 50 percent split (Loss)

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation Folds

☒ Percentage split %

Figure 74: Holt Winters time series with 50 percent split (Rank)

Time taken to test model on training split: 0 seconds

=== Summary ===

Correlation coefficient	0
Mean absolute error	8.65
Root mean squared error	10.0032
Relative absolute error	107.4226 %
Root relative squared error	103.2793 %
Total Number of Instances	5527

Figure 75: Result of time series with 50 percent split

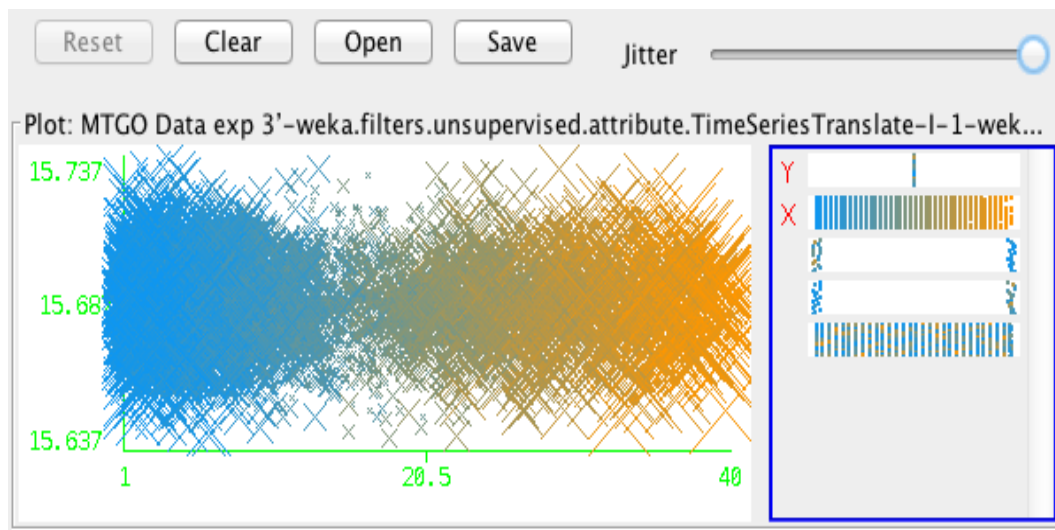


Figure 76: Visualisation of time series with 50 percent split (Wins)

Outcome

Initial exploration of exponential smoothing as shown in figures 57 to 64 did not show results that represented predictive data in a graph format, however the results displayed were rather interesting and seemed worth exploring. Upon conducting a formal exploration, time series seems to have been successful yielding predictive results of both wins and losses for over 5000 instances.

Figures 65 to 76 show what appears to be initial data (points shown in blue Xs) overlaid with predictive data (points shown in orange Xs). Many of these instances fall within the data collected and visualised in WEKA demonstrate a conformance with the already defined temporal data suggesting prediction for the attributes.

This approach could be incorrect for time series data due to the fact that it does not fall into the typical format required for temporal visualisation. There is a danger that this data is being viewed in a biased nature, therefore it would be prudent to conduct further exploration into time series analysis. Doing this would allow for a more complete and thorough experimentation of temporal data and would create results of more certainty.

7 Problems encountered

During the process of data collection and cleansing, there have been a number of problems encountered; upon identification, these concerns were documented accordingly. Below is a detailed specification of the problem areas and a solution enacted to solve it.

7.1 Raw data extraction

The initial data that was to be compiled into the customised data set required some analysis and careful thought before its inception. Upon thorough investigation into the raw data [1], it was identified that the web page presented to a user was constructed from multiple sources. This revealed a problem in that the data is not downloadable into a singular file and each particular unique event is segregated onto a separate web page comprised of these multiple sources. Shown in figures 77, 78 and 79 this has been done for general usability of the website where raw data is stored, but for the purposes of data collection this is a major issue to encounter.

On top of this, there was no comparable data to obtain before June 2014, which presented issues in limiting the amount of data that could be obtained to create the data set as a whole. It was imperative to select data within at least a 6 month bracket to allow for time series analysis to be conducted. This also aided in determining the granular level of detail for the data, which was selected to be daily events for every day over the specified time period to provide daily snapshots of data.

Due to the complex nature that the raw data has been presented in [1], it provided little options to conduct a prompt collection of the data. With this in mind, it became essential to avoid manually typing data for each record into the data set and to instead search for an automated solution. Removing manual data entry from the process would not only save time but also provide the means to collect large amounts of data faster and provide a greater resource of stored information to work with.

[NEW TO MAGIC?](#)
[GAME INFO](#)
[EVENTS](#)
[ARTICLES](#)
[COMMUNITY](#)
[MY ACCOUNTS](#)

[MTGO.COM](#)
[WHAT IS MAGIC ONLINE?](#)
[SCHEDULES](#)
[DECKLISTS](#)
[RESOURCES](#)
[CREATE ACCOUNT](#)

DECKLISTS

Check out the latest deck tech from Magic Online events with our event coverage.

FROM
mm/dd/yyyy

TO
mm/dd/yyyy

SORT BY
Newest

SEARCH

ARTICLE - CONSTRUCTED STANDARD EVENT #8124909 ON 04/06/2015 IN SCHEDULED EVENTS APRIL 7, 2015

STANDARD DAILY

Decklists Standings Draft viewer See the draft viewer for this event here! Decklists nikouni (4-0) ...

ARTICLE - CONSTRUCTED PAUPER EVENT #8124906 ON 04/06/2015 IN SCHEDULED EVENTS APRIL 7, 2015

PAUPER DAILY

Decklists Standings Draft viewer See the draft viewer for this event here! Decklists cspickle (4-0) ...

ARTICLE - CONSTRUCTED MODERN EVENT #8124912 ON 04/06/2015 IN SCHEDULED EVENTS APRIL 7, 2015

MODERN DAILY

Decklists Standings Draft viewer See the draft viewer for this event here! Decklists _megafone_ (4-0) ...

ARTICLE - CONSTRUCTED LEGACY EVENT #8124904 ON 04/06/2015 IN SCHEDULED EVENTS APRIL 7, 2015

LEGACY DAILY

Decklists Standings Draft viewer See the draft viewer for this event here! Decklists jacetmsst (4-0) ...

ARTICLE - CONSTRUCTED LEGACY EVENT #8124904 ON 04/06/2015 IN SCHEDULED EVENTS APRIL 7, 2015

LEGACY DAILY

Decklists Standings Draft viewer See the draft viewer for this event here! Decklists jacetmsst (4-0) ...

Figure 77: Web page to search daily events

Decklists		
NIKOUNI (4-0) STANDARD DAILY #8124909 ON 04/06/2015		
DECKLIST	STATS	SAMPLE HAND
SORT BY: Overview		
Planeswalker (3) 3 Xenagos, the Reveler	Sorcery (5) 3 Crater's Claws 2 Roast	Land (24) 6 Forest 3 Haven of the Spirit Dragon 2 Mana Confluence 3 Mountain 2 Rugged Highlands 4 Temple of Abandon 4 Wooded Foothills
Creature (24) 2 Ashcloud Phoenix 4 Courser of Kruphix 2 Dragonlord Atarka 4 Elvish Mystic 4 Stormbreath Dragon 4 Sylvan Caryatid 4 Thunderbreak Regent	Instant (4) 4 Draconic Roar	60 Cards
Sideboard (15) 1 Roast 2 Arbor Colossus 2 Destructive Revelry 2 Hornet Nest 3 Mistcutter Hydra 3 Seismic Rupture 2 Wild Slash		

Figure 78: Raw web data

Rank	Name	Points	OMWP	GWP	OGWP
1	nikouni	12	0.6875	1.0000	0.5803
2	Rvng	12	0.6458	0.8000	0.6028
3	_megafone_	12	0.6250	0.8000	0.6614
4	xsarq	12	0.5833	0.8000	0.5042
5	DiscoVick	12	0.5417	0.8889	0.5028
6	MagicLair	12	0.5417	0.8000	0.5083
7	Arianne	12	0.5417	0.7273	0.5306
8	M_Z	12	0.5208	0.8000	0.4881
9	Nachopl	9	0.7083	0.7500	0.5864
10	shamanicgreen	9	0.6875	0.6667	0.6250

Figure 79: Raw web data (cont.)

7.2 Custom web crawler

Phillips [2] aided greatly in the creation of a bespoke and customised software solution that allowed for the automation of obtaining the relevant data required for the creation of a bespoke data set. One of the web crawlers biggest strengths is that it can select and store a few thousand records in the space of an hour, which was a significant improvement upon manual entry that generated at most 260 records in the same time frame.

While this may seem the most ideal solution, some limitations were agreed to be applied upon the software [2]. This is because the web crawler with more work would be capable of not only collecting the data, but to also conduct data cleansing before storage in a .CSV file. This would have been a convenient solution, however automatic data cleansing would need to be programmatically instructed within the web crawler to allow the software to identify specified patterns within the data and to more importantly both classify and correct errors. This approach would have taken a lot of time to produce and for automating the data collection process, it was critical that the web crawler extracted the data in its raw format only.

Despite Phillips [2] best efforts, the web crawler did require some small alterations upon completion. Before alteration of code, the web crawler was not able to facilitate the extraction of some specified data elements that were needed for the completed data set. Based on the code that was already in place the solution was straight forward to rectify and provided the software additional functionality to obtain the data required.

Due to the web presentation of the raw data Phillips [2] suggestion of automatically incrementing a date value at the end of a specified hyperlink was not suitable in collecting the data automatically. This is because each individual web page for every event that required harvesting was unfortunately not predictable as anticipated. To work around this issue, when collecting data the web link of the page had to be manually copied and hard coded into the software solution. Using this method also meant that a date stamp would also need to be hard coded manually before running the program.

7.3 WEKA 3.7.12

While WEKA [3] has provided the means required to conduct data experimentation it has not been without its fault. The data needed for this report required a lot of cleansing in order to get the data into WEKA accordingly, which took

time and effort to achieve.

The version that was chosen for analysis was 3.7.12 in particular the developers version that should have allowed for association and time series analysis in a more practical way. Upon further inspection of the algorithms present within 3.7.12 it was identified that the version did not include the required algorithms originally proposed for association and time series. This was a significant setback to the project and required quick thought to rectify the situation.

Clustering was selected as a substitution for association as it was observed in exploration 1 that the cluster points could still be utilised to answer the research questions and classification was selected to conduct time series analysis. These were not the most ideal selections for the explorations required.

WEKA [3] also suffers from a severe memory heap issue resulting in system crashes and a limit of how far data can be processed. This proved a big problem initially with time series when originally attempting to forecast the data as standard practice dictates. Unfortunately the analysis did not go the way it was intended due to these issues, but some results and phenomena have been observed through the use of differing methods from conventional standards.

7.4 Pre-processing of data

Collection of data using the customised web crawler did present cleansing issues before data was ready for mining analysis. As shown in figure 80, the event number field was one of these matters encountered. When the data was extracted from the web page via the software solution, excess information had been pulled into the field that was not needed. In order to make the field consistent when imported into the data mining software it was mandatory that the data be trimmed down to be acceptable. This did present another issue however as many of the fields that needed to be tailored were of varying lengths. To solve this the formulae utilised was altered to conduct these adjustments and facilitate the matter.

Another issue encountered was identified upon the removal of duplications from the wins and losses columns. Figure 81 shows that the predicted values of 4 and 3 for wins along with 0 and 1 for losses had been accounted for, however another value of 66 wins and 3 losses was shown. This came across as a complete outlier and an error that had occurred during the web crawlers initial procurement of the data. Interestingly, this error was consistent within records identified by the player name Kodi as shown in figure 82. Upon observation of the surrounding records within figure 83 and investigation at the raw data source, it was iden-

tified that the record should have read 3 wins and 1 losses; this was corrected accordingly with no further discrepancies identified.

#7143393
#7143420 on 06/02/20

Figure 80: Event Number pre-processing issue

	A	B
1	4	0
2	3	1
3	66	3

Figure 81: Removal of Win/Loss duplicates

8031	39	'kodi'	66	3	'#7620385'	'Standard Daily'	29/10/2014	6	14	5	10	0	0	25	0
------	----	--------	----	---	------------	------------------	------------	---	----	---	----	---	---	----	---

Figure 82: Record of abnormality identified

8097	20	'robert15184'	3	1	'#7589650'	Standard Daily	28/10/2014	3	21	4	8	0	0	24	0
8098	21	'JaceTheBodySculptor'	3	1	'#7589650'	Standard Daily	28/10/2014	1	35	0	0	0	1	23	0
8099	22	'gokky'	3	1	'#7589650'	Standard Daily	28/10/2014	5	16	3	10	1	0	25	0
8100	23	'rambo88'	3	1	'#7589650'	Standard Daily	28/10/2014	1	23	4	8	0	0	24	0
8101	24	'Marselo'	3	1	'#7589650'	Standard Daily	28/10/2014	0	2	2	25	4	0	27	0
8102	25	'che_rocha'	3	1	'#7589650'	Standard Daily	28/10/2014	4	19	4	9	0	0	24	0
8103	26	'Waterd'	3	1	'#7589650'	Standard Daily	28/10/2014	6	11	4	11	0	3	25	0
8104	27	'kodi'	66	3	'#7589650'	Standard Daily	28/10/2014	6	14	5	10	0	0	25	0
8105	28	'portaro89'	3	1	'#7589650'	Standard Daily	28/10/2014	2	16	0	17	0	1	24	0
8106	29	'blusideoflife'	3	1	'#7589650'	Standard Daily	28/10/2014	0	2	3	24	4	0	27	0
8107	30	'transmongifying'	3	1	'#7589650'	Standard Daily	28/10/2014	6	11	4	12	0	0	25	2
8108	31	'L1X0'	3	1	'#7589650'	Standard Daily	28/10/2014	0	25	4	3	3	2	23	0
8109	32	'Rukif'	3	1	'#7589650'	Standard Daily	28/10/2014	2	13	2	18	0	1	24	0
8110	33	'SFlaren'	3	1	'#7589650'	Standard Daily	28/10/2014	0	21	2	10	1	8	18	0
8111	34	'Dawnelemental'	3	1	'#7589650'	Standard Daily	28/10/2014	2	17	4	9	0	3	25	0

Figure 83: Secondary location of irregularity

8 Reflection

The analysis conducted in this report clearly shows that exploration 2 could be proven to correlate with the outlined association. Data within exploration 2 shows a high consistency of connotation with the attributes suggesting a high likelihood that the data holds a relation. It could be conceived that players combining these kinds of cards choose them based on the combos and synergy that they provide. To explore this however would require far more intrinsic data of each card type and specified properties to provide an even finer granularity to the data.

Exploration 1 could be argued that a relation exists if combined with more attributes. In comparison to exploration 2, the data lacks only 1 attribute but the analysis has not shown proof that the association rule applies to the data. That said if more attributes were to be combined with exploration 1 it could be construed that the data in question could have high likelihood of association.

Time series analysis conducted seems to show predictive data of potential wins and losses for the next month of events, however the result is inconclusive at this time. Certainty cannot be applied to the temporal analysis due to the fact that data result has not been presented in the typical standard. An attempt has been made to achieve time series analysis in a visual form different from the conventional standard but further investigation would be needed to provide certainty of the predictive results in a more understandable format.

Despite setbacks with the creation and analysis of the data, a custom built dataset has been developed and the explorations conducted have produced some results allowing for the research questions to be answered. Further analysis would be required and additional data could be collected to provide a more finite granularity of data for exploration.

9 Research

References

- [1] **Wizards of the Coast.** *Magic the Gathering decklist compendium*
Online: <http://magic.wizards.com/en/gameinfo/products/magiconline/decklists>,
Hasbro, 2014, Accessed: 30th October 2014.
- [2] **Phillips, Jaymie L.** *Private Communication*, Univeristy of Wales Trinity
Saint David's: Swansea Metropolitan Campus, 2014.
- [3] **University of Waikato.** *WEKA: Data mining software in Java* Online:
<http://www.cs.waikato.ac.nz/ml/weka/>, Univeristy of Waikato, 2014, Ac-
cessed: 2nd November 2014.

Bibliography

- **Wizards of the Coast.** *What is Magic: the Gathering?*,
Online: <http://magic.wizards.com/en/new-to-magic>, Accessed: 1st Novem-
ber 2014.
- **Li, D.** *Improvements in Data Mining Association Rules Algorithm*, Inter-
national Journal of Database Theory and Application, Vol 8; No 2, 2015.
- **Gahirwal, M; Vijayalakshmi, M.** *Inter Time Series Sales Forecasting*,
Vivekanand Education Societys Institute of Technology, 2013.