PREDICTING THE WINNING TEAM IN A PREMIER LEAGUE GAME

DATA TRANSFORMATION & INTEGRATION

Cathal Hughes & Russell Brady – Group 40

OUR DATASETS

- Premier League Games Datasets – 2000/01 to 2017/18 – from football-data.co.uk – 18 separate datasets
- This dataset contains information about every premier league game from this time period.
- The number of attributes in these datasets ranged from 28 to 65.
- Lots of irrelevant data – in–game stats, bookies odds

Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	B365H	B365D	B365A	BWH	BWD	BWA
E0	11/08/2017	Arsenal	Leicester	4	3	H	2	2	D	M Dean	27	6	10	3	9	12	9	4	0	1	0	0	1.53	4.5	6.5	1.5	4.6	6.75
E0	12/08/2017	Brighton	Man City	0	2	A	0	0	D	M Oliver	6	14	2	4	6	9	3	10	0	2	0	0	11	5.5	1.33	11	5.25	1.3
E0	12/08/2017	Chelsea	Burnley	2	3	A	0	3	A	C Pawson	19	10	6	5	16	11	8	5	3	3	2	0	1.25	6.5	15	1.22	6.5	12.5
E0	12/08/2017	Crystal Pa	Huddersfi	0	3	A	0	2	A	J Moss	14	8	4	6	7	19	12	9	1	3	0	0	1.83	3.6	5	1.8	3.5	4.75
E0	12/08/2017	Everton	Stoke	1	0	H	1	0	H	N Swarbri	9	9	4	1	13	10	6	7	1	1	0	0	1.7	3.8	5.75	1.7	3.6	5.5
E0	12/08/2017	Southamp	Swansea	0	0	D	0	0	D	M Jones	29	4	2	0	10	13	13	0	2	1	0	0	1.62	4	6.5	1.57	4	6
E0	12/08/2017	Watford	Liverpool	3	3	D	2	1	H	A Taylor	9	14	4	5	14	8	3	3	0	3	0	0	6	4.2	1.62	6	4.2	1.55
E0	12/08/2017	West Bror	Bournemo	1	0	H	1	0	H	R Madley	16	9	6	2	15	3	8	2	3	1	0	0	2.4	3.3	3.3	2.4	3.2	3.1
E0	13/08/2017	Man Unite	West Ham	4	0	H	1	0	H	M Atkinso	22	9	6	1	19	7	11	1	2	2	0	0	1.3	5.75	12	1.28	5.5	11
E0	13/08/2017	Newcastle	Tottenhar	0	2	A	0	0	D	A Marrine	6	18	3	6	6	10	5	7	1	2	1	0	5.5	4	1.7	5.25	3.8	1.67
E0	19/08/2017	Bournemo	Watford	0	2	A	0	0	D	R East	6	19	2	7	6	14	8	5	1	3	0	0	2	3.6	4	1.95	3.5	4
E0	19/08/2017	Burnley	West Bror	0	1	A	0	0	D	M Atkinso	20	8	0	1	11	11	5	5	1	0	0	1	2.63	3.2	3	2.55	3	3.1
E0	19/08/2017	Leicester	Brighton	2	0	H	1	0	H	L Probert	14	5	4	2	8	10	6	2	1	0	0	0	1.73	3.8	5.5	1.75	3.6	5
E0	19/08/2017	Liverpool	Crystal Pa	1	0	H	0	0	D	K Friend	23	4	13	1	12	13	4	2	1	3	0	0	1.36	5.25	9.5	1.34	5.25	9
E0	19/08/2017	Southamp	West Ham	3	2	H	2	1	H	L Mason	14	16	5	8	18	10	7	2	1	1	0	1	1.75	3.8	5.25	1.7	3.8	5
E0	19/08/2017	Stoke	Arsenal	1	0	H	0	0	D	A Marrine	11	18	4	6	6	11	2	9	0	0	0	0	4.5	3.8	1.85	4.5	3.75	1.78
E0	19/08/2017	Swansea	Man Unite	0	4	A	0	1	A	J Moss	6	17	1	8	12	11	3	5	1	1	0	0	11	5	1.36	9.25	4.75	1.36
E0	20/08/2017	Huddersfi	Newcastle	1	0	H	0	0	D	C Pawson	7	13	3	5	13	10	7	3	3	4	0	0	2.5	3.25	3.2	2.4	3.25	3.1
E0	20/08/2017	Tottenhar	Chelsea	1	2	A	0	1	A	A Taylor	18	9	6	2	14	21	14	3	3	3	0	0	2.1	3.6	3.75	2	3.5	3.75
E0	21/08/2017	Man City	Everton	1	1	D	0	1	A	R Madley	19	7	6	2	7	9	7	1	1	2	1	1	1.33	5.75	10	1.33	5.25	9.25
E0	26/08/2017	Bournemo	Man City	1	2	A	1	1	D	M Dean	9	19	3	8	13	14	2	5	5	4	0	1	11	6	1.3	9.25	5.25	1.33
E0	26/08/2017	Crystal Pa	Swansea	0	2	A	0	1	A	A Marrine	16	7	3	3	9	6	1	1	4	1	0	0	1.95	3.5	4.33	1.9	3.5	4.2
E0	26/08/2017	Huddersfi	Southamp	0	0	D	0	0	D	S Attwell	16	6	6	3	10	10	5	4	0	1	0	0	3.9	3.4	2.1	3.9	3.2	2.1
E0	26/08/2017	Man Unite	Leicester	2	0	H	0	0	D	M Oliver	22	11	7	4	8	7	9	3	1	2	0	0	1.33	5.5	11	1.34	5	9.75
E0	26/08/2017	Newcastle	West Ham	3	0	H	1	0	H	N Swarbri	16	8	8	3	17	11	7	5	1	3	0	0	2.3	3.4	3.4	2.25	3.25	3.4

WHY DID WE CHOOSE THIS WORKSHOP?

- Our initial dataset had no attributes that we could use directly to make a prediction.
- Techniques outlined in this workshop would allow us to create relevant attributes – aggregation/attribute construction
- As the data was in raw form – potentially needed to use normalisation
- We had 18 datasets – these need to be integrated.

POSSIBLE ATTRIBUTES

- We removed attributes that we didn't want to include in the prediction – in-game stats, bookies odds
- We still had no clear attributes to make a prediction with.
- Data transformation techniques - generate some attributes.
- Aggregation - Date, Full Time Result, Goals Scored etc – Home Team Points, Away Team Points
- Attribute Construction – Generate Form

Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR
E0	10/8/2018	Man United	Leicester	2	1	H
E0	11/8/2018	Bournemouth	Cardiff	2	0	H
E0	11/8/2018	Fulham	Crystal Palac	0	2	A
E0	11/8/2018	Huddersfield	Chelsea	0	3	A
E0	11/8/2018	Newcastle	Tottenham	1	2	A

```python
def get_goals_scored(playing_stat):
    # Create a dictionary with team names as keys
    teams = {}
    for i in playing_stat.groupby('HomeTeam').mean().T.columns:
        teams[i] = []

    # the value corresponding to keys is a list containing the match location
    for i in range(len(playing_stat)):
        HTGS = playing_stat.iloc[i]['FTHG']
        ATGS = playing_stat.iloc[i]['FTAG']
        teams[playing_stat.iloc[i].HomeTeam].append(HTGS)
        teams[playing_stat.iloc[i].AwayTeam].append(ATGS)

    # Create a dataframe for goals scored where rows are teams and cols are m
    GoalsScored = pd.DataFrame(data=teams, index = [i for i in range(1,39)]).
    GoalsScored[0] = 0
    # Aggregate to get uptil that point
    for i in range(2,39):
        GoalsScored[i] = GoalsScored[i] + GoalsScored[i-1]
    return GoalsScored

# Gets the goals conceded agg arranged by teams and matchweek
def get_goals_conceded(playing_stat):
    # Create a dictionary with team names as keys
    teams = {}
    for i in playing_stat.groupby('HomeTeam').mean().T.columns:
        teams[i] = []

    # the value corresponding to keys is a list containing the match location
    for i in range(len(playing_stat)):
        ATGC = playing_stat.iloc[i]['FTHG']
        HTGC = playing_stat.iloc[i]['FTAG']
        teams[playing_stat.iloc[i].HomeTeam].append(HTGC)
        teams[playing_stat.iloc[i].AwayTeam].append(ATGC)

    # Create a dataframe for goals scored where rows are teams and cols are m
    GoalsConceded = pd.DataFrame(data=teams, index = [i for i in range(1,39)]
    GoalsConceded[0] = 0
    # Aggregate to get uptil that point
    for i in range(2,39):
        GoalsConceded[i] = GoalsConceded[i] + GoalsConceded[i-1]
    return GoalsConceded


def get_gss(playing_stat):
    GC = get_goals_conceded(playing_stat)
    GS = get_goals_scored(playing_stat)

    j = 0
    HTGS = []
    ATGS = []
    HTGC = []
    ATGC = []

    for i in range(380):
        ht = playing_stat.iloc[i].HomeTeam
        at = playing_stat.iloc[i].AwayTeam
        HTGS.append(GS.loc[ht][j])
        ATGS.append(GS.loc[at][j])
        HTGC.append(GC.loc[ht][j])
        ATGC.append(GC.loc[at][j])

        if ((i + 1)% 10) == 0:
            j = j + 1

    playing_stat['HTGS'] = HTGS
    playing_stat['ATGS'] = ATGS
    playing_stat['HTGC'] = HTGC
    playing_stat['ATGC'] = ATGC

    return playing_stat
```

DATA TRANSFORMATION - AGGREGATION

- Carry out transformation across all 18 datasets.
- Then integrate them.
- Easier to create attributes year by year.
- Arduous Process – as each attribute needed to abide by the date
- HTGD ATGD Code

```python
# Apply to each dataset
stats1 = get_gss(stats1)
stats2 = get_gss(stats2)
stats3 = get_gss(stats3)
stats4 = get_gss(stats4)
stats5 = get_gss(stats5)
stats6 = get_gss(stats6)
stats7 = get_gss(stats7)
stats8 = get_gss(stats8)
stats9 = get_gss(stats9)
stats10 = get_gss(stats10)
stats11 = get_gss(stats11)
stats12 = get_gss(stats12)
stats13 = get_gss(stats13)
stats14 = get_gss(stats14)
stats15 = get_gss(stats15)
stats16 = get_gss(stats16)
stats17 = get_gss(stats17)
stats18 = get_gss(stats18)
```

```python
def get_matches(playing_stat):
    # Create a dictionary with team names as keys
    teams = {}
    for i in playing_stat.groupby('HomeTeam').mean().T.columns:
        teams[i] = []

    # the value corresponding to keys is a list containing the match result
    for i in range(len(playing_stat)):
        if playing_stat.iloc[i].FTR == 'H':
            teams[playing_stat.iloc[i].HomeTeam].append('W')
            teams[playing_stat.iloc[i].AwayTeam].append('L')
        elif playing_stat.iloc[i].FTR == 'A':
            teams[playing_stat.iloc[i].AwayTeam].append('W')
            teams[playing_stat.iloc[i].HomeTeam].append('L')
        else:
            teams[playing_stat.iloc[i].AwayTeam].append('D')
            teams[playing_stat.iloc[i].HomeTeam].append('D')

    return pd.DataFrame(data=teams, index = [i for i in range(1,39)]).T

def get_form(playing_stat,num):
    form = get_matches(playing_stat)
    form_final = form.copy()
    for i in range(num,39):
        form_final[i] = ''
        j = 0
        while j < num:
            form_final[i] += form[i-j]
            j += 1
    return form_final

def add_form(playing_stat,num):
    form = get_form(playing_stat,num)
    h = ['M' for i in range(num * 10)]  # since form is not available for n MW (
    a = ['M' for i in range(num * 10)]

    j = num
    for i in range((num*10),380):
        ht = playing_stat.iloc[i].HomeTeam
        at = playing_stat.iloc[i].AwayTeam

        past = form.loc[ht][j]               # get past n results
        h.append(past[num-1])                    # 0 index is most recent

        past = form.loc[at][j]               # get past n results.
        a.append(past[num-1])                   # 0 index is most recent

        if ((i + 1)% 10) == 0:
            j = j + 1

    playing_stat['HM' + str(num)] = h
    playing_stat['AM' + str(num)] = a


    return playing_stat
```

DATA TRANSFORMATION – ATTRIBUTE CONSTRUCTION

- Form generated from FTR
- New feature is constructed
- Similar to aggregation, but not a summarization

DATA INTEGRATION – PREMIER LEAGUE STANDINGS

	DiffLP
FTR_A	0.297604
FTR_D	0.041648
FTR_H	-0.30365
HTP	-0.44759
ATP	0.447116
HM1_D	0.016018
HM1_L	0.145605
HM1_W	-0.16606
HM2_D	0.020329
HM2_L	0.125732
HM2_W	-0.14142
HM3_D	0.014648
HM3_L	0.155469
HM3_W	-0.16943
AM1_D	-0.03333
AM1_L	-0.12749
AM1_W	0.152112
AM2_D	-0.01434
AM2_L	-0.14591
AM2_W	0.161384
AM3_D	-0.00774
AM3_L	-0.14423
AM3_W	0.149627
HTGD	-0.47039
ATGD	0.469929
DiffFormPt	-0.38983
DiffLP	1

- Dataset with Premier League Standings for each team in last 18 years.
- Difference in league position – previous year
- Dealing with Null values
- Correlation analysis (Pearson) - ensure we were not introducing a redundancy into our dataset.

Team	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Arsenal	2	2	1	2	1	2	4	4	3	4	3	4	3	4	4	3	2	5	6
Aston Vill	6	8	8	16	6	10	16	11	6	6	6	9	16	15	15	17	20		
Birmingham				13	10	12	18		19		9	18							
Blackburn			10	6	15	15	6	10	7	15	10	15	19						
Blackpool												19							
Bolton			16	17	8	6	8	7	16	13	14	14	18						
Bournemouth																	16	9	12
Bradford	17	20																	
Brighton																			15
Burnley											18					19		16	7
Cardiff															20				
Charlton		9	14	12	7	11	13	19											
Chelsea	5	6	6	4	2	1	1	2	2	3	1	2	6	3	3	1	10	1	5

```
Standings = pd.read_csv(loc + "EPLStandings.csv")
Standings.set_index(['Team'], inplace=True)
Standings = Standings.fillna(20)

def get_last(playing_stat, Standings, year):
    HomeTeamLP = []
    AwayTeamLP = []
    for i in range(380):
        ht = playing_stat.iloc[i].HomeTeam
        at = playing_stat.iloc[i].AwayTeam
        HomeTeamLP.append(Standings.loc[ht][year])
        AwayTeamLP.append(Standings.loc[at][year])
    playing_stat['HomeTeamLP'] = HomeTeamLP
    playing_stat['AwayTeamLP'] = AwayTeamLP
    return playing_stat

stats1 = get_last(stats1, Standings, 0)
stats2 = get_last(stats2, Standings, 1)
stats3 = get_last(stats3, Standings, 2)
stats4 = get_last(stats4, Standings, 3)
stats5 = get_last(stats5, Standings, 4)
stats6 = get_last(stats6, Standings, 5)
stats7 = get_last(stats7, Standings, 6)
stats8 = get_last(stats8, Standings, 7)
stats9 = get_last(stats9, Standings, 8)
stats10 = get_last(stats10, Standings, 9)
stats11 = get_last(stats11, Standings, 10)
stats12 = get_last(stats12, Standings, 11)
stats13 = get_last(stats13, Standings, 12)
stats14 = get_last(stats14, Standings, 13)
stats15 = get_last(stats15, Standings, 14)
stats16 = get_last(stats16, Standings, 15)
stats17 = get_last(stats17, Standings, 16)
stats18 = get_last(stats18, Standings, 17)
```

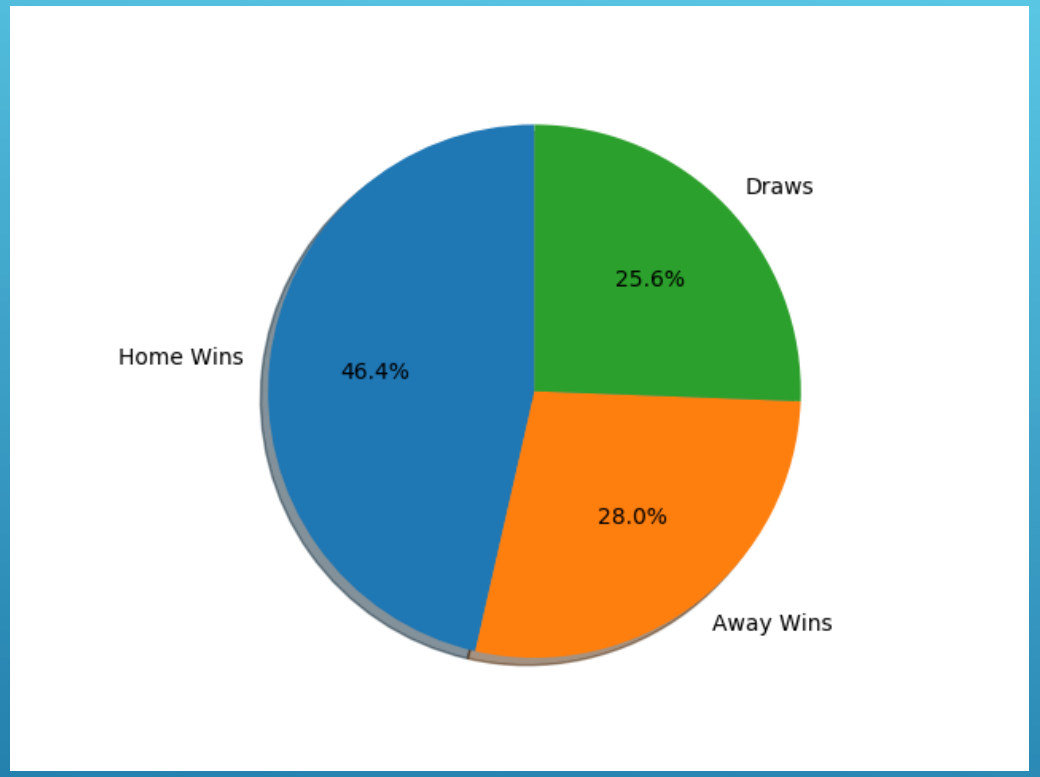
DATA INTEGRATION – PREMIER LEAGUE RESULTS 2000/01 – 2017/18

- We need to convert 18 datasets to 1 dataset.
- Schema Integration/Object Matching – tricky
- Dates in different datasets were formatted differently
- Dropped irrelevant columns

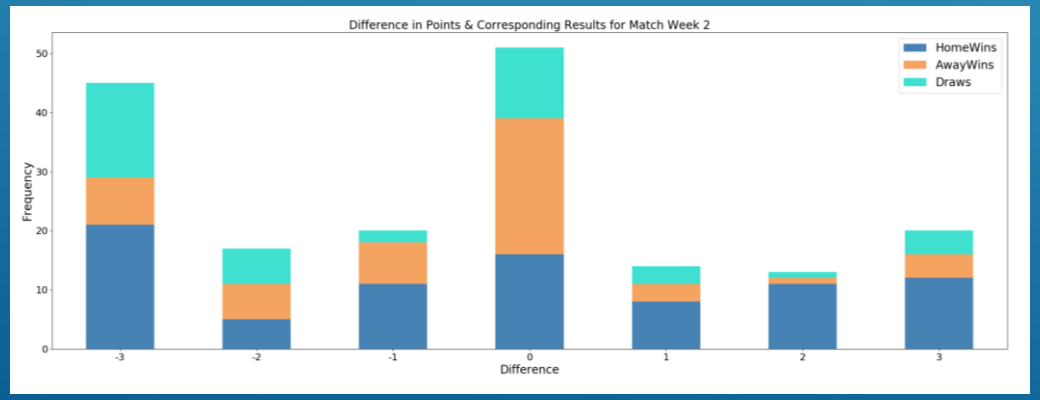
	FTR	HTP	ATP	HM1	HM2	HM3	AM1	AM2	AM3	HTGD	ATGD	DiffPts	DiffFormP	DiffLP
30	H	5	4	D	D	W	D	W	L	2	1	1	1	-16
31	D	3	1	L	L	W	D	L	L	-2	-3	2	2	-2
32	H	4	4	L	D	W	D	W	L	0	1	0	0	-3
33	D	3	2	L	L	W	D	L	D	-1	-1	1	1	5
34	D	4	6	D	L	W	W	W	L	0	3	-2	-2	3
35	A	6	6	W	W	L	W	W	L	0	1	0	0	3
36	D	2	4	D	L	D	L	D	W	-1	1	-2	-2	4
37	H	5	4	D	W	D	W	D	L	1	-1	1	1	-12
38	H	4	2	D	L	W	L	D	D	-1	-2	2	2	-2
39	D	4	6	D	W	L	W	W	L	1	3	-2	-2	15
40	D	6	7	L	W	W	D	L	W	-2	2	-1	-1	11
41	A	4	5	L	W	D	W	L	D	-2	-1	-1	-1	14
42	H	8	3	W	D	W	D	D	L	2	-1	5	5	-7
43	H	7	6	W	D	L	W	L	W	1	-2	1	1	-16
44	H	8	4	W	D	D	D	L	L	8	-2	4	4	-6
45	A	5	4	D	L	D	L	D	W	1	0	1	1	-1
46	D	9	5	W	W	W	D	D	L	3	0	4	4	6
47	D	3	4	D	D	L	D	L	L	-1	-1	-1	-1	-4

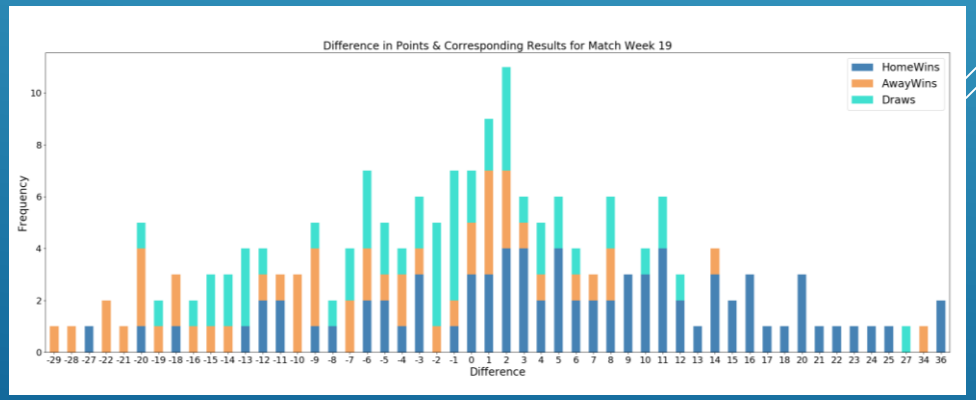
```
def get_date(date):
    if date == '':
        return None
    else:
        return dt.strptime(date, '%d/%m/%y').date()

def get_date_other(date):
    if date == '':
        return None
    else:
        return dt.strptime(date, '%d/%m/%Y').date()


data1.Date = data1.Date.apply(get_date)
data2.Date = data2.Date.apply(get_date)
data3.Date = data3.Date.apply(get_date_other)        # The date format for this dataset is different
data4.Date = data4.Date.apply(get_date)
```

ATTRIBUTE ANALYSIS



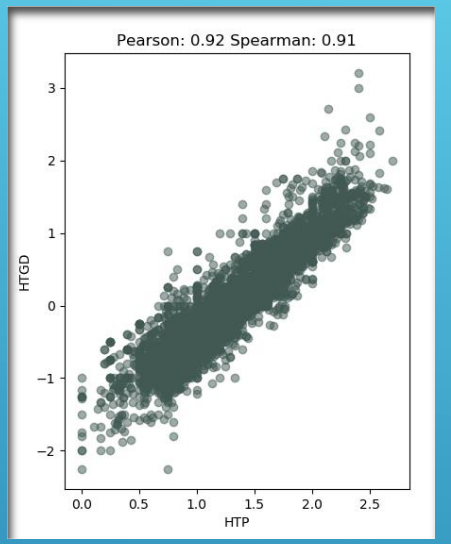
	P(W/H)	P(L/H)	P(D/H)	P(W/A)	P(L/A)	P(D/A)	P(W)	P(L)	P(D)	TotalMatches
LWW	0.219153	0.099448	0.110497	0.171271	0.246777	0.152855	0.390424	0.346225	0.263352	543
WWD	0.302575	0.126609	0.130901	0.130901	0.178112	0.130901	0.433476	0.304721	0.261803	466
WDD	0.220447	0.162939	0.118211	0.124601	0.258786	0.115016	0.345048	0.421725	0.233227	313
DDW	0.184669	0.097561	0.114983	0.139373	0.278746	0.184669	0.324042	0.376307	0.299652	287
DWD	0.226994	0.150307	0.141104	0.168712	0.214724	0.098160	0.395706	0.365031	0.239264	326
WDW	0.265979	0.070103	0.101031	0.202062	0.197938	0.162887	0.468041	0.268041	0.263918	485
DWW	0.250545	0.078431	0.100218	0.246187	0.193900	0.130719	0.496732	0.272331	0.230937	459
WWW	0.279605	0.082237	0.098684	0.239035	0.165570	0.134868	0.518640	0.247807	0.233553	912
WDL	0.283753	0.157895	0.155606	0.109840	0.196796	0.096110	0.393593	0.354691	0.251716	437
DLL	0.237255	0.178431	0.115686	0.098039	0.264706	0.105882	0.335294	0.443137	0.221569	510
LLW	0.177852	0.109060	0.115772	0.161074	0.275168	0.161074	0.338926	0.384228	0.276846	596
DLW	0.139738	0.128821	0.093886	0.168122	0.305677	0.163755	0.307860	0.434498	0.257642	458
LWD	0.238532	0.149083	0.149083	0.121560	0.238532	0.103211	0.360092	0.387615	0.252294	436
DLD	0.202614	0.205882	0.120915	0.124183	0.232026	0.114379	0.326797	0.437908	0.235294	306
LDD	0.181507	0.140411	0.133562	0.136986	0.260274	0.147260	0.318493	0.400685	0.280822	292
WWL	0.296642	0.138060	0.151119	0.111940	0.199627	0.102612	0.408582	0.337687	0.253731	536
WLW	0.202206	0.086397	0.090074	0.176471	0.272059	0.172794	0.378676	0.358456	0.262868	544
DDL	0.300940	0.181818	0.122257	0.106583	0.200627	0.087774	0.407524	0.382445	0.210031	319
DWL	0.264398	0.185864	0.175393	0.054974	0.201571	0.117801	0.319372	0.387435	0.293194	382
WLD	0.223278	0.125891	0.121140	0.130641	0.258907	0.140143	0.353919	0.384798	0.261283	421
LDW	0.199495	0.101010	0.101010	0.161616	0.277778	0.159091	0.361111	0.378788	0.260101	396
DDD	0.274510	0.137255	0.117647	0.147059	0.220588	0.102941	0.421569	0.357843	0.220588	204
WLL	0.244288	0.170475	0.173989	0.110721	0.196837	0.103691	0.355009	0.367311	0.277680	569
LLL	0.194690	0.201011	0.152971	0.082174	0.264223	0.104930	0.276865	0.465234	0.257901	791
LDL	0.228682	0.195736	0.127907	0.091085	0.246124	0.110465	0.319767	0.441860	0.238372	516
LLD	0.201258	0.148847	0.111111	0.113208	0.303983	0.121593	0.314465	0.452830	0.232704	477
LWL	0.248788	0.187399	0.176090	0.072698	0.213247	0.101777	0.321486	0.400646	0.277868	619





CORRELATION COEFFICIENT

- Analysis of relationship between all attributes after individual attribute analysis
- Redundancy detection
- Removal of unnecessary attributes



	FTR_A	FTR_D	FTR_H	HTP	ATP	HM1_D	HM1_L	HM1_W	HM2_D	HM2_L	HM2_W	HM3_D	HM3_L	HM3_W	AM1_D	AM1_L	AM1_W	AM2_D	AM2_L	AM2_W	AM3_D	AM3_L	AM3_W	HTGD	ATGD	DiffFormPt	DiffLP
FTR_A	1	-0.36298	-0.58102	-0.19896	0.239781	-0.02261	0.139537	-0.12381	-0.00678	0.073648	-0.06622	0.01836	0.047415	-0.06429	-0.07408	-0.13874	0.198917	0.002596	-0.06484	0.063504	-0.02111	-0.06068	0.079128	-0.20182	0.243284	-0.20454	0.297604
FTR_D	-0.36298	1	-0.54748	-0.06228	-0.00507	0.103958	-0.02675	-0.069	-0.01024	0.011334	-0.00201	-0.00758	0.032234	-0.02549	0.097759	-0.04815	-0.03995	0.013331	-0.01644	0.004498	-0.0029	0.00122	0.001423	-0.06892	0.002306	-0.04073	0.041648
FTR_H	-0.58102	-0.54748	1	0.233082	-0.21092	-0.0705	-0.10195	0.171459	0.015033	-0.07604	0.061225	-0.00987	-0.07074	0.08	-0.01886	0.166658	-0.14375	-0.01398	0.072596	-0.06096	0.021496	0.053433	-0.07231	0.241446	-0.2205	0.219271	-0.30365
HTP	-0.19896	-0.06228	0.233082	1	0.019031	-0.04157	-0.32258	0.373556	-0.07199	-0.3034	0.361659	-0.06472	-0.315	0.375104	-0.00794	0.0165	-0.00887	-0.00766	-0.0161	0.023372	-0.01526	0.010071	0.003863	0.916189	-0.00776	0.515646	-0.44759
ATP	0.239781	-0.00507	-0.21092	0.019031	1	0.017061	-0.01559	0.000299	-0.00166	-0.00608	0.00744	-0.01308	0.00971	0.002128	-0.06715	-0.31323	0.360728	-0.0573	-0.32593	0.383648	-0.04707	-0.32523	0.364233	0.003808	0.915148	-0.49999	0.447116
HM1_D	-0.02261	0.103958	-0.0705	-0.04157	0.017061	1	-0.4885	-0.42374	-0.00948	-0.00612	0.014432	0.008767	-0.00999	0.002076	0.010741	-0.0235	0.013131	0.003716	-0.00937	0.006119	-0.01537	0.012685	0.001382	-0.01075	0.012423	-0.03952	0.016018
HM1_L	0.139537	-0.02675	-0.10195	-0.32258	-0.01559	-0.4885	1	-0.58336	0.0385	0.046306	-0.07967	-0.0011	0.075505	-0.07482	-0.00856	-0.00815	0.015401	0.004909	0.01563	-0.02038	-0.00714	0.006193	0.000343	-0.31643	-0.01484	-0.26194	0.145605
HM1_W	-0.12381	-0.069	0.171459	0.373556	0.000299	-0.42374	-0.58336	1	-0.03114	-0.04237	0.069266	-0.00702	-0.06908	0.07574	-0.00111	0.030328	-0.02821	-0.00856	-0.0075	0.015458	0.021722	-0.01824	-0.00164	0.338476	0.003838	0.308696	-0.16606
HM2_D	-0.00678	-0.01024	0.015033	-0.07199	-0.00166	-0.00948	0.0385	-0.03114	1	-0.43173	-0.46587	-0.02076	0.000286	0.018565	-0.00722	0.009218	-0.0025	0.004274	-0.0086	0.004824	0.016975	-0.01214	-0.00337	-0.04222	0.000936	-0.04953	0.020329
HM2_L	0.073648	0.011334	-0.07604	-0.3034	-0.00608	-0.00612	0.046306	-0.04237	-0.43173	1	-0.59701	0.029086	0.02989	-0.05643	0.012936	-0.00866	-0.00308	-0.0053	-0.00147	0.006344	-0.00186	0.01428	-0.01244	-0.29159	-0.00958	-0.23678	0.125732
HM2_W	-0.06622	-0.00201	0.061225	0.361659	0.00744	0.014432	-0.07967	0.069266	-0.46587	-0.59701	1	-0.01007	-0.02958	0.038844	-0.00627	0.000299	0.005245	0.001399	0.009087	-0.01051	-0.01328	-0.00321	0.015205	0.323598	0.008568	0.276326	-0.14142
HM3_D	0.01836	-0.00758	-0.00987	-0.06472	-0.01308	0.008767	-0.0011	-0.00702	-0.02076	0.029086	-0.01007	1	-0.45677	-0.44937	-0.00146	0.005304	-0.00382	-0.00228	0.008749	-0.0068	0.029088	0.000792	-0.02713	-0.03402	-0.01016	-0.03762	0.014648
HM3_L	0.047415	0.032234	-0.07074	-0.315	0.00971	-0.00999	0.075505	-0.06908	0.000286	0.02989	-0.02958	-0.45677	1	-0.58944	0.001533	-0.00595	0.004375	-4.1E-05	0.037149	-0.03771	0.007685	-0.02101	0.013819	-0.31282	0.008505	-0.23995	0.155469
HM3_W	-0.06429	-0.02549	0.08	0.375104	0.002128	0.002076	-0.07482	0.07574	0.018565	-0.05643	0.038844	-0.44937	-0.58944	1	-0.00022	0.00116	-0.00093	0.002115	-0.04525	0.044041	-0.03413	0.020384	0.010758	0.345033	0.000681	0.275126	-0.16943
AM1_D	-0.07408	0.097759	-0.01886	-0.00794	-0.06715	0.010741	-0.00856	-0.00111	-0.00722	0.012936	-0.00627	-0.00146	0.001533	-0.00022	1	-0.41515	-0.48313	0.005536	-0.01006	0.005151	-0.03369	0.028434	0.002394	-0.00424	-0.04205	0.055277	-0.03333
AM1_L	-0.13874	-0.04815	0.166658	0.0165	-0.31323	-0.0235	-0.00815	0.030328	0.009218	-0.00866	0.000299	0.005304	-0.00595	0.00116	-0.41515	1	-0.59596	0.008314	0.05177	-0.06022	0.007337	0.048047	-0.05416	0.003883	-0.29823	0.256279	-0.12749
AM1_W	0.198917	-0.03995	-0.14375	-0.00887	0.360728	0.013131	0.015401	-0.02821	-0.0025	-0.00308	0.005245	-0.00382	0.004375	-0.00093	-0.48313	-0.59596	1	-0.01289	-0.04094	0.053406	0.022673	-0.07134	0.050006	1.8E-06	0.324128	-0.29543	0.152112
AM2_D	0.002596	0.013331	-0.01398	-0.00766	-0.0573	0.003716	0.004909	-0.00856	0.004274	-0.0053	0.001399	-0.00228	-4.1E-05	0.002115	0.005536	0.008314	-0.01289	1	-0.46815	-0.44031	-0.01257	0.008147	0.003329	0.001652	-0.02693	0.040293	-0.01434
AM2_L	-0.06484	-0.01644	0.072596	-0.0161	-0.32593	-0.00937	0.01563	-0.0075	-0.0086	-0.00147	0.009087	0.008749	0.037149	-0.04525	-0.01006	0.05177	-0.04094	-0.46815	1	-0.58725	0.039422	0.038405	-0.07368	-0.01939	-0.32152	0.228154	-0.14591
AM2_W	0.063504	0.004498	-0.06096	0.023372	0.383648	0.006119	-0.02038	0.015458	0.004824	0.006344	-0.01051	-0.0068	-0.03771	0.044041	0.005151	-0.06022	0.053406	-0.44031	-0.58725	1	-0.02854	-0.04649	0.071817	0.018192	0.35135	-0.26873	0.161384
AM3_D	-0.02111	-0.0029	0.021496	-0.01526	-0.04707	-0.01537	-0.00714	0.021722	0.016975	-0.00186	-0.01328	0.029088	0.007685	-0.03413	-0.03369	0.007337	0.022673	-0.01257	0.039422	-0.02854	1	-0.44558	-0.46514	-0.0233	-0.0274	0.044734	-0.00774
AM3_L	-0.06068	0.00122	0.053433	0.010071	-0.32523	0.012685	0.006193	-0.01824	-0.01214	0.01428	-0.00321	0.000792	-0.02101	0.020384	0.028434	0.048047	-0.07134	0.008147	0.038405	-0.04649	-0.44558	1	-0.58524	0.021925	-0.30838	0.249322	-0.14423
AM3_W	0.079128	0.001423	-0.07231	0.003863	0.364233	0.001382	0.000343	-0.00164	-0.00337	-0.01244	0.015205	-0.02713	0.013819	0.010758	0.002394	-0.05416	0.050006	0.003329	-0.07368	0.071817	-0.46514	-0.58524	1	-0.00057	0.329755	-0.28705	0.149627
HTGD	-0.20182	-0.06892	0.241446	0.916189	0.003808	-0.01075	-0.31643	0.338476	-0.04222	-0.29159	0.323598	-0.03402	-0.31282	0.345033	-0.00424	0.003883	1.8E-06	0.001652	-0.01939	0.018192	-0.0233	0.021925	-0.00057	1	-0.00179	0.45774	-0.47039
ATGD	0.243284	0.002306	-0.2205	-0.00776	0.915148	0.012423	-0.01484	0.003838	0.000936	-0.00958	0.008568	-0.01016	0.008505	0.000681	-0.04205	-0.29823	0.324128	-0.02693	-0.32152	0.35135	-0.0274	-0.30838	0.329755	-0.00179	1	-0.46129	0.469929
DiffFormPt	-0.20454	-0.04073	0.219271	0.515646	-0.49999	-0.03952	-0.26194	0.308696	-0.04953	-0.23678	0.276326	-0.03762	-0.23995	0.275126	0.055277	0.256279	-0.29543	0.040293	0.228154	-0.26873	0.044734	0.249322	-0.28705	0.45774	-0.46129	1	-0.38983
DiffLP	0.297604	0.041648	-0.30365	-0.44759	0.447116	0.016018	0.145605	-0.16606	0.020329	0.125732	-0.14142	0.014648	0.155469	-0.16943	-0.03333	-0.12749	0.152112	-0.01434	-0.14591	0.161384	-0.00774	-0.14423	0.149627	-0.47039	0.469929	-0.38983	1

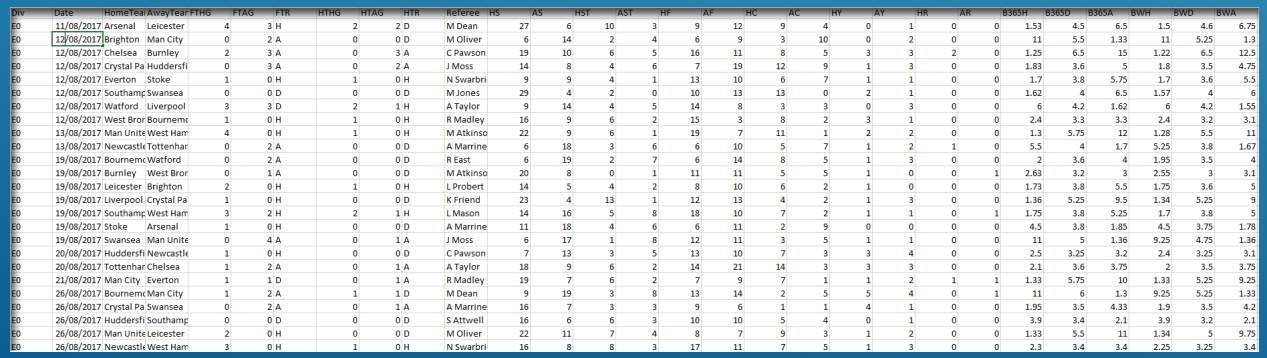
NORMALISATION ?

- We implemented normalisation on continuous data
- Attributes of different units – goals, points, league position
- If you have outliers in your data set, normalizing your data will certainly scale the "normal" data to a very small interval
- Used Decimal Scaling
- Should we normalise?

CONCLUSION

- What we learned:
  - Attributes in dataset may not be relevant on their own
  - Attribute construction and aggregation reveals relationships and discover missing information
  - Normalisation may not be the best solution
  - Attributes may be redundant – Goal Diff and Points
- Implementing transformation and integration techniques has completely changed our dataset

Initial Dataset:



Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR	B365H	B365D	B365A	BWH	BWD	BWA
E0	11/08/2017	Arsenal	Leicester	4	3	H	2	2	D	M Dean	27	6	10	3	9	12	9	4	0	1	0	0	1.53	4.5	6.5	1.5	4.6	6.75
E0	12/08/2017	Brighton	Man City	0	2	A	0	0	D	M Oliver	6	14	2	4	6	9	3	10	0	2	0	0	11	5.5	1.33	11	5.25	1.3
E0	12/08/2017	Chelsea	Burnley	2	3	A	0	3	A	C Pawson	19	10	6	5	16	11	8	5	3	3	2	0	1.25	6.5	15	1.22	6.5	12.5
E0	12/08/2017	Crystal Pa	Huddersfi	0	3	A	0	2	A	J Moss	14	8	4	6	7	19	12	9	1	3	0	0	1.83	3.6	5	1.8	3.5	4.75
E0	12/08/2017	Everton	Stoke	1	0	H	1	0	H	N Swarbri	9	9	4	1	13	10	6	7	1	1	0	0	1.7	3.8	5.75	1.7	3.6	5.5
E0	12/08/2017	Southamp	Swansea	0	0	D	0	0	D	M Jones	29	4	2	0	10	13	13	0	2	1	0	0	1.62	4	6.5	1.57	4	6
E0	12/08/2017	Watford	Liverpool	3	3	D	2	1	H	A Taylor	9	14	4	5	14	8	3	3	0	3	0	0	6	4.2	1.62	6	4.2	1.55
E0	12/08/2017	West Bror	Bournemo	1	0	H	1	0	H	R Madley	16	9	6	2	15	3	8	2	3	1	0	0	2.4	3.3	3.3	2.4	3.2	3.1
E0	13/08/2017	Man Unite	West Ham	4	0	H	1	0	H	M Atkinso	22	9	6	1	19	7	11	1	2	2	0	0	1.3	5.75	12	1.28	5.5	11
E0	13/08/2017	Newcastle	Tottenhar	0	2	A	0	0	D	A Marrine	6	18	3	6	6	10	5	7	1	2	1	0	5.5	4	1.7	5.25	3.8	1.67
E0	19/08/2017	Bournemo	Watford	0	2	A	0	0	D	R East	6	19	2	7	6	14	8	5	1	3	0	0	2	3.6	4	1.95	3.5	4
E0	19/08/2017	Burnley	West Bror	0	1	A	0	0	D	M Atkinso	20	8	0	1	11	11	5	5	1	0	0	1	2.63	3.2	3	2.55	3	3.1
E0	19/08/2017	Leicester	Brighton	2	0	H	1	0	H	L Probert	14	5	4	2	8	10	6	2	1	0	0	0	1.73	3.8	5.5	1.75	3.6	5
E0	19/08/2017	Liverpool	Crystal Pa	1	0	H	0	0	D	K Friend	23	4	13	1	12	13	4	2	1	3	0	0	1.36	5.25	9.5	1.34	5.25	9
E0	19/08/2017	Southamp	West Ham	3	2	H	2	1	H	L Mason	14	16	5	8	18	10	7	2	1	1	0	1	1.75	3.8	5.25	1.7	3.8	5
E0	19/08/2017	Stoke	Arsenal	1	0	H	0	0	D	A Marrine	11	18	4	6	6	11	2	9	0	0	0	0	4.5	3.8	1.85	4.5	3.75	1.78
E0	19/08/2017	Swansea	Man Unite	0	4	A	0	1	A	J Moss	6	17	1	8	12	11	3	5	1	1	0	0	11	5	1.36	9.25	4.75	1.36
E0	20/08/2017	Huddersfi	Newcastle	1	0	H	0	0	D	C Pawson	7	13	3	5	13	10	7	3	3	4	0	0	2.5	3.25	3.2	2.4	3.25	3.1
E0	20/08/2017	Tottenhar	Chelsea	1	2	A	0	1	A	A Taylor	18	9	6	2	14	21	14	3	3	3	0	0	2.1	3.6	3.75	2	3.5	3.75
E0	21/08/2017	Man City	Everton	1	1	D	0	1	A	R Madley	19	7	6	2	7	9	7	1	1	2	1	1	1.33	5.75	10	1.33	5.25	9.25
E0	26/08/2017	Bournemo	Man City	1	2	A	1	1	D	M Dean	9	19	3	8	13	14	2	5	5	4	0	1	11	6	1.3	9.25	5.25	1.33
E0	26/08/2017	Crystal Pa	Swansea	0	2	A	0	1	A	A Marrine	16	7	3	3	9	6	1	1	4	1	0	0	1.95	3.5	4.33	1.9	3.5	4.2
E0	26/08/2017	Huddersfi	Southamp	0	0	D	0	0	D	S Attwell	16	6	6	3	10	10	5	4	0	1	0	0	3.9	3.4	2.1	3.9	3.2	2.1
E0	26/08/2017	Man Unite	Leicester	2	0	H	0	0	D	M Oliver	22	11	7	4	8	7	9	3	1	2	0	0	1.33	5.5	11	1.34	5	9.75
E0	26/08/2017	Newcastle	West Ham	3	0	H	1	0	H	N Swarbri	16	8	8	3	17	11	7	5	1	3	0	0	2.3	3.4	3.4	2.25	3.25	3.4

Final Dataset:

	FTR	HTP	ATP	HM1	HM2	HM3	AM1	AM2	AM3	HTGD	ATGD	DiffFormP	DiffLP
30	H	1.25	1	D	D	W	D	W	L	0.5	0.25	0.25	-16
31	D	0.75	0.25	L	L	W	D	L	L	-0.5	-0.75	0.5	-2
32	H	1	1	L	D	W	D	W	L	0	0.25	0	-3
33	D	0.75	0.5	L	L	W	D	L	D	-0.25	-0.25	0.25	5
34	D	1	1.5	D	L	W	W	W	L	0	0.75	-0.5	3
35	A	1.5	1.5	W	W	L	W	W	L	0	0.25	0	3
36	D	0.5	1	D	L	D	L	D	W	-0.25	0.25	-0.5	4
37	H	1.25	1	D	W	D	W	D	L	0.25	-0.25	0.25	-12
38	H	1	0.5	D	L	W	L	D	D	-0.25	-0.5	0.5	-2
39	D	1	1.5	D	W	L	W	W	L	0.25	0.75	-0.5	15
40	D	1.2	1.4	L	W	W	D	L	W	-0.4	0.4	-0.2	11