FOOTBALL DATA ANALYSIS

Introduction:

Football or Soccer for our American friends is more than just a game. It is played by **250 million** players in over **200 countries** making it the most popular sport. Each of these countries has a domestic league of their own in which teams compete for being labelled the best football team of that country.

Being the staunch football fan I am, I decided to investigate the most popular domestic league of the world ((English Premier League) for factors that can influence the outcome of any match.

This was my first proper data analysis project and I'm not necessarily aiming to use the results obtained to make a model or a prediction system. This project just aims to satisfy my obsession with football as well as get my hands dirty working as a Data Analyst.

Data Set:

The data sets have been attached as separate files in the repository. A file containing the description of the data set has also been attached. Some columns related to betting statistics are missing from the data but playing statistics are there.

What I did and How I approached it:

Questions Posed & Answers

1) How much of a factor is home advantage?
All teams in an EPL season play 19 matches at home and 19 matches at away grounds. I wanted to know whether the teams playing at home had a higher chance of winning than teams playing away.

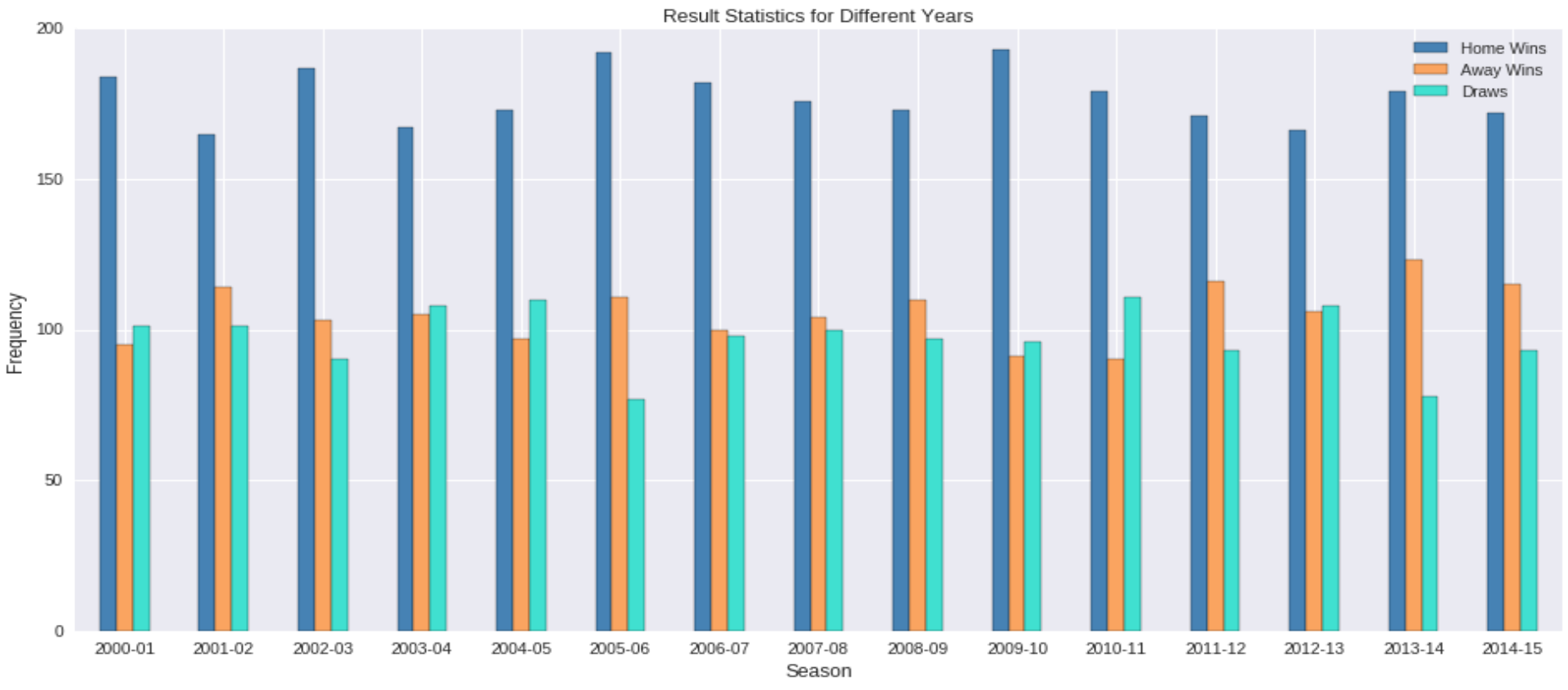
Steps Taken:

1. All the data of the past 15 years was combined and grouped according to the Full Time Result (FTR).
2. The length of each group was evaluated to get Home wins, Away wins and Draws.

Results:

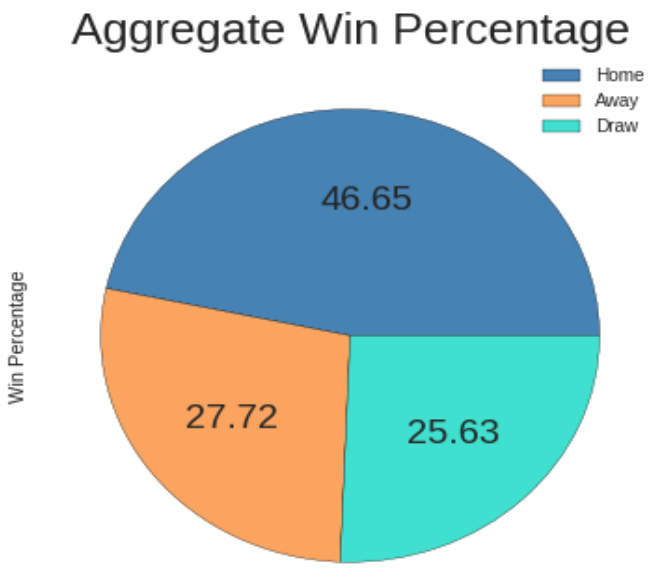
The following table and graph illustrate the number of Home wins, Away wins and Draws in a particular EPL season.

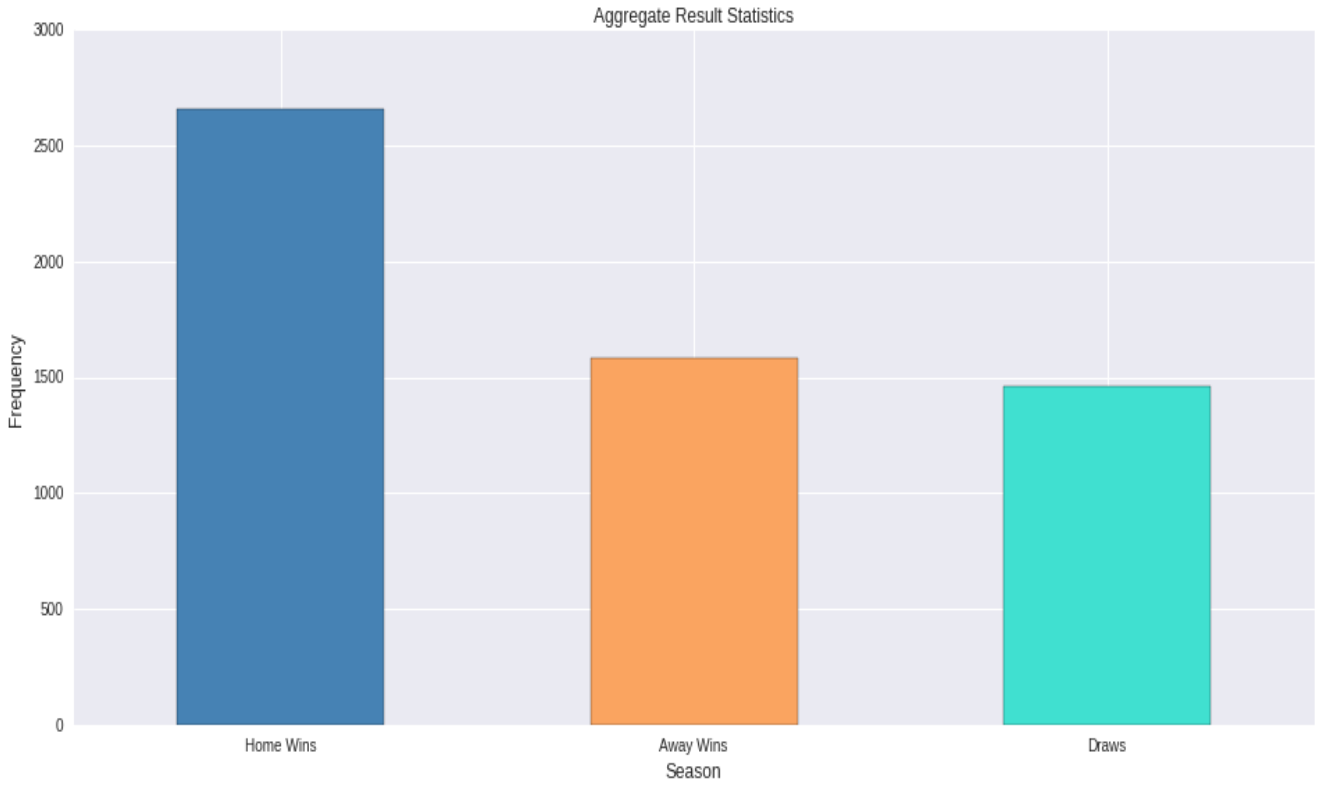
Year	Home Wins	Away Wins	Draws	Total
2000-01	184	95	101	380
2001-02	165	114	101	380
2002-03	187	103	90	380
2003-04	167	105	108	380
2004-05	173	97	110	380
2005-06	192	111	77	380
2006-07	182	100	98	380
2007-08	176	104	100	380
2008-09	173	110	97	380
2009-10	193	91	96	380
2010-11	179	90	111	380
2011-12	171	116	93	380
2012-13	166	106	108	380
2013-14	179	123	78	380
2014-15	172	115	93	380



The following table and graph illustrates the total number of Home wins, Away Wins and Draws in the past 15 years. The aggregate win percentage for the Home side as well as Away side has been represented as a pie chart.

	Home Wins	Away Wins	Draws	Total
Overall	2659	1580	1461	5700





Conclusions:

It is clear from the above data that in each of the EPL seasons, the home teams has won way more matches than the away team. This might be because of the following reasons:

- Football is a team sport and the crowd affects the morale of the team as a whole. While playing at home, the crowd is cheering for the home side and this helps them to perform better.
- Familiarity with the pitch and the weather conditions may also be helping the home team to perform better.
- Home teams don't have to travel long distances to reach the stadiums, whereas the visiting teams might have to travel the entire breadth of the country resulting in fatigue for the visiting team.

2) How do the results of the past matches predict the outcome of the next game?

The results of the past 5 games or past 3 games have traditionally been considered a good enough metric to determine how a team will perform in the current game. Based on the past ‘n’ games we can get an idea of how the team is going to perform. For e.g. - if a team has lost its last 5 matches, it is very likely that it will either lose or draw its next match.

Steps taken:

1. First of all, a separate dictionary was created in which the keys were team names and values were their results for the entire season.
2. Then ‘n’ results were taken at a time for each team starting from Matchweek 1. This way we obtained combinations of match results and the subsequent result.
3. The combinations were then made into keys of a dictionary and values as the subsequent result of each combination
4. The values were then analysed to get probabilities of winning/losing/drawing relative to that combination.

5. This whole process was repeated for all the datasets of the past 15 years and results were grouped by the respective combinations.

Results Obtained:

P (W) - Probability of winning the succeeding match. P
(L) - Probability of losing the succeeding match.
P (D) – Probability of drawing the succeeding match.
The total matches' column refers to the total number of matches preceding which a particular combination was observed

For combination of past 5 games the probabilities were to be found as follows:

Sorted by highest probability of winning (Top 30)

Combinations	P(W)	P(L)	P(D)	Total Matches
WDDWW	0.68	0.12	0.2	25
DWWDW	0.62963	0.074074	0.296296	27
WWDWW	0.627907	0.186047	0.186047	86
WWWWD	0.616279	0.232558	0.151163	86
WWLWD	0.611111	0.222222	0.166667	54
WDDDL	0.6	0.266667	0.133333	15
WWWDW	0.576471	0.247059	0.176471	85
WWWDL	0.575	0.275	0.15	40
WWWWW	0.5625	0.197917	0.239583	192
WDWLD	0.541667	0.25	0.208333	24
WWDLD	0.533333	0.333333	0.133333	15
DWWWL	0.531915	0.212766	0.255319	47
DLWDW	0.529412	0.205882	0.264706	34
WLWWW	0.523077	0.261538	0.215385	65
LDWWW	0.522727	0.136364	0.340909	44
WDWWW	0.520408	0.27551	0.204082	98
DLWWW	0.52	0.3	0.18	50
DDWDW	0.52	0.24	0.24	25
LWDWW	0.517857	0.303571	0.178571	56
DWWWW	0.511628	0.255814	0.232558	86
WDWDW	0.511628	0.232558	0.255814	43
LWWWW	0.505882	0.294118	0.2	85
WLDDD	0.5	0.272727	0.227273	22
DDDLL	0.5	0.3	0.2	20
WDWDL	0.5	0.277778	0.222222	36
DDDDL	0.5	0.3125	0.1875	16
WWDDD	0.5	0.25	0.25	24
WWDLW	0.5	0.314815	0.185185	54
WWDDL	0.5	0.25	0.25	36
DLDDD	0.5	0.416667	0.083333	12

Sorted by highest probability of losing (Top 30)

Combinations	P(W)	P(L)	P(D)	TotalMatches
DLLLD	0.210526	0.684211	0.105263	38

DDDDW	0.153846	0.615385	0.230769	13
DLDWL	0.222222	0.592593	0.185185	27
LLDLW	0.08	0.58	0.34	50
LWDLD	0.3	0.575	0.125	40
DWDWL	0.285714	0.571429	0.142857	28
LLLDW	0.270833	0.5625	0.166667	48
LDDWD	0.318182	0.545455	0.136364	22
WLDLW	0.216216	0.540541	0.243243	37
LDDDD	0.384615	0.538462	0.076923	13
DDWDD	0.235294	0.529412	0.235294	17
LDLWD	0.382353	0.529412	0.088235	34
LLWLD	0.288136	0.525424	0.186441	59
DDLWL	0.25	0.522727	0.227273	44
LDDLD	0.26087	0.521739	0.217391	23
WLLLL	0.213333	0.52	0.266667	75
DWLLD	0.259259	0.518519	0.222222	27
LDDDW	0.172414	0.517241	0.310345	29
WWLDL	0.241379	0.517241	0.241379	29
WLLDL	0.33871	0.516129	0.145161	62
LLDLD	0.268293	0.512195	0.219512	41
DLDLW	0.27907	0.511628	0.209302	43
LDLLW	0.327273	0.509091	0.163636	55
LWLLW	0.315789	0.508772	0.175439	57
DWLDL	0.305556	0.5	0.194444	36
DWDLL	0.3	0.5	0.2	30
LDLWW	0.3	0.5	0.2	20
LDWDW	0.230769	0.5	0.269231	26
LLLLL	0.257143	0.5	0.242857	140
LWWDD	0.3	0.5	0.2	30

Sorted by highest probability of drawing (Top 30)

Combinations	P(W)	P(L)	P(D)	TotalMatches
DLLDD	0.178571	0.321429	0.5	28
LWDDW	0.2	0.3	0.5	30
DDLLD	0.210526	0.315789	0.473684	19
DDDWL	0.275862	0.275862	0.448276	29
LWWLD	0.294118	0.264706	0.441176	34
WDLDW	0.322581	0.258065	0.419355	31
DWLWD	0.318182	0.272727	0.409091	22
DLDWD	0.318182	0.272727	0.409091	22
WLLLW	0.34	0.26	0.4	50
DWWLD	0.4	0.2	0.4	20
DWWWD	0.422222	0.2	0.377778	45
LDWDL	0.3125	0.3125	0.375	32
DLWDL	0.25	0.375	0.375	40
DDDDD	0.375	0.25	0.375	8
LDDWL	0.3125	0.3125	0.375	32
LLDWW	0.342857	0.285714	0.371429	35
LWDDD	0.473684	0.157895	0.368421	19
WLLDD	0.4	0.233333	0.366667	30
LWDWL	0.212121	0.424242	0.363636	33
LLDDL	0.363636	0.272727	0.363636	33
WLLWD	0.305556	0.333333	0.361111	36
DWDLW	0.309524	0.333333	0.357143	42
DWLWL	0.419355	0.225806	0.354839	31
LLDWL	0.333333	0.313725	0.352941	51
LLLWL	0.222222	0.428571	0.349206	63
DWDWD	0.307692	0.346154	0.346154	26
WDDWL	0.275862	0.37931	0.344828	29
DDLLW	0.342105	0.315789	0.342105	38
LWWWD	0.394737	0.263158	0.342105	38
WWWLW	0.47561	0.182927	0.341463	82

For combination of past 3 games the probabilities were to be found as follows:

Sorted by highest probability of winning

Combination	P(W)	P(L)	P(D)	Total Matches
WWW	0.513907	0.251656	0.234437	755
DWW	0.491979	0.270053	0.237968	374
WDW	0.48	0.2625	0.2575	400
WWD	0.440415	0.30829	0.251295	386
DDD	0.426136	0.346591	0.227273	176
DDL	0.423077	0.342308	0.234615	260
WWL	0.408163	0.356009	0.235828	441
WDL	0.407008	0.342318	0.250674	371
LWW	0.400881	0.337004	0.262115	454
WLW	0.372294	0.363636	0.264069	462
DWD	0.369811	0.392453	0.237736	265
LWD	0.364384	0.380822	0.254795	365
WLD	0.354467	0.383285	0.262248	347
LDW	0.353846	0.396923	0.249231	325
WDD	0.346304	0.420233	0.233463	257
DLL	0.342298	0.427873	0.229829	409
WLL	0.340909	0.380165	0.278926	484
LLW	0.336714	0.391481	0.271805	493
LWL	0.333959	0.390244	0.275797	533
DLD	0.329457	0.453488	0.217054	258
LDL	0.328671	0.435897	0.235431	429
LDD	0.327935	0.376518	0.295547	247
DLW	0.306329	0.435443	0.258228	395
LLD	0.304569	0.441624	0.253807	394
DWL	0.299694	0.397554	0.302752	327
DDW	0.289796	0.404082	0.306122	245
LLL	0.283951	0.459877	0.256173	648

Sorted by highest probability of losing

Combination	P(W)	P(L)	P(D)	Total Matches
LLL	0.283951	0.459877	0.256173	648
DLD	0.329457	0.453488	0.217054	258
LLD	0.304569	0.441624	0.253807	394
LDL	0.328671	0.435897	0.235431	429
DLW	0.306329	0.435443	0.258228	395
DLL	0.342298	0.427873	0.229829	409
WDD	0.346304	0.420233	0.233463	257
DDW	0.289796	0.404082	0.306122	245
DWL	0.299694	0.397554	0.302752	327
LDW	0.353846	0.396923	0.249231	325
DWD	0.369811	0.392453	0.237736	265
LLW	0.336714	0.391481	0.271805	493
LWL	0.333959	0.390244	0.275797	533
WLD	0.354467	0.383285	0.262248	347
LWD	0.364384	0.380822	0.254795	365
WLL	0.340909	0.380165	0.278926	484
LDD	0.327935	0.376518	0.295547	247
WLW	0.372294	0.363636	0.264069	462
WWL	0.408163	0.356009	0.235828	441

DDD	0.426136	0.346591	0.227273	176
WDL	0.407008	0.342318	0.250674	371
DDL	0.423077	0.342308	0.234615	260
LWW	0.400881	0.337004	0.262115	454
WWD	0.440415	0.30829	0.251295	386
DWW	0.491979	0.270053	0.237968	374
WDW	0.48	0.2625	0.2575	400
WWW	0.513907	0.251656	0.234437	755

Sorted by highest probability of drawing

Combination	P(W)	P(L)	P(D)	Total Matches
DDW	0.289796	0.404082	0.306122	245
DWL	0.299694	0.397554	0.302752	327
LDD	0.327935	0.376518	0.295547	247
WLL	0.340909	0.380165	0.278926	484
LWL	0.333959	0.390244	0.275797	533
LLW	0.336714	0.391481	0.271805	493
WLW	0.372294	0.363636	0.264069	462
WLD	0.354467	0.383285	0.262248	347
LWW	0.400881	0.337004	0.262115	454
DLW	0.306329	0.435443	0.258228	395
WDW	0.48	0.2625	0.2575	400
LLL	0.283951	0.459877	0.256173	648
LWD	0.364384	0.380822	0.254795	365
LLD	0.304569	0.441624	0.253807	394
WWD	0.440415	0.30829	0.251295	386
WDL	0.407008	0.342318	0.250674	371
LDW	0.353846	0.396923	0.249231	325
DWW	0.491979	0.270053	0.237968	374
DWD	0.369811	0.392453	0.237736	265
WWL	0.408163	0.356009	0.235828	441
LDL	0.328671	0.435897	0.235431	429
DDL	0.423077	0.342308	0.234615	260
WWW	0.513907	0.251656	0.234437	755
WDD	0.346304	0.420233	0.233463	257
DLL	0.342298	0.427873	0.229829	409
DDD	0.426136	0.346591	0.227273	176
DLD	0.329457	0.453488	0.217054	258

Conclusions:

- Past form for past 5 games as well as 3 games can be a good metric to predict the outcome of the next game.
- Judging purely on the basis of past ‘n’ games can produce incorrect results. This may be because a team might have had a run of easy fixtures running to a tough game.
- Past form alone can’t successfully predict games. It should be combined with some other metric which can evaluate the strength of the opposition to get the best results.

3) How does the difference in points between the two teams predict the outcome of the match?

Steps taken:

1. From (2) we had obtained all the results of all teams in a given season. These results were converted to numeric terms.
2. A Win was substituted by 3, Draw by 1 and Loss by 0.
3. All the points were then summed up.
4. For finding the difference in points between any two teams, we went to the corresponding matchweek and subtracted their respective accumulated points.
5. The points differences were obtained by the following formula:

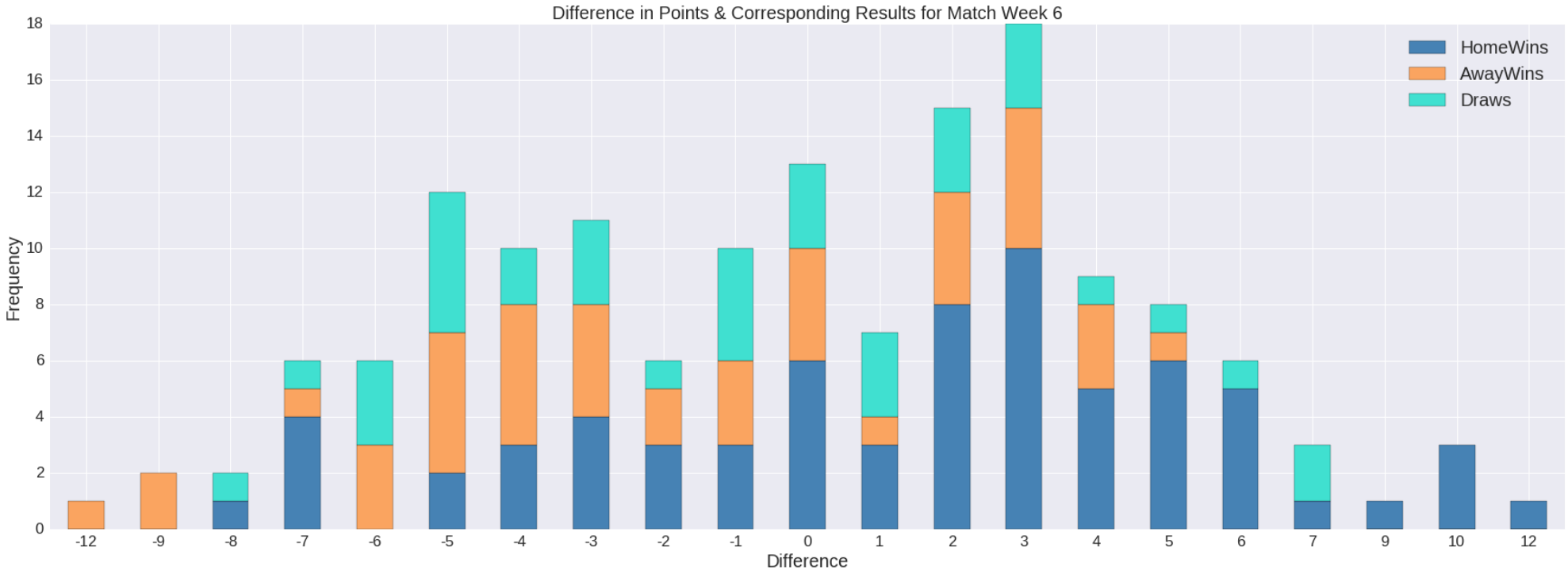
$$\textit{Point difference} = \textit{Home Team Points} - \textit{Away Teams Points}$$

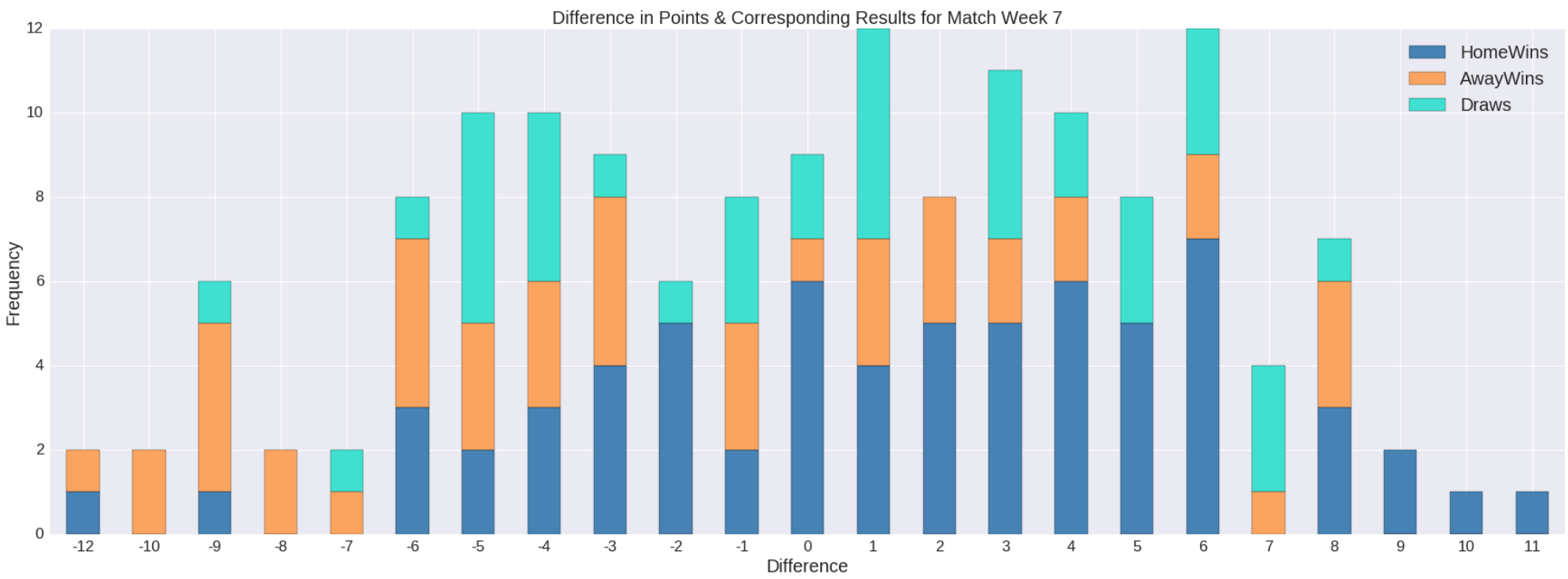
6. This whole process was repeated for all the datasets of the past 15 years and the results were then arranged by matchweek.

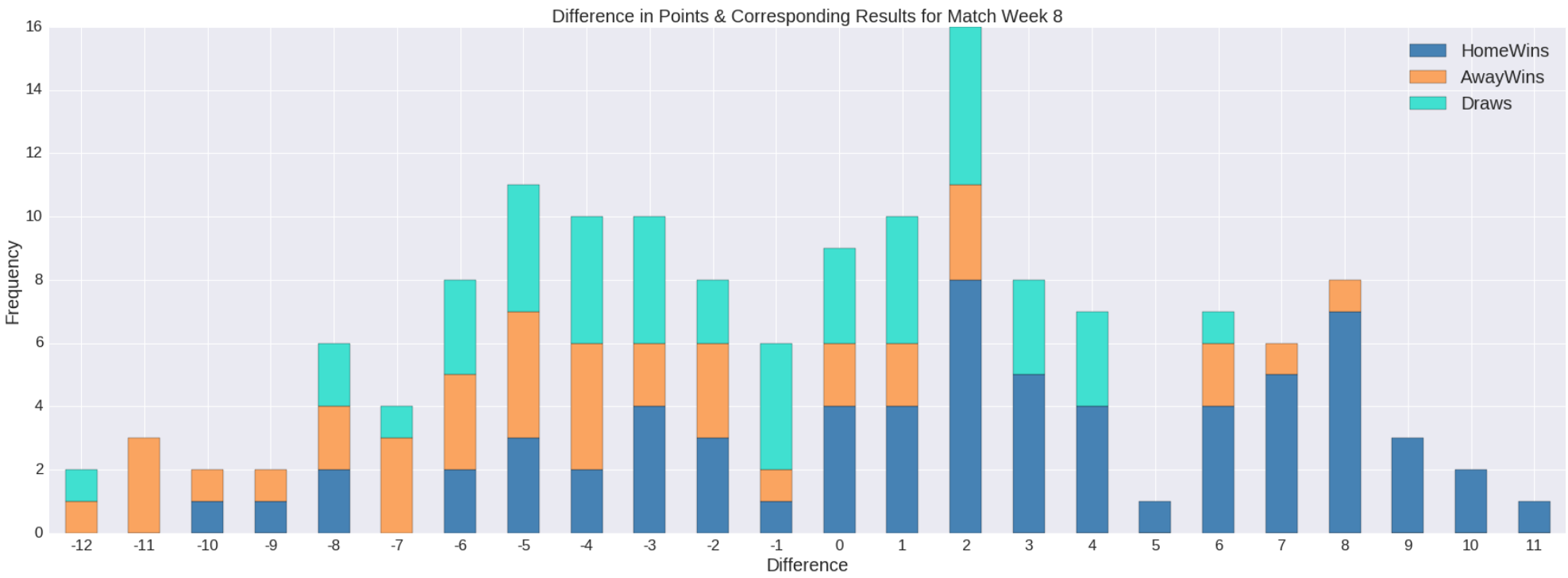
Results Obtained:

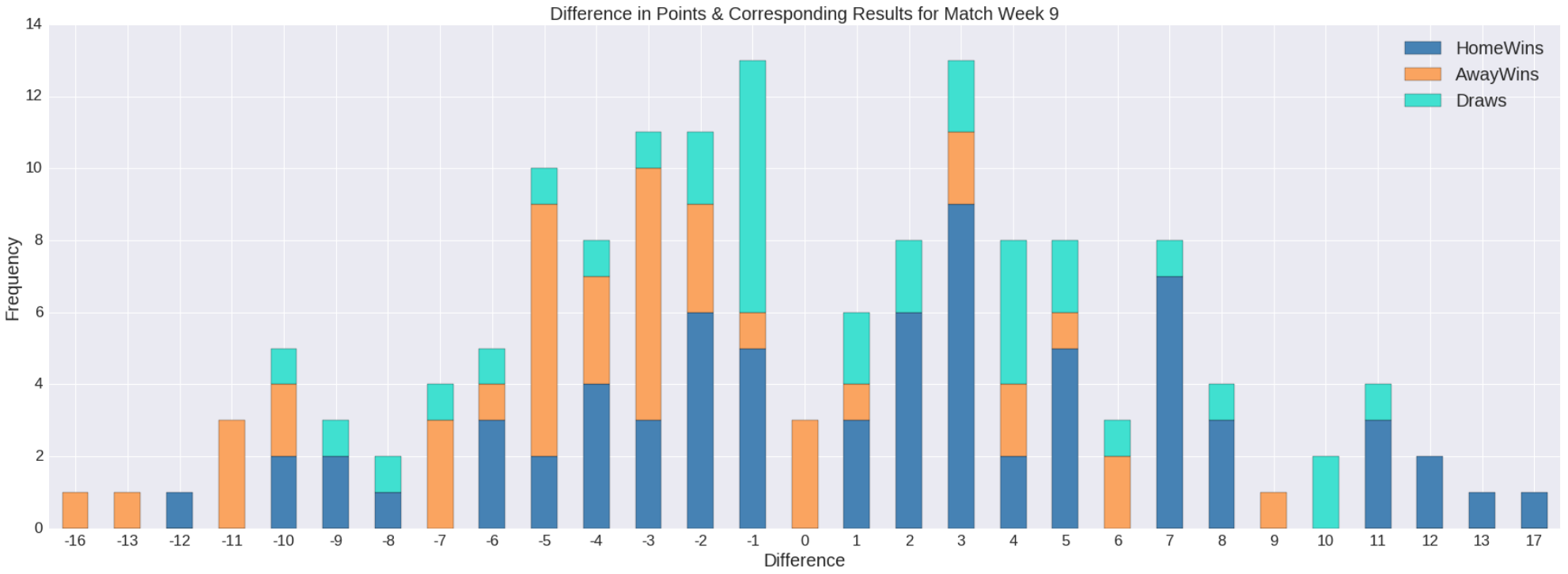
On x axis, the values to the right of zero represent those instances where the points for the Home team exceeded those of Away team's. The values to the left of zero represent instances where the points of Away team exceeded Home team's. A point difference of zero means that the Home team as well as the away team had equal points.

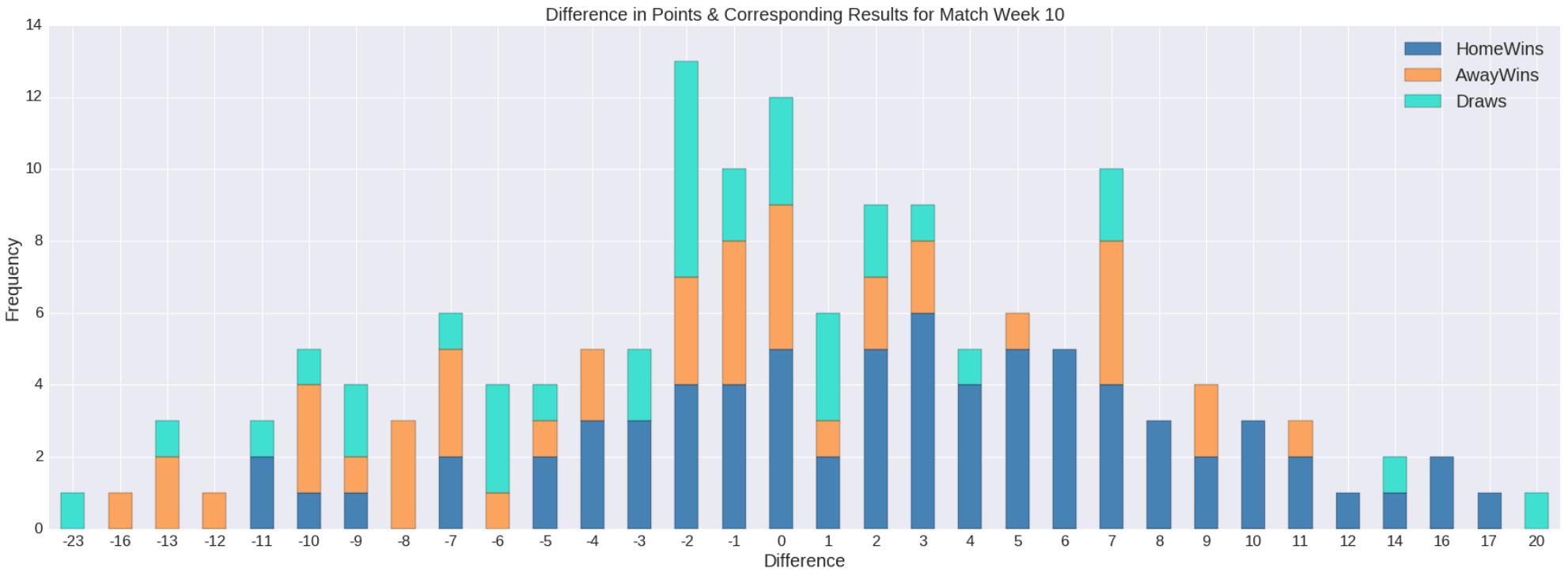
The graphs have been plotted only after Matchweek 5. This has been done because in the starting matchweeks, the gap between the top team and the bottom teams isn't much and so, it can give us a wrong picture.

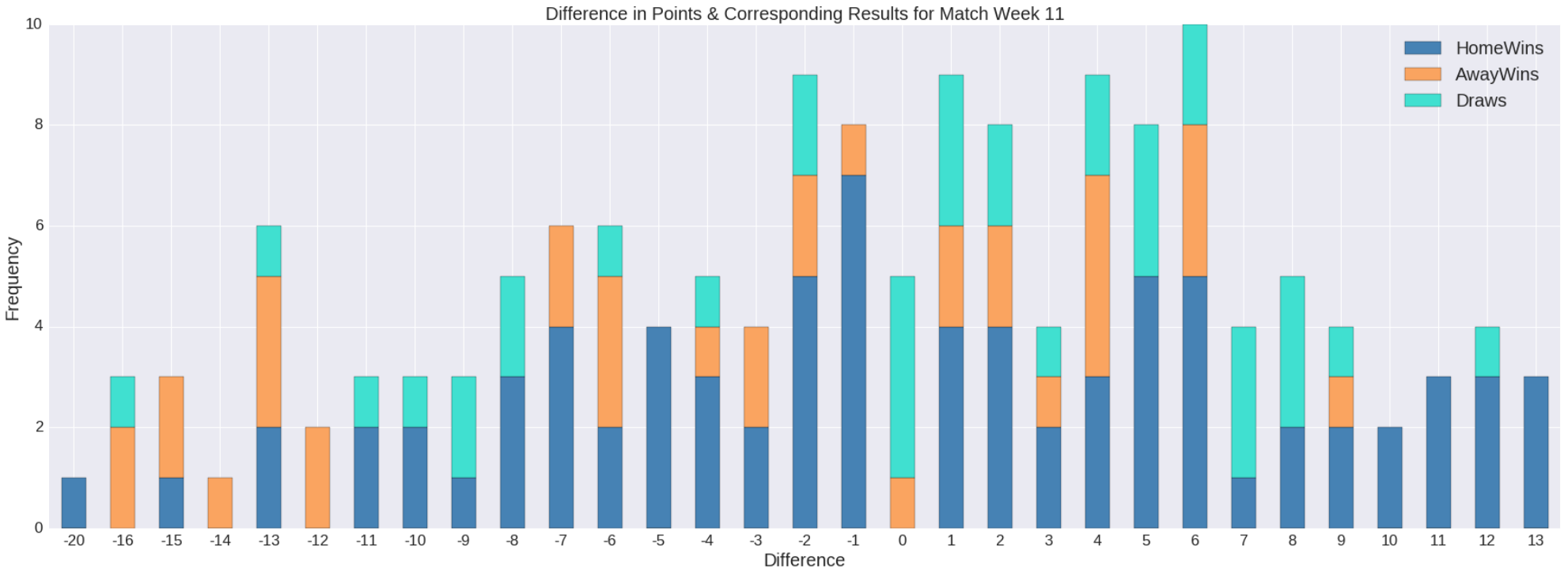


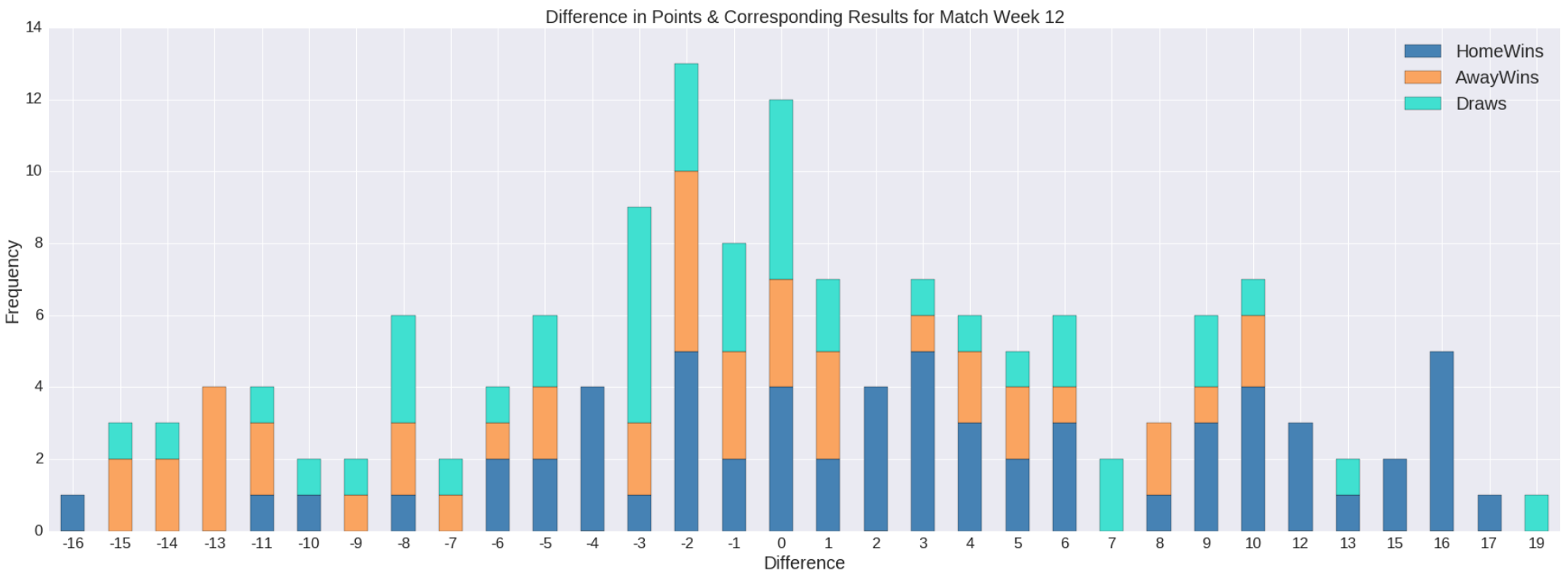


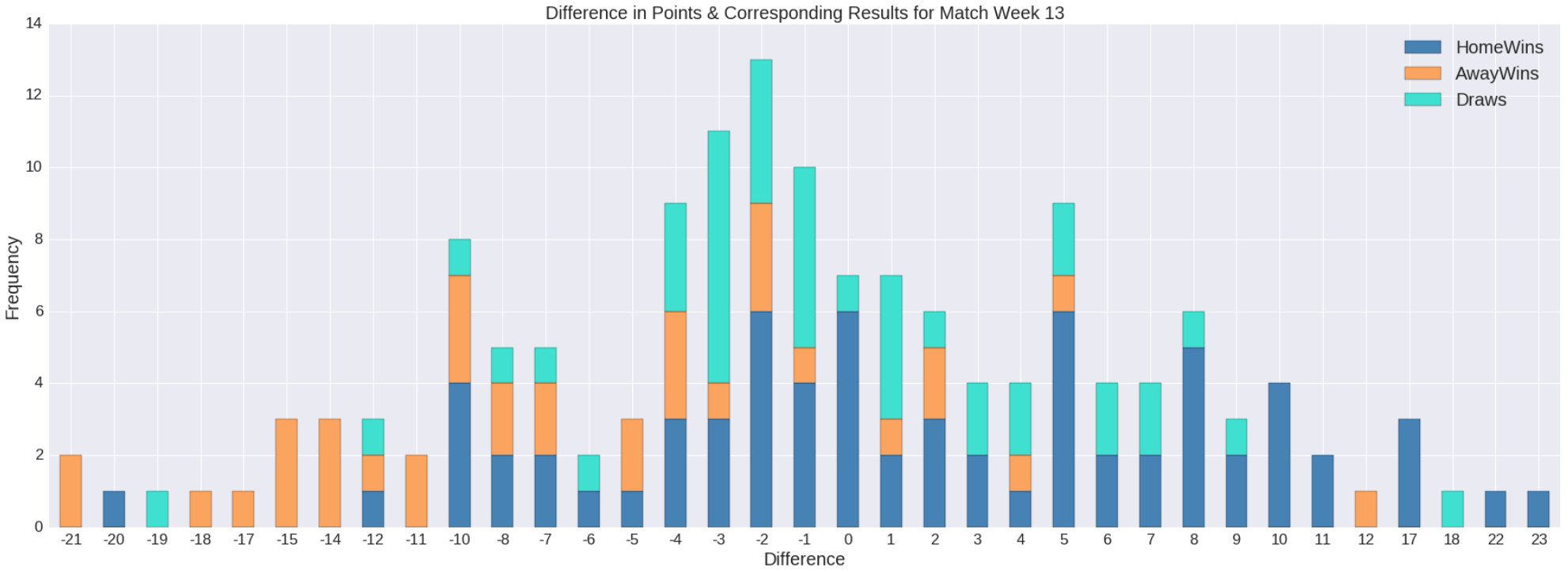


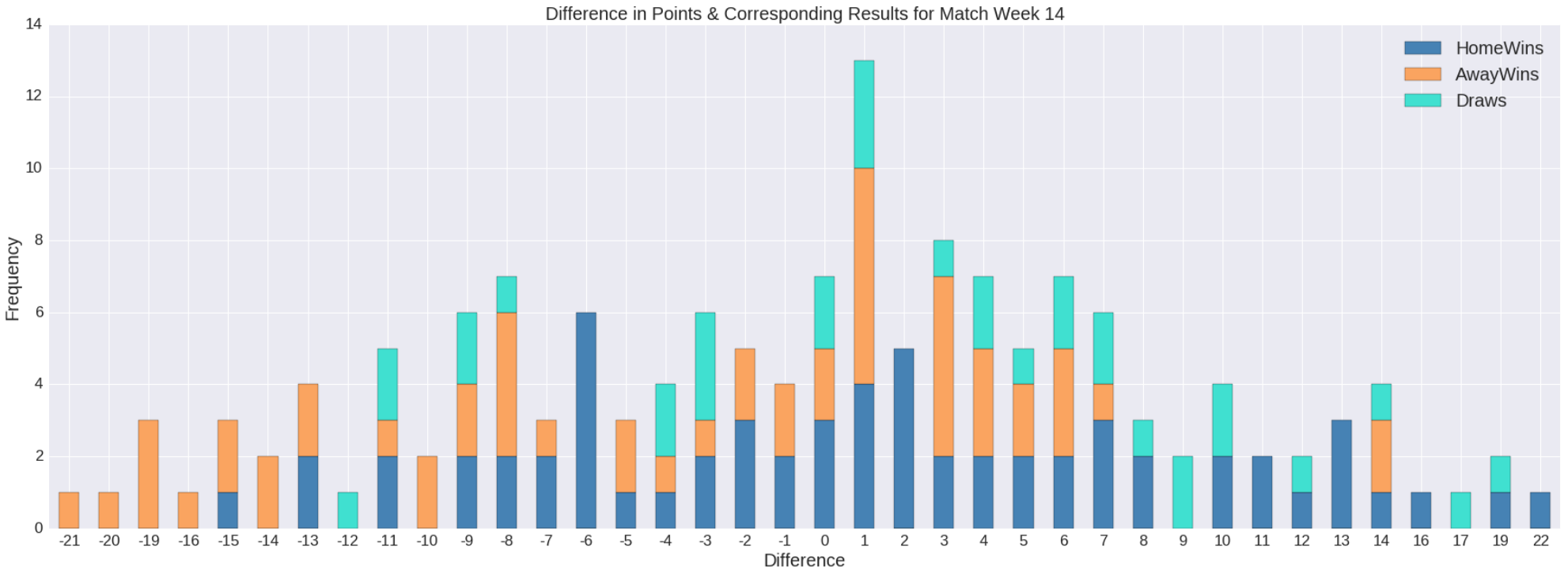

Difference in Points & Corresponding Results for Match Week 9
HomeWins
AwayWins
Draws
Frequency
Difference
0
2
4
6
8
10
12
14
-16 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 17

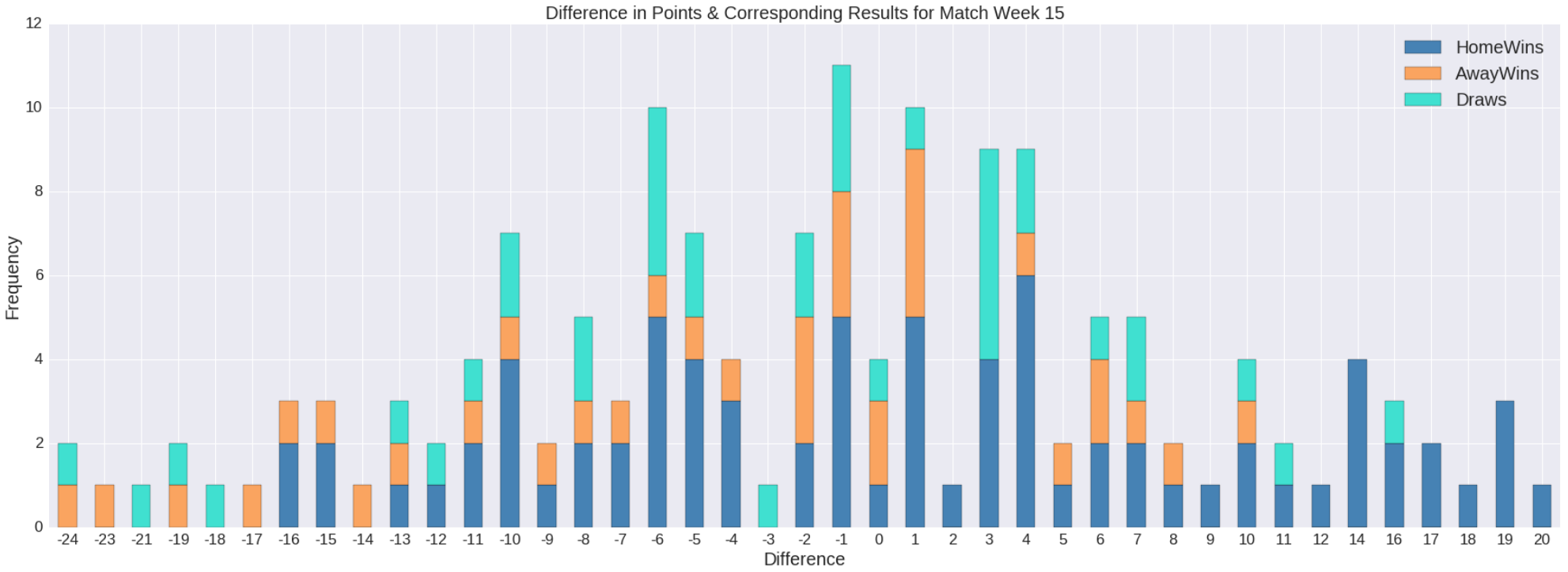

Difference in Points & Corresponding Results for Match Week 10
HomeWins
AwayWins
Draws
Frequency
Difference
0
2
4
6
8
10
12
14
-23 -16 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 14 16 17 20

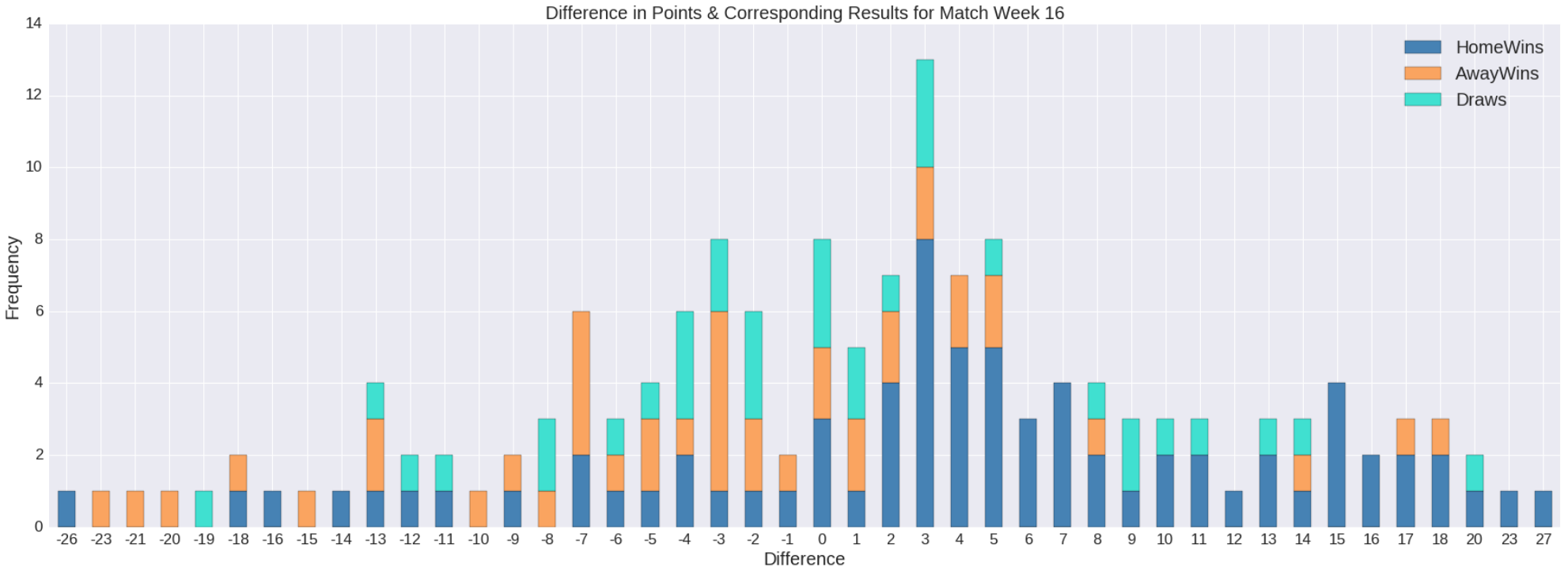

Difference in Points & Corresponding Results for Match Week 11
HomeWins
AwayWins
Draws
Frequency
Difference
0
2
4
6
8
10
-20 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13

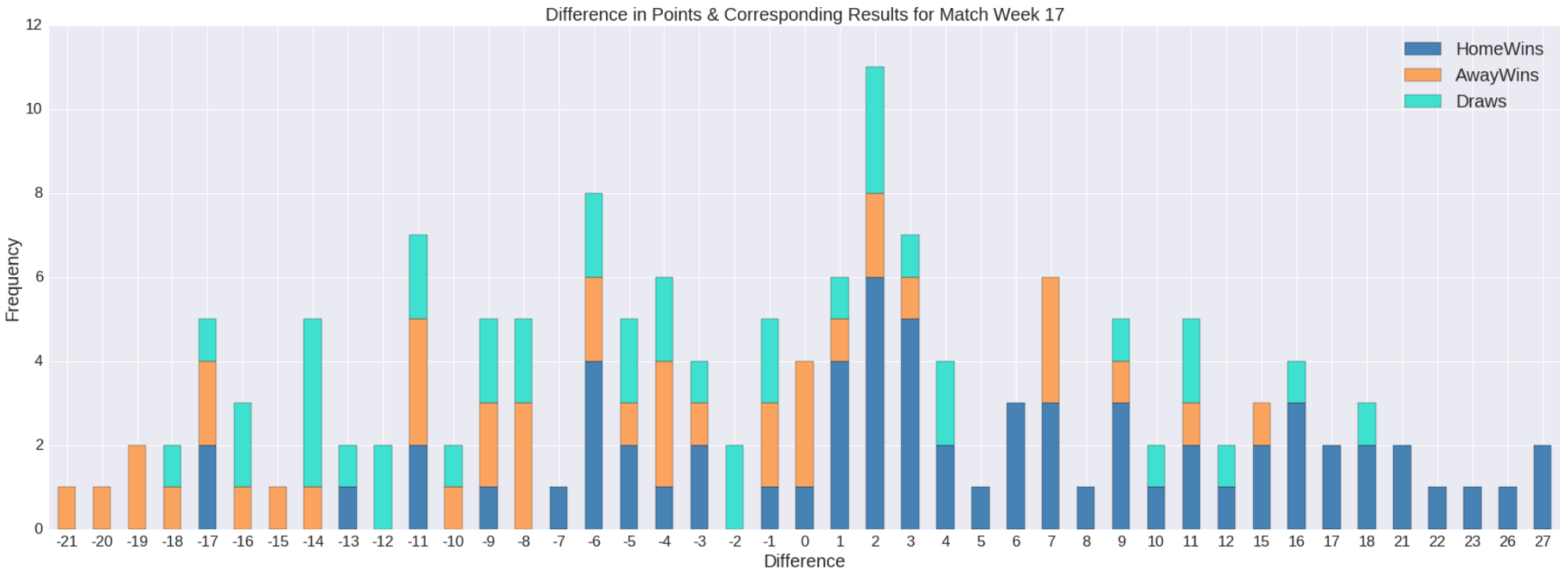

Difference in Points & Corresponding Results for Match Week 12
HomeWins
AwayWins
Draws
Frequency
Difference
0
2
4
6
8
10
12
14
-16 -15 -14 -13 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 12 13 15 16 17 19

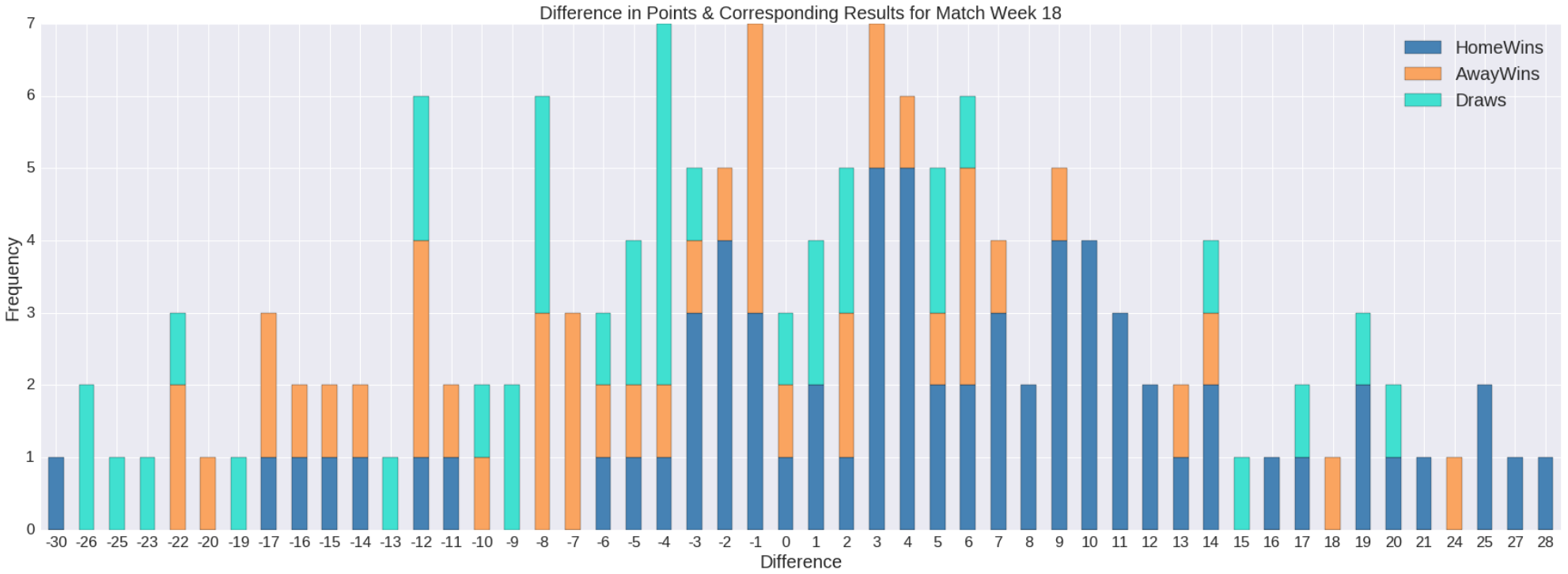

Difference in Points & Corresponding Results for Match Week 13
HomeWins
AwayWins
Draws
Frequency
Difference
-21 -20 -19 -18 -17 -15 -14 -12 -11 -10 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 17 18 22 23

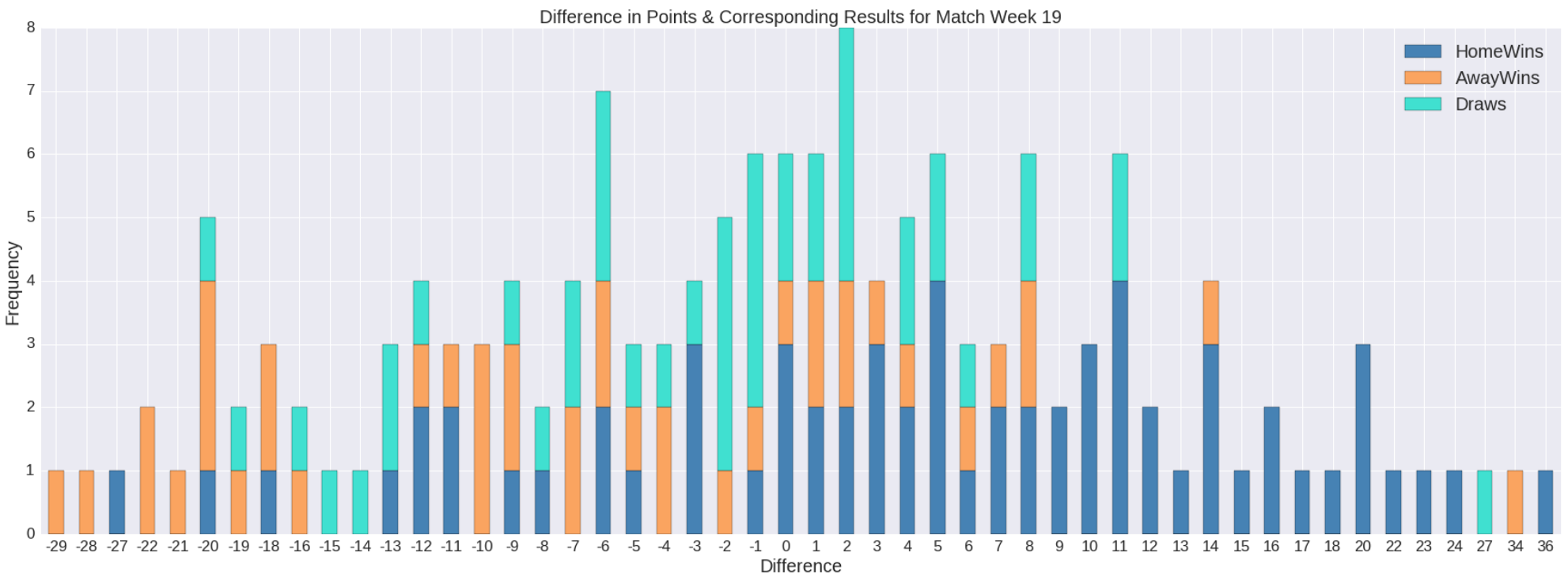

Difference in Points & Corresponding Results for Match Week 14
HomeWins
AwayWins
Draws
Frequency
Difference
-21 -20 -19 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17 19 22

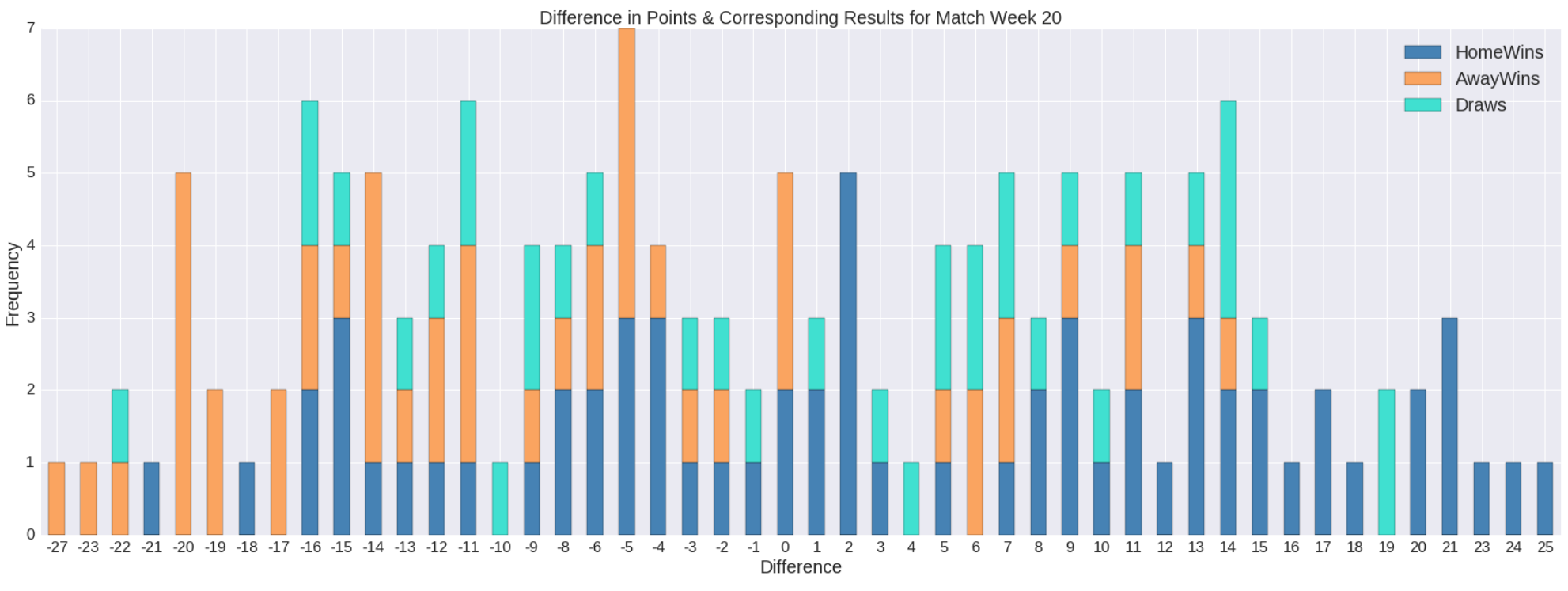

Difference in Points & Corresponding Results for Match Week 15
HomeWins
AwayWins
Draws
Frequency
Difference
-24 -23 -21 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 14 16 17 18 19 20

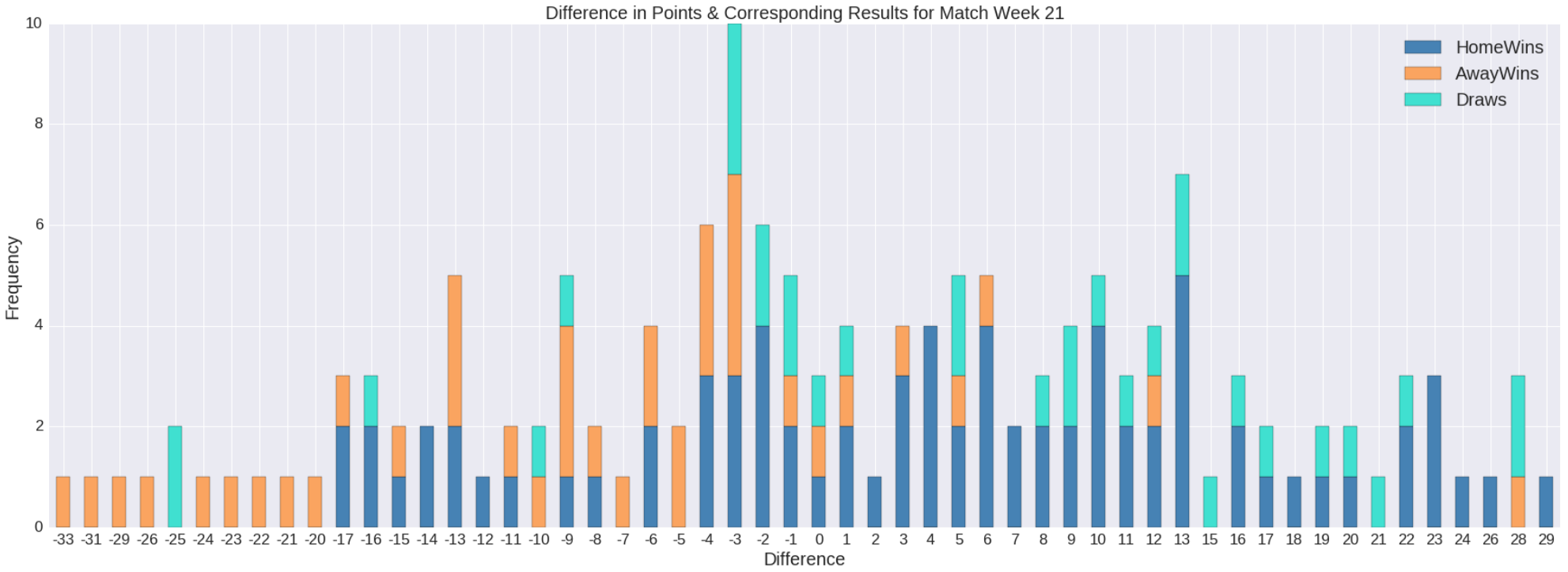

Difference in Points & Corresponding Results for Match Week 16
HomeWins
AwayWins
Draws
Frequency
Difference
-26 -23 -21 -20 -19 -18 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 23 27

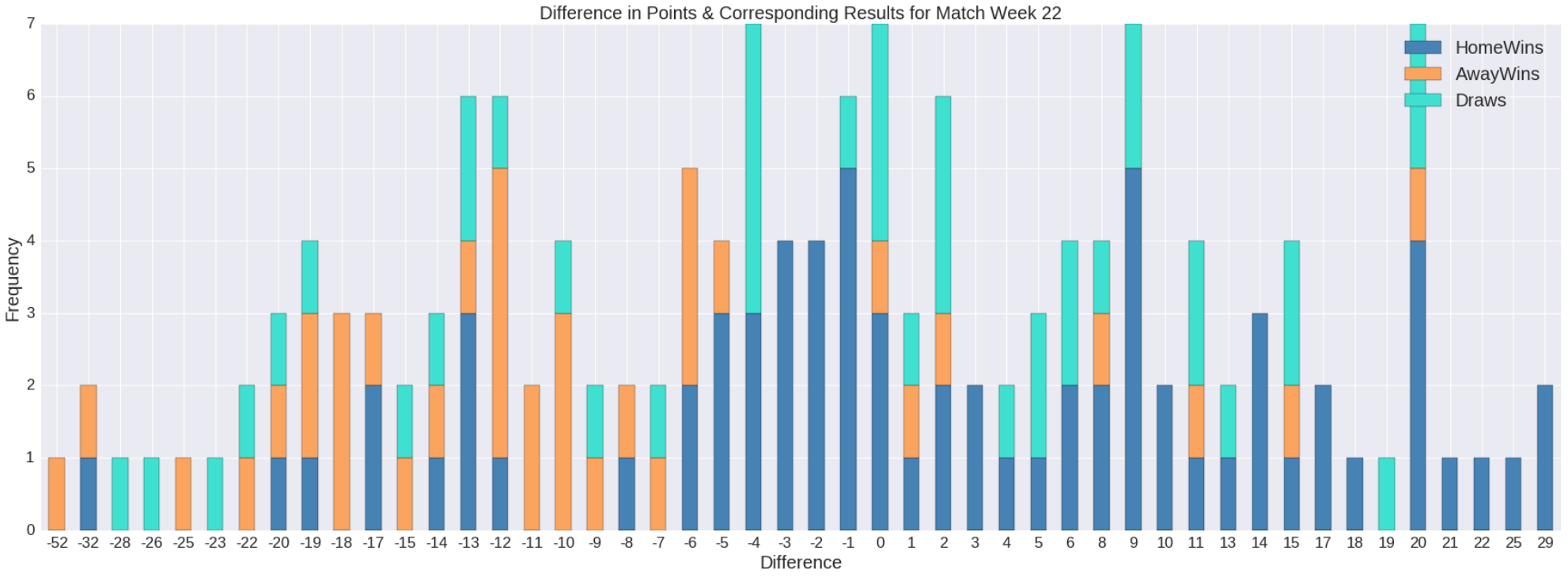

Difference in Points & Corresponding Results for Match Week 17
HomeWins
AwayWins
Draws
Frequency
Difference

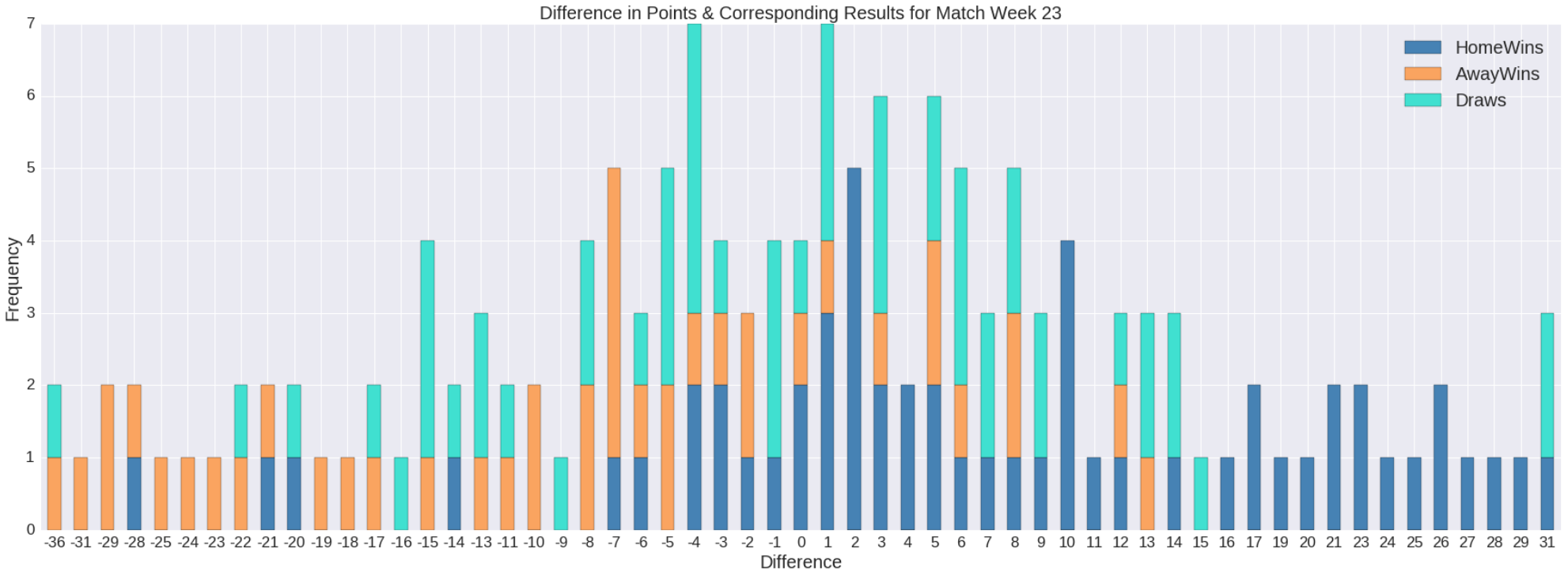

Difference in Points & Corresponding Results for Match Week 18
HomeWins
AwayWins
Draws
Frequency
Difference

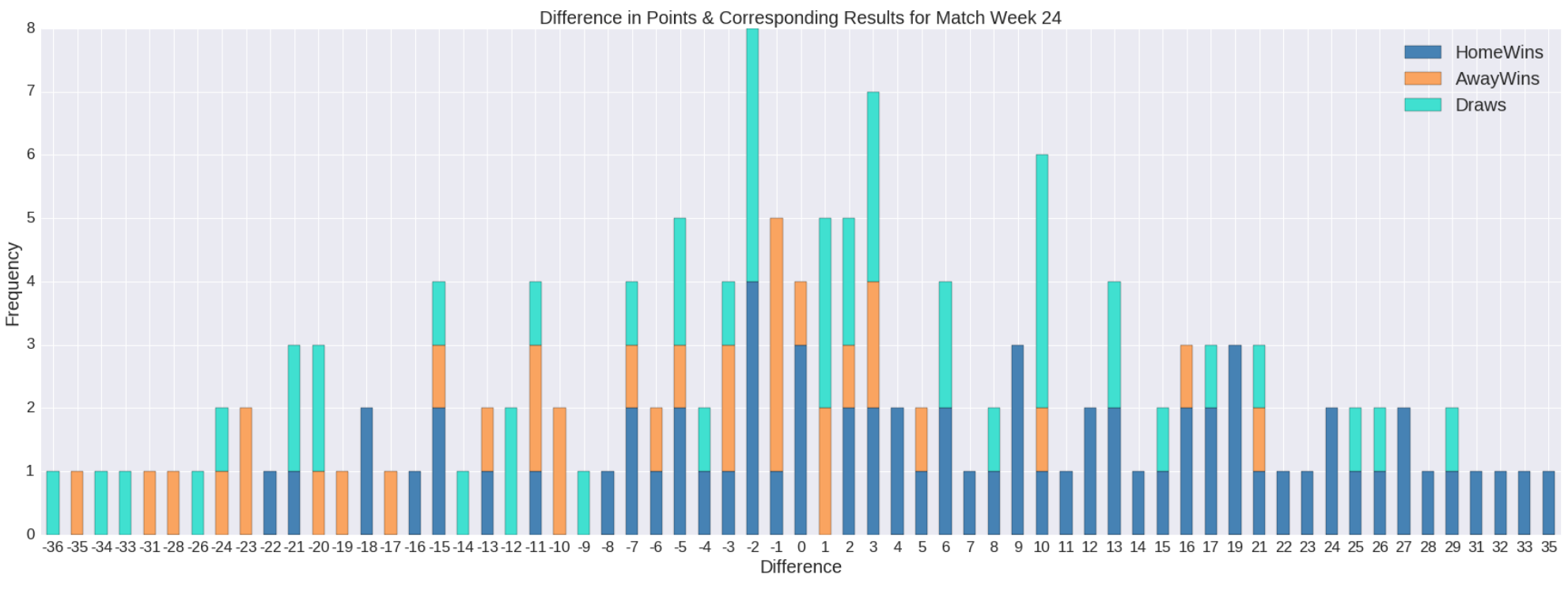

Difference in Points & Corresponding Results for Match Week 19
HomeWins
AwayWins
Draws
Frequency
Difference

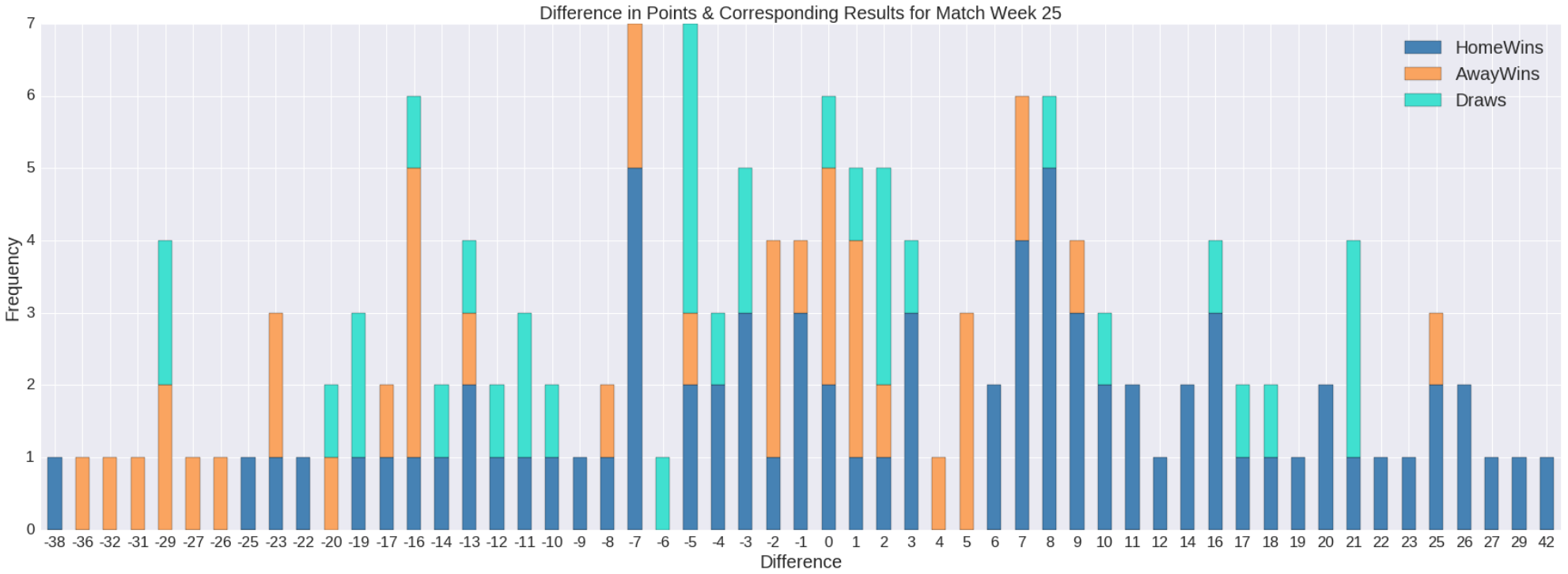

Difference in Points & Corresponding Results for Match Week 20
HomeWins
AwayWins
Draws
Frequency
Difference

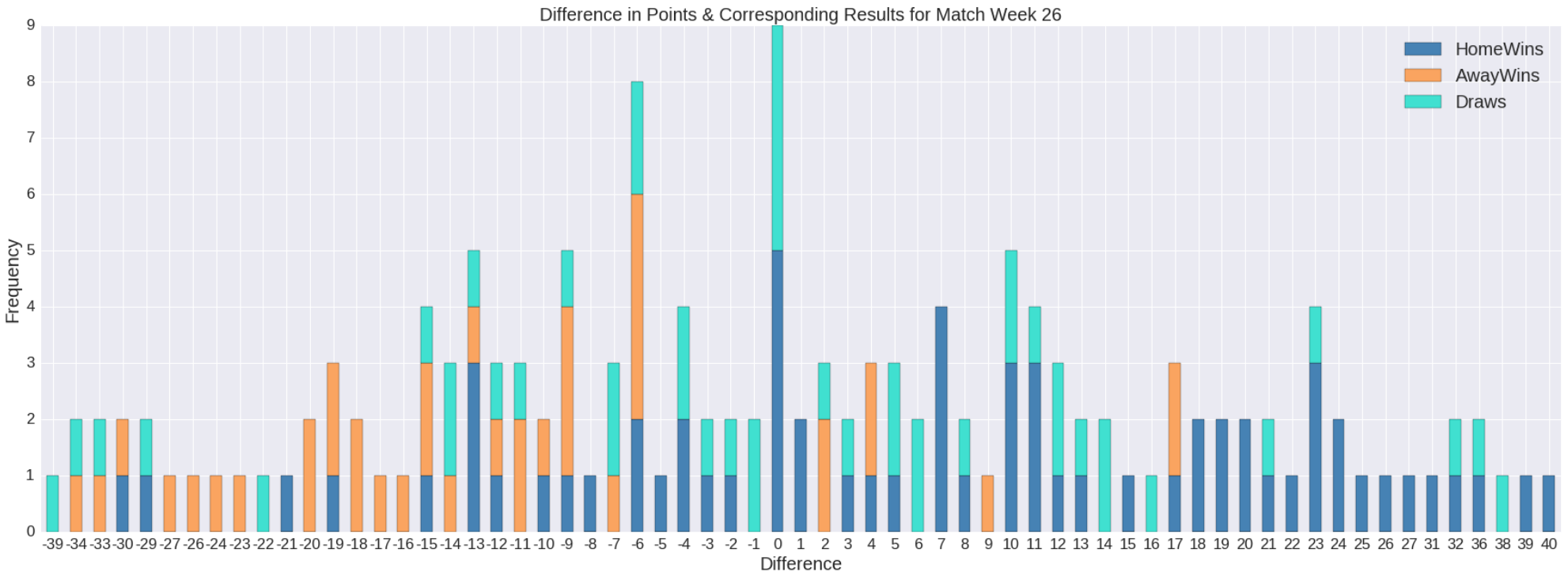

Difference in Points & Corresponding Results for Match Week 21
HomeWins
AwayWins
Draws
Frequency
Difference

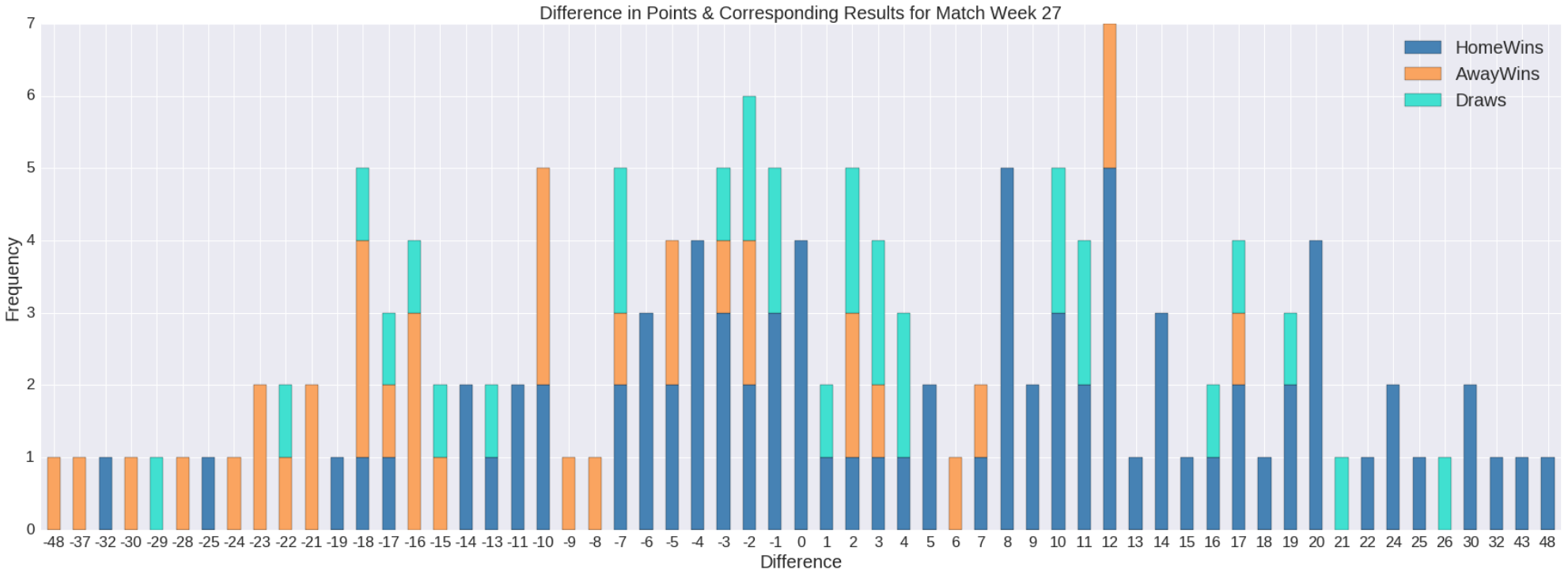

Difference in Points & Corresponding Results for Match Week 22
HomeWins
AwayWins
Draws
Frequency
Difference

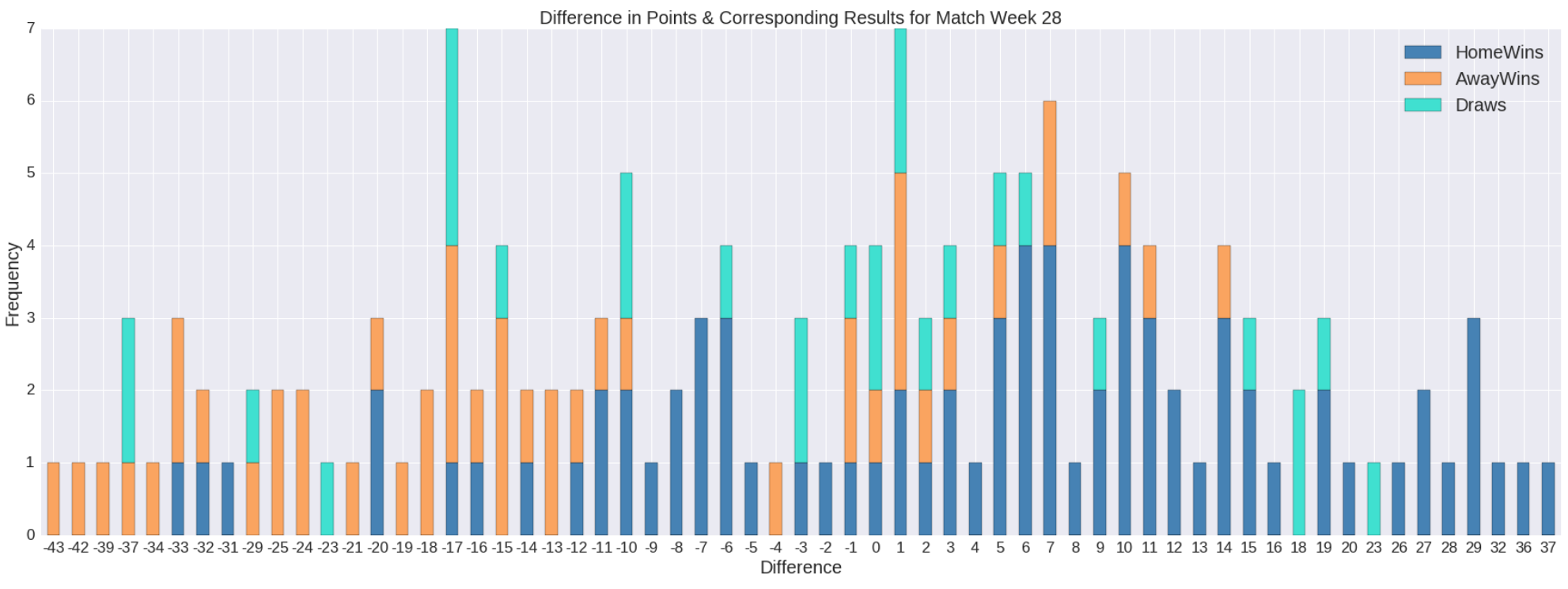

Difference in Points & Corresponding Results for Match Week 23
HomeWins
AwayWins
Draws
Frequency
Difference

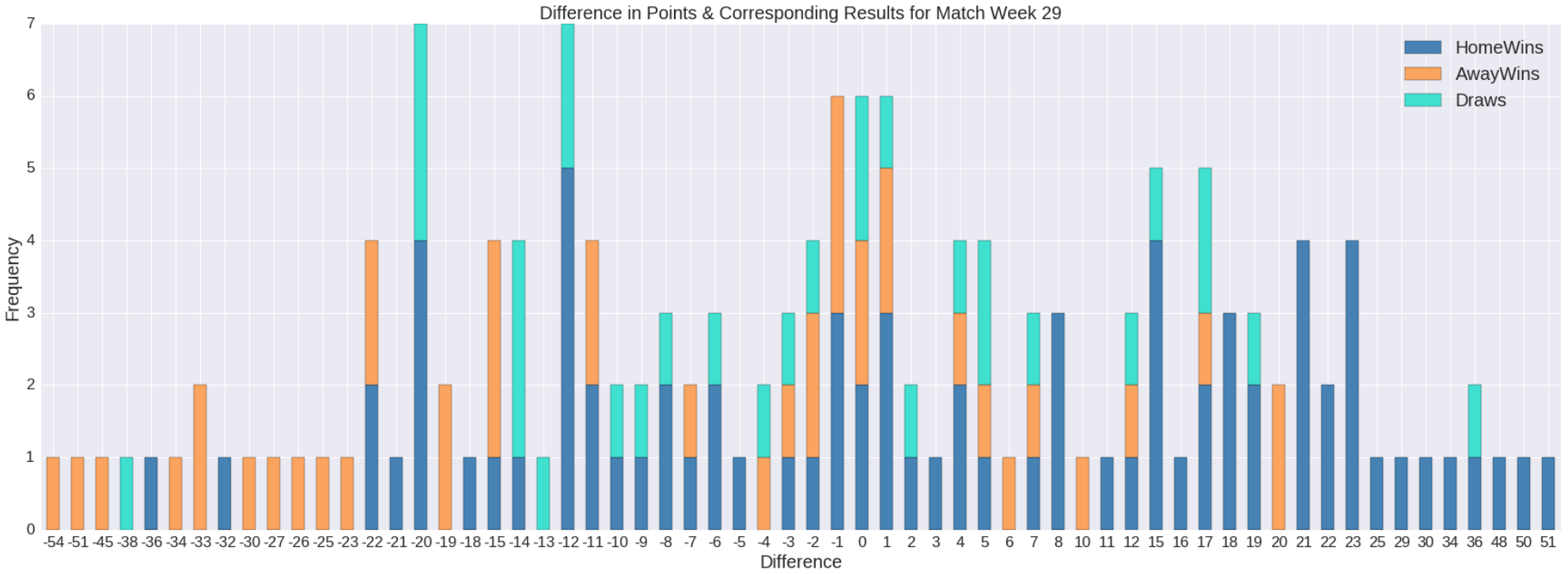

Difference in Points & Corresponding Results for Match Week 24
HomeWins
AwayWins
Draws
Frequency
Difference

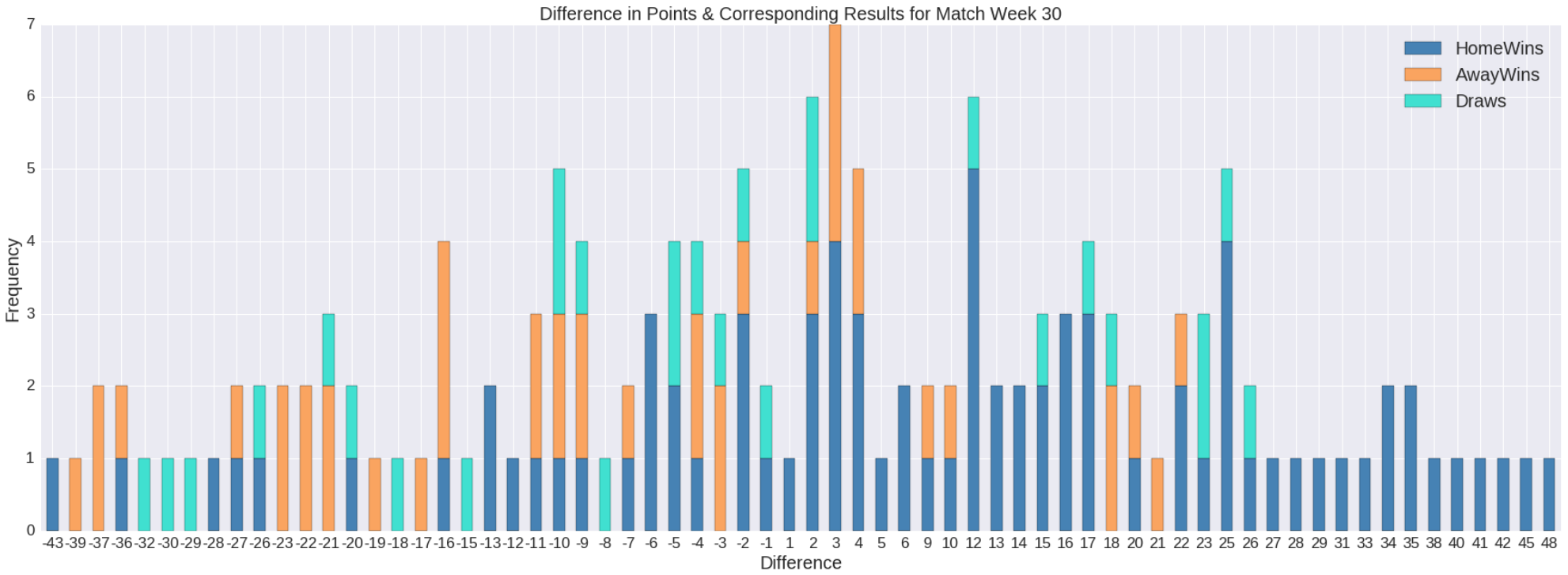

Difference in Points & Corresponding Results for Match Week 25
HomeWins
AwayWins
Draws
Frequency
Difference
-38 -36 -32 -31 -29 -27 -26 -25 -23 -22 -20 -19 -17 -16 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 14 16 17 18 19 20 21 22 23 25 26 27 29 42

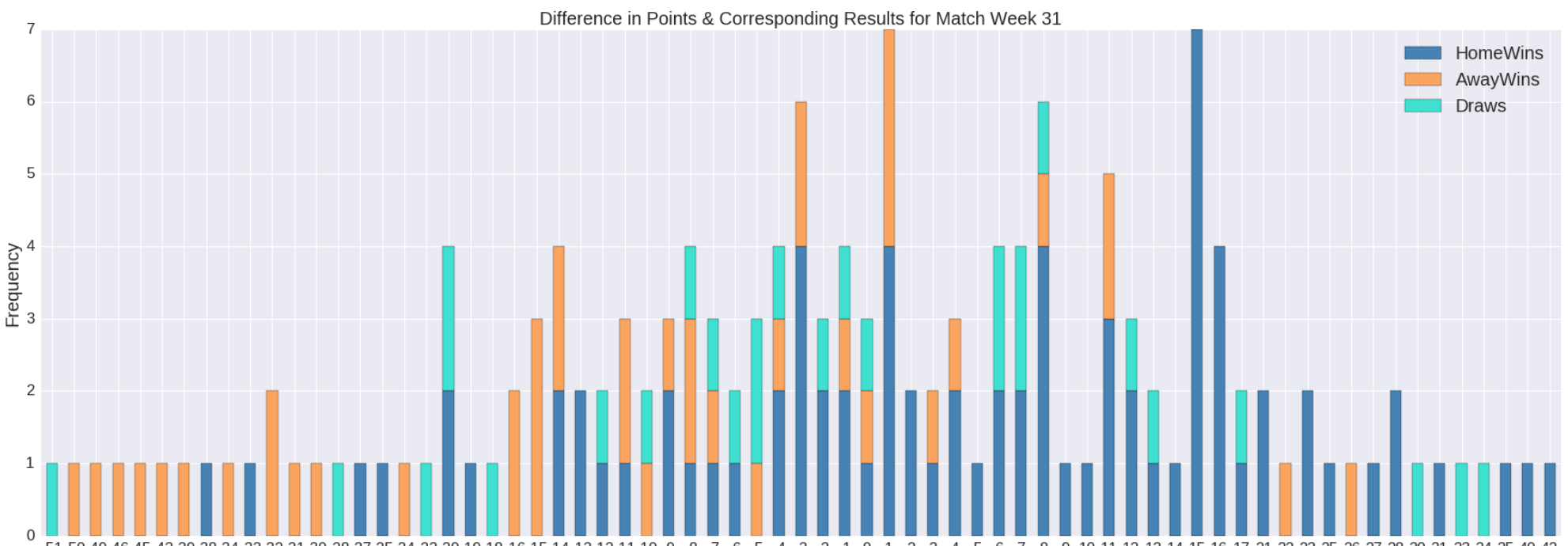

Difference in Points & Corresponding Results for Match Week 26
HomeWins
AwayWins
Draws
Frequency
Difference
-39 -34 -33 -30 -29 -27 -26 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 31 32 36 38 39 40


Difference in Points & Corresponding Results for Match Week 27
HomeWins
AwayWins
Draws
Frequency
Difference
-48 -37 -32 -30 -29 -28 -25 -24 -23 -22 -21 -19 -18 -17 -16 -15 -14 -13 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26 30 32 43 48

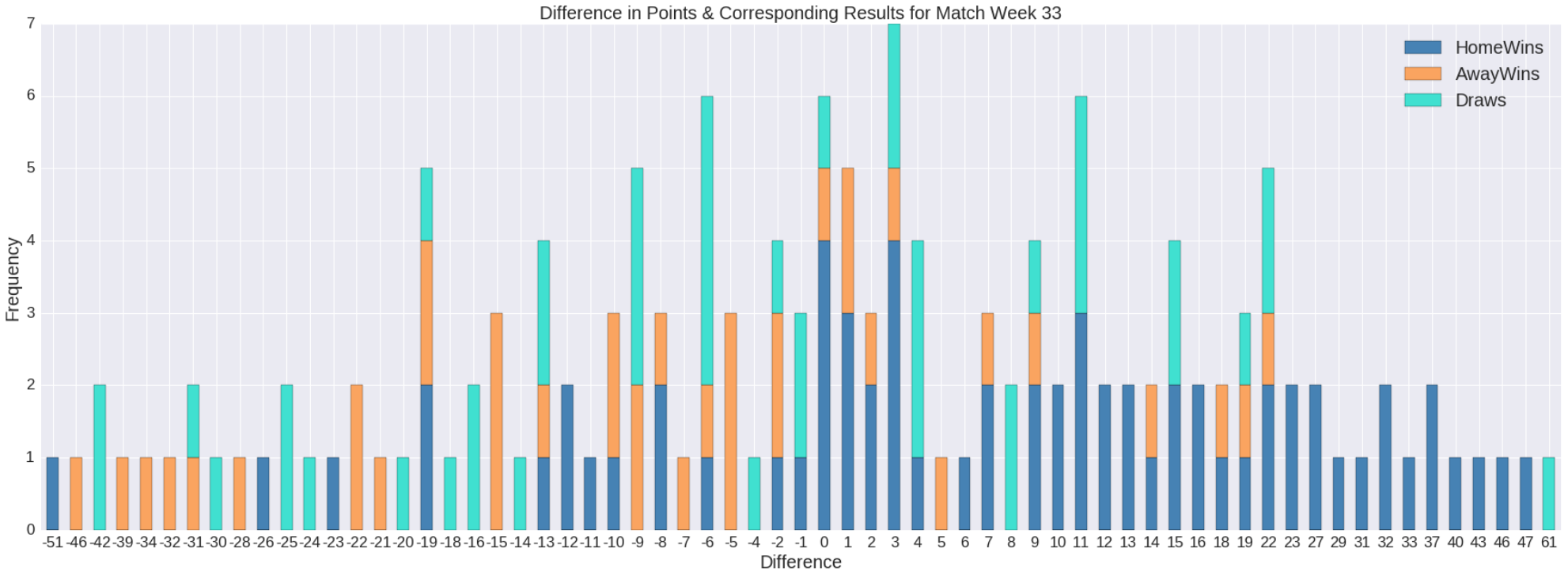

Difference in Points & Corresponding Results for Match Week 28
HomeWins
AwayWins
Draws
Frequency
Difference
-43 -42 -39 -37 -34 -33 -32 -31 -29 -25 -24 -23 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 19 20 23 26 27 28 29 32 36 37

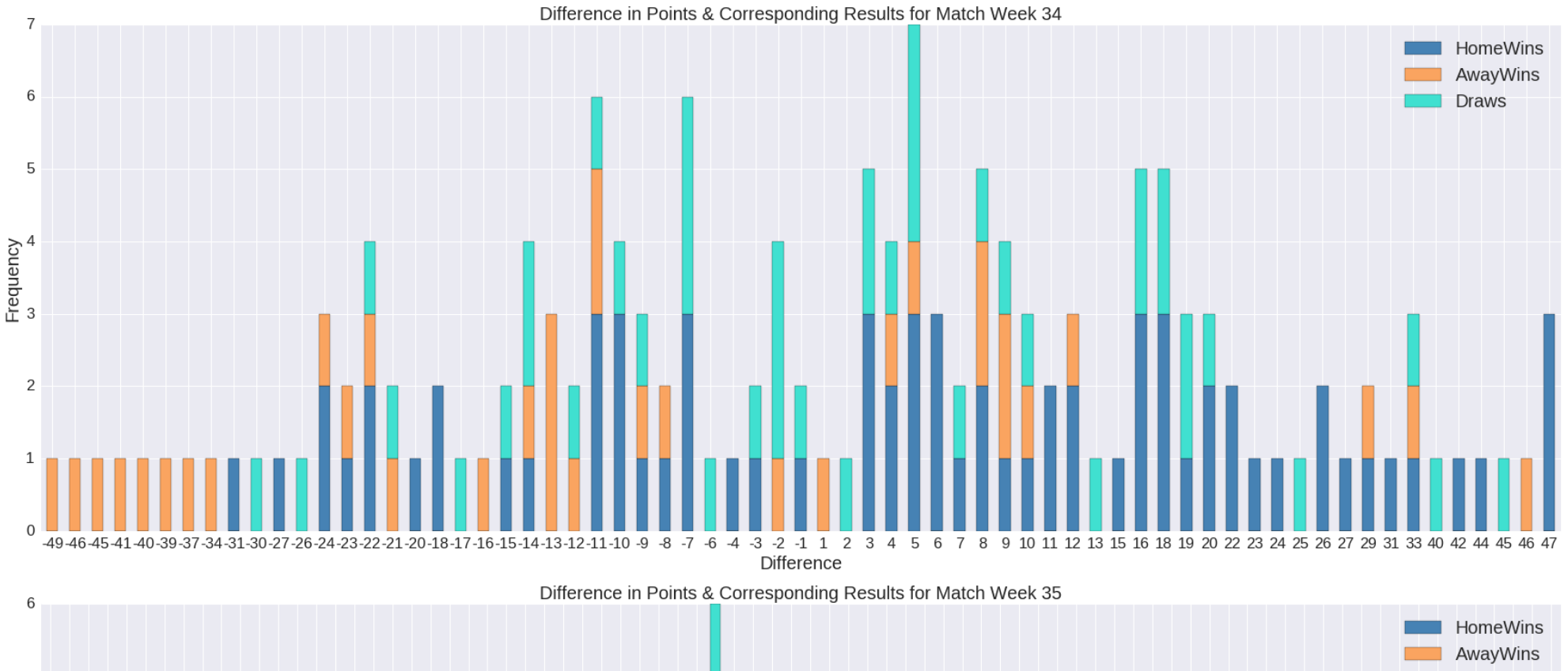

Difference in Points & Corresponding Results for Match Week 29
HomeWins
AwayWins
Draws
Frequency
Difference
-54 -51 -45 -38 -36 -34 -33 -32 -30 -27 -26 -25 -23 -22 -21 -20 -19 -18 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 10 11 12 15 16 17 18 19 20 21 22 23 25 29 30 34 36 48 50 51

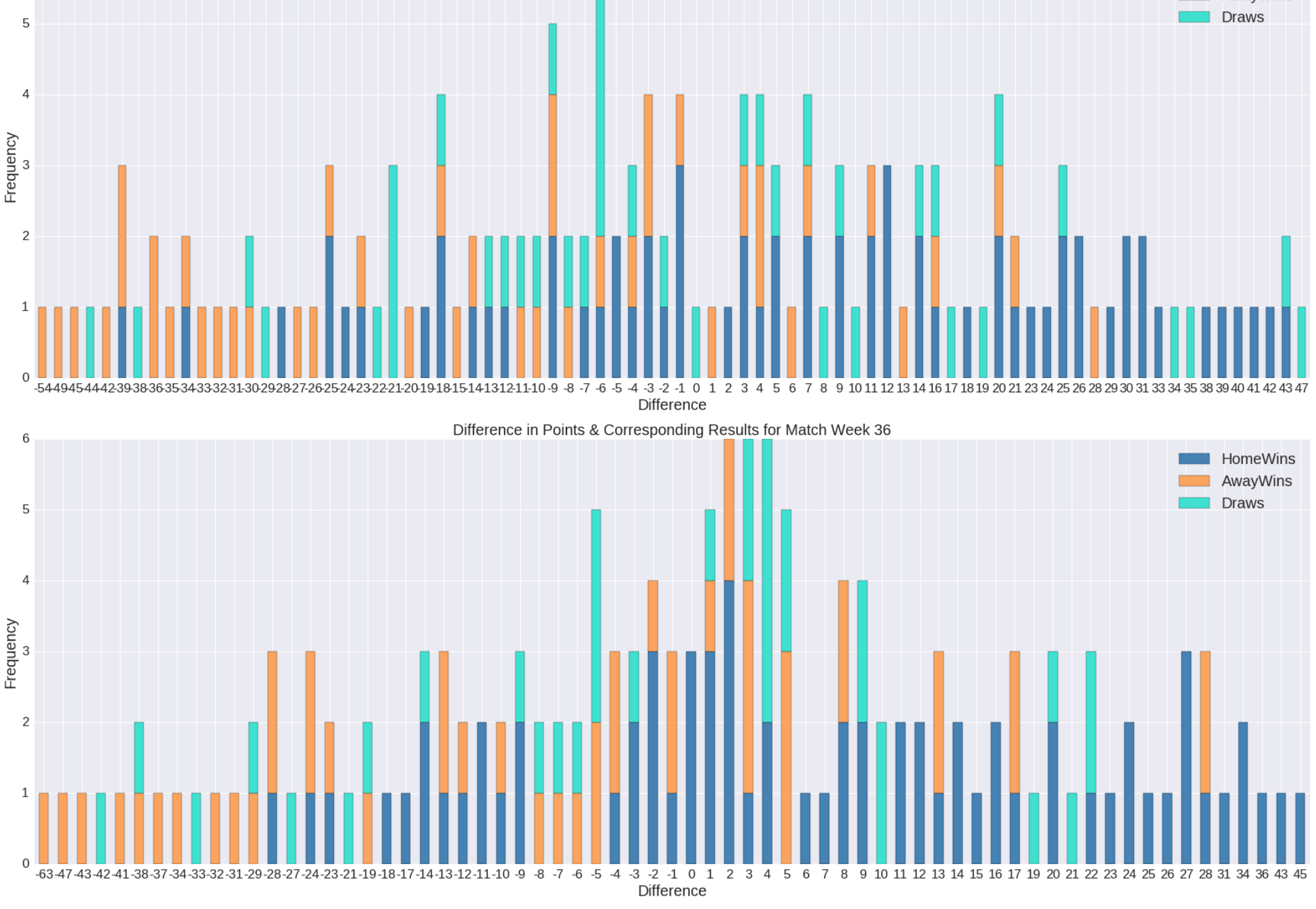

Difference in Points & Corresponding Results for Match Week 30
HomeWins
AwayWins
Draws
Frequency
Difference
-43 -39 -37 -36 -32 -30 -29 -28 -27 -26 -23 -22 -21 -20 -19 -18 -17 -16 -15 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 9 10 12 13 14 15 16 17 18 20 21 22 23 25 26 27 28 29 31 33 34 35 38 40 41 42 45 48


Difference in Points & Corresponding Results for Match Week 31
HomeWins
AwayWins
Draws
Frequency
Difference
-51 -50 -49 -46 -45 -43 -39 -38 -34 -33 -32 -31 -30 -28 -27 -25 -24 -23 -20 -19 -18 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 21 22 23 25 26 27 28 30 31 33 34 35 40 42

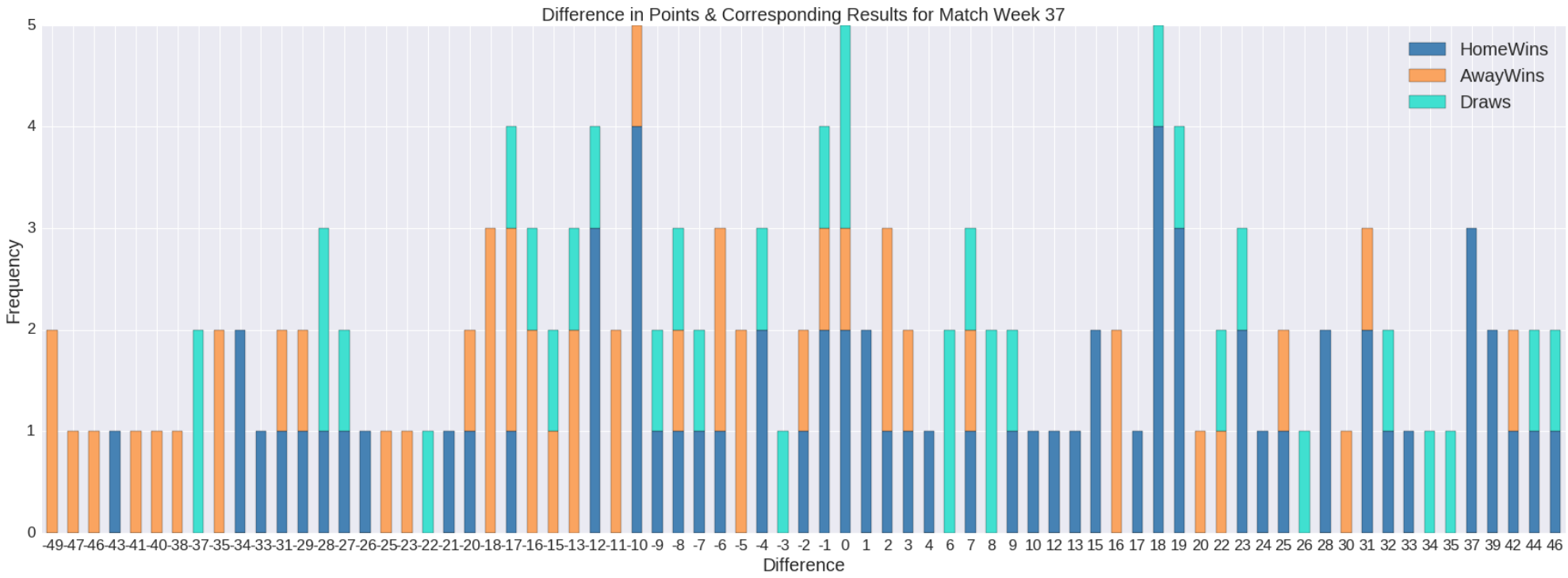

Difference in Points & Corresponding Results for Match Week 32
HomeWins
AwayWins
Draws
Frequency
Difference
-48 -47 -41 -39 -37 -36 -35 -34 -32 -30 -29 -27 -26 -25 -23 -22 -21 -19 -18 -17 -16 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 19 21 22 24 25 26 27 28 29 31 32 33 34 36 44 48

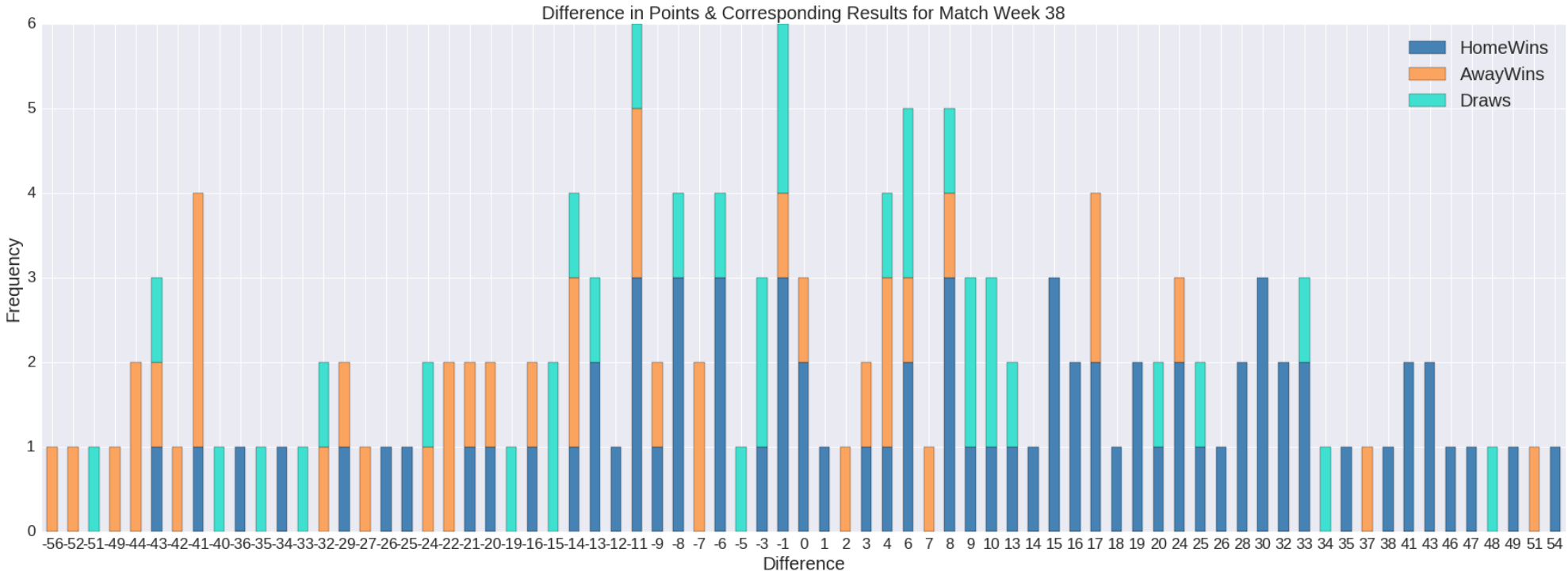

Difference in Points & Corresponding Results for Match Week 33
HomeWins
AwayWins
Draws
Frequency
Difference


Difference in Points & Corresponding Results for Match Week 34
HomeWins
AwayWins
Draws
Frequency
Difference


Difference in Points & Corresponding Results for Match Week 35
HomeWins
AwayWins
Draws
Frequency
Difference

Difference in Points & Corresponding Results for Match Week 36
HomeWins
AwayWins
Draws
Frequency
Difference





Conclusions:

- The graph becomes alternates between being bimodal and normal. This is because the points of most teams aren't much different. The extreme differences are given by top team vs. mid table team/ bottom team or mid table team facing the team at the bottom of the table. Away wins becomes more prominent in bars to the left of 0 and Home wins become more prominent to the right of 0 as we progress through the season.
- Comparatively home wins have a larger proportion to the right of zero than the proportion of away wins to the left of zero. This is because the probability of team is affected a lot by the location of the match, i.e., whether they're playing home or away. Refer to Question 1, for exact probabilities.

4) How does the past head to head record predict the outcome of a match?

Steps taken:

1. First of all, an empty dictionary was made which had team names for a particular season as keys and another dictionary as values. The subsequent dictionary had the names of opposition teams as keys and en empty list as respective values
2. This empty list was then filled by the results of the initial dictionary's team against the subsequent dictionary's team.
3. Wins/Losses/Draws for the initial dictionary's team were then filled.
4. The results were then aggregated for 'n' seasons. 5. Only those teams which were present in all 'n' season were taken into account. This was done for obtaining consistent results.

Results:

The following table illustrates the head to head record of past 4 seasons for a sample of teams along with the probabilities of winning/losing/drawing for the match of subsequent season against the same opposition.

P (W/H) – Probability of winning at home for the header team.
P (L/H) – Probability of losing at home for the header team.
P (D/H) – Probability of drawing at home for the header team.
P (W/A) – Probability of winning away for the header team.
P (L/A) – Probability of losing away for the header team.
P (D/A) – Probability of drawing away for the header team.
HR' Year – Home result for the header team for that particular year
AR' Year – Away result for the header team for that particular year

In the following tables the head to head records are for Seasons 2009-10 till 2013-14 and the results are for 2014-15.

Manchester City

	P(W/H)	P(L/H)	P(D/H)	HR (2014-15)	P(W/A)	P(L/A)	P(D/A)	AR (2014-15)
Arsenal	0.5	0.25	0.25	W	0.25	0.25	0.5	D
Aston Villa	1	0	0	W	0.5	0.25	0.25	L
Chelsea	1	0	0	L	0.25	0.5	0.25	L
Everton	0.25	0.5	0.25	W	0	1	0	W
Fulham	0.5	0	0.5	W	0.75	0	0.25	W
Liverpool	0.5	0	0.5	W	0	0.25	0.75	L
Man United	0.25	0.5	0.25	W	0.5	0.5	0	W
Stoke	1	0	0	W	0	0	1	D
Sunderland	0.75	0	0.25	D	0	0.75	0.25	L
Tottenham	0.75	0.25	0	W	0.25	0.5	0.25	W

Arsenal

	P(W/H)	P(L/H)	P(D/H)	HR (2014-15)	P(W/A)	P(L/A)	P(D/A)	AR (2014-15)
Aston Villa	0.75	0.25	0	L	0.5	0	0.5	W
Chelsea	0.25	0.5	0.25	D	0.25	0.75	0	L
Everton	0.5	0	0.5	D	0.75	0	0.25	L
Fulham	0.5	0	0.5	W	0.5	0.25	0.25	W
Liverpool	0.25	0.25	0.5	W	0.75	0	0.25	L
Man City	0.25	0.25	0.5	D	0.25	0.5	0.25	L
Man United	0.25	0.5	0.25	D	0	1	0	L
Stoke	1	0	0	W	0.25	0.25	0.5	L
Sunderland	0.5	0	0.5	W	0.5	0.25	0.25	W
Tottenham	0.75	0.25	0	W	0	0.75	0.25	W

Chelsea

	P(W/H)	P(L/H)	P(D/H)	HR (2014-15)	P(W/A)	P(L/A)	P(D/A)	AR (2014-15)
Arsenal	0.75	0.25	0	W	0.5	0.25	0.25	D
Aston Villa	0.5	0.25	0.25	W	0.5	0.25	0.25	L
Everton	0.5	0	0.5	W	0.25	0.75	0	L
Fulham	0.5	0	0.5	W	0.5	0	0.5	W
Liverpool	0.25	0.5	0.25	W	0.25	0.5	0.25	W
Man City	0.5	0.25	0.25	W	0	1	0	W
Man United	0.5	0.25	0.25	W	0.5	0.5	0	D
Stoke	1	0	0	W	0.5	0	0.5	L
Sunderland	0.75	0.25	0	L	1	0	0	W
Tottenham	0.5	0	0.5	W	0.25	0.25	0.5	D

Conclusions:

Based on the above we can see that the past head to head record for 5 seasons is not a very reliable metric. This can be because of the following reasons:

- Every EPL season is different. Some teams might perform better in a particular season and then be average for the next few years.
- Most teams are generally inconsistent throughout the season, i.e., they may perform better in the first quarter and then slack off in the next quarter. Because of this, there is no certain way to predict the outcome purely based on past head to head record.

Past head to head record can be really useful for predicting the performance of top teams against the mid table teams. Since the top teams tend to be consistent and the mid table teams inconsistent, past head to head record can be a good metric.

Similarly, past head to head record alone is unreliable but if combines with another metric like past form, it can give pretty accurate results.

Final Conclusion:

1) Home teams have a definite advantage over Away teams. On aggregate, home team win 46.65% matches compared to 27.72% matches won by the away teams.
2) Based on the past ‘n’ games, a team’s probability for winning/losing/drawing can be predicted. But it is best to combine it with some other metric, as this alone won’t suffice.

3) The difference in points can also be used for predicting results. Although, these predictions should be taken with a pinch of salt. EPL is famous for its upsets and any team no matter what its points tally is, is capable of defeating any other team.
4) The head to head record also gives a reasonable picture for predicting future results. Although, the predictions are much more valid for top teams than mid table teams. This is because top teams are much more consistent than mid table teams.

All the above metrics should not be considered alone. Instead while predicting a result, an appropriate combination of the above should be taken and based on that the result should be estimated. Only taking a single metric can give incorrect results.