REPORT FOR FINAL EXAM

SUBJECT: 11077 DECISION SUPPORT SYSTEMS AND BI

Done by: Manabayev Beket (ID: 24505)

Group: CSSE-191M

Title: A short research on under-/over-performance of popular football teams between seasons 2014/15 and 2018/19, and methods of identifying similar patterns and anomalies in play-styles

GLOSSARY

1. **XG** (also known as «expected goal» metric) is a statistical measure of quality of chances created and conceded;
2. **XG-difference** is the distinction between actual goals scored and expected goals from metric;
3. **NPXG** is a statistical measure, which indicates expected goals excluding penalties;
4. **XGA** is expected goals rate conceded by some squad within a tour;
5. **XGA-difference** is the distinction between actual missed goals and expected missed goals computed by metric;
6. **NPXGA** is a statistical measure, which shows expected goals allowed by a particular team;
7. **NPXGD** is the difference between expected goals from both observed team and its opponent, excluding penalties;
8. **PPDA** coefficient is a measure of passes done by opponent on the opposition half, before the observed team engages in pressure. To be precise, PPDA coefficient shows the willingness of a particular squad in terms of enemy pressure on the enemy's half. The less PPDA rate, the more aggressively does some team play;
9. **OPPDA** coefficient is a measure of passes done by the observed team in the opposition half, before the enemy engages in pressure. To be precise, OPPDA coefficient demonstrates how good is the particular squad at resisting enemy's attempts to engage in pressure and return the ball back. The higher OPPDA index, the more time does some team play without losing the ball under opponent's pressure;
10. **DC** is a statistical rate of passes completed within estimated 20 yards (18 meters) of goal, excluding crosses;
11. **ODC** is a statistical reflection of DC metric, but it concentrates on amount of passes done by opponent within estimated 20 yards of goal, excluding crosses;
12. **XPTS** is a statistical measure of expected points for each team;
13. **XPTS-difference** is the distinction between actual amount of earned points by the end of season and expected points predicted with the usage of metric.

DATASET

To begin with, I decided to choose the dataset created by Kaggle user Sergi Lehkyi named «Football Data: Expected Goals and Other Metrics», which consists of detailed information about Top European Leagues advanced statistics starting from season 2014/15 till 2018/19. In order to implement the dataset Sergi has scrapped available summary information from the web-site https://understat.com/[1] to precisely look at some numbers of all football squads of various professional leagues. Currently the dataset contains the following summary statistics by the end of each season for 6 UEFA Football leagues:

- La Liga (Spain);
- EPL (England & Wales);
- BundesLiga (Germany);
- Serie A (Italy);
- Ligue 1 (France);
- RFPL (Russia).

Among the default parameters in the dataset there are:

- Team's position;
- Squad title;
- Amount of matches completed;
- Amount of wins;
- Amount of draws;
- Amount of loses;
- Goals scored;
- Goals missed;
- Final points (end of season).

HYPOTHESIS

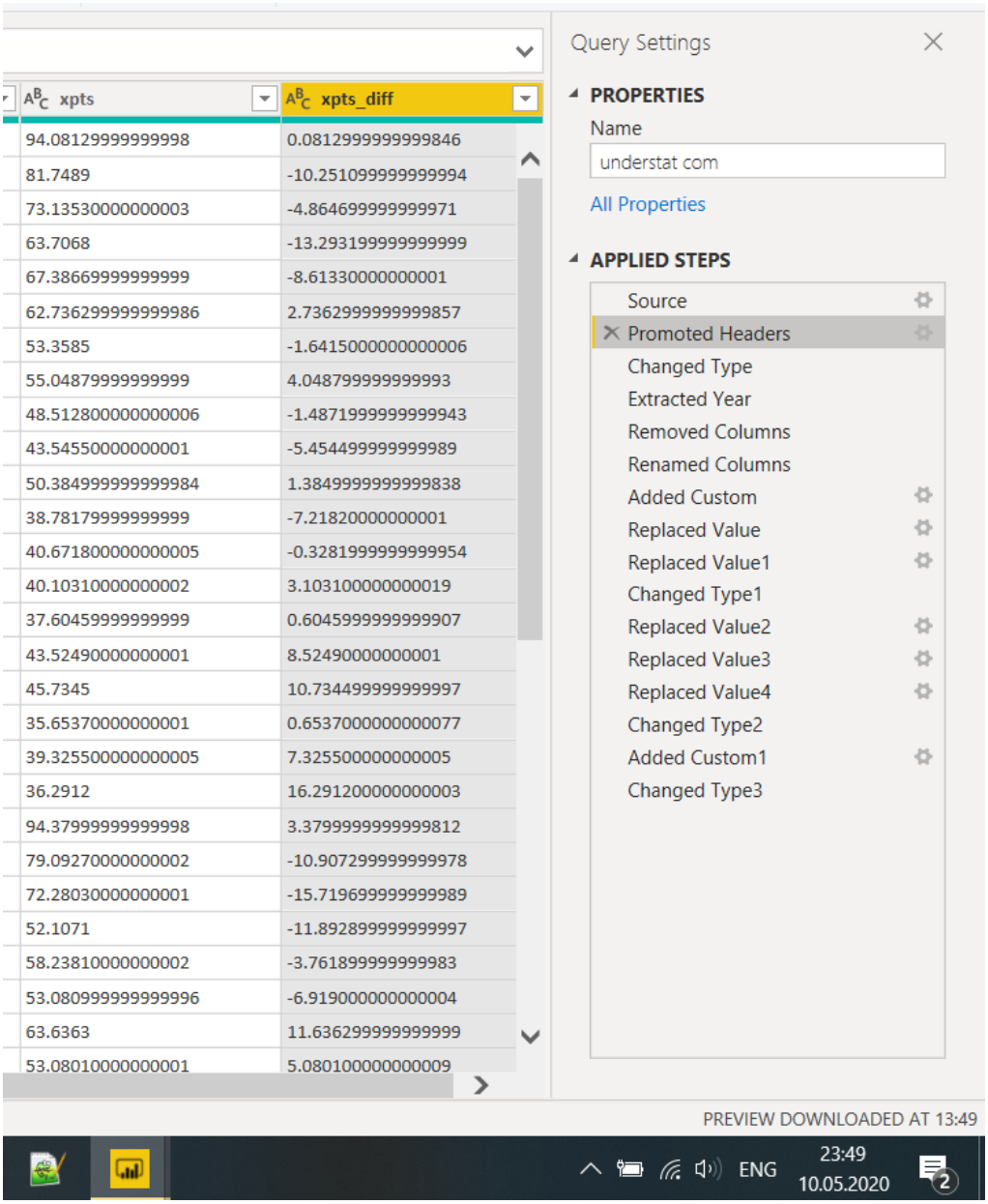
It is widely believed by many football enthusiasts that powerful English squads from EPL have started to dramatically over-perform in their home league, which plenty of fans connect to a smooth drop of other clubs' quality of play. So I have identified two statements that I am going to address in details later:

1. Squads from TOP5 EPL from year to year demonstrate a sharp increase of quantitative gap between strong football clubs and middle tier/outsiders. Thus, TOP5 EPL teams seem to over-perform frequently starting from season 2014/15.
2. Squads from the remaining TOP15 EPL are either existing in the middle of the table or in worst case compete with 2 random notorious teams in order to not be relegated to the EFL Championship. In other words, middle tier/outsiders experience a dramatic under-performance.

DATA ANALYSIS & EXPERIMENTS

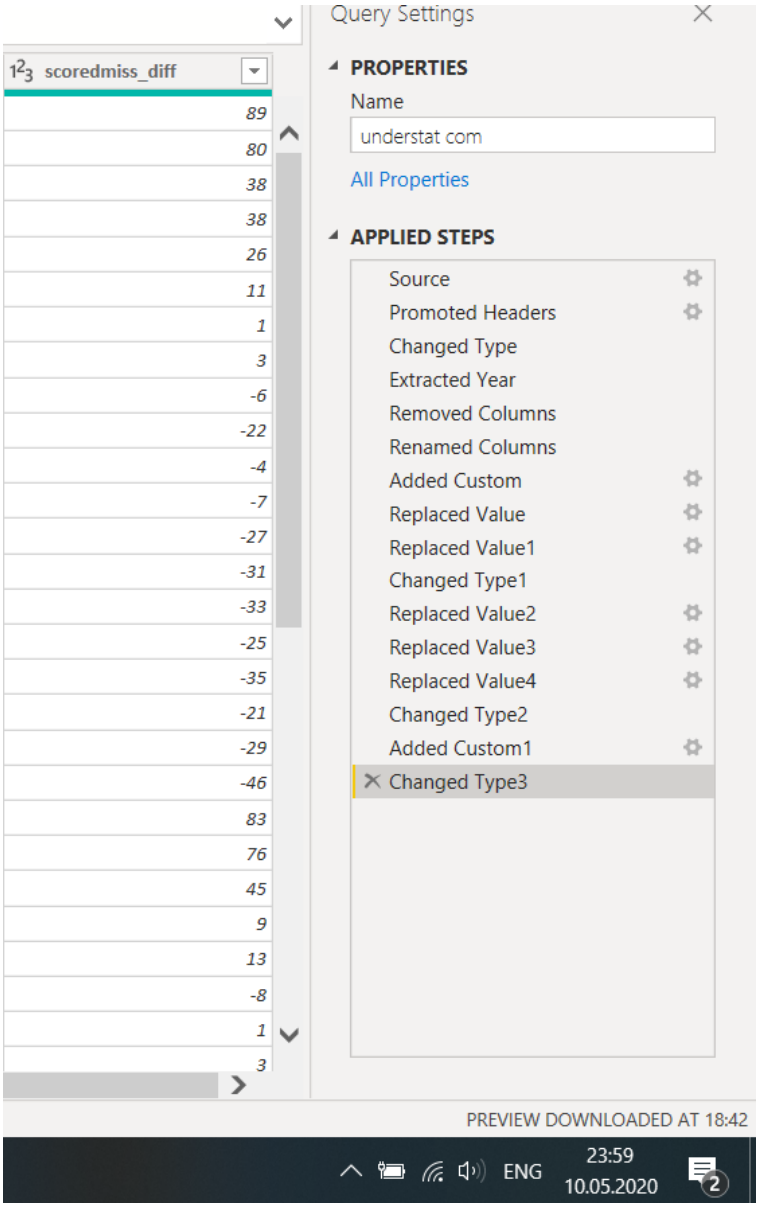
In order to either support these statements or refute them I chose Power BI Desktop, which provides enough instruments and tools for implementing plenty of graphs, charts, tables and other visuals.

Before digging into data from understat.com I had to perform some data cleaning, since initial information has various drawbacks. For example, most of floating-point numbers contained in the dataset is actually a simple text. The following picture below shows how the data looked like before transformation:



Picture 1 — Unformatted text strings.

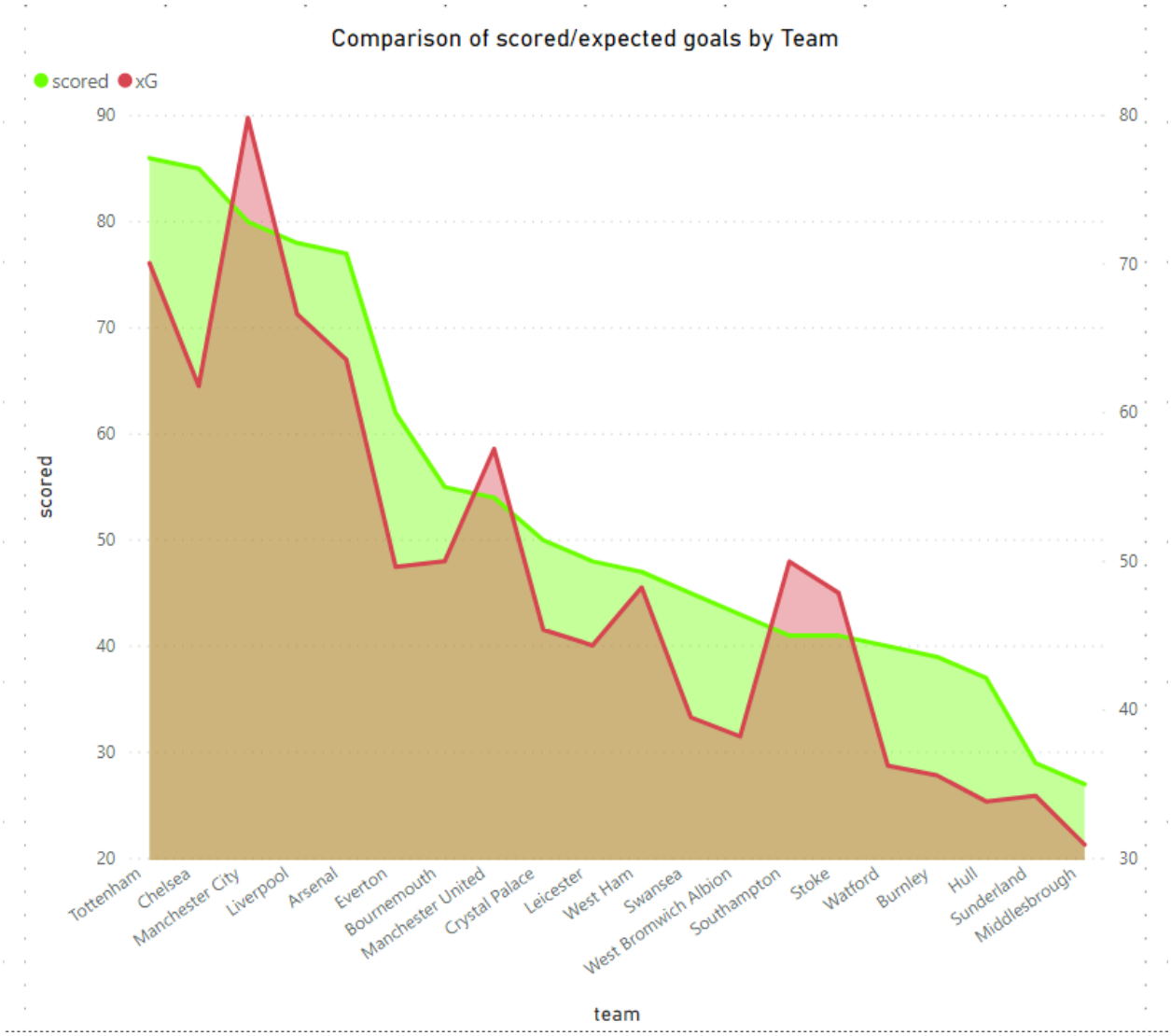
After converting all number-like strings into actual decimal numbers I decided to replace a column NPXGD, since it provides unnecessary data, which I couldn't find a proper application. Next, I perform conversion of date-like strings into actual date values such that I could easily specify the exact season, when I have to. Several empty columns have been renamed to avoid confusion. The "goal difference" column was introduced such that other users or third parties didn't have to compute the difference between actual goals scored and actual goals missed by a particular team.



Picture 2 — Applied steps section and the new column.

1.1 — Comparison of results of expected goals metric and real values

I have begun my short research by implementing a line graph, which provides information about scored/expected goals relation sorted by Squads. Quadratic blocks on the right side of Page 1 demonstrate an amount of matches completed by the end of season, the most points earned by champion and the least points earned by outsider.



Picture 3 — Comparison of EPL squads' scored/expected goals results at 2016.

Season	Squad Name	Goals scored	Expected goals scored	Scored/xG difference
2014 (3)	MU	62	54,21	7,79
	Tottenham	58	52,39	5,61
	Swansea	46	40,9	5,1
2015 (7)	Tottenham	69	63,42	5,52
	West Ham	65	54,4	10,6
	Liverpool	63	54,37	8,63
	Everton	59	53,98	5,02
	Sunderland	48	40,5	7,5
	Bournemouth	45	38,71	6,29
	Newcastle	44	37,61	6,39
2016 (6)	Tottenham	86	70,07	15,93
	Chelsea	85	61,80	23,2
	Liverpool	78	66,63	11,37
	Arsenal	77	63,58	13,42
	Everton	62	49,63	12,37
	Swansea	45	39,5	5,5
2017 (6)	MC	106	91,43	14,57
	Liverpool	84	77,49	6,51
	MU	68	59,04	8,96
	Leicester	56	50,29	5,71
	West Ham	48	36,80	11,2
	Bournemouth	45	39,99	5,01
2018 (3)	Liverpool	89	79,46	9,54
	Arsenal	73	64,8	8,2
	Tottenham	67	61,75	5,25

Table 1 — All over-performance cases (5+ goals gap) between seasons 2014/15 and 2018/19.

Since the first goal of this research is to investigate over-performance of TOP5 EPL clubs I decided to dedicate the following table above specifically for such squad types. When it comes to over-performance in terms of xG and scored goals, it means that a particular team has actually scored more than one could expect by applying xG metric.

Looking attentively into the goal scored/xG graphs of various time periods there is no doubt that in terms of amount of scored/expected goals there are very few cases, which can be characterized as an obvious over-performance. For the 5 year period only Tottenham — famous squad from London — and Liverpool from

Merseyside have presented a stable yet not always sharp over-performance in comparison with other clubs.

For example, in seasons 2014 and 2015 The Spurs scored/xG difference rate hasn't exceeded 5, 61 value. After that there was a dramatic increase in the difference rate had tripled and by the end of season 2016 the value was ~16, which is actually one of the most significant events happened within 5 years. It must be mentioned that Tottenham presented no over-performance signs in 2017 season, which might be an indication of players' tiredness after outstanding performance in the previous season. By the end of 2018 Tottenham only scored 5 more goals than it was expected with the use of xG metric.

Some may admit that Liverpool appeared 4 times at 5 year period. That is true. However, it is worth to note only seasons 2015 and 2016, at which The Reds scored 8,63 and 11,37 goals respectively.

Other interesting cases here are West Ham United and Chelsea. The Irons represent middle tier EPL clubs and seem to be the only football club in England, which is not a regular competitor in the upper part of EPL table, but at the same time West Ham presents some solid performance along with Swansea. Despite Swansea appeared as many times in the table as The Irons did, The Swans tend to *slightly* over-perform by scoring ~5 goals more than expected. On the other hand, The London club in case of over-performing, is able to score at least 10 goals more than it could be expected according to xG metric. Another club from London named Chelsea FC appears in the table only once, but with the greatest gap between scored and expected goals from 2014 to 2018 — 23.2 goals more than predicted by xG system.

Overall, stable gap increases have been demonstrated only by two football clubs: Tottenham and Liverpool (which started to get stronger since season 2017).

1.2 — Comparison of expected points and actual points rates

As it is obvious from the previous part, TOP5 EPL clubs have only one stable candidate, which can be referred to as “over-performing top-club”. This is Tottenham. To further extend my knowledge, I have created a line graph that compares actual earned points with predicted ones.

Season	Squad	True PTS	Expected PTS	PD
2014 (3)	Chelsea	87	75,32	11,68
	MU	70	63,03	6,97
	Swansea	56	43,32	12,68
2015 (3)	Leicester	81	68,94	12,06
	MU	66	56,44	9,56
	West Ham	62	49,8	12,2
2016 (4)	Chelsea	93	75,74	17,26
	Tottenham	86	75,37	10,63
	Liverpool	76	69,83	6,17
	Arsenal	75	62,12	12,88
2017 (3)	MC	100	91,09	8,91
	MU	81	62,33	18,67
	Burnley	54	41	13
2018 (5)	MC	98	90,64	7,36
	Liverpool	97	83,45	13,55
	Tottenham	71	61,44	9,56
	Arsenal	70	58,97	11,03
	West Ham	52	43,72	8,28

Table 2 — All over-performance cases (6+ points earned) between seasons 2014/15 and 2018/19.

Unfortunately, this comparison was not as detailed as the previous one due to fewer cases of squads performing noticeably better than it has been expected. First of all, I should mention multiple clubs that have a tendency to earn more points than other opponents. Noticeable signs of over-performance were spotted in case of Chelsea FC at seasons 2014 and 2016, but the squad’s results have returned to a normal state (at least on the end of 2018/19).

Liverpool FC’s difference between earned points and expected ones also became larger (from 6,17 in 2016 to 13,55 in 2018). Consequently, slight gap increase, which started in 2016, by the end of 2018 became much larger, indicating

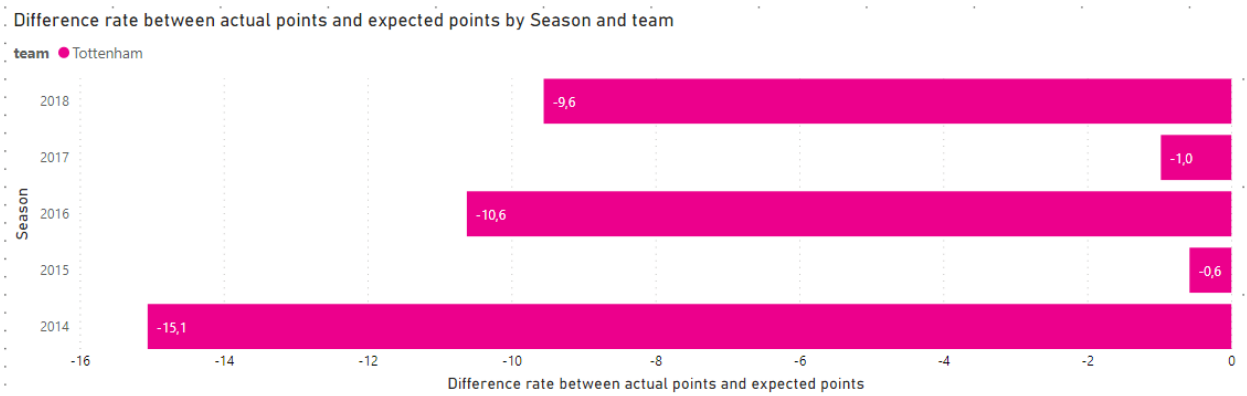
that Liverpool is either pretty luck squad or the team has been shifted towards team spirit and “last-minute” victories.

The most frequent cases in that comparison are again West Ham — The Irons, and Manchester United. Despite having “one-time” clubs such as Swansea in 2014 and Burnley in 2017, a solid squad from London demonstrates the best results among middle tier/outsider teams in EPL. For instance, in 2015 The Irons have earned 12 points more than the system has expected them to obtain. In 2018 the rate has reduced to 8.28 point gap.

On the other hand, Manchester United showed an undeniable upward trend. For example, in 2014 their PD rate was at 6,97 (almost 7) points. In the next year it increased on 2 and 2017 difference between expected/actual points achieved 18,67. This is the biggest map for 5 year period among all clubs in EPL.

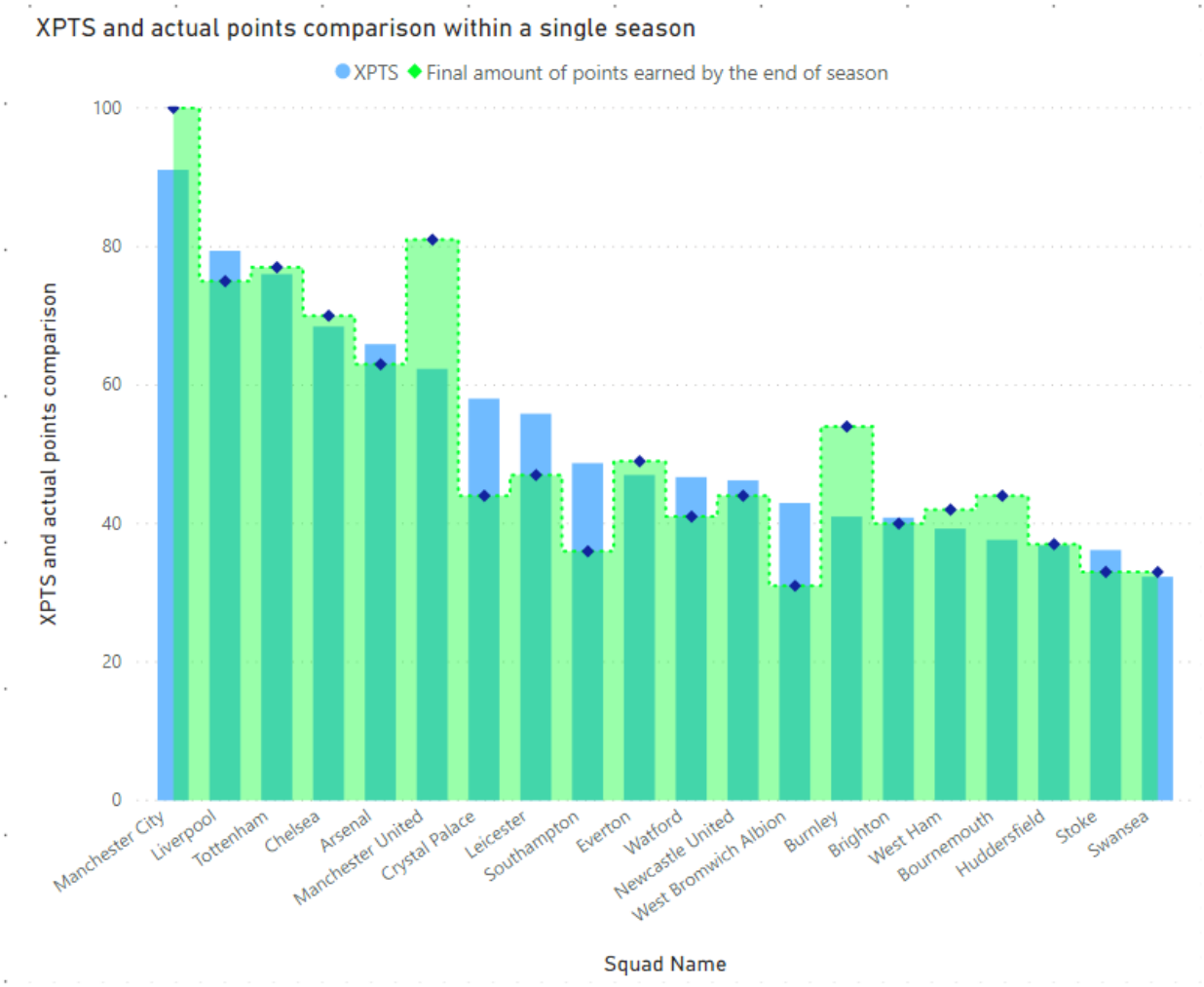
Overall, only 3 three clubs from high EPL tier presented very solid results in terms of over-performance by earned points. Those are Liverpool (hardly could be called a top-club considering later 2010s), Manchester United and Tottenham.

In addition, the stacked bar chart of Tottenham squad’s performance is given below:



Picture 4 — London top-club HAS NEVER showed under-performing rates from 2014/15 till 2018/19 seasons. All values are negative, which in case of XPTS metric mean that the squad is doing too well.

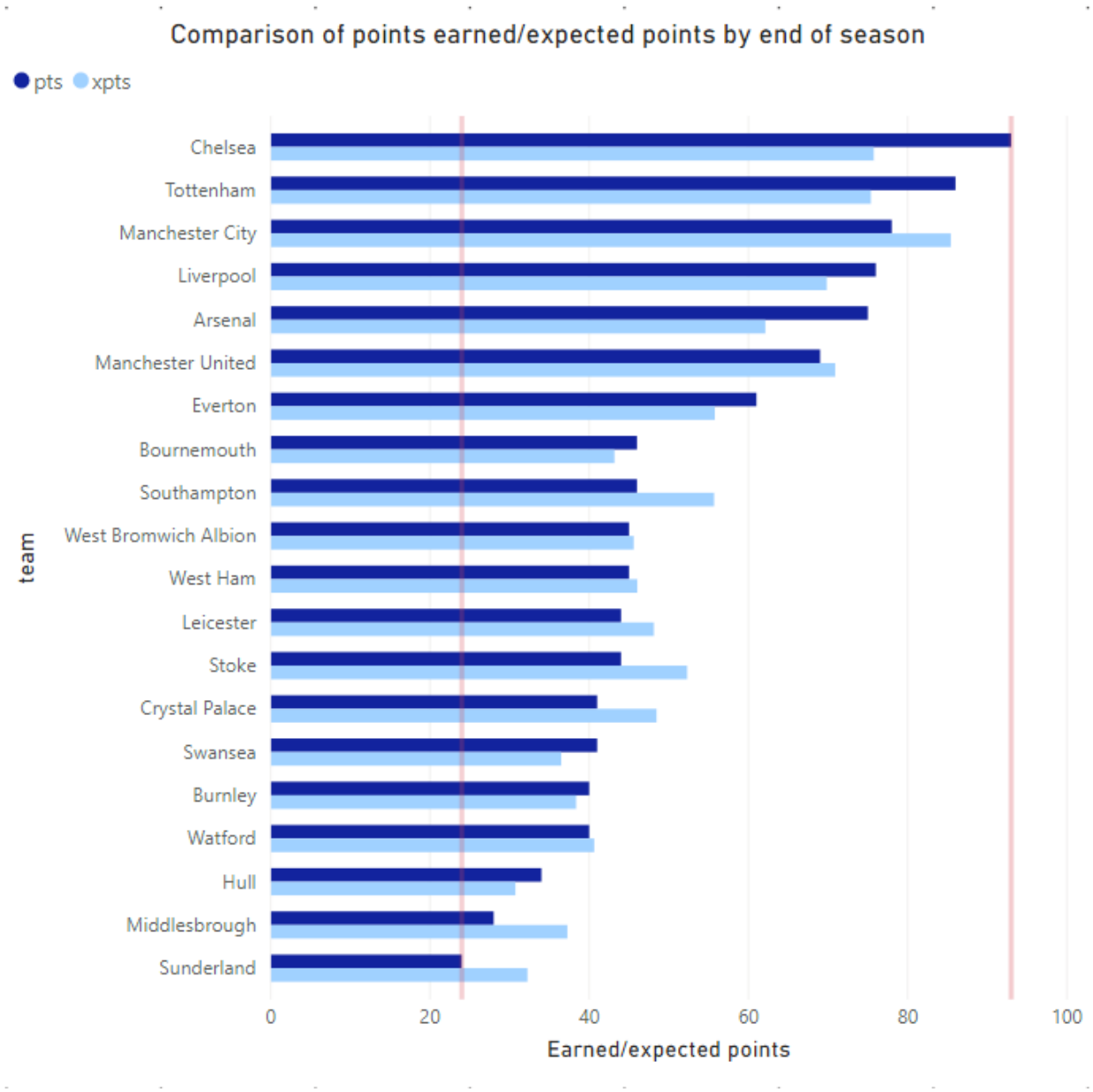
Another bar chart related to XPTS metric is more centered around showing the real difference between XPTS values and final amount of points earned by EPL teams at the end of particular season. An example below comprehensively describes which squads are doing too well, which are okay and which are seriously under-performing:



Picture 5 — Bar graph indicates MC, MU, Burnley and West Ham as the over-performing ones (earned points are painted in green), while teams such as Crystal Palace, (kind of) Leicester, Everton and West Bromvich Albion obtained less points than XPTS metric predicted them to.

To sum up all the given information, it is clearly seen that the first statement is only partially true for very few top-clubs (like Tottenham and Manchester United), which tend to over-perform more frequently and obtain more points on top of that.

However, not only TOP5 teams, but also middle tier and even some of outsiders showed that tendency for performing way better than expected is not related to some particular tier of teams in EPL. I also cannot fully support the second statement, which says “TOP15 teams seem to under-perform for the last years”, because in case of comparison of expected goals/actual goals rates, some squads from the bottom of English Premier League demonstrates a pure over-performance. For example, West Ham among all middle-tier clubs in England showed the greatest gaps on comparative graphs. Other similar clubs and outsiders while present slight over-performing rates, still prove that even a small squad may showcase pretty solid results in the long run.



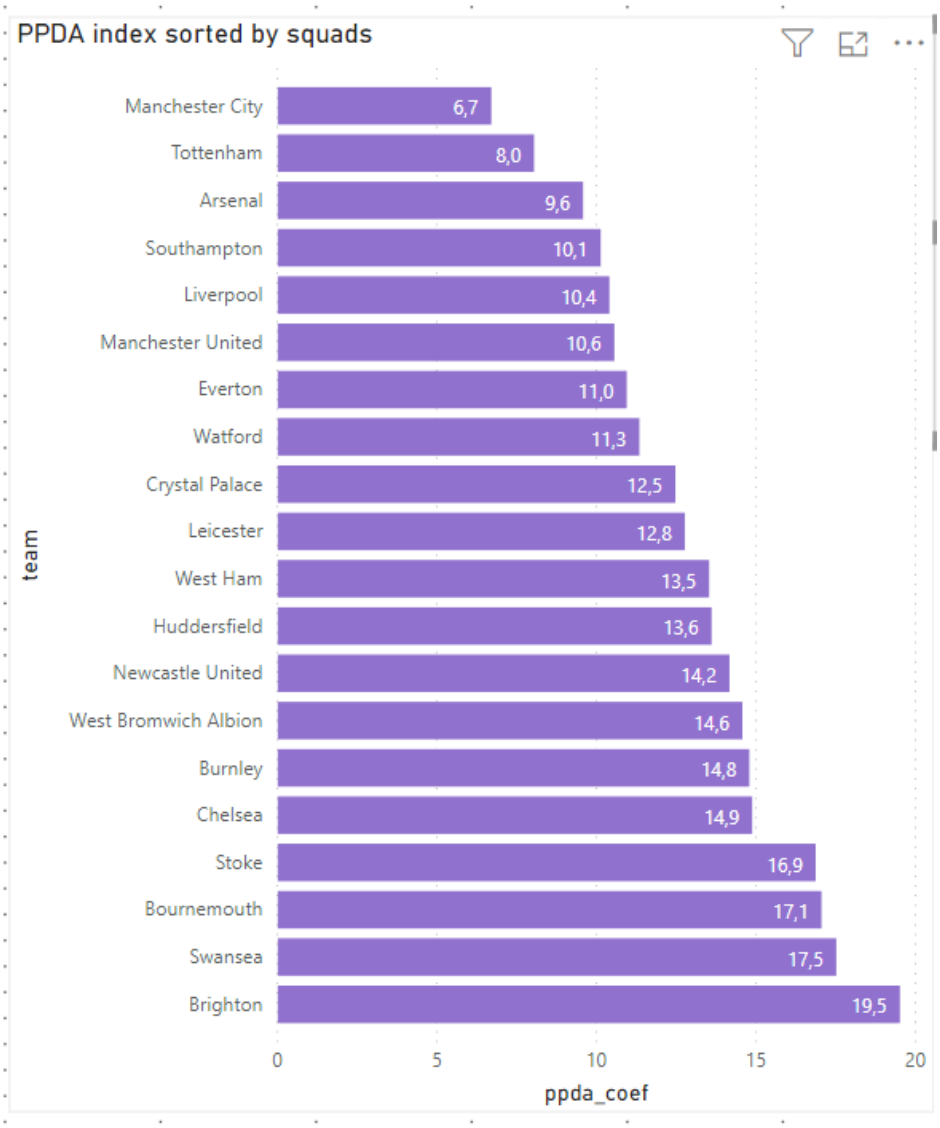
Picture 6 — Red lines show the best and worst squad in terms of earned points.

ADDITIONAL INFORMATION

1.3 — PPDA rates and how pressure intensity might affect squad's result in the long run

I suppose it is important to mention some interesting details (mostly reasons) about football teams' performance and how it is possible for some to do outstanding work on the field.

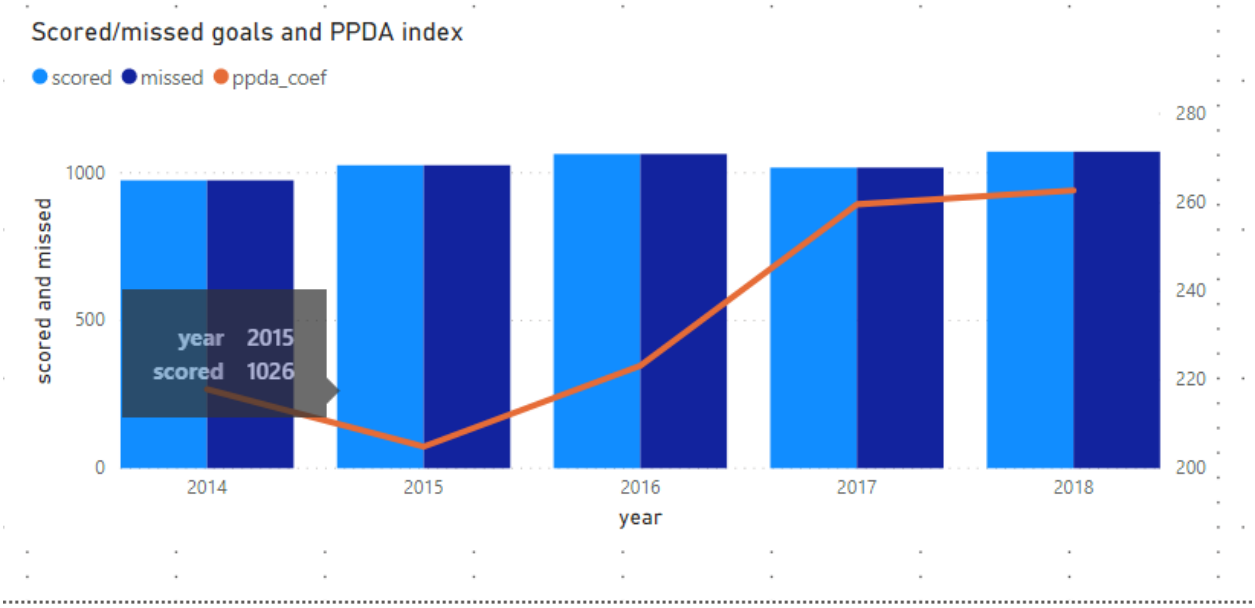
For instance, let's take a look on a PPDA rate for a season 2017/18. In fact, PPDA index is a great tool for defining the willingness of a particular team to engage into pressure on the half of opponent's field.



Picture 7 — PPDA results for different EPL squads in 2018/19 season.

The lowest PPDA rates belong to Man City, Arsenal and Tottenham squads. Citizens' numbers are quite impressive, because City players only allow ~6-7 passes on the opposition half, before they attempt a defensive move. By contrast, Tottenham players allow to do 1 more pass (their index is 8,05) and then Spurs trigger the pressure in order to counter-attack their opponent. Arsenal is only slightly worse than these 2 pressure monsters, having PPDA index of 9,58 that is still a great result in comparison with PPDA rates of other EPL squads.

Also there is pretty weird correlation, which is characterized by an overall recession of pressure intensity between 2014/15 and 2018/19 seasons.



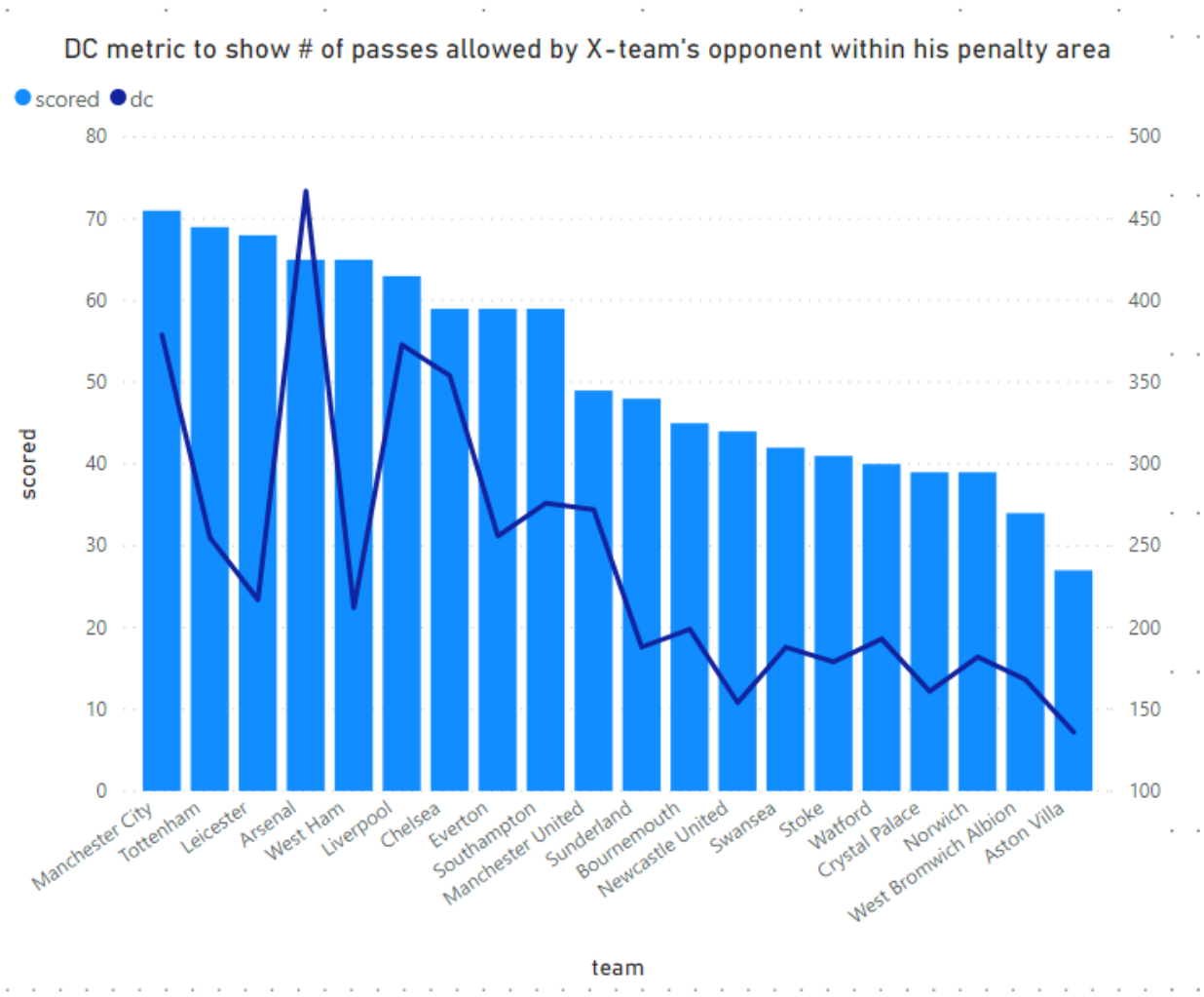
Picture 8 — PPDA index has increased between 2014 and 2018, which only means that professional football clubs are shifting away from high intensity pressure on the field and might concentrate on improvements and innovations (like engaging in pressure by trigger).

1.4 — DC and ODC metrics, passes allowed within penalty area

As it has been stated previously, **DC** is an quality metric that shows how good is some team at positional attacking. The higher DC rate is, the better team X performs in terms of **delivering the ball into opponent's penalty area**. For instance, Liverpool has DC index of 431, which means LFC players, while having

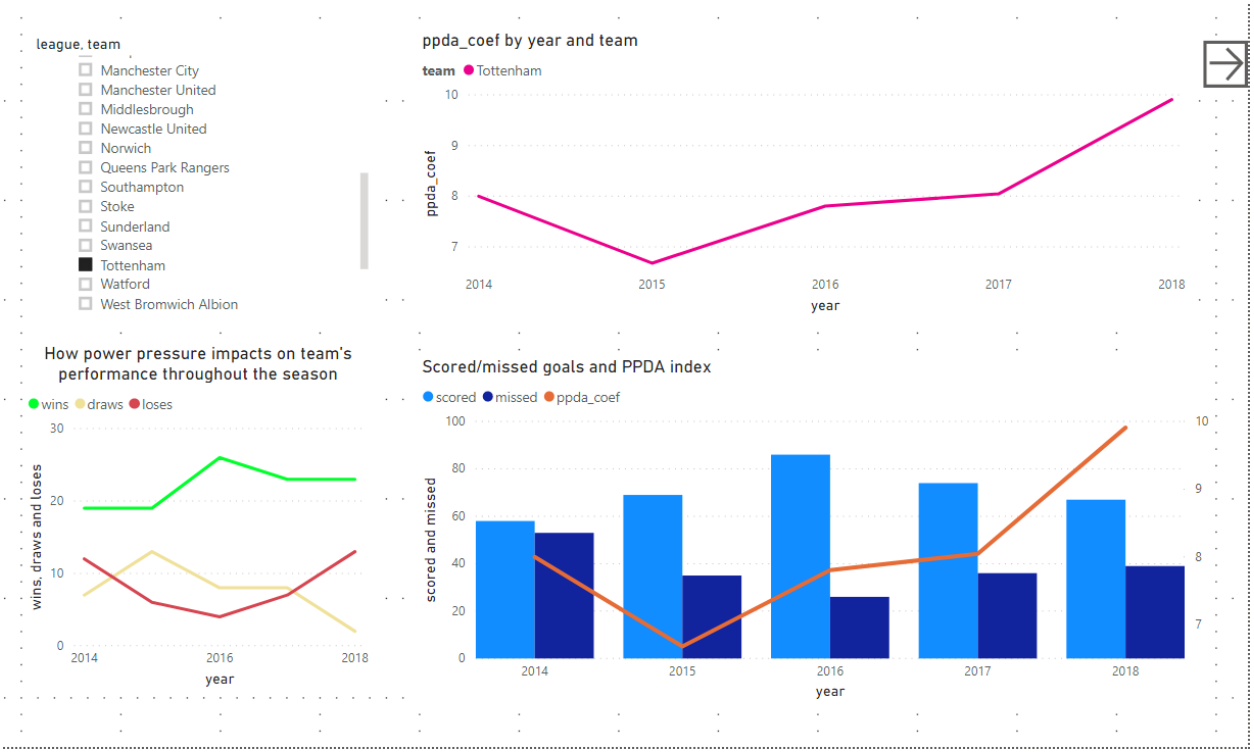
fewer passes in penalty area of opponents, have scored only **6 goals** less than 18/19 EPL champion Manchester City, who made 582 passes at the same conditions.

I would be unfair to not mention how statistically awesome has been Tottenham between 2014 and 2018 years. For instance, in 2018 Tottenham was only allowed to make 270 passes within 18 yards for a goal, and Spurs scored 67 goals by the end of 18/19 season. Considering Arsenal FC, it is understood that Gunners made 130 passes more than Tottenham players did, but only scored 73 (!) goals. The difference is only 6 goals. Spurs undoubtedly had shockingly monstrous rate of successfully executed scoring moments. This aspects of Tottenham's awesome play originates generally from seasons 2014/15 and 2015/16, at which Spurs also haven't had much of passes allowed by opponents, but still gained their position in EPL Top-club list, thanks to great usage of scoring moments. The line and stacked bar graph below provides data about this powerhouse from London:



Picture 9 — Spurs scored 69 goals being limited only to 255 passes before opponent attempts defensive action. By contrast, Manchester City scored 71 but required 379 passes within penalty area (18-meters).

I suppose that stable over-performance of Tottenham actually has its roots in the ability of a squad to effectively use scoring chances with less effort in terms of amount passes required.



Picture 10 — Tottenham's card showing an increase in PPDA rate (it means that Spurs' pressure intensity has been reduced slightly) and a noticeable rise of missed goals.

CONCLUSION

The proposed hypothesis, that contained 2 controversial statements yet popular in football community, proved to be wrong for the most part. Over-performance trend has been observed not only in case of TOP5 strongest football clubs in EPL, but also at some middle-tier and surprisingly enough outsiders. At the same time it also proves that some considerable part of medium/weak squads reaches expected numbers. Statistical approach and availability of charts, graphs and visuals helped me to better understand the problem of how some football clubs perform better than others or sometimes worse than they used to be seasons ago. I should admit that Power BI Desktop proved to be very reliable and efficient software for transferring advanced sports statistics into awesome comprehensive visuals, which might show hidden patterns or unexpected trends. I guess during the writing of master thesis I will try to apply Power BI Desktop software whenever possible in terms of data cleaning, data manipulation, transformation, adding new values, building interesting interactive graphs with multiple lines available and implementing multi-page reports.

REFERENCES

1. Understat.com — Advanced statistics and football quality metrics;
2. Rathke A. An examination of expected goals and shot efficiency in soccer //Journal of Human Sport and Exercise. – 2017. – Т. 12. – №. 2. – С. 514-529.
3. Brechot M., Flepp R. Dealing with randomness in match outcomes: how to rethink performance evaluation in european club football using expected goals //Journal of Sports Economics. – 2020. – Т. 21. – №. 4. – С. 335-362.
4. Herbinet C. Predicting football results using machine learning techniques //MEng thesis, Imperial College London. – 2018.
5. Van Haaren J., Davis J. Predicting the final league tables of domestic football leagues //Proceedings of the 5th international conference on mathematics in sport. – 2015. – С. 202-207.
6. Spearman W. Beyond expected goals //Proceedings of the 12th MIT sloan sports analytics conference. – 2018. – С. 1-17.
7. Eggels H., van Elk R., Pechenizkiy M. Explaining Soccer Match Outcomes with Goal Scoring Opportunities Predictive Analytics //MLSA@ PKDD/ECML. – 2016.
8. Gelade G. Evaluating the ability of goalkeepers in English Premier League football //Journal of quantitative analysis in sports. – 2014. – Т. 10. – №. 2. – С. 279-286.
9. Mackay N. Predicting goal probabilities for possessions in football //Master of Science. Vrije Universiteit Amsterdam. URl: https://beta.% 20vu.% 20nl/nl/Images/werkstuk-mackay% 5C_tcm235-849981.% 20pdf. – 2017.
10. Herbinet C. Predicting football results using machine learning techniques //MEng thesis, Imperial College London. – 2018.
11. Schauss P. Stupidity in football //The Aesthetics, Poetics, and Rhetoric of Soccer. – Routledge, 2018. – С. 141-160.
12. https://www.kaggle.com/slehkyi/football-why-winners-win-and-losers-lose