UNIVERSITI TEKNOLOGI MALAYSIA

Business Intelligence

Alternative Assessment

NAME	: Mohammad Amirun Haziq bin Mohammad Fadzli
MATRICS NUMBER	: A18CS0109
SUBJECT	: DBA
PROGRAM	: 3SCSP
Phone Num	: 016-7809910

Table of Contents

Table of Figures

Business Intelligence

BI Alternative Assessment Report

Introduction

English Premier League or EPL is the highest tier of football hierarchy in England's football. Founded in by FA England as FA Premier League on 20 February 1992 and changed its name to English Premier League later on. Using data source from Kaggle, match fixtures and result from season 2000-2001 until 2017-2018 can be found. The result shows the accumulation of the statistics for each game from time to time.

Match fixtures and results are beneficial for each match but the to get the overall picture of the season and points, an accumulation processes consisting of Extract, Transform, and Load can be used. ETL can be said as a framework to build a successful guideline in building a good workspace and getting the important result in mind.

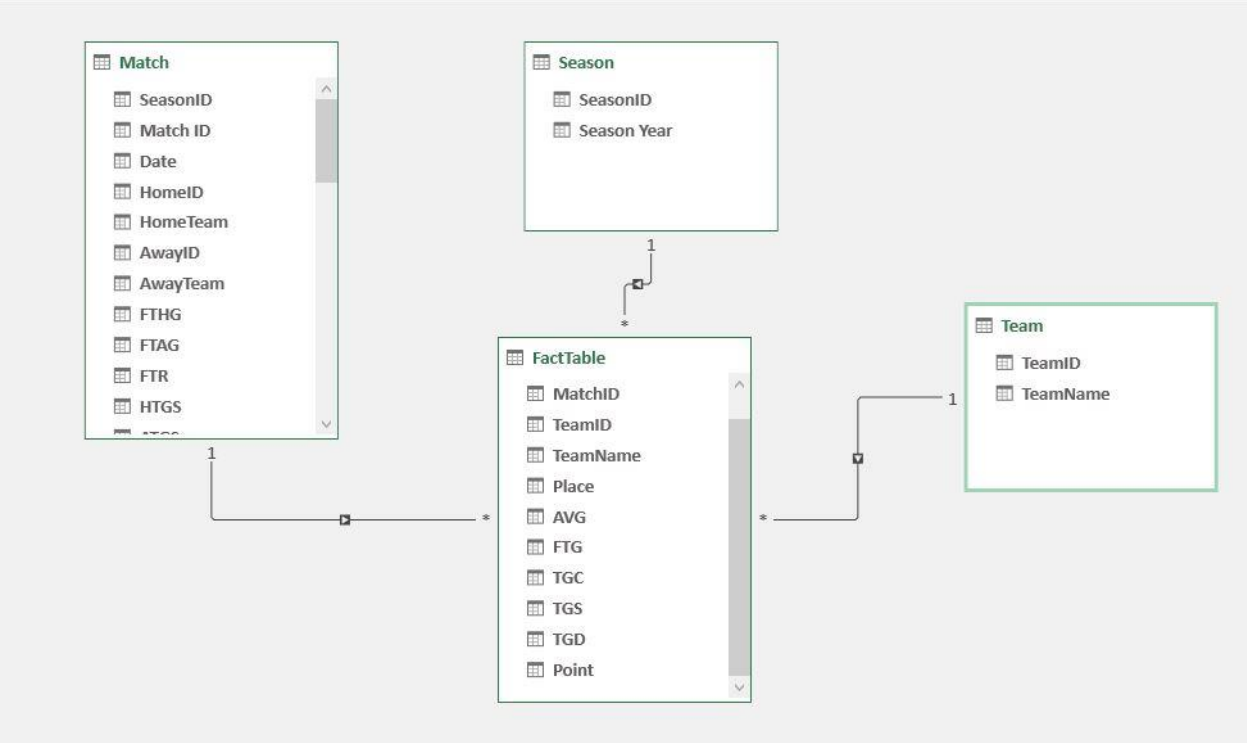
The objective of this report is to show the most prevalent result that can be found and may help in building a great team during the process gaining insight from the data. Using the data as means of measure is common and help big and small club establishing and improving themselves in the future. As per variables used in this analysis are such as Teams, Dates, Seasons (in Date dimension), Match Results and Stats, Home/ Away Result while compiled together to form a fact table that contain the information of match place, cumulative stats for a season, and their points.

Methodology and BI tool

- Extract: In this phase, we can see some data cleaning such as removing and trimming column happens.
- Transform: During this phase,
- Load: At this point,
- Tools Used: There are 2 major tools used which are Talend and MS Power BI:
    - Talend: As per data warehousing process, Talend played a major role in building a great workspace for that intention. Starting with connection to data source until producing output for the information needed, Talend have mechanisms to ensure the data warehousing can progress smoothly from starting point until finishing product.
    - MS Power BI: For data visualization part, MS Power BI have some great mechanisms in visualizing great data and information for a better understanding. The simplicity of Power BI brings great potential to visualizing information such as line chart, bar chart, and many more.

Data Warehouse

Relationship View



Data warehouse is a collection of data used mainly for queries and analytics. These helps organization in making a better business decision and build a systemic approach solving the organization's problems from time to time.

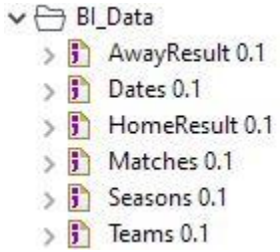


Figure 1: Putting file in metadata.

In this step, we will set the file path for the delimited file used and use it as a reference for a future usage. We will also ensure the type of delimited used, heading rows number, setting the schema for the data type, and creating purpose and comments as a guideline in the future.

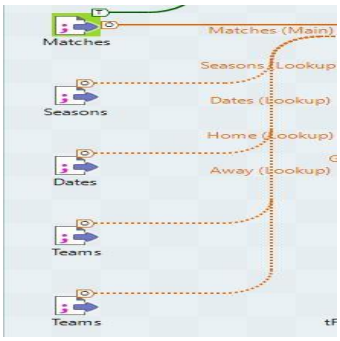


Figure 2:Linking Data sources to a tMap.

Then, we will be linking all the data sources to a single point of joining components, which is tMap. The type of data sources will be determined into two group, which are Main and Lookup tables. This process will be used in joining the data at tMap to get a better result based on our needs. The data sources can be easily dragged after we set the metadata file and use it for our convenience.

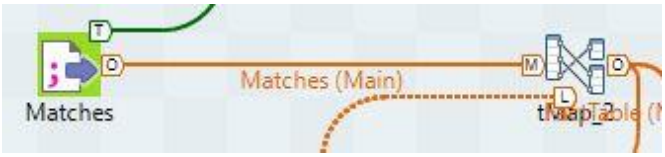


Figure 3:Joining the data sources to tMap.

After that, we will be joining all the data sources to the tMap. M will stand for Main table, L for Lookup table, while O for output. We can right-click at the connection if we want to change the type of connection. It is a good practice to ensure the connection is named with a better name. This will prevent any unnecessary misunderstanding.

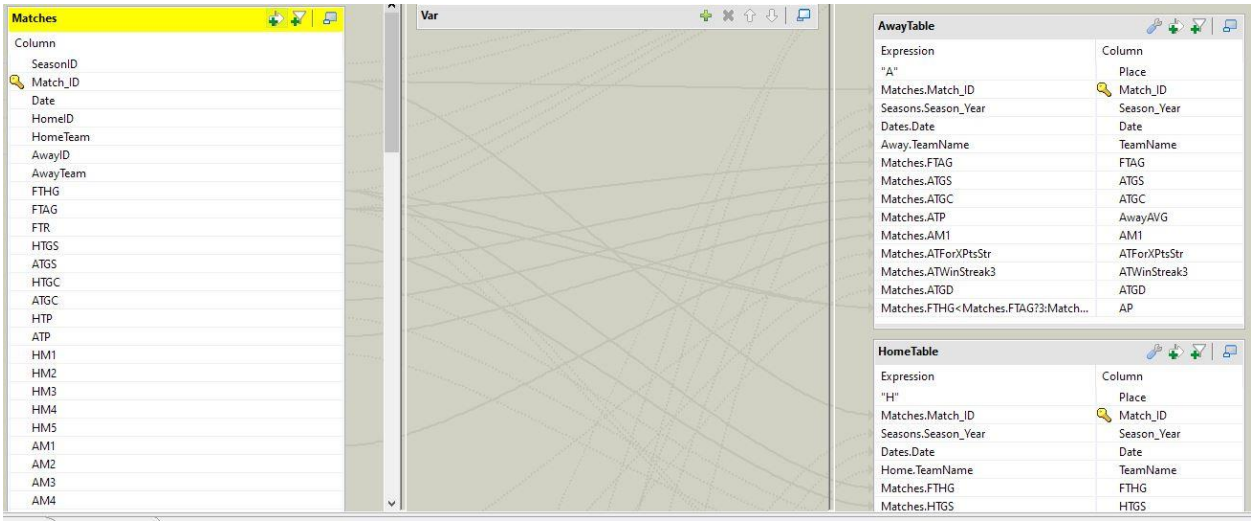


Figure 4: Separating the needed data to different table but same schema.

In our main table, we used Matches as a main table. The table used to consist of two different type of teams that played together in a match. One is Home Team, while another is Away Team. This data needs to be separated and set aside accordingly. There will be some data cleaning such as trimming and removing unnecessary column by having the needed data only in our table. We will also use inner join to join all the table together.

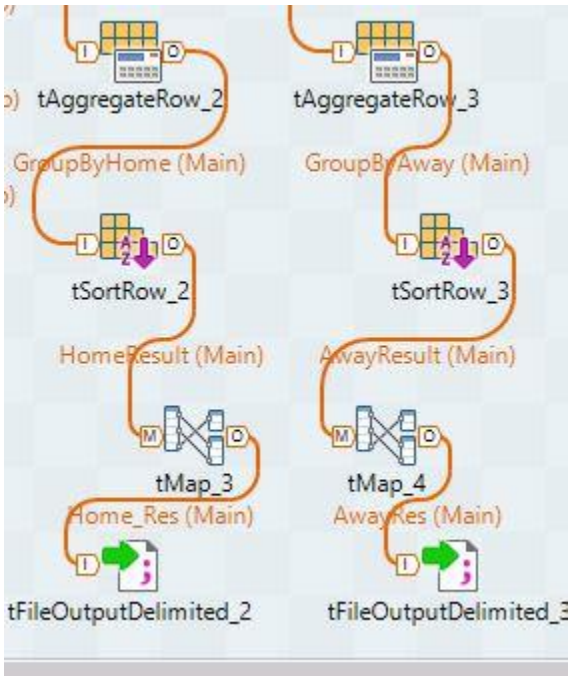


Figure 5: Use same operation for two table (Home and Away) with same schema.

In our case, the output from the data will have three different operations that executed separately but with the same method. First, tAggregateRow will be used to group the data and getting aggregation value such as COUNT, SUM, and AVG. This aggregation value will be being used to get a better insight for the overall picture. Next, is tSortRow to sort the row based on column mentioned. Lastly, tMap to ensure the data needed for the output table will being produced as intended.

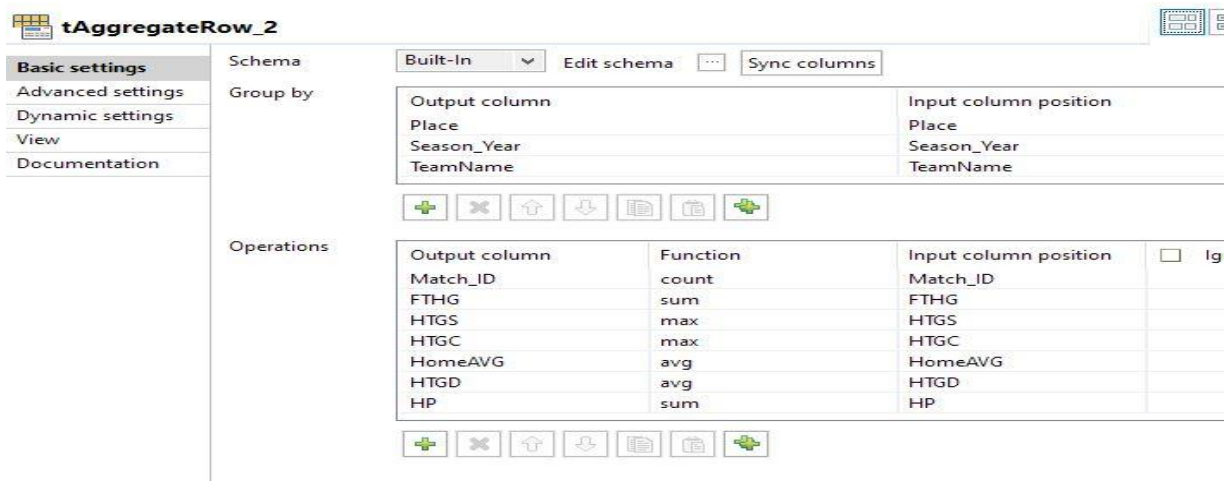


Figure 6: Group the table to get aggregated value.

tAggregateRow will group the data based on place, season year, and team name. Then, it will get the aggregated value for the analytics and visualizations purpose later on.

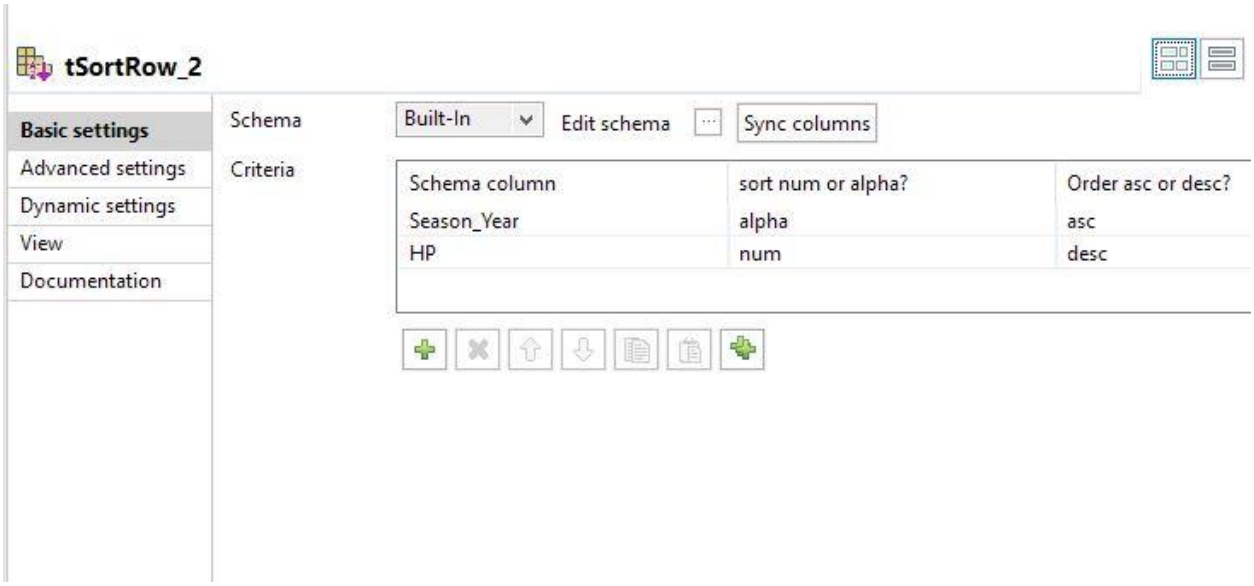


Figure 7: Sort Row based on column specified.

Next, we will sort the row based on season year and home point (HP). Started from the early season and sort the point based on the leader of the points. It can be used to ease our checking if the data is in correct order or in a better position.

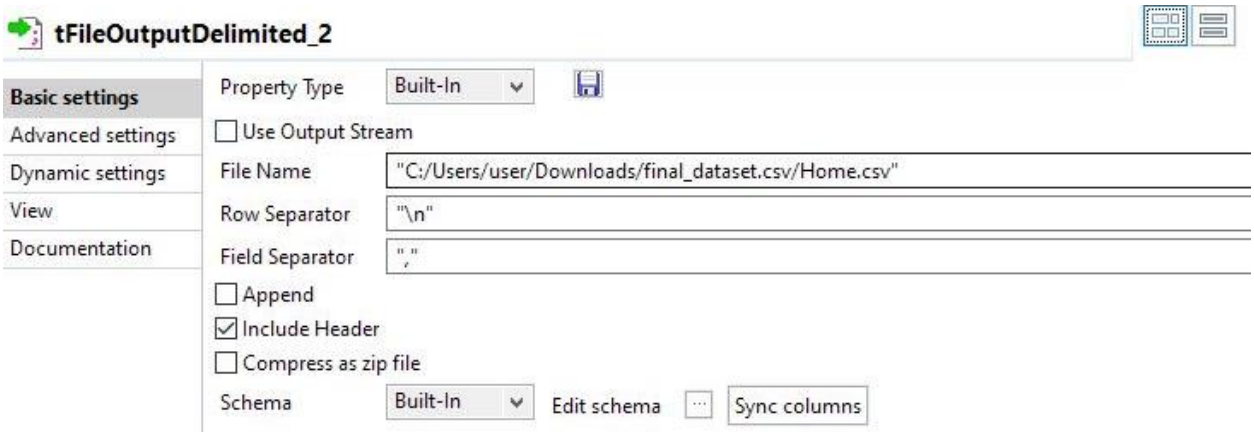


Figure 8: Output the table in specified path.

tFileOutputDelimited will produce the output in specified path. You can set the row separator, field separator, and option of header in the file. As I said before, the same procedure will occur to Away table also and having same processes but output the file in sale directory but different filename.

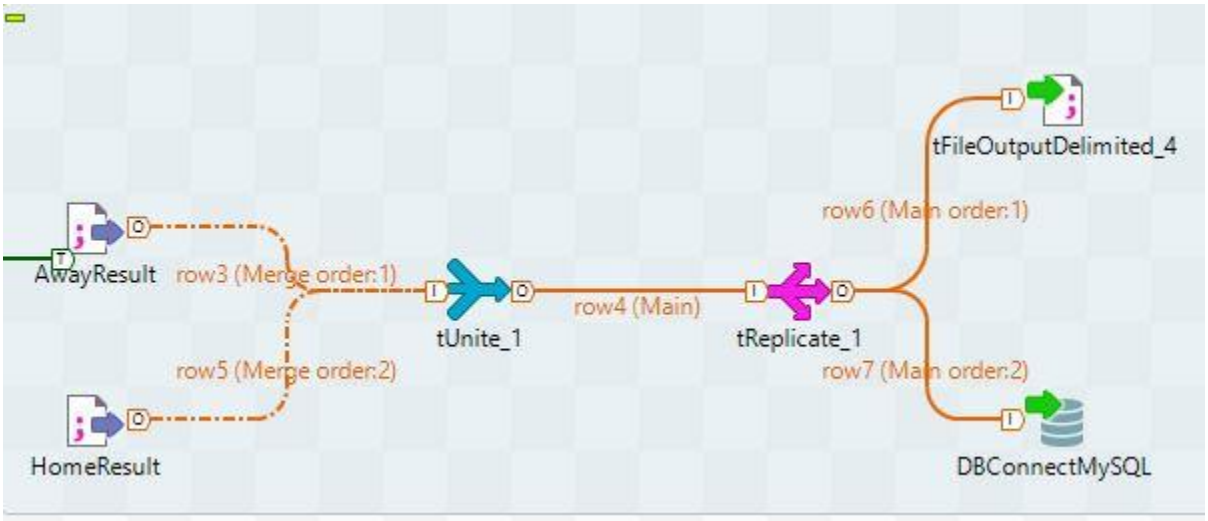


Figure 9: Subjob of merging and replicating result to two different outputs.

After we ensure the processes are progressing smoothly, we will merge the table together to have a fact table consisting of Away and Home result together. The process will be being proceeded by tUnite and followed by tReplicate to produce two different output types. First, delimited file and the latter one are database output file.

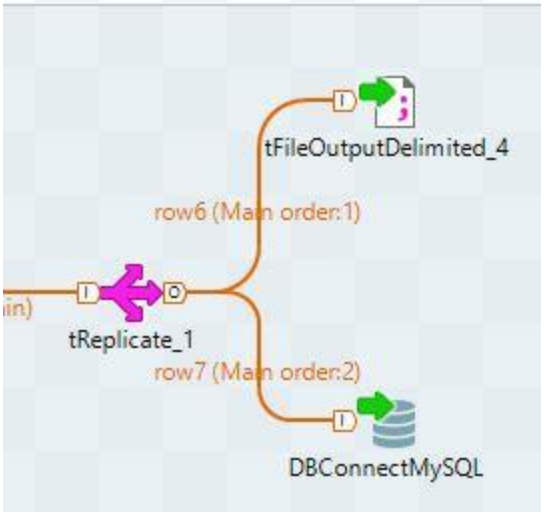


Figure 10: Producing delimited file and database output.

tReplicate will make two main connections to have a two same result but different destination. First, will be tFileOutputDelimited that will be being used to produce delimited file for visualizations purpose. While the latter for connecting to database and the requirement of the assessment.

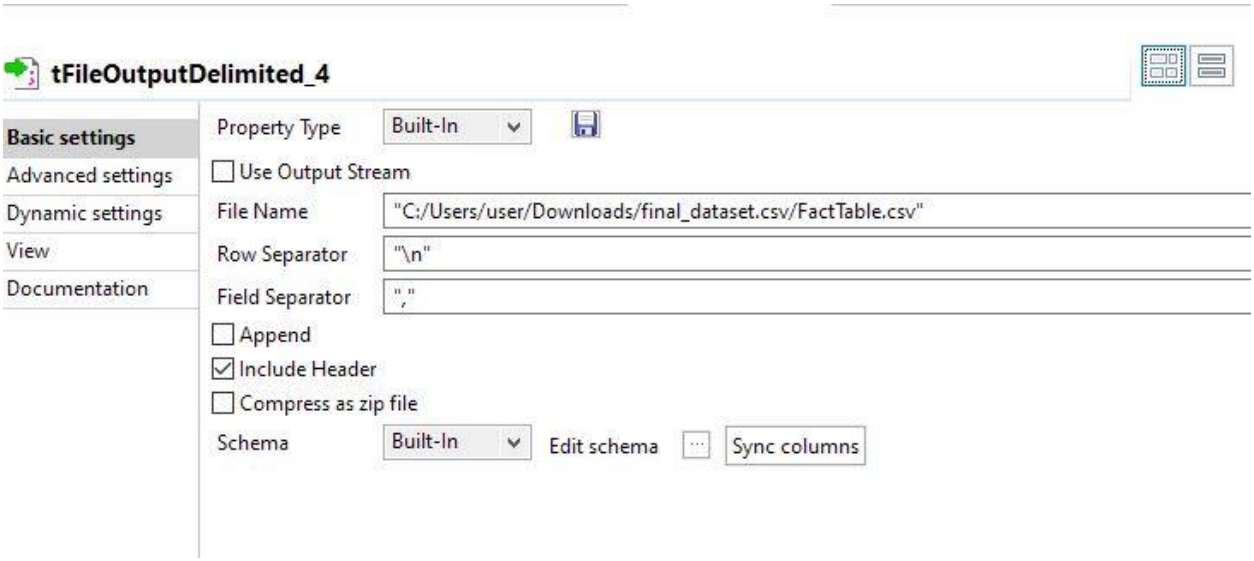


Figure 11: Setting fact table.

After the merging of the data, the fact table is finally produced. Set the filename and path correctly to ensure the Load process of data for visualizations are produced. As stated before, it also can set the row separator, field separator, and header inclusion. Make sure the schema is tally with the schema needed and the end product can be obtained.

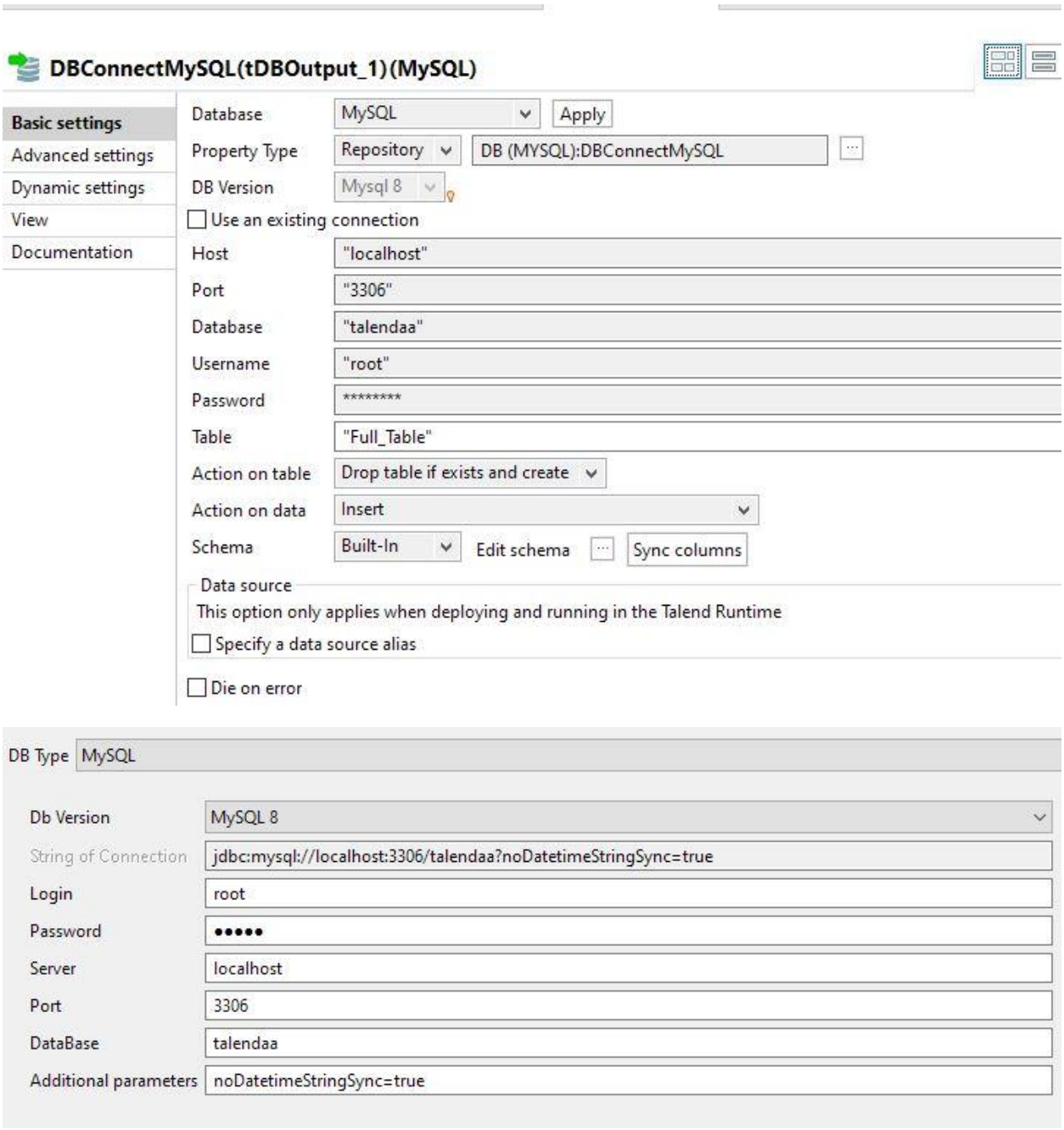


Figure 12: Connecting output to the database (MySQL).

To connect with database, we must set the connection of the database to the Talend properly. Make sure all the credentials are being inserted correctly. As for me, the server used is localhost and using port 3306 as the connector. Database also need to be created before moving on to the next step.

Output Table

Home (360 rows)

Place	Match_ID	Season_Year	TeamName	FTG	TGS	TGC	AVG	TGD	Point
H	19	2000-2001	Arsenal	45	59	34	1.5511655	1	48
H	19	2000-2001	Man United	49	77	26	1.9534248	1	47
H	19	2000-2001	Liverpool	40	65	37	1.4944032	1	43
H	19	2000-2001	Chelsea	44	62	41	1.1943333	1	42
H	19	2000-2001	Tottenham	31	44	53	1.1812152	0	39
H	19	2000-2001	Charlton	31	50	53	1.1912998	0	38
H	19	2000-2001	Ipswich	31	54	40	1.3451158	0	38
H	19	2000-2001	Leeds	36	61	42	1.3948858	0	36
H	19	2000-2001	Southampton	27	37	46	1.0497149	0	35
H	19	2000-2001	Newcastle	26	41	50	1.2832043	0	34
H	19	2000-2001	Sunderland	24	41	37	1.2844393	0	34
H	19	2000-2001	Leicester	28	34	46	1.4474567	0	34
H	19	2000-2001	Aston Villa	27	43	38	1.3316253	0	32
H	19	2000-2001	Derby	23	36	58	0.7705961	0	31
H	19	2000-2001	Everton	29	43	57	1.014262	0	26

Away (360 Rows)

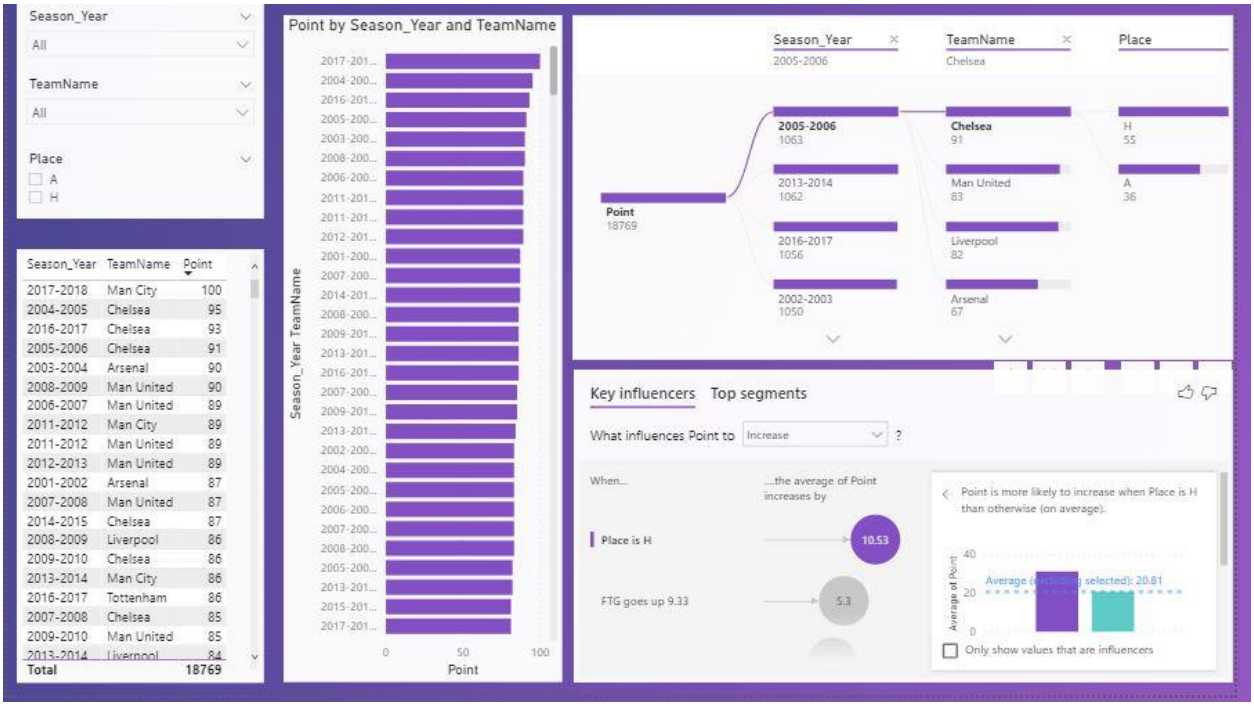
Place	Match_ID	Season_Y	TeamNam	FTG	TGS	TGC	AVG	TGD
A	19	2000-2001	Man Unite	30	78	28	2.042467	1
A	19	2000-2001	Leeds	28	54	39	1.452532	6
A	19	2000-2001	Ipswich	26	56	41	1.390681	-1
A	19	2000-2001	Liverpool	31	67	39	1.583042	6
A	19	2000-2001	Sunderlan	22	44	39	1.391891	-3
A	19	2000-2001	Middlesbr	26	41	40	0.848094	5
A	19	2000-2001	Arsenal	18	61	35	1.681835	-7
A	19	2000-2001	Aston Vill	19	46	40	1.256967	-4
A	19	2000-2001	Chelsea	24	66	44	1.319884	-1
A	19	2000-2001	Man City	21	39	61	0.889436	-13
A	19	2000-2001	West Ham	21	44	48	0.973319	-9
A	19	2000-2001	Newcastle	18	41	50	1.319083	-15
A	19	2000-2001	Southamp	13	35	45	1.032987	-13
A	19	2000-2001	Everton	16	40	54	0.998233	-16
A	19	2000-2001	Coventry	22	34	58	0.852207	-18
A	19	2000-2001	Charlton	19	48	50	1.318825	-19

Fact Table (Home + Away) –>same schema (360 + 360 rows)

1	Place	Match_ID	Season_Year	TeamName	FTG	TGS	TGC	AVG	TGD	Poin
2	A	19	2000-2001	Man United	30	78	28	2.0424669	11	3
3	A	19	2000-2001	Leeds	28	54	39	1.4525316	6	3
4	A	19	2000-2001	Ipswich	26	56	41	1.3906808	-1	2
5	A	19	2000-2001	Liverpool	31	67	39	1.583042	6	2
6	A	19	2000-2001	Sunderland	22	44	39	1.3918908	-3	2
7	A	19	2000-2001	Middlesbrough	26	41	40	0.8480941	5	2
8	A	19	2000-2001	Arsenal	18	61	35	1.681835	-7	2
9	A	19	2000-2001	Aston Villa	19	46	40	1.2569668	-4	2
714	H	19	2017-2018	Crystal Palace	29	43	55	0.61733824	0	2
715	H	19	2017-2018	Bournemouth	26	41	58	0.7870252	0	2
716	H	19	2017-2018	Burnley	16	35	37	1.4522959	0	2
717	H	19	2017-2018	Huddersfield	16	28	57	1.127762	0	2
718	H	19	2017-2018	Swansea	17	27	54	0.7785288	0	2
719	H	19	2017-2018	Stoke	20	32	65	0.82812244	0	2
720	H	19	2017-2018	Southampton	20	37	55	0.9219135	0	1
721	H	19	2017-2018	West Brom	21	30	54	0.8213299	0	1

Data Visualization

Data Visualizations are the process of transforming information to visual representation. This is to ensure the data can be easier to understand and see from different perspective and angle. It is in human nature to understand the info information using visual cue rather than boring text representation.



Season_Year	TeamName	Point
2017-2018	Man City	100
2004-2005	Chelsea	95
2016-2017	Chelsea	93
2005-2006	Chelsea	91
2003-2004	Arsenal	90
2008-2009	Man United	90
2006-2007	Man United	89
2011-2012	Man City	89
2011-2012	Man United	89
2012-2013	Man United	89
2001-2002	Arsenal	87
2007-2008	Man United	87
2014-2015	Chelsea	87
2008-2009	Liverpool	86
2009-2010	Chelsea	86
2013-2014	Man City	86
2016-2017	Tottenham	86
2007-2008	Chelsea	85
2009-2010	Man United	85
2013-2014	Liverpool	84
Total		18769

Figure 13: Page 1 - Home/Away Points Relationship

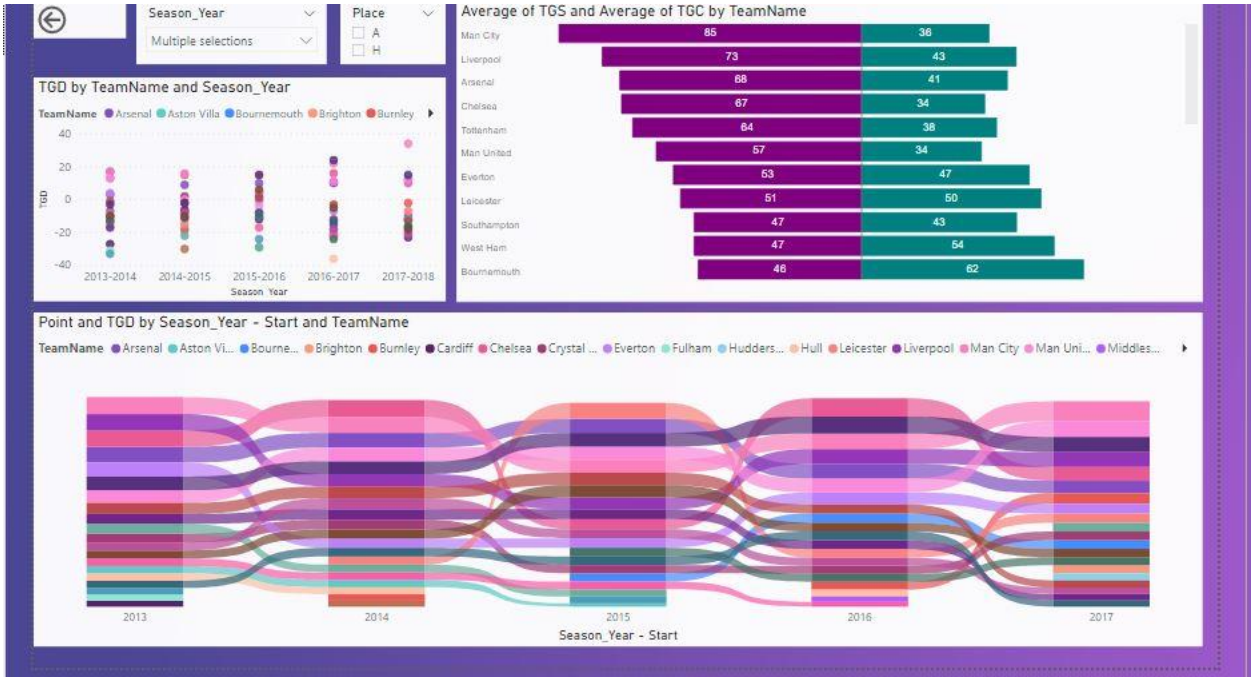


Figure 14: Page 2 - Total Goal Difference Factors

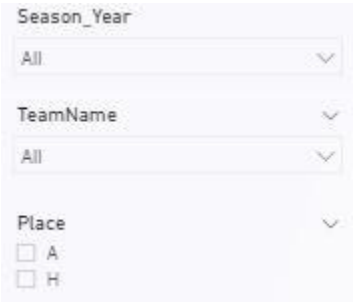


Figure 15: Filtering Important Features

Slicer are used to filter the value of the visualizations are within organization's needs. This can prevent cluttered graph or chart and only emphasizing important needed data. This process will ensure the decision makers keep track with the important features and used it appropriately according to business's needs. The filtering will apply to all elements in same dashboard.

Season_Year	TeamName	Point
2017-2018	Man City	100
2004-2005	Chelsea	95
2016-2017	Chelsea	93
2005-2006	Chelsea	91
2003-2004	Arsenal	90
2008-2009	Man United	90
2006-2007	Man United	89
2011-2012	Man City	89
2011-2012	Man United	89
2012-2013	Man United	89
2001-2002	Arsenal	87
2007-2008	Man United	87
2014-2015	Chelsea	87
2008-2009	Liverpool	86
2009-2010	Chelsea	86
2013-2014	Man City	86
2016-2017	Tottenham	86
2007-2008	Chelsea	85
2009-2010	Man United	85

Figure 16: Sorting Team According to Point

Table is a great measure to see the data with appropriate format and order. Using the use of sort and filter function, we can detect the teams that accumulated highest point and lowest point during a season, or a period of time. For example, during 2000-2001 until 2017-2018, we can find that the highest number of accumulated is by Man City during 2017-2018, for an amazing 100 points. Looking at the data, we can refer Man City during their glory days to understand more things that are needed to become more successful.

Best Practice: Put the most important features in the leftmost, and the measuring variable in the right most.

Interactive Element: Can be sorted and linked with Slicer.

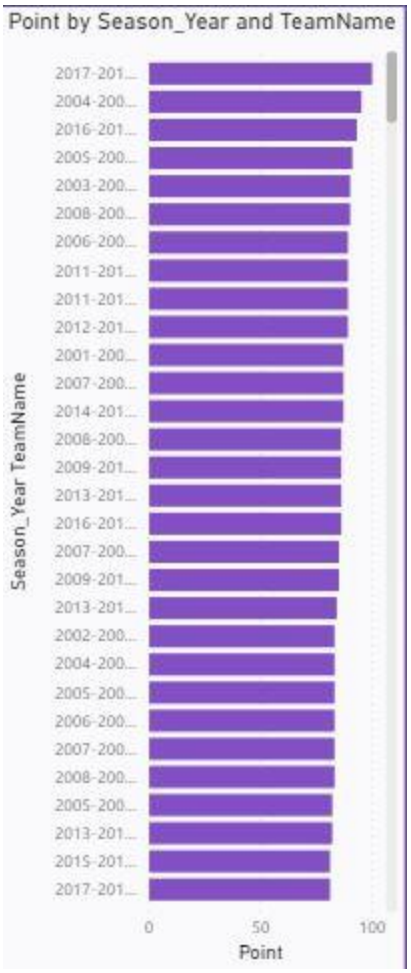


Figure 17: Comparing the Teams' Position Better with Visual Representation

Bar graph has same purpose with table in detecting which teams have the highest point during a season. It is a fair point to say that Bar graph is more appropriate in comparing the highest point accumulated by Team during a Season Year. The use of visual cue can illustrate the comparison between teams and gaining easier understanding how outstanding the performance of a team in a season.

Best Practice: Only insert the important values in the graph, sort it, and highlight the important values.

Interactive Element: Can be sorted, can use function drilling, and linked with slicer.

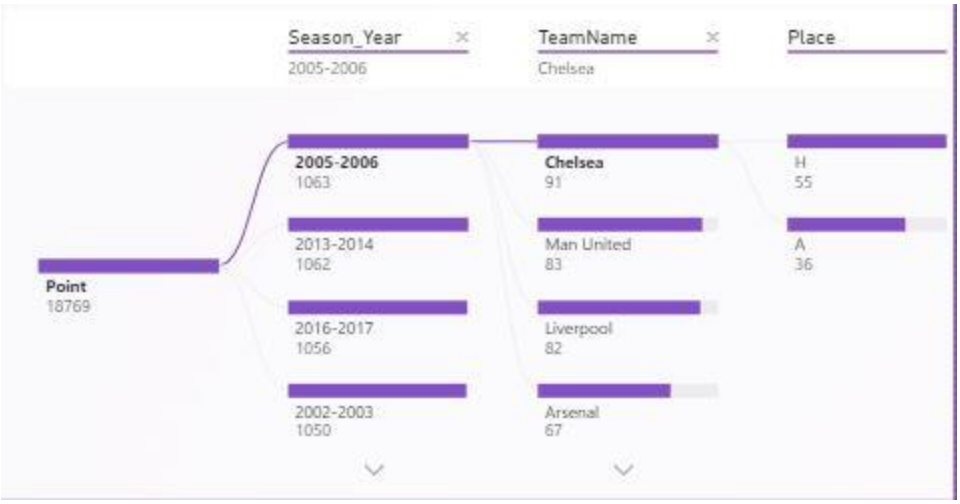


Figure 18: Finding Relationship Between Point to Place for Each Team and Season

Decomposition tree can expose to a team which play field that they need to improve. Drilling down until Place factor can be a measure whether it is needed a better performance in certain factor. For example, if a Team have higher amount points in Home, it maybe a factor of fans, total distance travelled, spirit, and home stadium controllable factor.

Best Practice: Put the accumulated in leftmost and put the hierarchy according to the feature's importance.

Interactive Element: Can be sorted, use drilling function, hide the less important features.

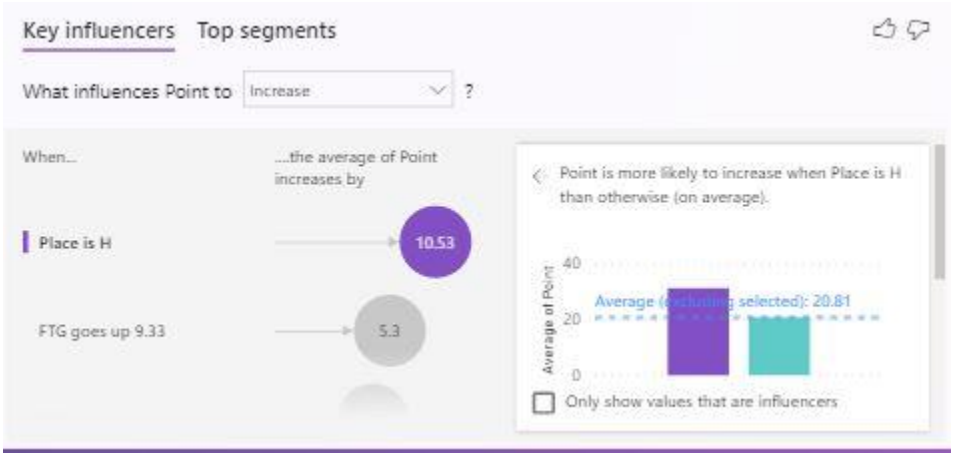


Figure 19: Key Influencer Point to Other Factors.

Key Influencers may be a cheat code in Power BI. It can use their built-in method to find the correlation and relationship between measured factor and other influence factor. As a conclusion, it can be a wildcard for the analyst to define any influential factor if they cannot see the relationship between factors as Key Influencers will detect influence factor to measured variable.

Figure 20: Finding Outlier For Total Goal Difference.

For scatter plot, we can find the mean or the mode of total goal difference by team and during certain season. If we consider the graph like a 3D model, we can see the number of total goal difference are resembling normal distribution. What we can acquire based on that are not the mode and the mean as the important part. It is actually the outlier of the teams during the seasons. For example, in 2016-2017, a team can be distinctively shown further from the crowd. This will bring us some question whether their number of total goals scored is high or low, or total goals conceded are high or low, or both factors are low resulting to resulted result. They need to empower their striking force or strengthen their defense to keep up with other teams. (If we can adjust in Power BI, we can group the data by the outlier groups)

Best Practice: Sort the X axis according to ascending order, use different color for each team and put legends to the graph.

Interactive Element: Linked with slicer

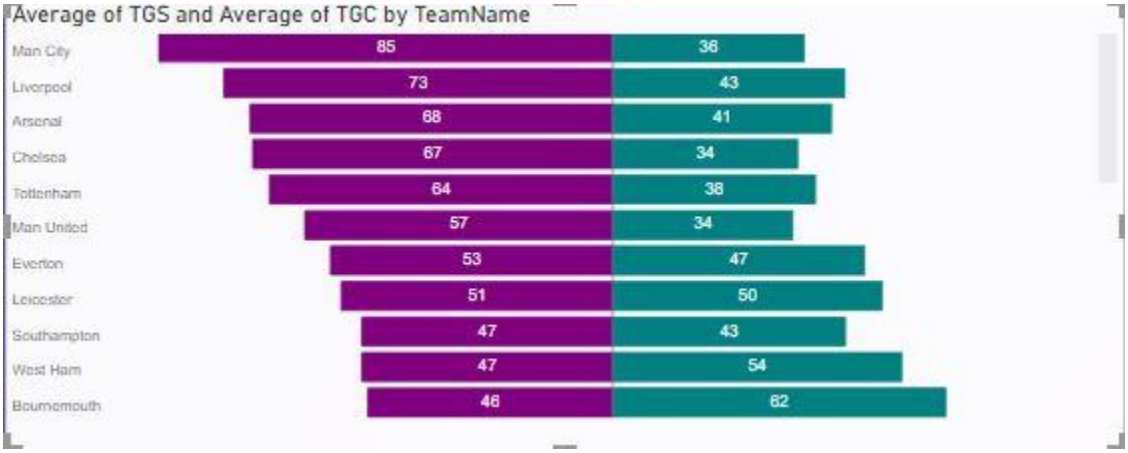


Figure 21: Measuring Relationship Between TGS and TGC

As stated before, Total Goal Difference or TGD plays a huge factor in teams' performance. With torpedo chart, we can see in the last 5 years, the average of Man City TGS are highest and it shown by their consistency for the last 5 years. Even though, Man City cannot boasts having the lowest TGC, but surely their TGD are one of the major factors making they the best teams during the last 5 years. Sir Alex Ferguson, one of the big names in football said:

"Attacking wins you games, but defense wins you title."

Even though, Man City did not have the best defense, but the little difference between defensive performance, and overwhelming power have successfully made them the best in EPL in the last 5 years.

Best Practice: sort one of the axis.

Interactive Element: Linked to slicer.

Figure 22: Consistency of Team During the Last 5 Seasons.

Ribbon chart is a great approach to measure the performance of a team during a period of time. As we see above, we can see the consistency of the team from a year to year. We can also other detail to determine whether the performance of a team is dipping in form or it is just an outlier. If we look Everton's point, we can see that their points are lowered but their rank only drop 1 rank. Which are quite impressive, a win match will gain them 3 points, and they drop 12 points which are 4 winning matches, but still managed to drop 1 rank only. This can be said either the leader is having very successful performance or the middle table teams are competing fiercely for their spots.

Best Practice: Sort the X axis in ascending order.

Interactive Element: Tooltips with information and linked to slicers.

Layout Decision: EPL's Theme Color is American Purple. Purple usually symbolizes of power and ambition. I believe EPL suit with the associated color. I also tried to fit in the chart with optimum space between each other. Therefore, there are some space between chart for ease of reading purpose.

Conclusion

As a conclusion, business intelligence decision making can be made easier with a good framework. The use of data warehouse and data visualization can make the decision making more concise and easier. The flexibility of perspective can make the decision flexible but on point according to business needs. It is a really good practice to ensure business make their decision making for future’s result.

Reference

FBref

- https://fbref.com/en/comps/9/stats/Premier-League-Stats#all_stats_standard

Proof of data usage permission:

- https://www.sports-reference.com/data_use.html

Kaggle Part:

- https://www.kaggle.com/saife245/english-premier-league