Project: Effect of on-field performance metrics on goal-scoring potential

Authors: Dong Li, Chee Henn Chin, Jaehyeok Kim, Mackenzie Linden

Research Question

What on-field performance metrics exhibited by AFL players has the strongest correlation with their goal-scoring potential and which model will predict this best?

Target Audience

The aim of this project is to investigate which of the on-field actions performed by AFL players throughout the season correlate strongly with the performance of the players in reference to their number of goals scored. Through a deeper understanding of these connections, this research will be beneficial to multiple groups surrounding the AFL, such as:

- team scouters (allowing for a better comprehension of which players are likely to be great goal scorers, even without knowledge of their goals scored)
- analysts (allowing for interpretation of the weighting of each stat)
- punters (determining who to bet on based off of previous stats)

Through this, the research will not only increase the appreciation for the aspect of data in AFL, but will allow for a greater grasp of the interconnectedness of every aspect of football.

Dataset

Our research was conducted based on a single primary dataset that was provided; a CSV file that listed the statistics of 685 AFL players across 23 different categories as of the 2022 AFL season.

The statistical categories analysed are as follows (totals for the whole season):

- GM - Games played - Games played by a player
- KI - Kicks - Kicks performed by a player
- MK - Marks - Successful marks performed by a player
- HB - Handballs - Handballs performed by a player
- DI - Disposals - Disposals performed by a player (a disposal is legally getting rid of the ball, via a handball or a kick)
- GL - Goals - Goals scored by a player
- BH - Behinds - Behinds scored by a player
- HO - Hit outs - Hit outs performed by a player (uncontested handballs)
- TK - Tackles - Tackles performed by a player

- RB - Rebound 50s - When a player moves the ball from their defensive 50 metre zone and back into the midfield
- IF - Inside 50s - When a player moves the ball from the midfield into their offensive 50 metre zone (only counted once per possession chain)
- CL - Clearances - When a player has the first effective disposal in a possession chain or a kick that clears a stoppage situation (an event such as an out-of-bounds that necessitates a centre bounce, ball-up, or throw-in)
- CG - Clangers - Kicks or handballs that give possession directly to the opposition
- FF - Free kicks for - When a player is granted a free kick by umpires after being interfered with
- FA - Free kicks against - When a player causes an infringement that results in a free kick being awarded to the opposition
- BR - Brownlow votes - Number of Brownlow votes received by a player throughout the season (The Brownlow medal is an award given to the "best and fairest" player, and the votes are given by the three umpires at the end of a game: 3 votes, 2 votes, and 1 vote respectively for the umpire's top 3)
- CP - Contested possessions - Possessions won by a player when the ball is in dispute (in competition with opposition players for the bal)
- UP - Uncontested possessions - Possessions gained by a player while under no physical pressure (usually from a teammate's disposal or an opposition player's clanger)
- CM - Contested marks - When a player makes a mark while under physical pressure of an opponent
- MI - Marks inside 50 - Marks made by a player inside their offensive 50 metre zone
- 1% - One percenters - Actions made by a player than benefit their team but are infrequent (such as smothers, blocking passes, etc.)
- BO - Bounces - When a player wins possession in a bounce (when the umpire bounces the ball against the ground at the middle of the field during a stoppage - specifically after a goal is scored or at the start of each quarter
- GA - Goal assist - When a player creates a goal by getting the ball to a teammate via a variety of methods, such as a disposal

As per the research question and aim, these metrics were compared to the goal scoring in order to model and analyse the correlation between them.

Pre-processing and data wrangling

In terms of pre-processing and wrangling to suit the goals for our research topic, we needed to be able to work on applying numerical operations onto our dataset. As the dataset was a semi-structured csv file it was smoothly able to be converted into a dataframe. Following this, with our intentions to work with the dataset to find correlations through these operations, we decided to remove the player names and player teams to keep the data strictly numerical. Some of the stats of the players were null since that player had not fulfilled the requirements to increment that stat (e.g. A player who had never scored a goal would have a null data in their associated stat for ‘GL’). As we needed to apply correlation functions onto our dataset, we could not work with null data, thus we replaced all null space with zeros in our csv file.

Methods

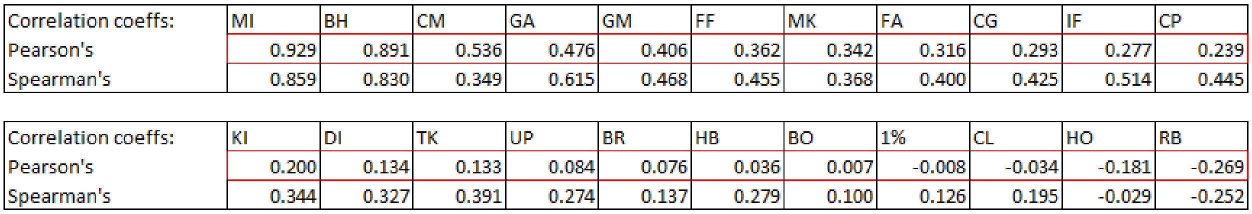
We based our modelling and evaluation primarily around k-nearest neighbours, linear regression and regression trees. Due to our aim in the project we placed heavy priority on the models having high accuracy; where false positives are more valuable than false negatives.

Preliminary Analysis

Before modelling and analysing the data, we determined that we need to be able to understand the basic correlations between each statistic with the goals scored. Through this, we would then be able to select which statistics would be suitable for a model based on the dataset.

Selection of stats

Through formatting our data in this way, we were able to produce Pearson correlation graphs **(see Fig. 1-23)** and coefficients as well as spearman's coefficients for each statistic in relation to ‘Goals scored’. Both of these correlation coefficient values produced a various range of values and for our research to encapsulate enough data we set a threshold value of 0.4, in which any statistics that had a pearson’s coefficient value x in which x > 0.4 or x < -0.4 would be included in the modelling our linear regression model. These choices were also generally supported by their spearman’s coefficient; a correlation coefficient that is able to correlate non-linear data as well. This choice of performance metrics would also allow us to filter out the performance metrics that would most likely be more heavily tied to non-goal scoring properties; e.g ones exhibited by defenders.



Correlation coeffs:	MI	BH	CM	GA	GM	FF	MK	FA	CG	IF	CP
Pearson's	0.929	0.891	0.536	0.476	0.406	0.362	0.342	0.316	0.293	0.277	0.239
Spearman's	0.859	0.830	0.349	0.615	0.468	0.455	0.368	0.400	0.425	0.514	0.445

Correlation coeffs:	KI	DI	TK	UP	BR	HB	BO	1%	CL	HO	RB
Pearson's	0.200	0.134	0.133	0.084	0.076	0.036	0.007	-0.008	-0.034	-0.181	-0.269
Spearman's	0.344	0.327	0.391	0.274	0.137	0.279	0.100	0.126	0.195	-0.029	-0.252

Figure 24 - Correlation coefficient values

KNN

First, KNN was selected to train the model to be able to predict the outcome between the 5 stats chosen and the goals. We evaluated the model using 5-fold cross-validation but it yielded a very poor 26% accuracy calculation. We also evaluated the same model using bootstrap cross-validation, however that also yielded an accuracy of 28%. Due to the accuracy being so low, it was evident that this model would not yield good predictions as it was not performing to a standard that would be able to suit this particular topic.

Linear regression

We then switched to a linear regression model since most of our data had a generally linear relationship with the goals scored (as seen in figures 1-5). We applied this model onto our dataset through supervised learning and by calculating the MSE (mean squared error) of the model, the model performed with an MSE=15.1 and an accuracy of 88.6%. This was a much higher accuracy and an overall far better performance than KNN was making this potentially a good model.

Regression tree

Following this, we created a regression tree using our data (Fig. 25**)** which in turn is able to categorise result data (in this case the number of goal scores) through a pathway of nodes and leaves (the other relevant correlated data). The regression tree had an MSE in the range of 13 to 14, overall outperforming linear regression in terms of MSE score.

Key results and discussion

We begin the analysis of our result by going over the notable statistics and the relevance of their correlation to the goal scoring ability of the players. The main focus being on the Pearson correlation values. Overall the models gave a positive correlation to most statistical categories, implying that more on-field actions in general leads to more goals scored.

The two statistics with the highest correlation are MI (Marks Inside 50)(Fig. 20) and BH (Behinds)(Fig. 7). However, these two are not very statistically relevant to our research question, as they do not represent any interesting data. MI and BH are directly related to the process of scoring a goal, such that an MI leads to a goal scoring opportunity, and a BH is an alternative outcome of such an opportunity to a goal. Which made the high correlation between goal scoring and MI and BH to be expected. FF and FA (Free Kicks For/Against) are also not very relevant either as they are generally unintentional events outside of a player’s control.

On the other hand, some other statistics prove to be much more relevant and interesting. For example, CM and CP (Contested Marks/Possessions) represent contested actions, and are both relatively strongly correlated with goal scoring ability. Most notably they are significantly more correlated with goals than their general or uncontested counterparts MK, UP and HO(refer to **Dataset** for abbreviation), which implies that more aggressive actions further up the field leads to more goals scored. This could tie into explain how CG (Clangers) has a relatively high correlation to goal scoring ability. CGs are inherently mistakes by players, and we assumed that would be detrimental to a player’s goal scoring ability. However, since goal scoring players are likely to be under pressure more often, this is a probable cause for the correlation of CGs and

goals. Interestingly, KI (Kicks) is not that strongly correlated with goals scored. This implies that goal scoring players aren’t generally kicking the ball the most. This could be explained by the positions of the goal scorers, which are likely to be mostly forwards, and thus don’t spend as much time needing to move the ball long distances across the field compared to a midfielder or defender.
Unfortunately, player positions aren’t a part of our data set so we cannot make any solid assumptions beyond speculation regarding position relevance. Three out of four of our negatively correlated statistics (1%, CL, RB) are defensive actions, which logically follows our previous assumption regarding offensive/aggressive actions pertaining to goal scoring. The most unexpected outcome of our analysis was BR (Brownlow Votes), which we found to have a very minor positive correlation with goals scored, when we expected the correlation to be relatively high. The methodology of the Brownlow voting system could explain this. The votes are awarded by the three umpires at the end of the game. If we assume that umpires pay most of their attention to players handling the ball, it is likely that there would be a bias towards midfielders since they spend the most time around the ball compared to other positions, as they are involved in both the defence and offence.

When applying our models onto the dataset, the models only work successfully under the conditions that we created where nonlinear and irrelevant data is discarded. Our models performed incredibly well when we were able to determine which data should be incorporated and which should be outright rejected. Our accuracy was considerably high at 88.6% and the model was performing strongly. Because of this, we can further support our claims that the selected statistics have the strongest and an explainable correlation to the amount of goals scored.

Evaluation

In terms of the dataset, a future investigation would benefit greatly from knowing the positions of each player. Through this, we can then get better correlation data between each position, as those who are in a more offensive position would definitely be more likely to score goals. Furthermore, there are certain correlation graphs that were created that indicated different relationships were present that were more than just linear - a few exponential and one hyperbolic. Knowing these can still contribute to our data set and should be accommodated for in future research. Finally, a lot of the correlation between some statistics and the goals scored were very explanatory, however some statistics were surprisingly found to have a lesser correlation than expected. This could be addressed using other models to determine if the models that had been used were a cause of this.

Appendix:

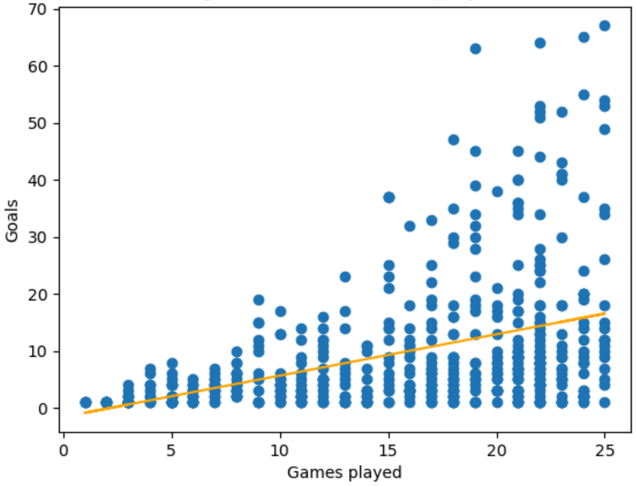


Figure 1: Goals vs Games played

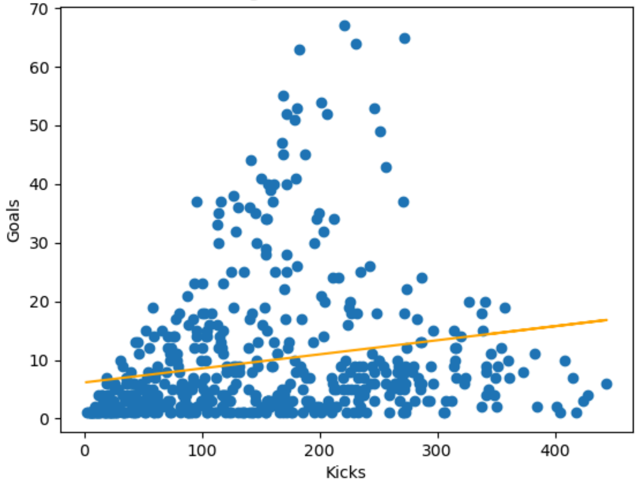


Figure 2: Goals vs Kicks

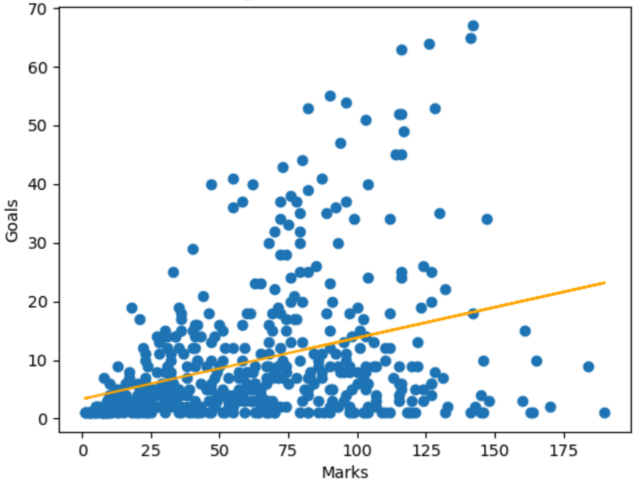


Figure 3: Goals vs Marks

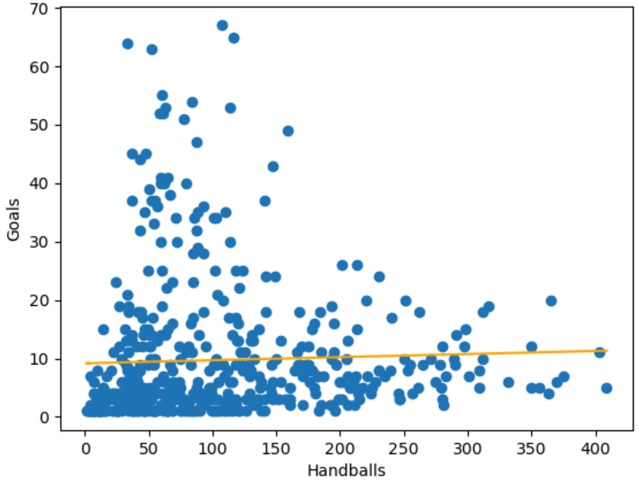


Figure 4: Goals vs Handballs

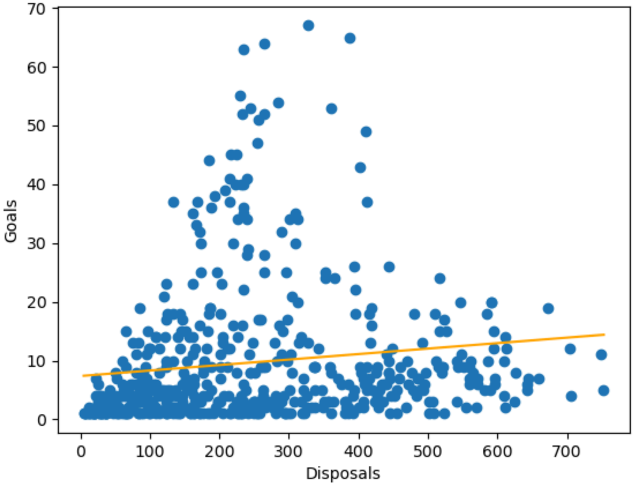


Figure 5: Goals vs Disposals

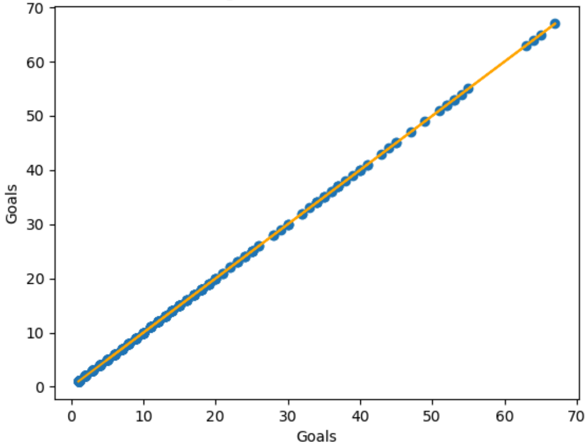


Figure 6: Goals vs Goals

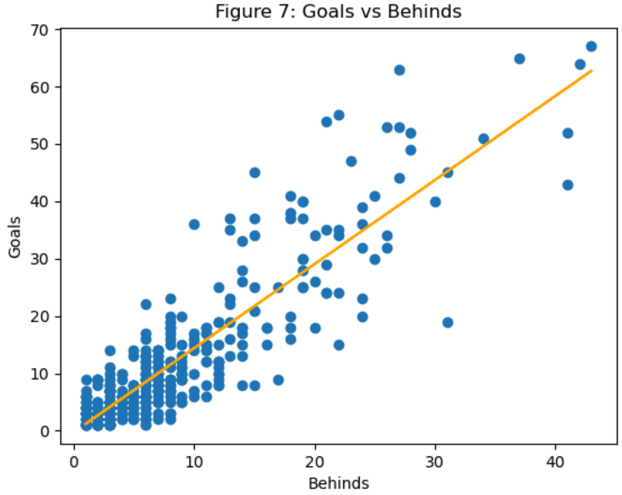


Figure 7: Goals vs Behinds

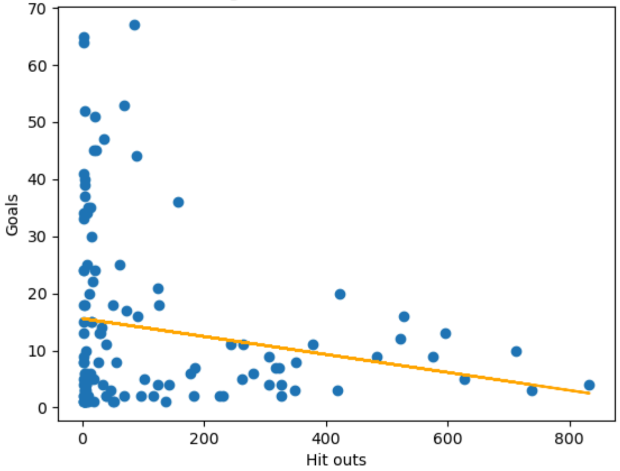


Figure 8: Goals vs Hit outs

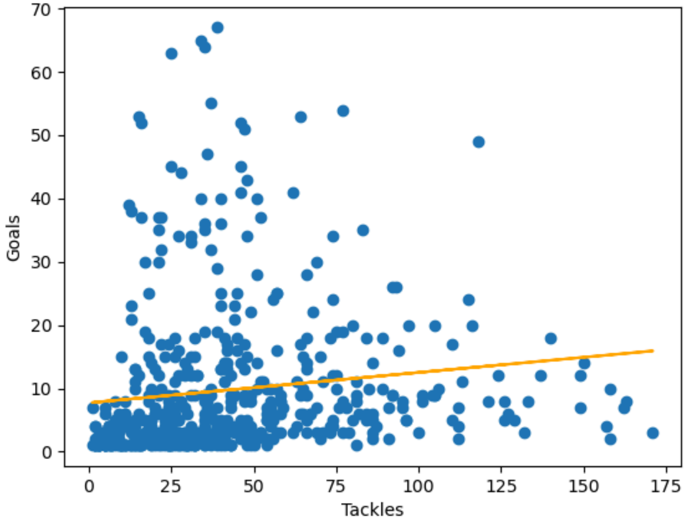


Figure 9: Goals vs Tackles

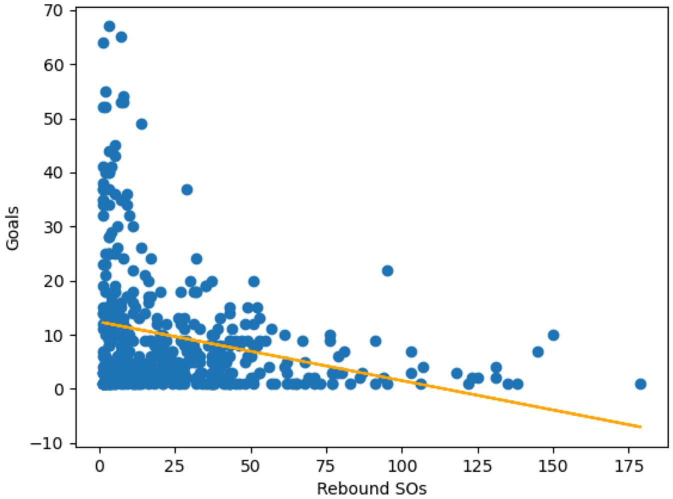


Figure 10: Goals vs Rebound SOs

Figure 11: Goals vs Inside SOs

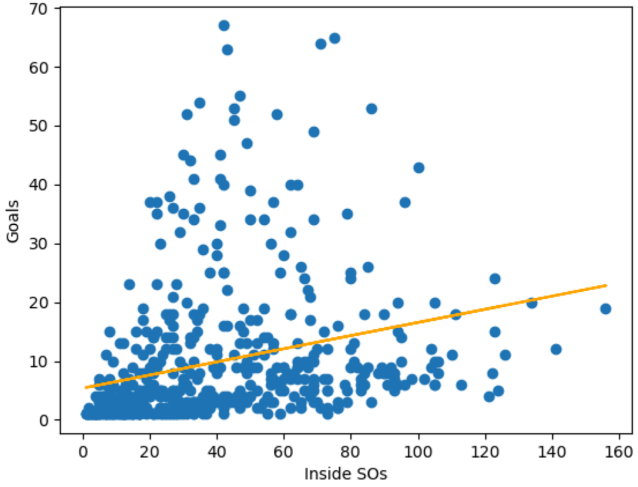


Figure 12: Goals vs Clearances

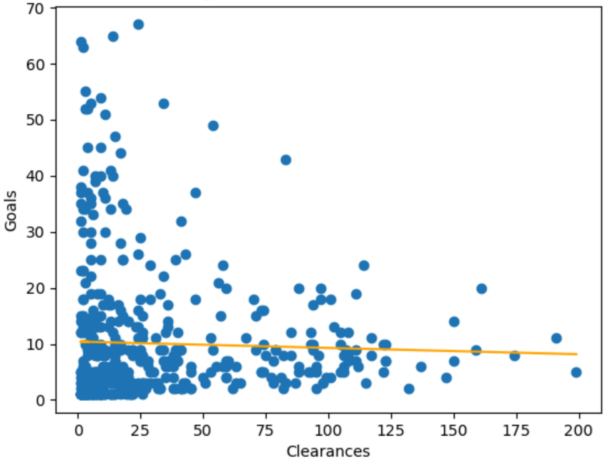


Figure 13: Goals vs Clangers

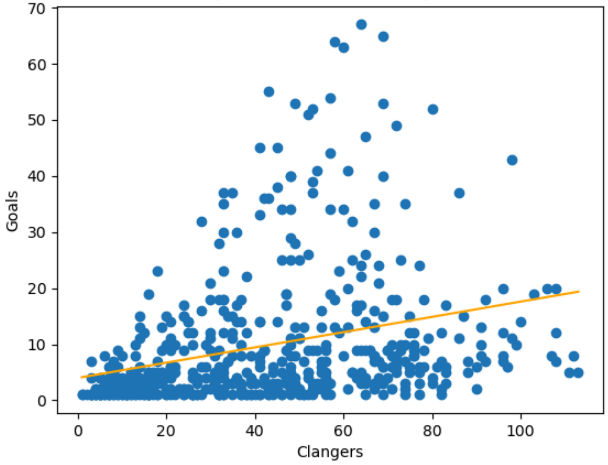


Figure 14: Goals vs Free kicks for

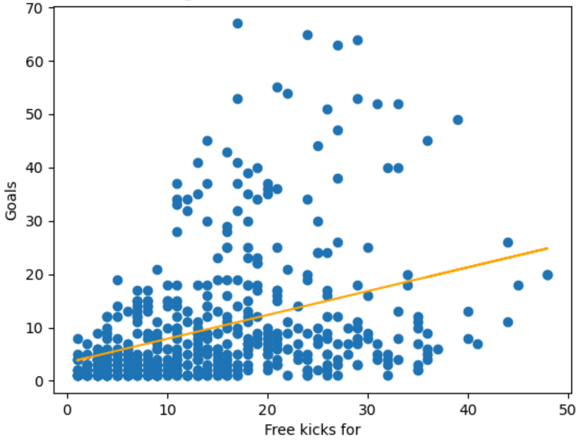


Figure 15: Goals vs Free kicks against

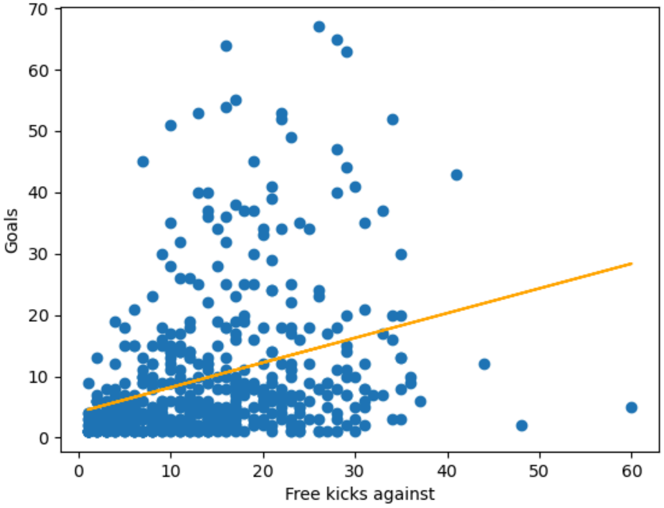


Figure 16: Goals vs Brownlow votes

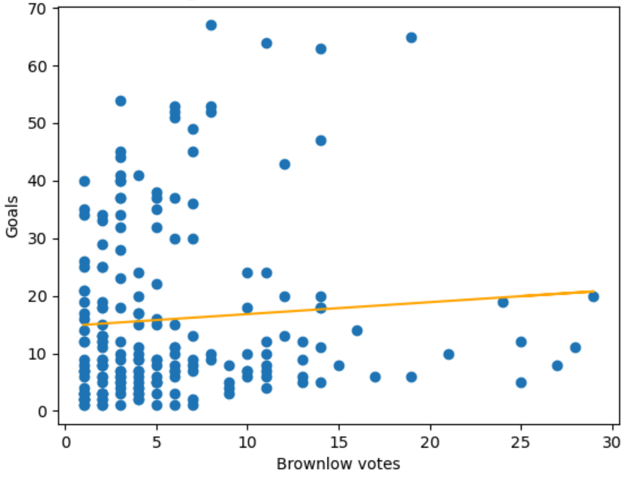


Figure 17: Goals vs Contested possessions

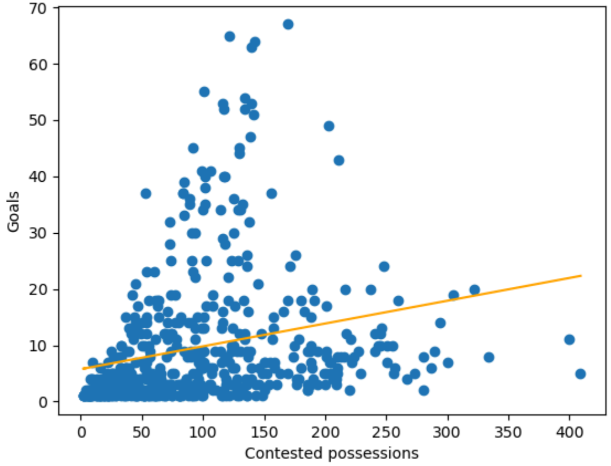


Figure 18: Goals vs Uncontested possessions

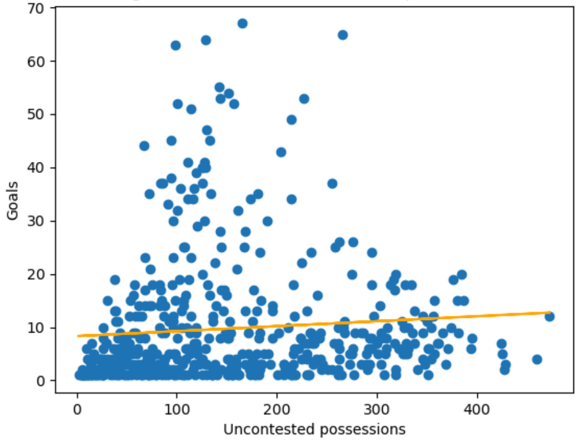


Figure 19: Goals vs Contested marks

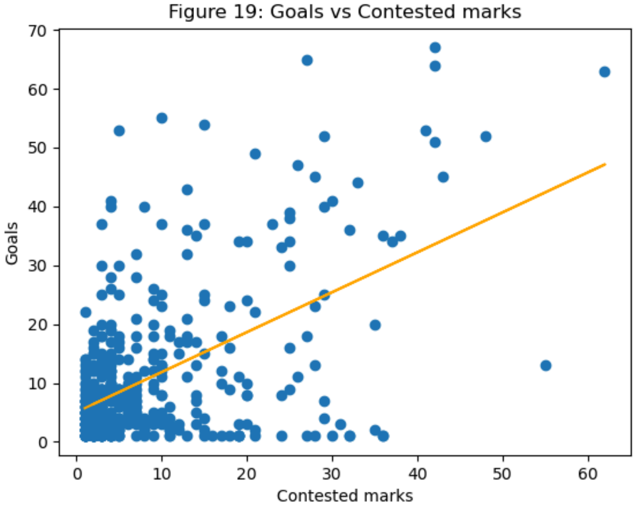


Figure 20: Goals vs Marks inside SO

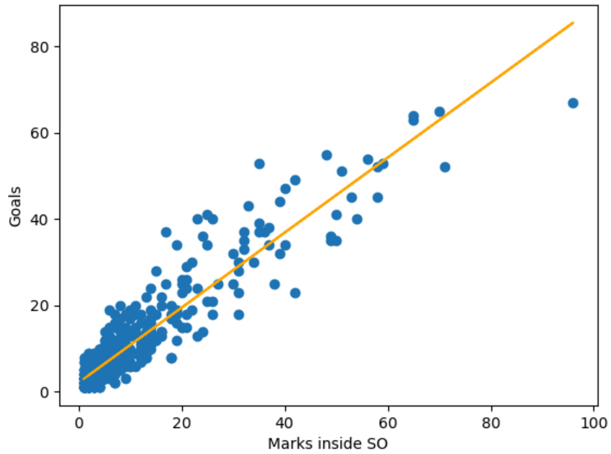


Figure 21: Goals vs One percenters

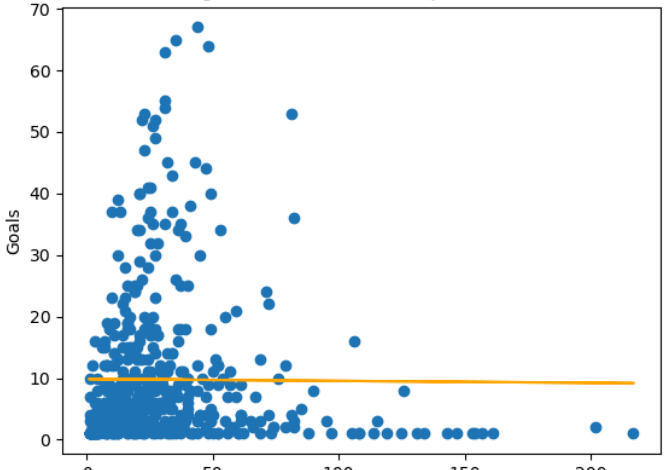
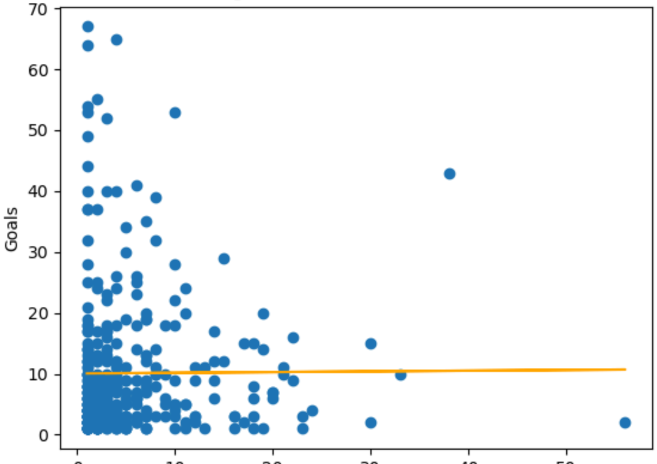
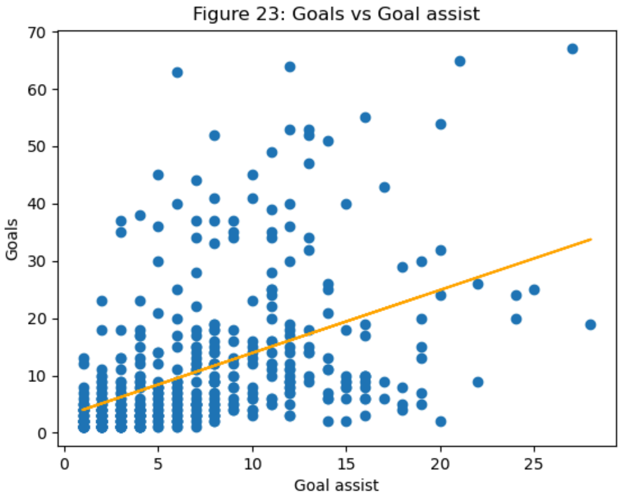


Figure 22: Goals vs Bounces





References:

afltables.com. (n.d.). *AFL Tables - Player Statistics - 2022*. [online] Available at: https://afltables.com/afl/stats/2022a.html.

www.championdata.com. (n.d.). *Glossary – AFL – Champion Data*. [online] Available at: https://www.championdata.com/glossary/afl/.