Executive Summary

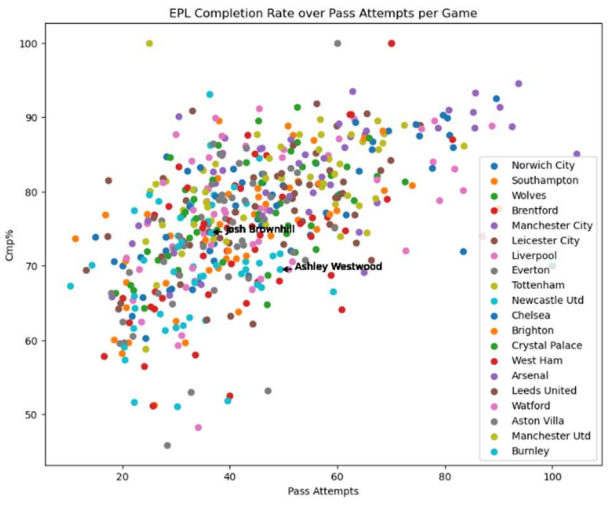
Abstract:

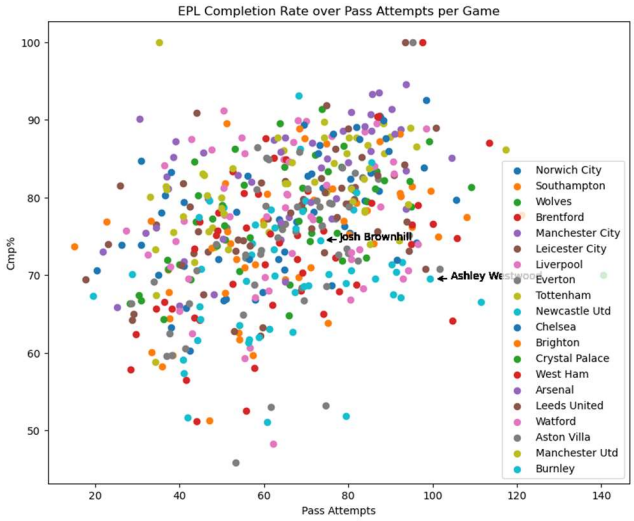
This project takes 2021-2022 English Premier League player-by-player passing data and showcases ways to better analyze the overall data by breaking it into smaller pieces. The main ways highlighted include breakdowns by team, by position, and by type of pass, each offering distinct conclusions and applications. The project mainly interacts with the class through its emphasis on data context and the importance of using data responsibly, paying attention to the dataset as a whole so that one can better understand how to break down and extract information from the data. It also touches on the Simpson paradox, in that insights can be found from smaller subsections of the data which otherwise wouldn't be able to be found in the dataset as a whole.

Teams:

One split looked at in this project was teams – specifically how team performance can obscure talented players on lower teams. To the right is a list of all the teams in the Premier League, and their average passes attempted per player per game. This list roughly mirrors the standings of the Premier League, and to a soccer fan, this phenomenon is fairly intuitive: better teams tend to have more possession of the ball in games, for a variety of reasons, but the underlying result as it relates to this dataset is that players on better teams have more opportunities to pass, simply because they have the ball more often. That's showcased on the poster by comparing two players on Burnley, at the bottom of this list, to the Premier League as a whole. On their team, they stand out as two of the most active passers, whereas in the league as a whole, they're obscured by the masses. One such way of correcting for this, as seen below, is by correcting the amount of passes attempted. In the graphs below, the left graph uses the original data. The right graph, meanwhile, corrects passing attempts by setting the common denominator for all teams to be the same as Manchester City's by simple multiplication. By looking at the two highlighted players, you can get a sense of how the graph shifts and changes because of this.

Team	Average Attempted Passes Per Player Per Game	Max Avg Attempts per Game
Manchester City	65.22	104.44
Chelsea	59.38	89.59
Liverpool	58.78	88.71
Tottenham	49.81	72.39
Leicester City	48.85	75.43
Brighton	48.45	90.00
Arsenal	46.90	67.39
West Ham	46.75	81.35
Wolves	46.45	100.00
Manchester Utd	46.37	83.46
Leeds United	43.92	66.25
Crystal Palace	43.32	72.56
Aston Villa	41.06	60.00
Brentford	37.89	61.44
Norwich City	37.26	60.34
Southampton	36.11	59.86
Watford	35.75	51.60
Everton	34.66	54.00
Newcastle Utd	34.63	59.17
Burnley	32.41	49.37





Positions:

While the dataset provides positions for players, the original dataset will occasionally list players as "MF, FW" or "FW, MF" or similar, to denote that the player has multiple positions across the season. In addition, how exactly a position is played often depends on game states, team tactics, formations, and other such schematics. To get clearer results when defining players by position, I decided to use k-means clustering, which split the dataset into four groups. As it turned out, most positions are played fairly similarly to each other, with 39 of 41 goalkeepers being clustered into group 0 (95%) and 108 of 111 goalkeepers being clustered into group 1 (97%). Similarly, most defenders were slotted into group 2, and midfielders were split between groups 1, 2, and 3. Again, this makes sense to someone who knows soccer – while strikers will solely attack, defenders are more nuanced, with many teams utilizing wingbacks who are listed as defenders but often times move forward to pitch into the attack. Meanwhile, midfielders are jacks of all trades, and while there are pure midfielders, they'll oftentimes be most defined by their prefixes, "defensive" or "attacking," in that they'll act very defensive or attack-minded, which explains their fairly even split across the board.

Pass Types:

By splitting players into various positions, we're then able to get much more interesting information out of preliminary graphics such as the ones below. For example, the Final Third graphic, while intuitive, in that midfielders will generally have the most chances to make progressive passes, tells us that in the Premier League strikers tend to sit forward more often, acting as target forwards to receive passes, instead of dropping into the midfield, as a sort of "false nine" provisional midfielder. Meanwhile, we can look to graphics on the distance of the pass, such as the long passes graphic, to find more unintuitive results. While the goalkeeper spread makes sense, in that goalkeeper punts and goal kicks are often not completed, the graph shows midfielders as the next group with the largest amount of long balls, instead of defenders, as might be expected. It does, however, further support the observation on target forwards from the key passes graph, as it implies that midfielders are often playing long balls over the top and into the corners, for strikers and wingers to run on to.