Sports: Flexible Schedule | Exploratory Analysis

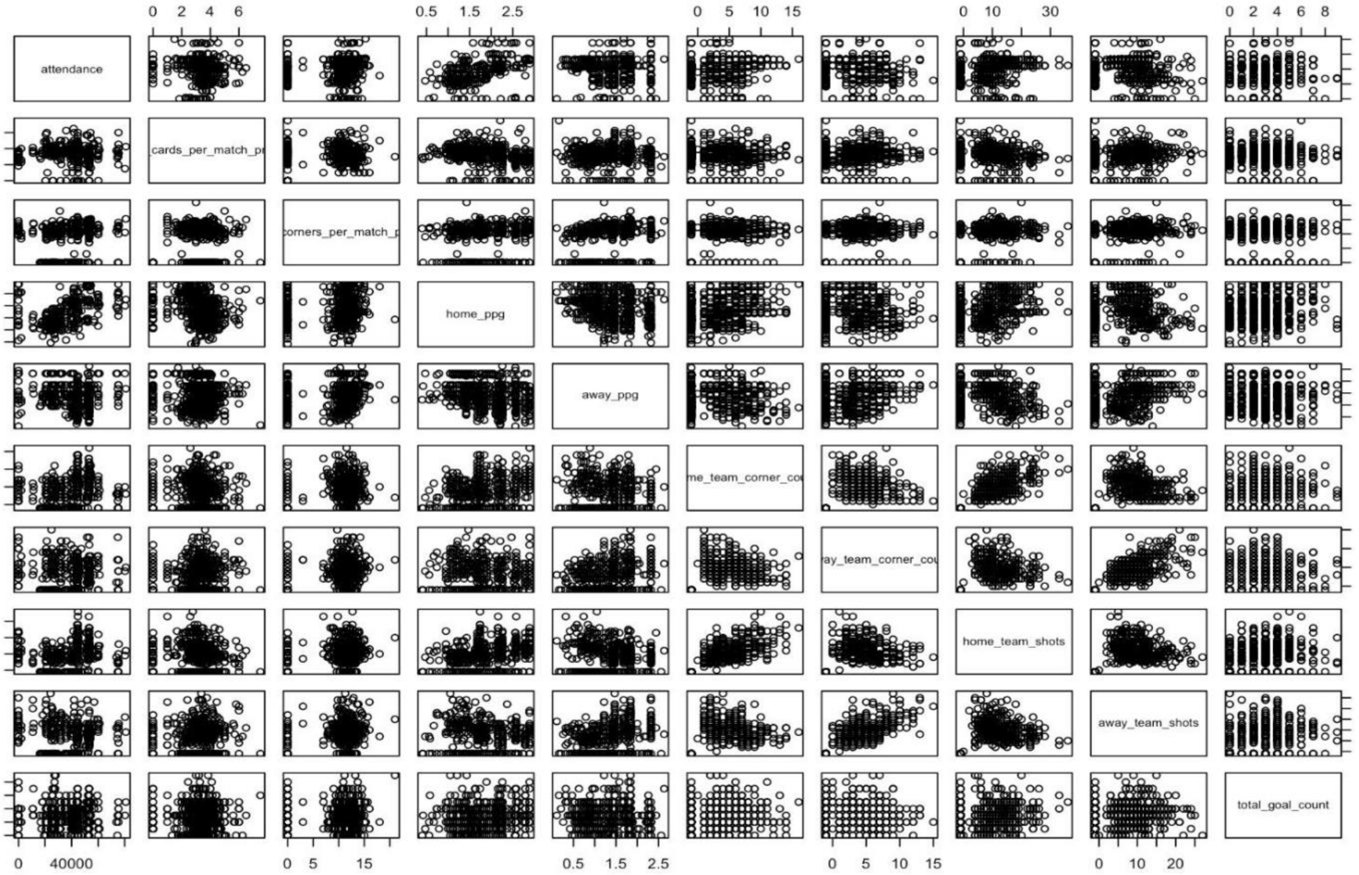
Approached the data set to analyse and describe the different variables that will be using to construct the model. Utilized histograms, five number summaries for my univariate analysis methods and scatter matrices, box plots, correlation matrices, and correlation tests for bivariate analysis. Discovered interesting insights to use down the road for building the model.

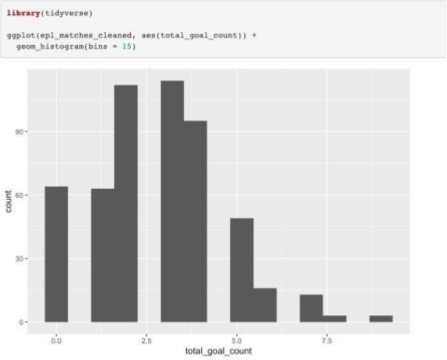
One insight found that I thought was interesting was in regards to the attendance count's effect on team form. Using ppg, which acts as a moving average for table points achieved per game, as a variable for determining team form the strongest team having a value close to 3 and the weakest having a value close to 0. While this had little effect for away teams, when teams were playing at home the attendance showed to be one of my stronger performance predictors ~0.41.

Explored whether a categorical variable had an effect on my anticipated DV. A boxplot with respect to what the summary of total_goal_count by referee is included. The referee assignment does not appear to have a huge effect on total goals scored, however specific referees do have interesting things that caught my eye. Martin Atkinson's sub 2.5 IQR, with a mean of 2.5 may indicate a lack of high scoring results when he is the referee. Andre Marriner seems to have the opposite IQR flip, perhaps indicating a higher scoring trend for games in which he is the ref.

While not a visualization of a specific variable, I also ran a cor() function across all of the numerical variables to see how each of them currently correlates to total_goal_count. That column of that function call is included and will be used to identify IVs in the data set that I will highlight and begin to work with in my next steps. I found it interesting that there was ultimately not a large amount of correlation between these variables, with the highest value being 0.689 for home_team_goal_count (which one might expect would be a much larger predictor towards total_goal count). Another variable in the correlation analysis that was interesting was home_shots_on_target, with a correlation coefficient of 0.1572. I expect this variable to serve my model well in the future. This variable and home_team_corner_count was expected to be highly correlated with goals scored and it is surprising to us that the correlation appears to be pretty weak.

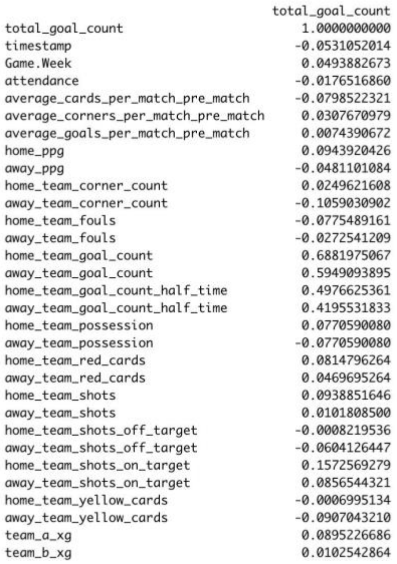
Visualizations inclusive of items mentioned above are below:



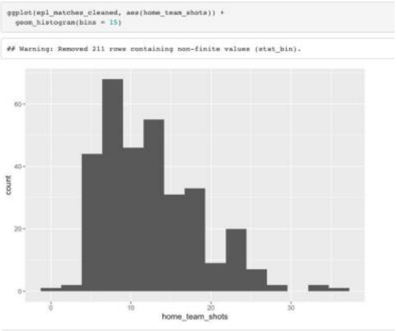


```
library(tidyverse)

ggplot(epl_matches_cleaned, aes(total_goal_count)) +
  geom_histogram(bins = 15)
```

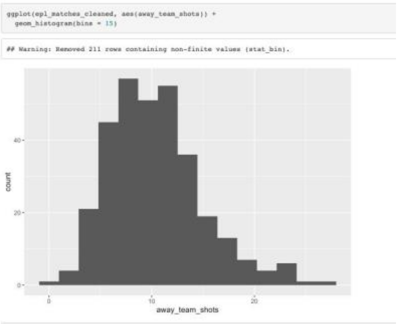


```
                                       total_goal_count
total_goal_count                           1.0000000000
timestamp                                 -0.0531052014
Game.Week                                  0.0493882673
attendance                                -0.0176516860
average_cards_per_match_pre_match         -0.0798522321
average_corners_per_match_pre_match        0.0307670979
average_goals_per_match_pre_match          0.0074390672
home_ppg                                   0.0943920426
away_ppg                                  -0.0481101084
home_team_corner_count                     0.0249621608
away_team_corner_count                    -0.1059030902
home_team_fouls                           -0.0775489161
away_team_fouls                           -0.0272541209
home_team_goal_count                       0.6881975067
away_team_goal_count                       0.5949093895
home_team_goal_count_half_time             0.4976625361
away_team_goal_count_half_time             0.4195531833
home_team_possession                       0.0770590080
away_team_possession                      -0.0770590080
home_team_red_cards                        0.0814796264
away_team_red_cards                        0.0469695264
home_team_shots                            0.0938851646
away_team_shots                            0.0101808500
home_team_shots_off_target                -0.0008219536
away_team_shots_off_target                -0.0604126447
home_team_shots_on_target                  0.1572569279
away_team_shots_on_target                  0.0856544321
home_team_yellow_cards                    -0.0006995134
away_team_yellow_cards                    -0.0907043210
team_a_xg                                  0.0895226686
team_b_xg                                  0.0102542864
```



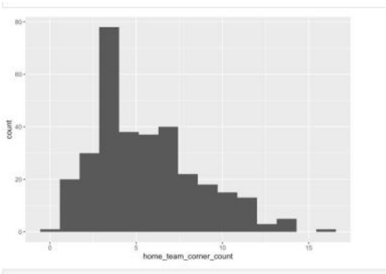
```r
ggplot(epl_matches_cleaned, aes(home_team_shots)) +
  geom_histogram(bins = 15)
```

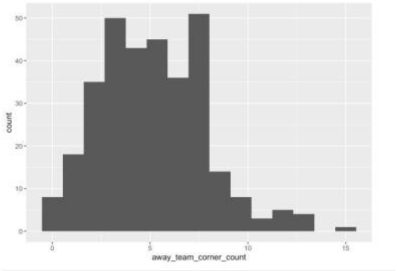
```
## Warning: Removed 211 rows containing non-finite values (stat_bin).
```

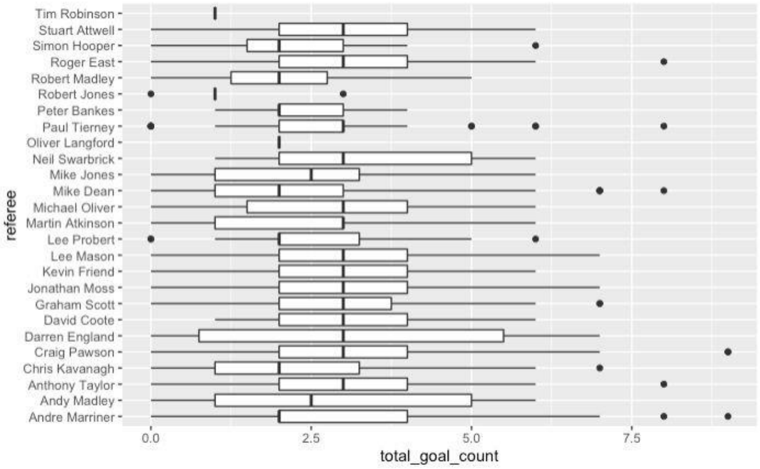


```r
ggplot(epl_matches_cleaned, aes(away_team_shots)) +
  geom_histogram(bins = 15)
```

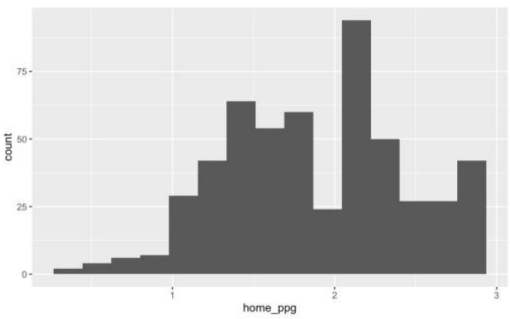
```
## Warning: Removed 211 rows containing non-finite values (stat_bin).
```

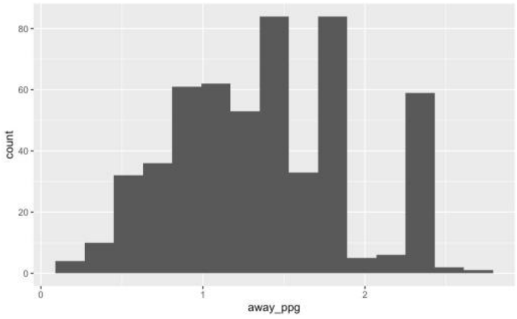






Possible new DVs:





	Attendance	Avg. Cards/Match	Avg Corners/ Match	Aveg Goals/ Match	Total Goal count	Average Cards Per Match
Min	0	0	0.0	0.0	0.0	0.0
1st Quartile	30,147	2.808	0.0	2.4	2.0	2.808
Median	44,115	3.335	10.005	2.75	3.0	3.335
3rd Quartile	51,253	3.830	12.0	3.083	4.0	3.830
Max	80,827	7.50	21.0	6.0	9.0	7.50

Based on the five number summary of the six variables selected above, I can infer that the size of the attendance and goals scored per match, whether by home team or away team, has a positive correlation. The total goal count has increased as the crowd size increases. Based on this observation, I believe that as the crowd size increases per match, so does the number of goals scored by both teams per match. There may be other variables that impact the number of goals scored by the teams, such as time of possession and number of fouls. For the experiment, I will use multiple regression that includes these six variables as well as time of possession and number of fouls to model goals scored. And compare if time of possession and fouls have any correlation on the goals scored. I also want to look at whether being home and away has any advantage, so I will also look at the goals scored by home team and away team in correlation with the crown size.

Insights and Future Experimentation

I am still hoping that the data I have can yield results for predicting one variable based on several other variables, but the preliminary analyses are revealing a couple of challenges for that process. I am seeing very low correlation and may need to revisit the processing steps and perhaps move in a different direction. In terms of specific insights from the visualization I have run, I am seeing a slight correlation between shots and goal count, cards_per_match_prematch and goal count, which is seen in the scatterplot matrix and will form the basis for my regression analysis moving forward. My overall insight appears to be "there is limited correlation towards" total_goal_count in the current form of the dataset, which is not very useful at the moment. This is all leading us to looking for other things I may want to predict and opening up new analysis. The biggest insight I came to was that home_ppg and away_ppg seemed to have enough of a difference in value, that it might be possible to run regression analysis on the full data set (not limited to Liverpool) to try to determine what variables were contributing to that difference. I were seeing that attendance, home and away corner counts, home and away shots were all correlating to home and away ppg in that scatterplot matrix. This was the seed for that possible new path forward.

I were slightly thrown by the apparent lack of correlation between the data I have and total_goal_count, so are looking forward towards other useful analysis/experimentation. One possible item I will build towards is an experiment on the full data set to answer the questions: "Does playing a match at your home field actually give you an advantage? If so, which variables account for this advantage?" This came up as I saw trends in home_ppg ("points per game") vs. away_ppg and wondered what might contribute to the difference in the shape of those data. Home seemed to be slightly higher than away, so I will build regression models to show "why." I am also still interested in modeling total_goal_count using multiple regression to fully exhaust that as a possibility. The final experiment I may pursue is an analysis of the categorical variable of referees involving regression with several dummy variables in the hopes of showcasing a predictive pattern on total_goal_count and possibly fouls or cards issued based on who is the referee of a match.