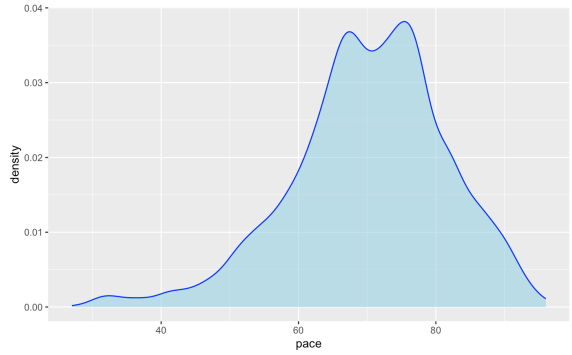
Group S Project Summary
Sarah Fatihi, Sophie Diop, James Street

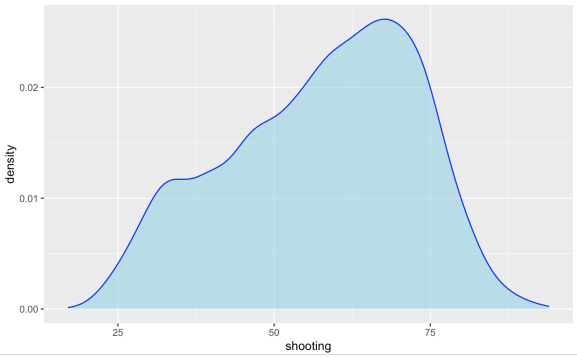
Project Aim: We want to study FIFA Stats in comparison to live soccer data. Our goal is to predict transfer market value from FIFA stats. Are FIFA stats accurate in predicting transfer market value? What are the best and worst predictors of transfer market value? FIFA uses six core predictors in determining a soccer players ability, are these predictors the best at determining transfer market value?

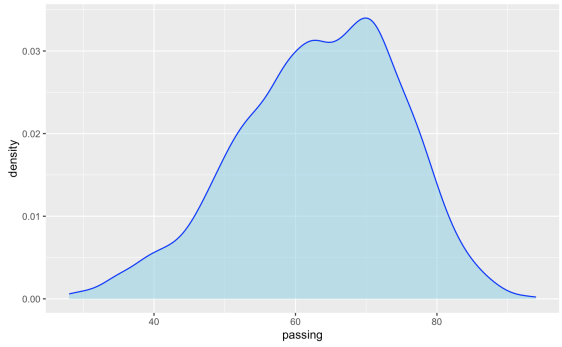
Data Source: The FIFA Football Players dataset, https://www.kaggle.com/datasets/rehandl23/fifa-24-player-stats-dataset, is a comprehensive collection of information about football (soccer) players from around the world. This dataset offers a wealth of attributes related to each player, making it a valuable resource for various analyses and insights into the realm of football, both for gaming enthusiasts and real-world sports enthusiasts. The data is based on the game FIFA 24 which contains information on soccer players from over 19,000 fully licensed players, 700 teams, and 30 leagues in the year 2023-2024.We are planning on slimming down this data to only include players in the premier league. Our financial dataset which contains information of players transfer value was scraped from: "https://www.transfermarkt.co.uk/premier-league/transfers/wettbewerb/GB1/plus/?saison_id=2023&s_w=&leihe=0&intern=0&intern=1". The site is made for football fans and contains transfer related football stats and history statistics for over 100 men's and women's club and national team competitions. This transfer data was taken from the most recent summer transfer window before the 2023/2024 season, and included all players who were signed by premier league teams within this period for a transfer fee.

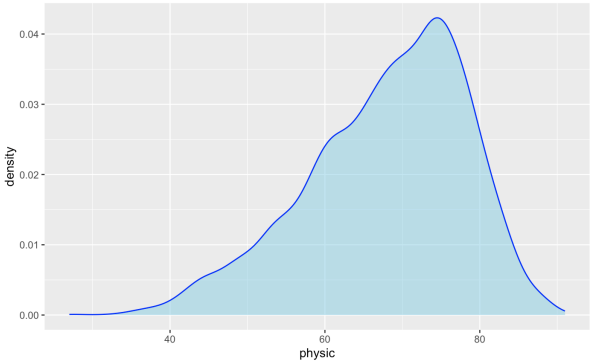
Variables: We have both categorical and quantitative variables as our explanatory variables. These are purchasing club (categorical), in-game Passing (quantitative), in-game Shooting (quantitative), in-game Dribbling (quantitative), in-game Defending (quantitative), in-game Physicality (quantitative), in-game Pace (quantitative). All of our quantitative in-game stats are on a scale of 0-100 determined by the people working at FIFA.

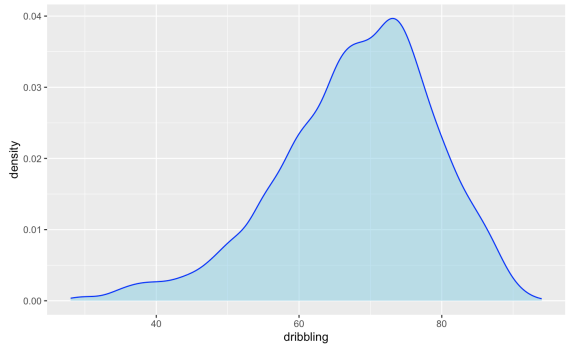
Univariate Analysis: In order to use these variables within our project we first need to examine the data to discover the shape, spread, and distribution of each explanatory variable. Visualizations to assess normality of distributions of our quantitative variables can be created using the R command gf_histogram. This operation allows us to see the distribution of our variable values presented in a histogram. From this point we can analyze if they have a normal distribution, and if not, what we can do to solve this issue. We will then use the favstats command to determine valuable information like mean, median, IQR's, standard deviation and normality. For our one categorical variable (Club), we will use a density plot to see how many players each club signed as well as the average price of those players. Additionally, we will perform a favstats command and a gf_histogram command of our explanatory variable in transfer price. This will allow us to analyze the mean, median, IQR's, standard deviation and normality. Of our variables, we found that most displayed relatively normal distributions with a general slight right skew to every variable. Physicality was the most right-skewed distribution. All of our variables were also unimodal, apart from defending which seems to be bimodal. As a result of the right-skew we would use the median value instead of the mean and the IQR instead of the standard deviation. These values for the Pace variable are a median of 71 and an IQR of 14, Shooting has a median of 59 and an IQR of 22.25, Passing has a median of 64 and an IQR of 16, Dribbling has a median of 69 and an IQR of 14, Defending has a median of 61 and an IQR of 33, Physicality has a median of 70 and an IQR of 25.

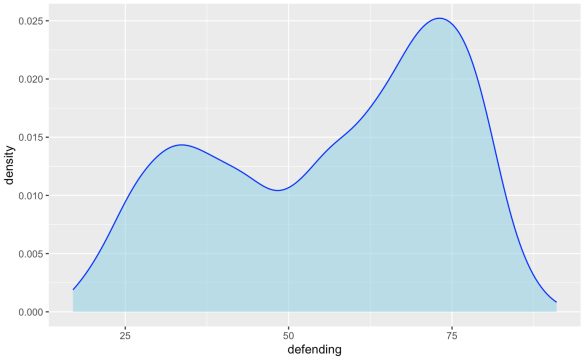


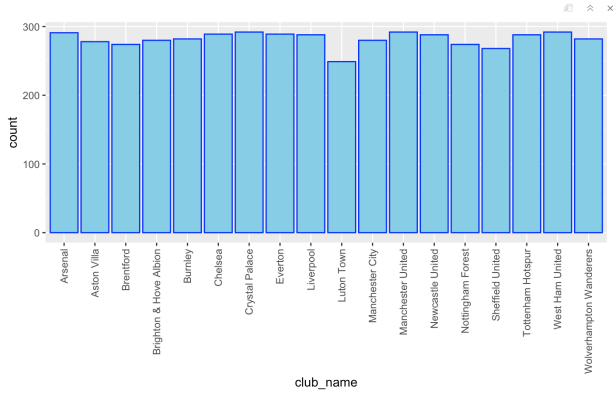






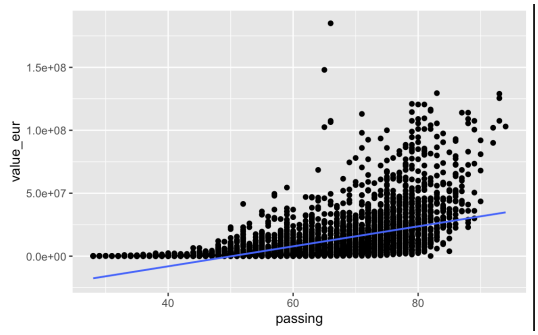


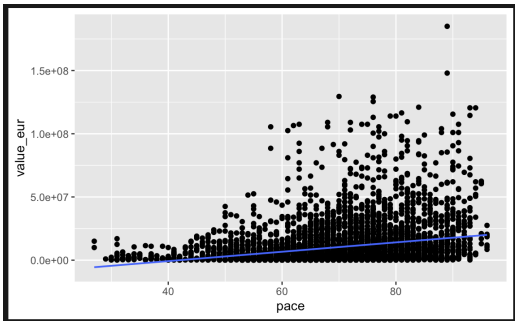


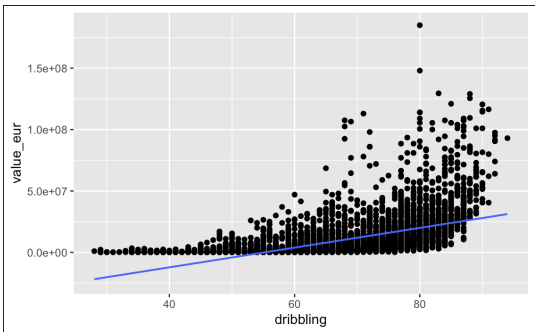


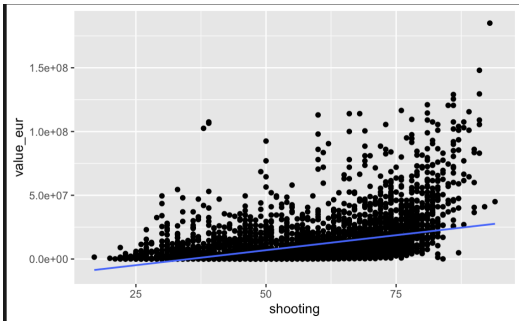
Bivariate Analysis: In this second step, we want to individually compare each of the six explanatory variables to our response variable of Transfer Price. For transfer_price (quantitative) vs our quantitative variables: short_pass, shot_power, dribbling, stand_tackle, and sprint_speed, we plan to use various scatterplots for bivariate analysis. For transfer_price (Quantitative) vs club, which is a qualitative variable,

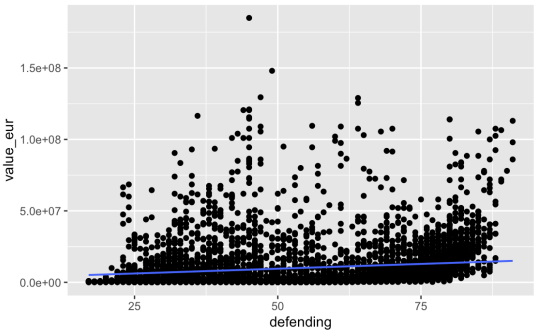
we plan to use a box plot for bivariate analysis. We will also create a GGpairs plot to identify which explanatory variables have the strongest relationship with transfer_price using the provided correlation coefficients and scatter plots. In this plot we are looking for the explanatory variables with clear linear relationships and high correlation coefficients with transfer price. Multicollinearity is also a concern we would like to address, and the correlation coefficients between the explanatory variables are a good starting point. However, we also want to conduct VIFs for all the explanatory variables to get a definitive check on multicollinearity. We will remove at least one of the variables if the VIF test shows a value of at least 5. The variables we choose to keep in our model will be used in multiple linear regression. Here, we would check the conditions of MLR to ensure our model is suitable for analysis. The scatter plots for every explanatory variable with transfer value seem to show weak positive correlation. Passing and Dribbling seem to have the strongest correlation with value with defending having the weakest correlation. None of these regression lines seem to be good fits however, with large residuals, especially on the higher ends of each explanatory variable. Some also appear to show a slight inclination of a possible exponential relationship not well categorized by a straight regression line. Shooting, Dribbling and Passing are the variables which show that a curved line of best fit may be more representative of the relationships. We also notice some correlation between our explanatory variables as shown in our ggpairs visual. Some of our variables are strongly correlated with one another, most notably Passing and Dribbling which have a correlation coefficient of 0.855 and dribbling and shooting which have a correlation coefficient of 0.782. This multicollinearity is something which we have made a note of and will attempt to tackle going forwards in our analysis of our data, and eventual findings.

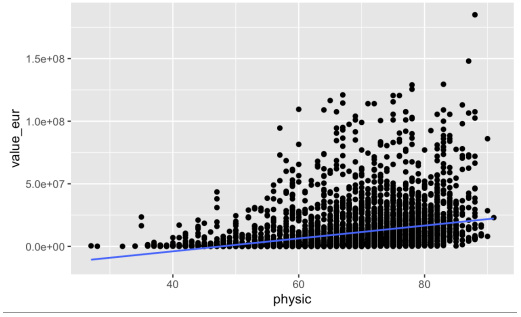


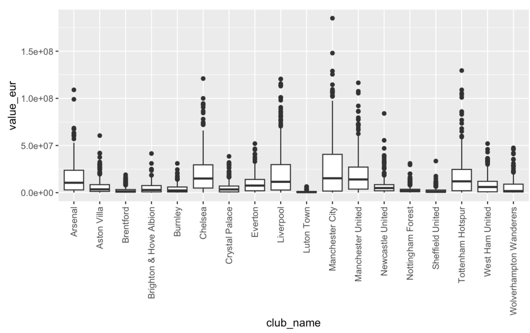


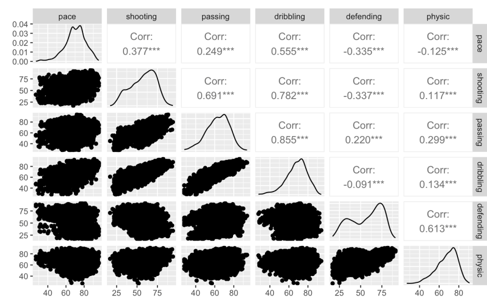




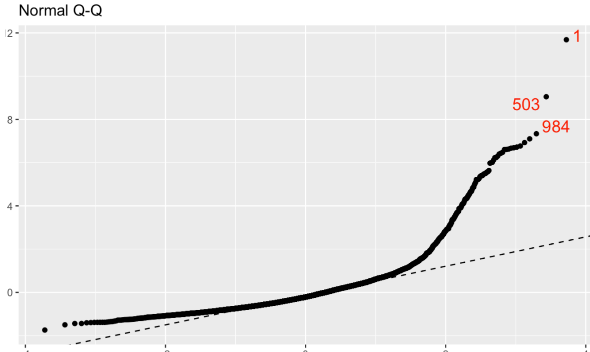


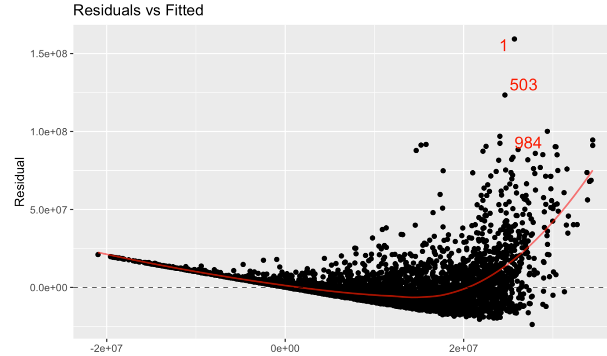






Multiple Linear Regression: The first thing we did was test a kitchen sink model using all six predictors. To ensure that we were using the best or most parsimonious model for predicting the value of the soccer players, we ran a best subsets regression for variable selection. The best subsets method revealed that the kitchen sink model, containing all the predictors, is the model that best predicts the values of our response variable value_eur. This model had the lowest Cp and highest R^2. Yet, when we looked at the VIFs, we were concerned about dribbling and passing as their VIFs were greater than 5 and they were highly correlated with a correlation coefficient 0.855. We then decided to remove dribbling as a predictor as in the best subsets model, passing was chosen as the best predictor to determine the value of the player. After running another best subsets regression for variable selection, the best model included pace, shooting, passing, and physic as the best predictors for value_eur as this model had the lowest Cp and highest R^2. Then, we checked the regression conditions for this data, which included linearity, independent, equal variance, and normal distribution of errors. We used the Normal Q-Q plot and residuals vs fitted plots. The normal Q-Q plot was not linear and the points greatly deviated from the line at one tail. Also, the points in the Residuals vs Fitted Plots did not appear random and there was not equal variance. Therefore, we concluded that the conditions for regression were not met and only the independence condition is satisfied. We then tried to meet the regression conditions for our final model by using log, square root, and power transformation on our response variable, value_eur, and our quantitative predictors, pace, shooting, passing, and physic. We were still unable to meet the conditions for regression when using these transformations, so there is some issue in the data itself. So, we settled on the model including pace, shooting, passing, and physic as the best predictors for value_eur, and we will leave the data as is because simplicity is usually the best default in this situation. The current regression equation is: predicted value_eur = -69518366 + 226714(pace) + 74219(shooting) + 579126(passing) + 339827(physic).






```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -69518366    1907559   -36.4  < 2e-16 ***
pace           226714      18618    12.2  < 2e-16 ***
shooting        74219      18546     4.0  6.4e-05 ***
passing        579126      24391    23.7  < 2e-16 ***
```

	rsq <dbl>	cp <dbl>	pace <chr>	shooting <chr>	passing <chr>	defending <chr>	physic <chr>
1 (1)	0.289853	401.97553			*		
2 (1)	0.315353	207.41497			*		*

Two-Way ANOVA: We were not able to proceed as we did not have two categorical explanatory variables. The categorical variable that we do have overwhelmed R so we are looking into a way we can narrow the clubs down without ruining the integrity of the model.

```
Error in stop_if_high_cardinality(data, columns, cardinality_threshold) :                    Show Traceback
 Column 'club_name' has more levels (47) than the threshold (15) allowed.
 Please remove the column or increase the 'cardinality_threshold' parameter. Increasing the cardinality_threshold may produce long
 processing times
```

Conclusion and Future Analysis: The best subset model we chose consisted of a four predictor model using pace, shooting, passing, and physic. This model produced an r-squared of .342, which means this regression model accounts for 34.2% of the variance in value_eur. Although the model showed significant predictors, the p-values were less than alpha = 0.05, we failed to meet the conditions. We can continue to try and transform variables and try to bootstrap and randomize in order to try and meet the regression conditions. The nature of the data may be such that these transformations may not do much to fix the data and have it meet regression conditions. If this is the case, we will leave the data as is and conclude that the model should not be used.