Capstone Project 1 Milestone Report

Introduction, Problem and Client

The NBA is widely regarded as the best basketball league in the world, boasting the highest level of play from the best players in the world. In the last decade or so, there has been an influx of 'european big men' (centers originating from EU countries) into the NBA. Great passing ability and sound fundaments of the game are aspects commonly attributed to EU centers. This has made drafting players harder as an increasing number are coming from oversees.

The NBA Draft is an annual event during which teams from around the NBA draft young players into their teams. Simply put those players can either be from a college or high school in the US, or international players. Not all players in the NBA were drafted, however the large majority enter the league through the draft. Naturally then, the teams that participate in the draft want to ensure that their player selection is optimal for their team. NBA scouts facilitate in this process and use various methods to evaluate talent.

To make the jobs of NBA coaches and scouts easier, we will create models that will predict seasonal averages for the most common NBA statistics (Points per game, Assists per game etc). We will also compare US and EU centers and find the benefits/downsides of each category (we chose to compare only these 2 groups, as no other geographical group is large enough to make reliable statistical inferences possible). Equipped with this information, NBA teams (including its staff) will make better decisions regarding which players to draft, and what to expect from them as they progress through their careers.

Initial Data Acquisition and Wrangling

We had to scrape data from Basketball Reference (*www.basketball-reference.com*) as the website didn't have an API. To make the comparison as unbiased as possible, we scraped NBA center data from the 1980 season onwards (3pt shot introduced in 1980). Some player overview data as well as the URL of each player could be found on the website corresponding to the first letter of the surname (so the URL containing data about LeBron James could be found at the URL *www.basketball-reference.com/players/j)*. We cycled over each letter (excluding X, as there was no X URL) and scraped the URLs of all NBA centers. Since the websites also included the start of the player's career and their position, we were able to filter out pre 1980 seasons and select only the players at the center (C) position.

Once we attained the URLs of each player, we could access the information on their webpage. The data we extracted was in the first table on the webpage, where each row represented a season of that player's career, and each column a season average statistic such as Points per Game. We stored the scraped data in a dataframe. To that dataframe we also appended extra information that might be useful later such as the Weight, Height, 'Year in League' and Draft Rank (if a player was undrafted, we assigned a -1 label as a signpost for later wrangling steps) as columns. Finally, we determined the origin of the player (US or EU) and added it as a column also. We did this by scraping the birth place of the player and running it by a list of EU countries and US states to determine which group to place the player into. If the player was not from an EU country, or a US state, we labelled him as 'Neither'. We did this for all players and append all the dataframes to one 'master' dataframe where each row is a season played by an NBA center from the US or EU, and each column is a feature. We exported the dataframe to a .csv file.

Wrangling the ‘Player Data’ Dataframe

We read in the dataframe that we exported to a .csv file in the previous section, replaced the -1 values as NaN and corrected the data types of the columns in the dataframe. There were also multiple rows with the ‘Neither’ label in the ‘US or EU’ column, some of these were really EU entries, but misclassified due to differences in spelling. We considered players originating in Canada as part of the US category as fundamentally we want to infer differences in the upbringing of the players, and not the geographical regions (also there was a substantial amount of data from Canadian born players).

Having fixed the misclassified rows, we removed players who were not from the US or EU. This was justifiable, as the number of such rows was very small in comparison to the whole dataframe. There were instances where almost entire rows of data were null values. These corresponded to cases where players did not play during that season (injury or playing overseas are a few examples for the reasons why). Again, the number of null rows was not substantially large, and by the same logic as applied above, we removed these rows.

Finally, we dropped a few low information columns (3 Point %, Free Throw %), reset the index of the dataframe, and converted the ‘%’ columns from ratios to percentages. It’s important to note that we did not deal with un-drafed player data yet (rows where ‘Draft Rank’ column values are NaN) as this group, unlike the ones before, makes up a sizeable amount of data (around 10%) and represents a legitimate group of players. However, since we don’t know if ‘Draft Rank’ will be a feature in our predictive models, we can’t deal with these values yet. Furthermore, it doesn’t make sense to infer draft ranks for un-drafted players. As before, we exported the dataframe, ready for EDA.

Other Potential Data Sets

Other potential data sets include any other information that could have been scraped from basketball-reference, however in most cases this would represent the same information but transformed slightly (for example instead of reading in a table of ‘per game’ data, we read in ‘per 36 minutes’ data). The other option would have been to include advanced statistics, however most of these can be derived from the information we already have in our dataframe and as such would not provide much use in terms of predicting power.

EDA and Inferential Statistics Results

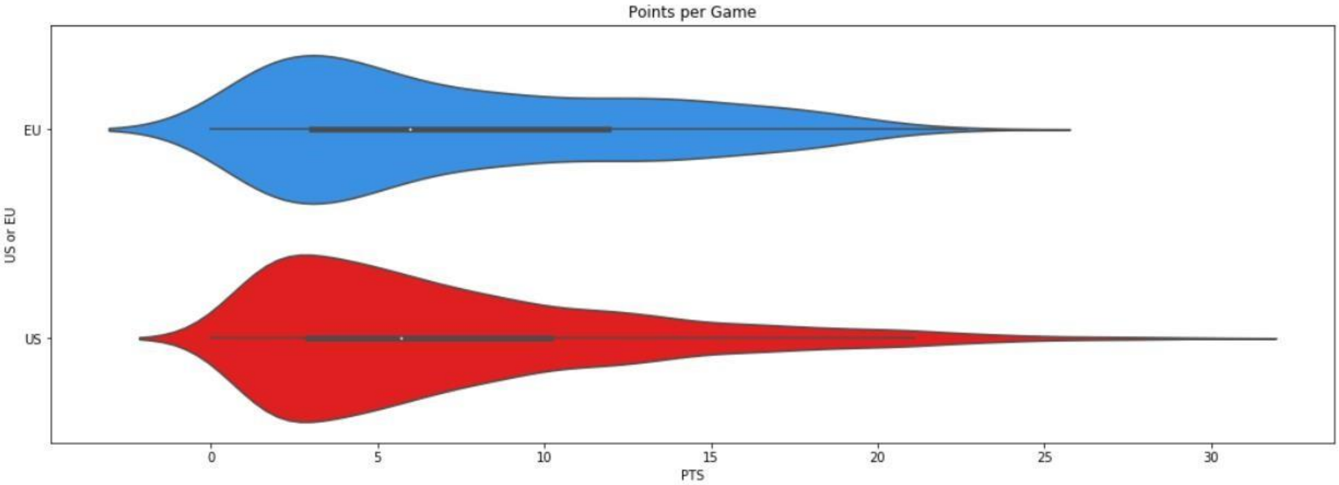
General Overview

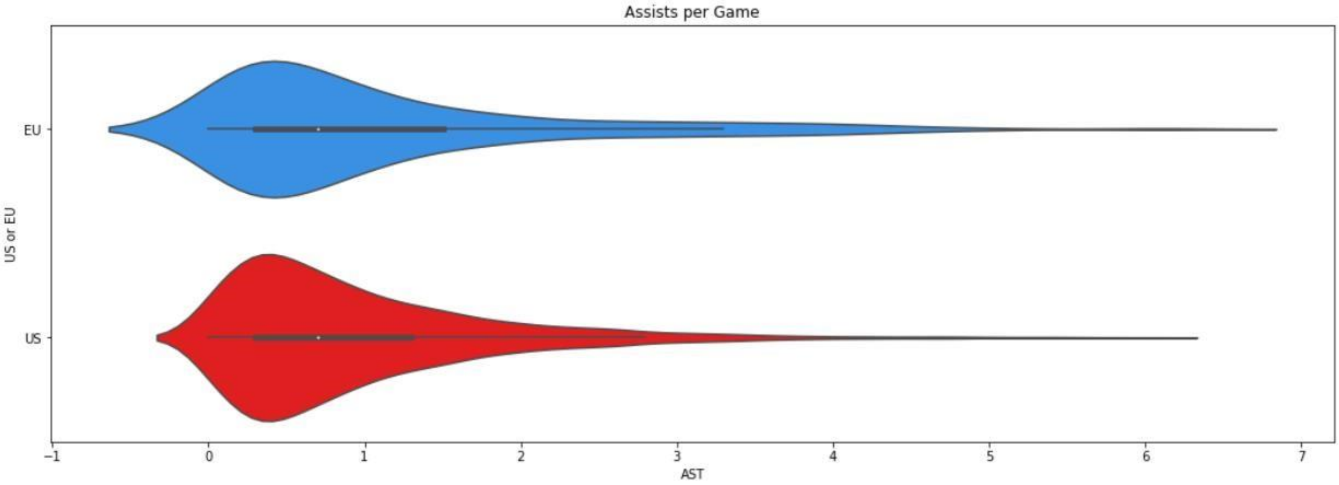
Given that the goal of the Capstone Project was to find out the differences between US and EU NBA centers and create predictive models for (Points per Game, Assists per Game, Rebounds per Game etc) we began Exploratory Data Analysis (EDA) with a general data overview. It was important to note that since we scraped the data of ALL the seasons played by NBA centers since 1980 (3pt shot introduced in 1980) the dataset, by definition, contained the data from the whole population.

We found that 3804 seasons were played by players originating in the US, and 592 by their EU counterparts. A time series plot of the number of players from each category over the years was plotted and confirmed (as expected) that the ratio of EU players in the NBA has grown since the year 2000. This was an important plot, as not only were we able to test the hypothesis that the ratio of EU players has grown but the increasing ratio allowed us to better compare the 2 groups (from a statistical basis).

US vs EU Centers

I chose a violin plot to visualise US and EU players in the most common descriptive statistics (Points, Assists, Rebounds etc) as not only are we able to compare summary statistics between groups (box plot inside the violin), but we are also able to see the distribution of the data (edge of the violin). As mentioned above, given that this dataset represents the populations of EU NBA centers and US NBA centers, by having the violin plots side to side, we made direct comparisons between the 2 groups for each statistic.





Above are a few of the Violin Plots from the EDA Jupyter Notebook. We should keep in mind that given the larger proportion of US players than EU players, ‘outliers’ are potentially more likely.

With regards to Points per Game (PPG), both groups are very similarly matched with the median and the lower quartile being almost identical. However, the EU players have a larger interquartile range, meaning that they are more evenly distributed with regards to how many points they earn. Furthermore the 95% confidence interval (CI) stretches further for the EU players. So, while there are a few seasons where US players averaged between 27.5 and 33 PPG, the EU players (proportionally speaking), on average, tend to earn slightly more points per game. The difference is minimal however. Perhaps this is because for EU players to apply to the NBA draft, they must be that much better than the other EU players to even consider it. This might explain the flatter distribution when compared to the US players, for whom the NBA is the natural choice when going pro and so you get a lot more 'average to below average' players who, had they not been born in the US (with all other things being equal) might not have applied for the draft.

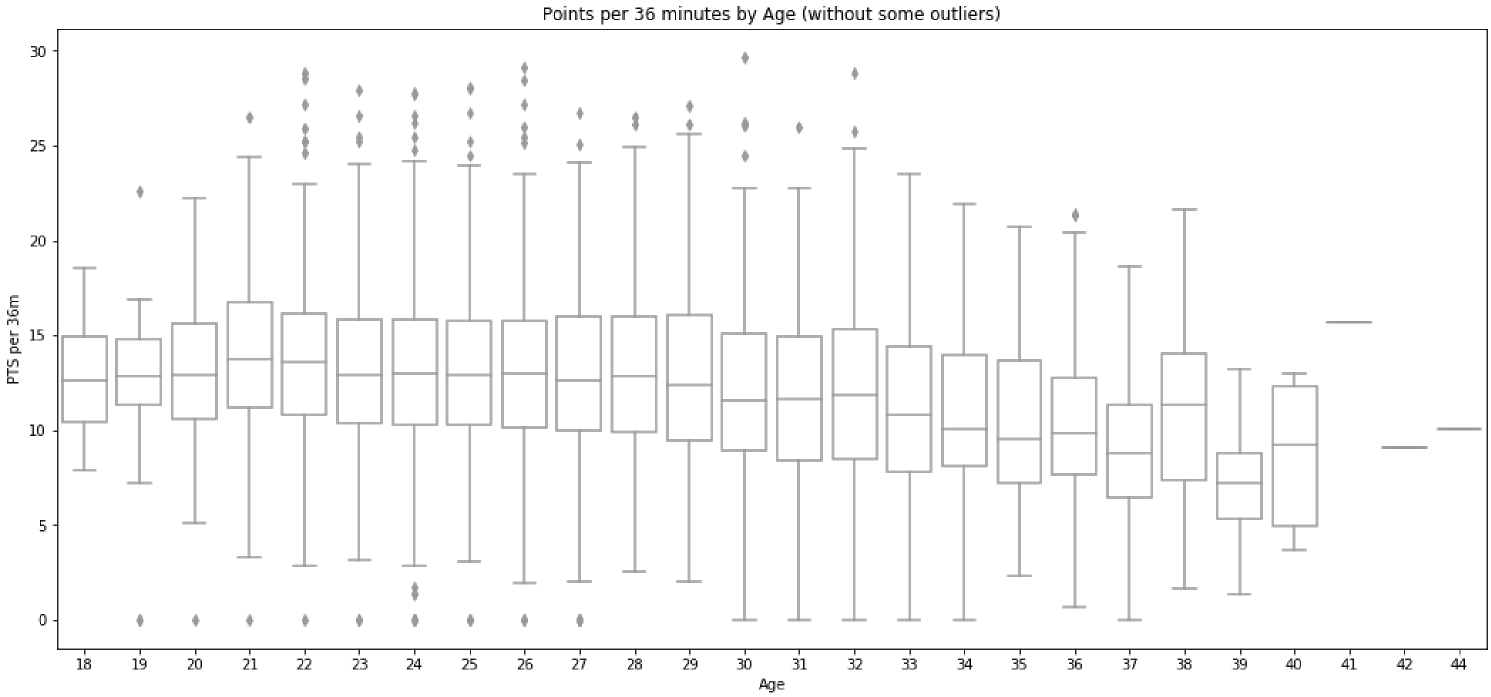
The Assists violin plot showed a similar story. The distributions are almost identical with the upper quartile and the 95% CI higher for the EU centers, but as before, not by much. Again, there are a few 'anomalies' in both categories stretching the violin to the right. This is perhaps contrary to the notion that 'European Big Men' are better passers. Similar dissimilarity is seen the violin plots for Rebounds, Steals, Blocks, Turnovers and Personal Fouls.

Thus, the main takeaway (perhaps even surprisingly so) is that with regards to the statistical measures, no significant differences were found between the groups. This meant that from a statistical basis US and EU players (as populations) are the same, even the distributions of values were almost identical. Thus, we concluded, perhaps surprisingly, that US and EU players (in totality of all the seasons played in the NBA) are the same (from a statistical standpoint). At that point we had effectively, and uninterestingly, answered a key question of the project *'What is difference between US and EU centers playing in the NBA, and what are their strengths and weaknesses'.*

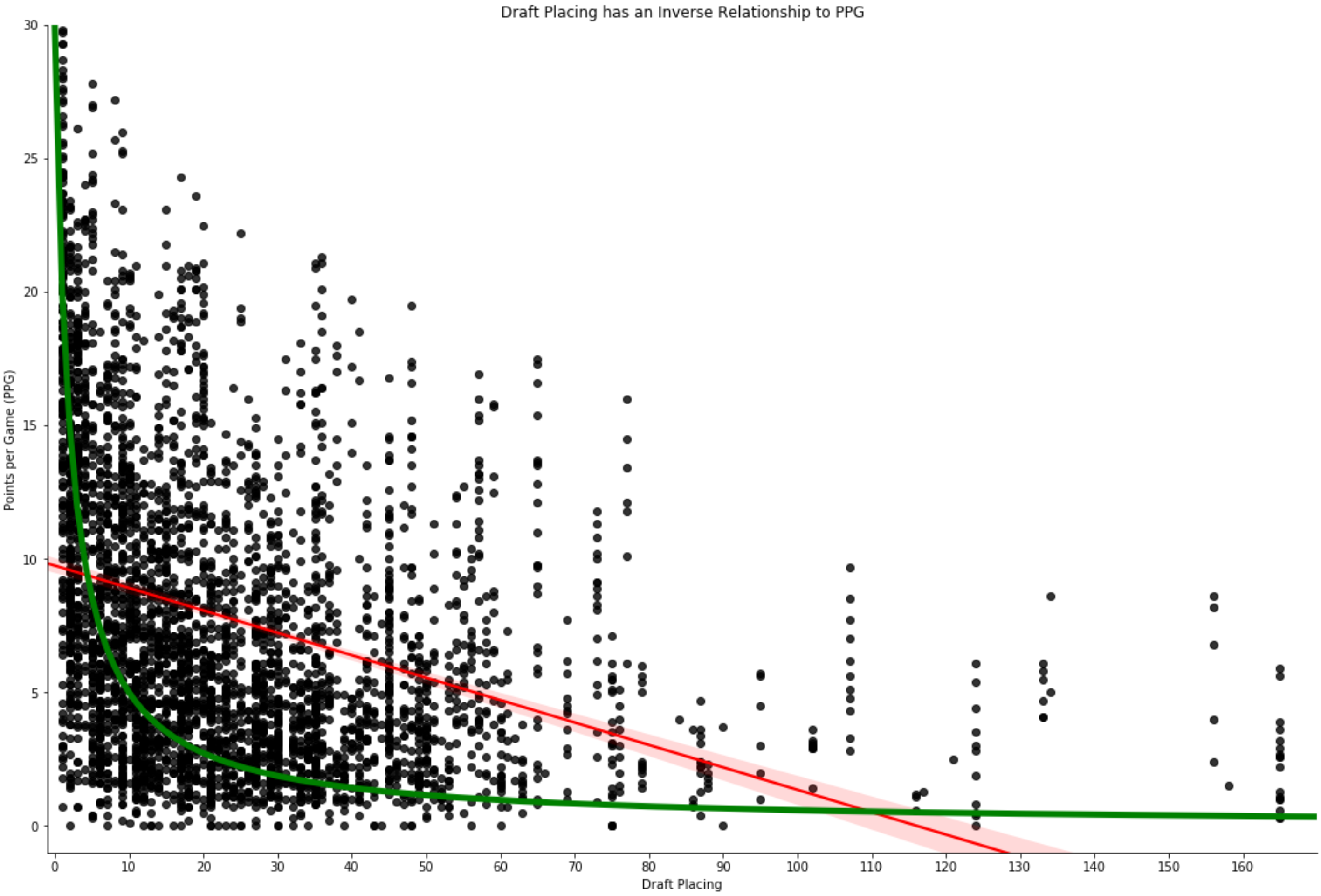
It's important to note, however, that while numerically the populations are the same, the ages at which EU and US players reach peak performance for example cannot be inferred. However, this does mean that we can disregard the 'US or EU' categorical variable (which classifies is the season was played by a player from the US or EU).

Feature Selection/Extraction

We first covered the relationship with the 'Age' column. There was a clear relationship, however given that we had aggregate data instead of discrete player career data, we could not easily track the progression of each individual player. To get a better idea of the relationship we viewed 'Age' as a series of box plots standardized by playing time (per 36 minutes) and found a decline in player efficiency with age.



We also dealt with 'Draft Placing' and 'eFG%' (Effective Field Goal %) columns and found non-linear relationships. However, we recognised the downside of heatmaps as a form of 'semi-automatic' feature selection given that heatmaps are only indicative of linear relationships, and we found that most of the features that we (rightfully) assumed would have some relationship with PTS had a nonlinear relationship, which the heatmap could not make visible.



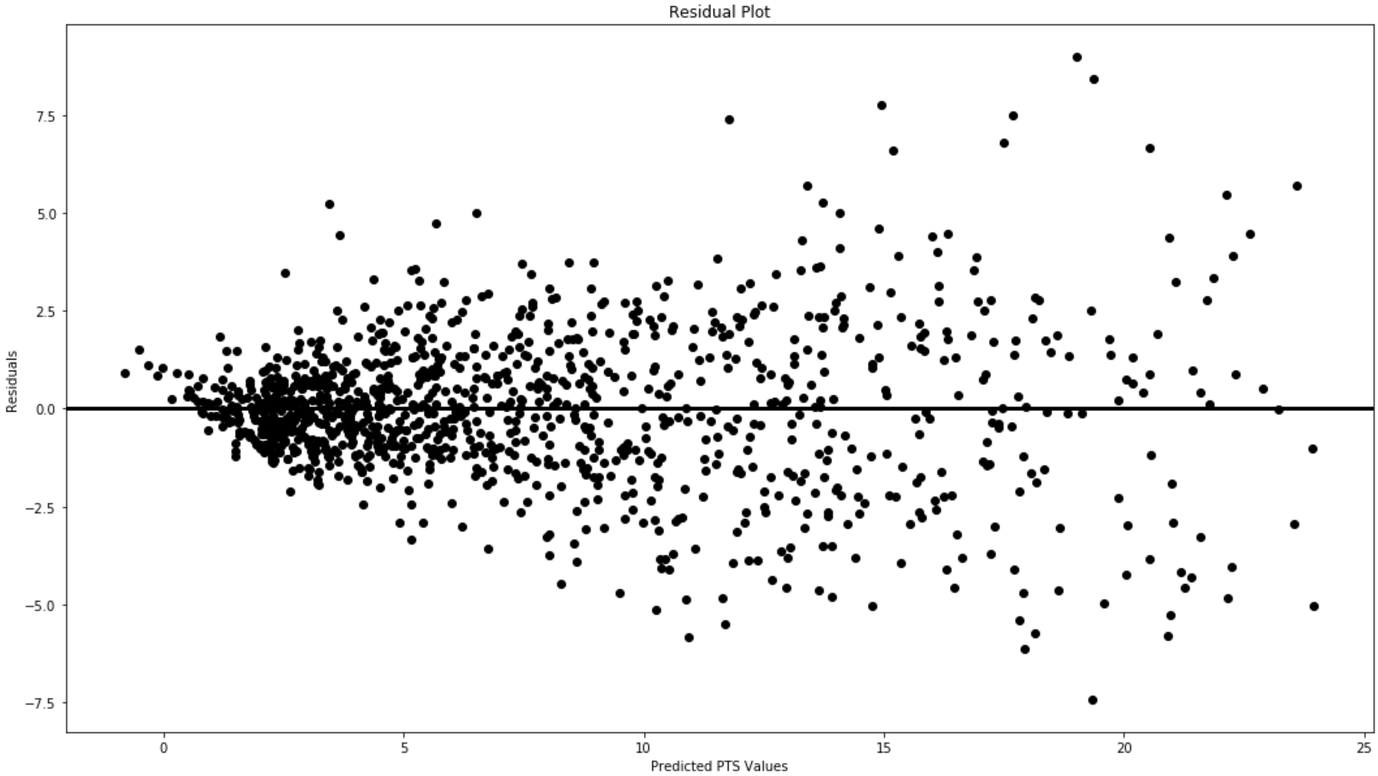
The conclusion then, in terms of feature selection, was that plotting scatter plots for all dependant variables and all independent variables is an inefficient time-consuming process (which is why we did not repeat it for the other dependant variables). Instead, we opted to begin training and testing of our linear regression models before spending an endless amount of time on feature selection (especially so given that there are functions that perform feature selection).

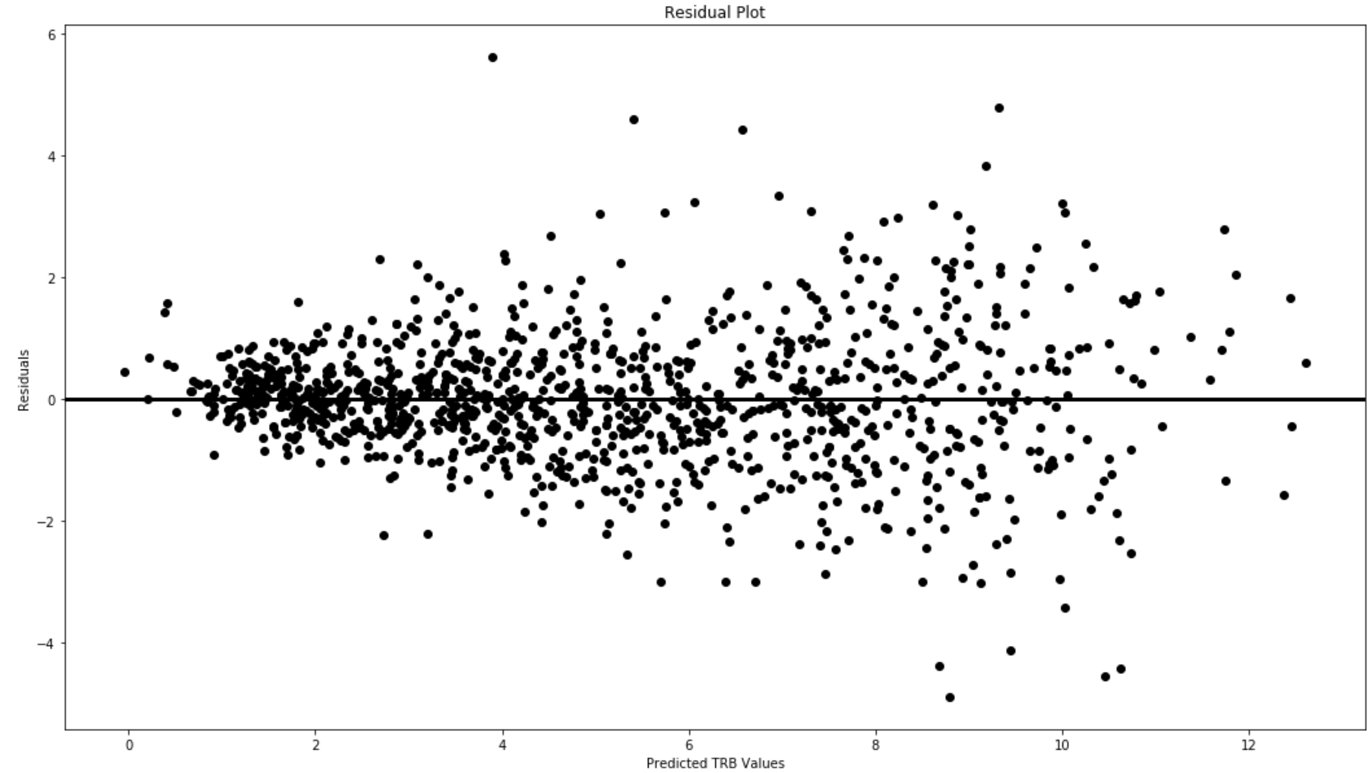
Hypotheses were formed and ready to be tested, however we found out that the data was not normally distributed (thus breaking the assumption of the Hypothesis Test). Also, as discussed before, given that we were working with population data (by definition), hypothesis tests are not necessary.

Building the Predictive Models/Machine Learning

Firstly, we split the data into 2 groups: drafted and undrafted players so we could create models for both separately. We built custom functions to reduce the number of lines of code written. Some of these functions served the purpose of plotting residual plots, others were used to easily predict values from unseen data in a simple way.

We went through many iterations of building a model, consulting the residual plot, and making changes with the aim to reduce the mean squared error and achieve homoscedasticity. The Points per Game and Rebounds per Game models performed much better than their counterparts (in terms of the mean squared error, considering the range in values of the statistics) both for drafted and undrafted player data, however heteroscedasticity was still present even after transforming the features and attempting weighted least squares regression, as we can see from the residual plots below.





Conclusion

We found that, in the end, there is no statistical difference (as a whole) between NBA centers from the US or EU in the major statistical categories and that european players are making up an everincreasing ratio of the NBA player base. The predictive models are able to fulfil the purpose of the project, which was to allow the user (fan or scout of an NBA team for example) to get predictions for the main performance statistics, however only the models for Points and Rebounds are sufficiently accurate to be used in a real-world scenario, however this depends on the required accuracy of the user and thus this decision is best made by the user. Perhaps by introducing more data and tinkering with the models heteroscedasticity could be resolved and the accuracy of the other models improved.

How do I get predictions using your models?

In the ideal case, any individual regardless of technical background would be able to get predictions by simply entering the data points required by the regression model. However, given that no 'userfriendly' API exists this restricts the users to those with basic Python knowledge. By using the *get_predictions()* function one can attain predictions for any of the models in the notebook.

Beyond the Project

The project focused on centers only, clearly, we could extend the real-world value of our models by adapting them to predict for players at any position in basketball. This would require a modification to the scraping algorithm amongst other things but would provide users with much more information.