Linear Regression Final Project

Catie Cronister, Ricky Zhang, Huidong Xu

I: Description of the dataset

This data set is from Kaggle and the data within it was scraped from the website basketball real gm. It contains data on 14582 different basketball players (however the same athlete is represented multiple times as their stats are split by season). The original data set is 53798 rows with 31 columns and has 6 categorical variables and 25 numerical variables. The descriptions of the variables are as follows:

League	Which basketball league the athlete played in for the given season, categorical
Season	The basketball season the player achieved the reported stats, categorical
Stage	This tells us if the stats were achieved in the regular season, in the play-offs or was the player in an international league, categorical
Team	Which professional team was the player a member of, categorical
GP	Games played in the reported season/stage by the player, numerical
MIN	Minutes played in the reported season/stage by the player, numerical
FGM	Field Goals made in the reported season/stage by the player, numerical
FGA	Field Goals attempted in the reported season/stage by the player, numerical
3PM	Three pointers made in the reported season/stage by the player, numerical
3PA	Three pointers attempted in the reported season/stage by the player, numerical
FTM	Free throws made in the reported season/stage by the player, numerical
FTA	Free throws attempted in the reported season/stage by the player, numerical
TOV	Turnovers committed in the reported season/stage by the player, numerical

PF	Personal fouls committed in the reported season/stage by the player, numerical
ORB	Count of offensive rebounds in the reported season/stage by the player, numerical
DRB	Count of defensive rebounds in the reported season/stage by the player, numerical
REB	Total rebounds in the reported season/stage by the player, this is the sum of ORB+DRB, numerical
AST	Assists in the reported season/stage by the player, numerical
STL	Total steals in the reported season/stage by the player, numerical
BLK	Total blocks in the reported season/stage by the player, numerical
PTS	Points scored in the reported season/stage by the player, numerical
Birth Year	Year the player was born, numerical
Birth Month	Month the player was born, numerical
Birthdate	Birthday of the player, numerical
Height	Height of player (feet' inches''), numerical
Height (cm)	Height of player in centimeters, numerical
Weight	Weight of players in pounds, numerical
Weight (kg)	Weight of players in kilograms, numerical
Nationality	Nationality of the player, categorical
High School	Which high school the player attended, categorical

The final dataset we decided to work with (after EDA and model diagnostics) was 7232 rows and 17 columns with three categorical variables and fourteen categorical variables.

II: Research Problem, Methodology and EDA

Basketball is a game of evolution and diversification: today's game is very different from 10 years ago; different leagues have their own playing styles; one player's performance can vary hugely in regular season and playoffs. Therefore, we were very interested in seeing how these factors affect players' key performance index – Points scored per game. We'll mostly be using Multiple Linear Regression approach to find answers to this problem.

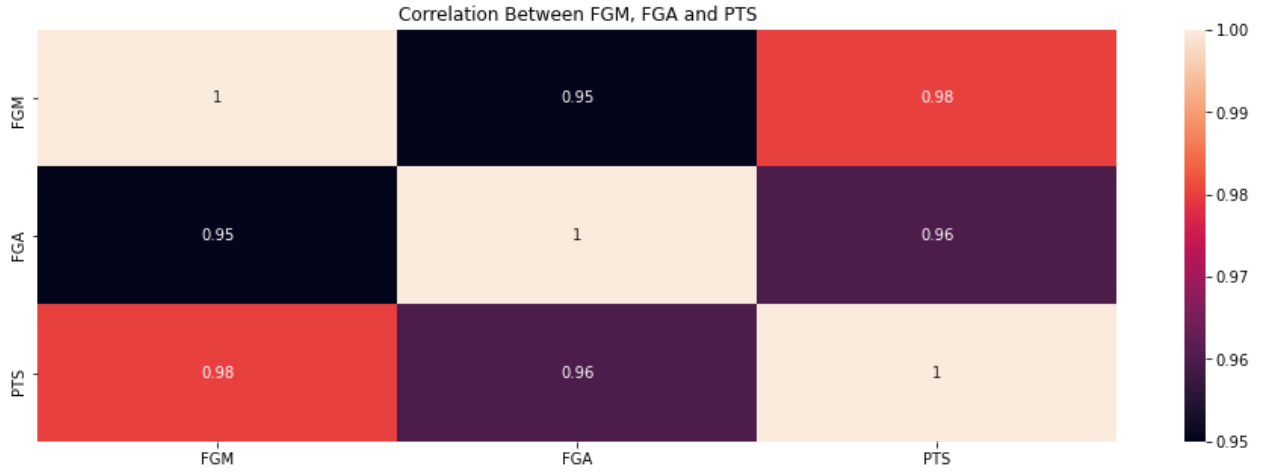
We began with filtering out players who played very few minutes and scored zero points in a season, as they were very likely to be injured. Then, in order to better interpret results of regression and make more meaningful conclusions, we decided to focus on the seasons between 2012-2020 and on some of the "top" leagues: NBA, Euroleague, Spanish, Turkish, and Italian leagues. We knew this would also help make the categorical variables more manageable later in our analysis.

After selecting the subset of data to focus on, we performed some initial check of data tidiness. We found the data already very structured and clean, but there's one problem that might damage the creditworthiness of our regression results – all players' statistics are season totals, but different leagues play significantly different amounts of games per season. Therefore, we decided to standardize the statistics for each player. To work around this problem we transformed the total stats for each player into stats per game. This allowed us to compare points across the same relative level independent of which league the player was a part of.

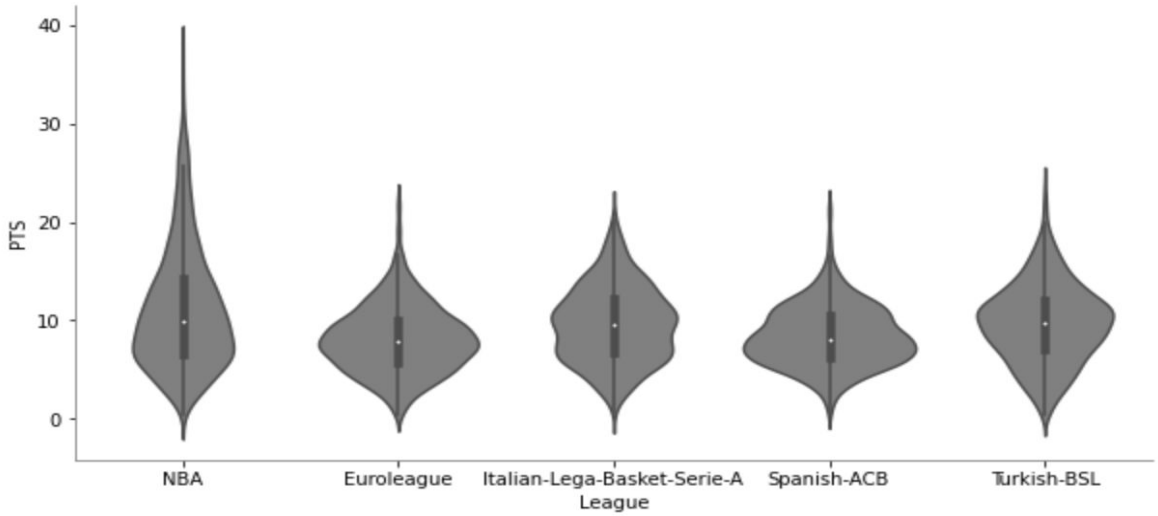
We also noticed that smf.ols function was having trouble with the columns that had a number in their name, so we renamed them (i.e. 3PM→ ThreePM).

As we continued with our exploratory analysis, we decided to delete some variables. For example, birthdays were irrelevant so we dropped them. We kept height_cm, weight_kg but dropped height & weight as they were telling us the same information. After that, we dropped high school as there were too many null values and we dropped nationality as there were too many levels of that categorical variable and keeping it in the model would jeopardize interpretation. Additionally, we deleted the player name and team as this was simply a way to identify the players and we felt was not giving us useful information about the points scored by the player.

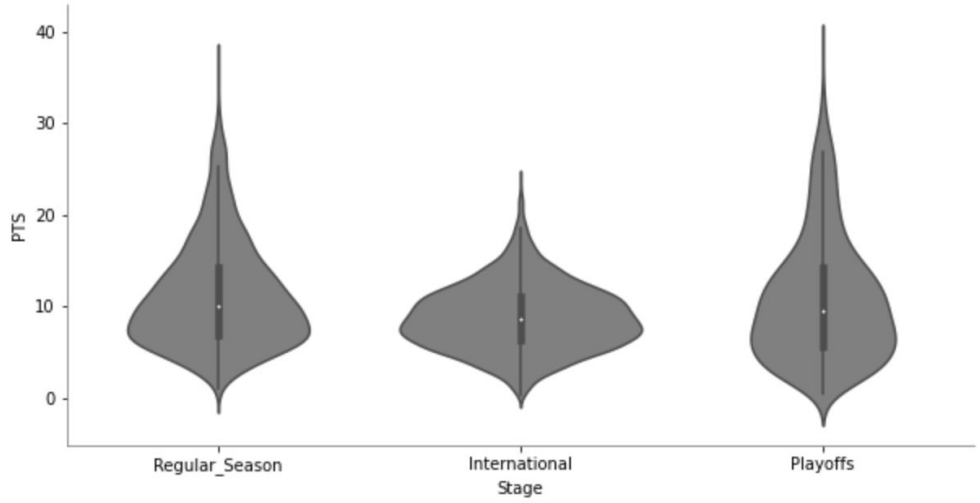
We then wanted to look for correlations between the predictor variables and the dependent variable. We quickly noticed that FGM, FGA and points were highly correlated. This makes sense as how many points you score is dependent on how many shots you take and how many baskets you make. Thus we dropped FGM, FGA.



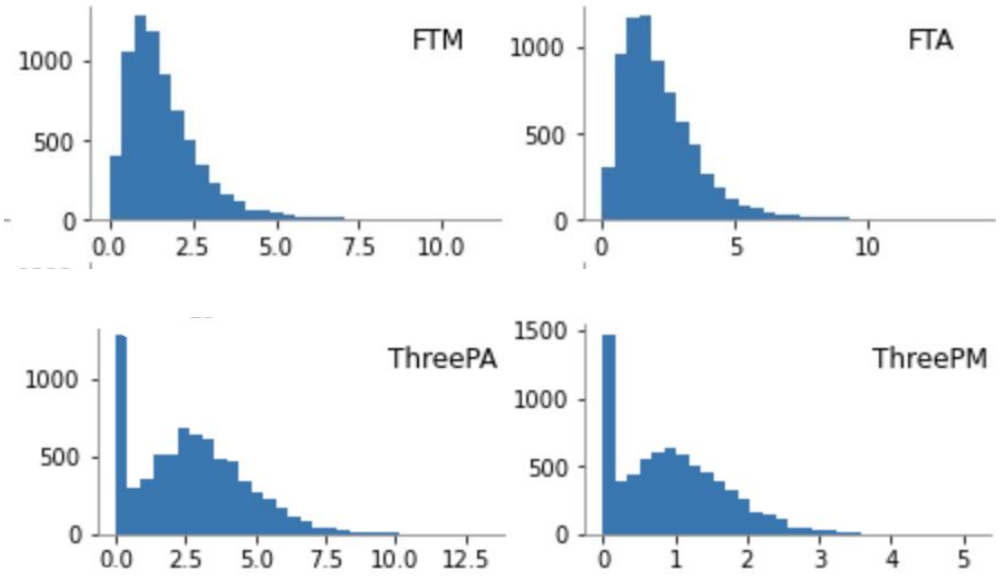
Additionally, we try to make some decisions about which categorical variables should be included in our model. We made violin plots to see if the distribution of the response variable PTS will vary based on the levels of the categorical variable. And we found that when we grouped by the categorical variable, League, the distributions of PTS of different groups would be different. For example, when the League is set to NBA, the PTS will have a much larger range and its largest value is approximately 40. For other levels of League, however, their largest values are less than 25. So the League seems to be a relevant factor to fit the model and predict the PTS.



For the categorical variable Stage, we also discovered that the distribution of PTS varied based on the levels of Stage. So, Stage might be a good choice to include in our model.



We also discovered that some categorical variables have pretty similar distributions, for example, FTM and FTA, and ThreePA and ThreePM. This is not a surprise because, as we mentioned in our description of the dataset, they are describing similar features.



III: Regression Analysis

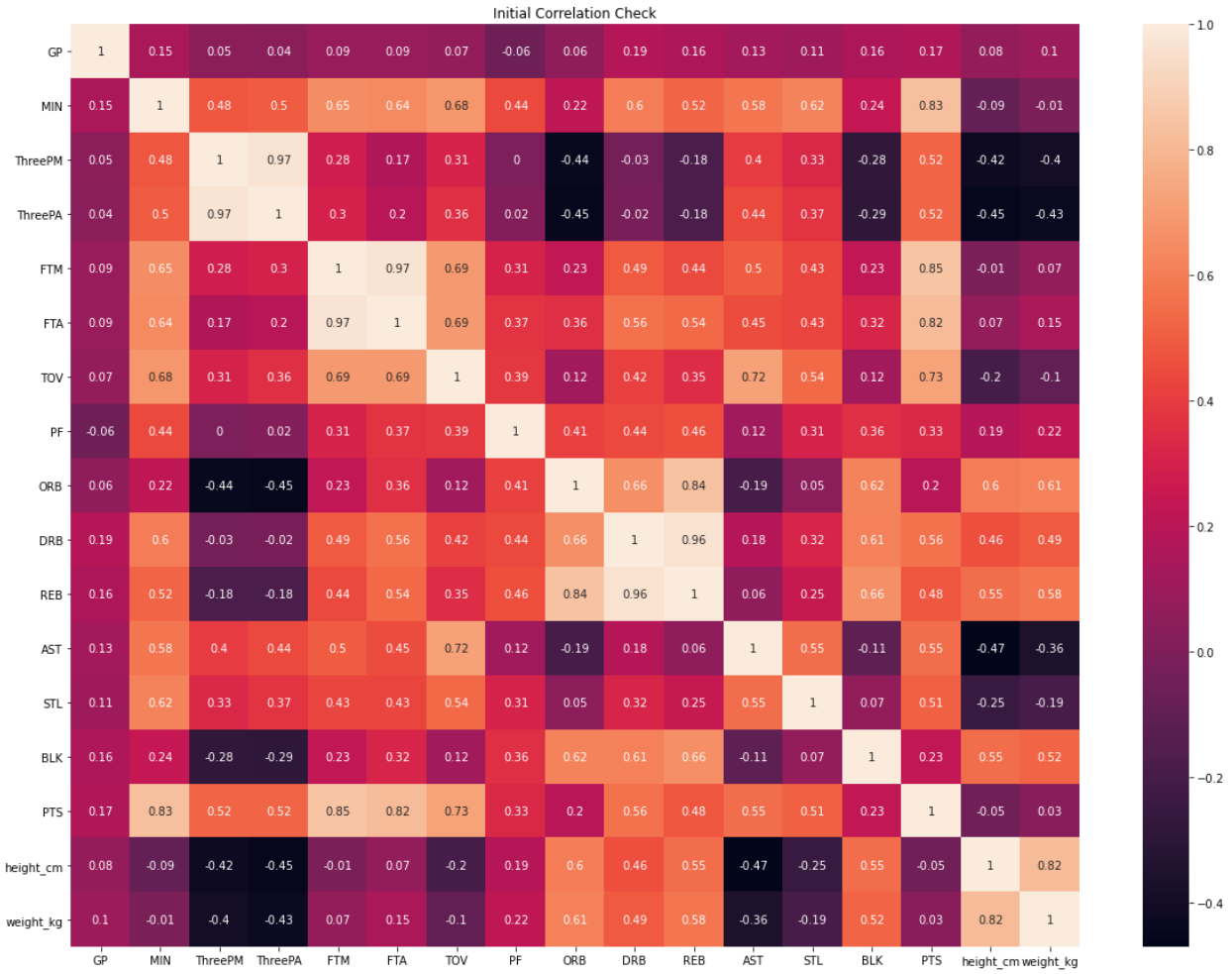
From here, we begin our multiple linear regression and we fit our first model using the method of least squares in python:

PTS~
C(League)+C(Season)+C(Stage)+GP+MIN+ThreePM+ThreePA+FTM+FTA+TOV+PF+ORB+DRB+REB+AST+STL+BLK+height_cm+weight_kg

When we look at the summary table we see that we have a p-value for our F-stat of approximately 0, so we can make the claim that at least one of our predictors is significant. Another way to interpret this is that the full model is significantly better than the reduced model with just the intercept. However, we do not spend much time on this model as we have a strong belief that model diagnosis will greatly change these results. So we decided to begin our model diagnosis process.

IV: Model Diagnosis

To begin our model diagnosis, we checked the multicollinearity of the model as we had an a priori notion it would be an issue. We did a quick correlation check between the predictors to get the following matrix:



As one can see, there are some high correlations in this matrix such as FTM and MIN, FTM and FTA. This was a good indicator that multicollinearity is present in our model so we decided to calculate the VIF scores for each of the predictors -- including our categorical variables and we get the following results:

	VIF Factor	features
0	1.604997e+03	Intercept
1	1.687274e+00	C(League)[T.Italian-Lega-Basket-Serie-A]
2	inf	C(League)[T.NBA]
3	1.808102e+00	C(League)[T.Spanish-ACB]
4	1.531581e+00	C(League)[T.Turkish-BSL]
5	1.758790e+00	C(Season)[T.2013 - 2014]
6	1.773674e+00	C(Season)[T.2014 - 2015]
7	1.765780e+00	C(Season)[T.2015 - 2016]
8	1.807465e+00	C(Season)[T.2016 - 2017]
9	1.877832e+00	C(Season)[T.2017 - 2018]
10	1.929894e+00	C(Season)[T.2018 - 2019]
11	2.366100e+00	C(Season)[T.2019 - 2020]
12	inf	C(Stage)[T.Playoffs]
13	inf	C(Stage)[T.Regular_Season]
14	1.108607e+01	GP
15	4.780737e+00	MIN
16	1.694146e+01	ThreePM
17	1.901469e+01	ThreePA
18	2.313308e+01	FTM
19	2.531964e+01	FTA
20	4.119870e+00	TOV
21	1.612578e+00	PF
22	2.989048e+04	ORB
23	1.201506e+05	DRB
24	2.289741e+05	REB
25	3.705196e+00	AST
26	1.926678e+00	STL
27	2.173313e+00	BLK
28	3.993517e+00	height_cm
29	3.527725e+00	weight_kg

As we can see, multiple variables are highly impacted by multicollinearity such as: ORD, DRB, REB, STAGE (both playoffs and regular season), League, etc… We would first deal with ‘NBA’ in C(League) and ‘Playoffs’ & ‘Regular_Season’ in C(Stage) as they have infinite VIF value, which implies perfect correlation.

In fact, this result does not intimidate us, as in our previous EDA, we found that 'Playoffs' and ‘Regular_Season' in 'Stage' are terms only for the NBA. In other words, all rows with Stage = ‘Playoffs’ or ‘Regular_Season’ have ‘League’ = ‘NBA’, and vise versa. No wonder they show such high correlation. This led us to the conclusion to drop the predictor stage.

When we calculate the VIF scores again without stage, we see much better results. However, ORB, DRB and REB are still large numbers, but, this makes sense as REB is simply ORB+DRB.

	VIF Factor	features
0	1593.781556	Intercept
1	1.626865	C(League)[T.Italian-Lega-Basket-Serie-A]
2	2.712913	C(League)[T.NBA]
3	1.647200	C(League)[T.Spanish-ACB]
4	1.511858	C(League)[T.Turkish-BSL]
5	1.758231	C(Season)[T.2013 - 2014]
6	1.771248	C(Season)[T.2014 - 2015]
7	1.765356	C(Season)[T.2015 - 2016]
8	1.806417	C(Season)[T.2016 - 2017]
9	1.873253	C(Season)[T.2017 - 2018]
10	1.912542	C(Season)[T.2018 - 2019]
11	2.062258	C(Season)[T.2019 - 2020]
12	1.385407	GP
13	4.765383	MIN
14	16.876450	ThreePM
15	18.958515	ThreePA
16	23.132155	FTM
17	25.317810	FTA
18	4.077495	TOV
19	1.611623	PF
20	29890.201279	ORB
21	120149.300162	DRB
22	228971.661665	REB
23	3.694909	AST
24	1.926638	STL
25	2.173247	BLK
26	3.993417	height_cm
27	3.527577	weight_kg

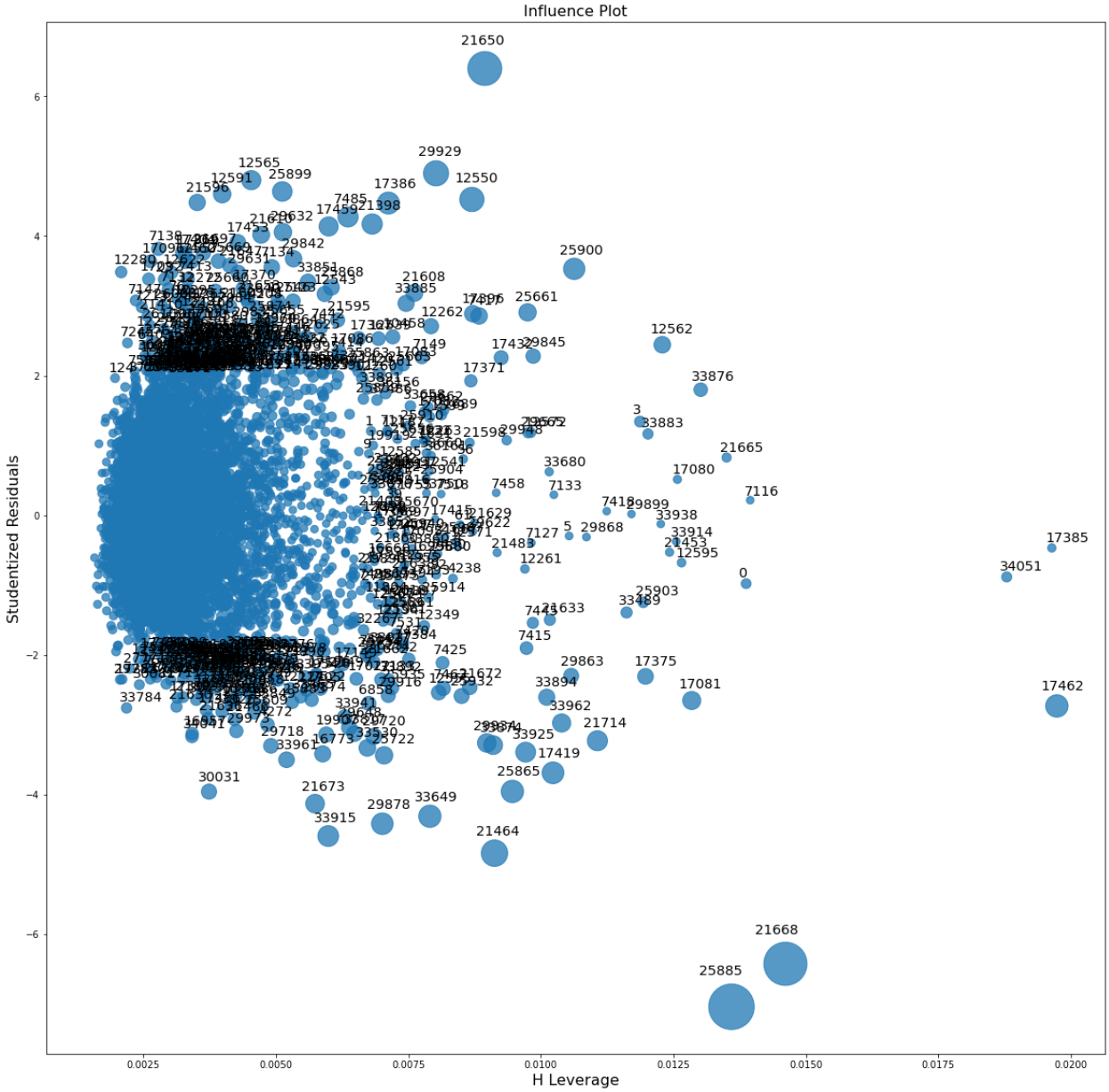
We wanted to keep DRB as it's most correlated to PTS according to correlation matrix. Other than that, 3PM and 3PA & FTM and FTA are also quite correlated pairs. We decided to keep 3PA and FTA, because similar to FGM, 3PM and FTM are very indicative of how many goals a player score -- they literally tell you a subset of how many points a player scored.

After we make these changes and calculate the VIF once more, we see the following.

	VIF Factor	features
0	1577.894558	Intercept
1	1.616599	C(League)[T.Italian-Lega-Basket-Serie-A]
2	2.642572	C(League)[T.NBA]
3	1.641077	C(League)[T.Spanish-ACB]
4	1.499849	C(League)[T.Turkish-BSL]
5	1.755876	C(Season)[T.2013 - 2014]
6	1.766982	C(Season)[T.2014 - 2015]
7	1.764361	C(Season)[T.2015 - 2016]
8	1.804801	C(Season)[T.2016 - 2017]
9	1.871504	C(Season)[T.2017 - 2018]
10	1.902326	C(Season)[T.2018 - 2019]
11	2.055736	C(Season)[T.2019 - 2020]
12	1.365821	GP
13	4.581121	MIN
14	2.108504	ThreePA
15	2.499077	FTA
16	4.045459	TOV
17	1.589288	PF
18	3.523952	DRB
19	3.546811	AST
20	1.898163	STL
21	2.115933	BLK
22	3.981719	height_cm
23	3.484250	weight_kg

All of the VIF scores are less than 4 so we can say that there is a light impact on multicollinearity for some of the predictors, but this is not worrisome, so we have taken care of any multicollinearity issues.

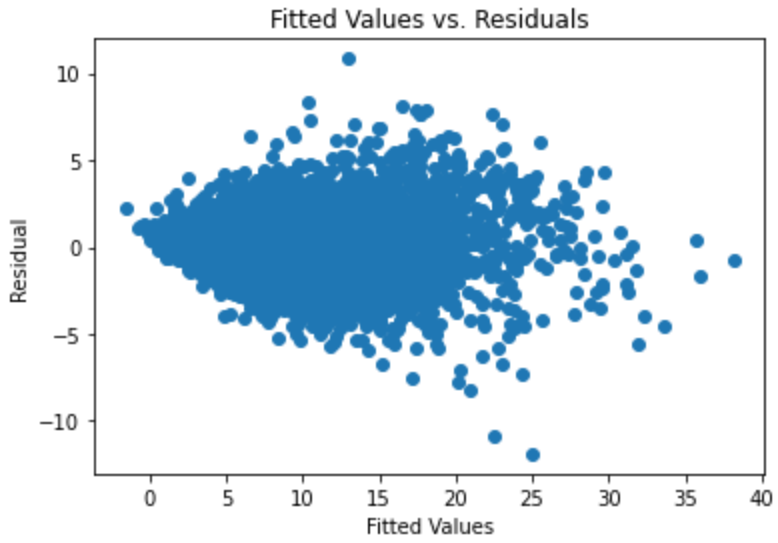
We then fit the model again with the remaining predictors, and we wanted to check for influential points. To get an initial idea of the points, we created a graphical representation of the influences where the size of points are given by cook's dist volume, the x-axis is the leverage and the y-axis is the externally studentized residuals. The points with big size and out of the (-3,3) should raise a flag.



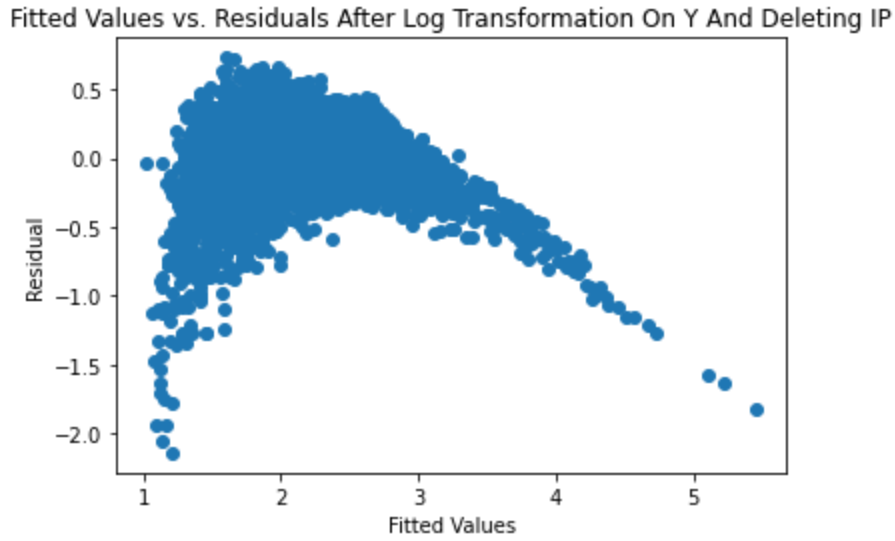
When we calculate the externally studentized residuals and see if they are above the threshold of t(.975,n-p-1) we end up with 408 observations above the cutoff. Similarly, we wanted to identify influential points using Cook's method . Using the cutoff of 4/n, we identified 477 influential points. We then took the intersection of these two lists to identify important influential points and end up with a list of 367 points.

As for now, we will keep the dataset as it is; if we find that heteroscedasticity exists, we will try deleting the influential points to see if it helps.

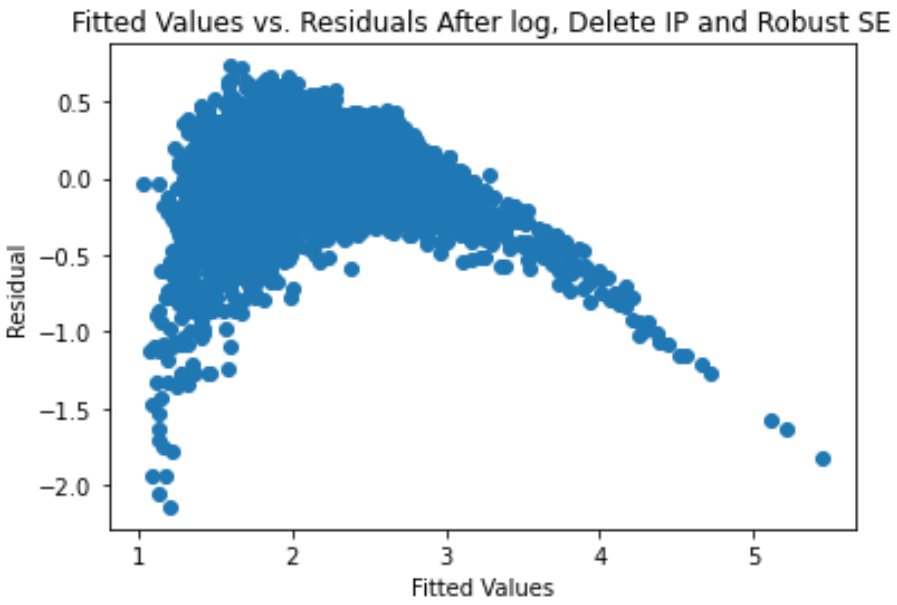
Our next step was to examine heteroscedasticity. An initial look at the fitted values vs residuals showed us a sort of funnel shape.



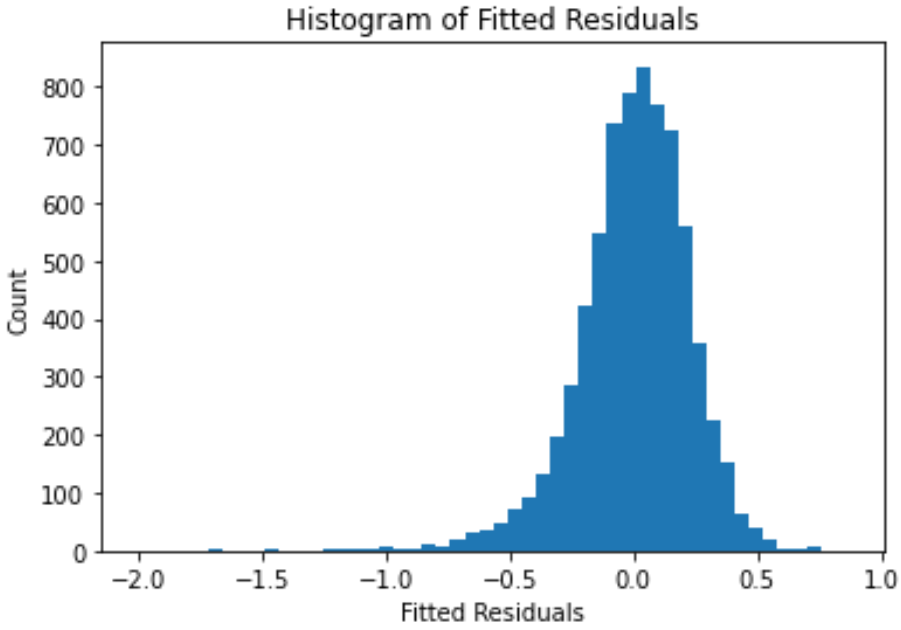
This was backed up by a Breusch-Pagan test where we get a p-value of 3.055e-216 so we fail to reject the null hypothesis that heteroscedasticity does not exist. Since heteroscedasticity exists, we took the natural log of PTS and dropped the 367 influential points mentioned above and refit the model. This substantially helped the problem, our p value increased to 4.8966e-66, however heteroscedasticity still exists.



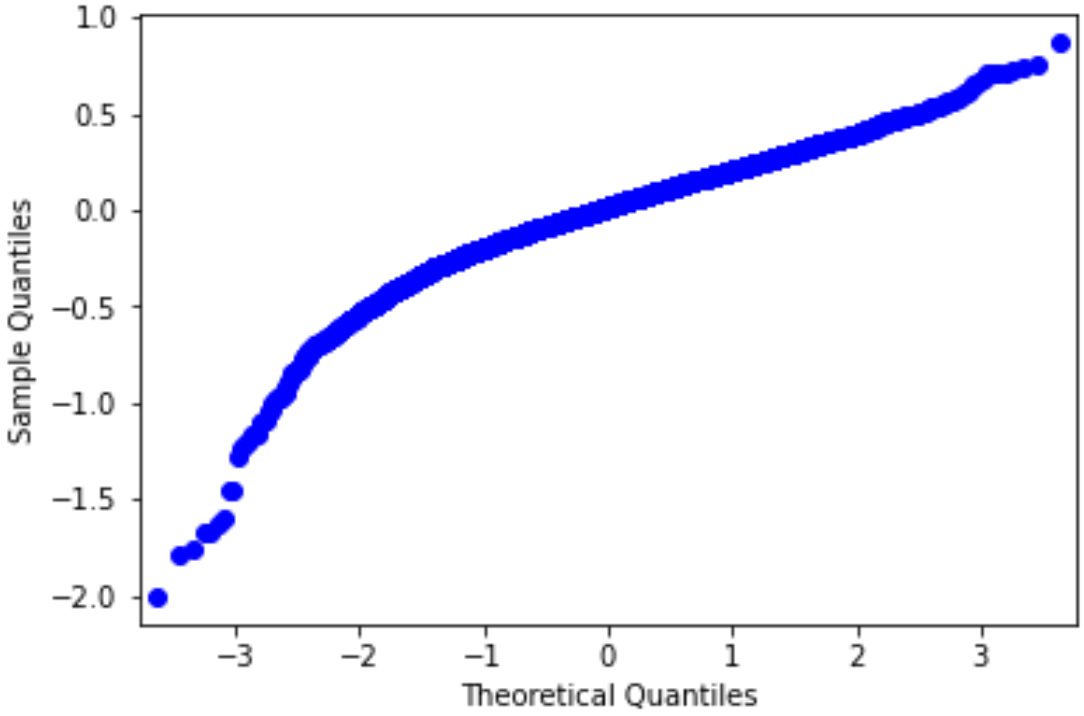
We then tried to use robust standard error to fit the model; nonetheless, heteroscedasticity still exists according to Breusch-Pagan test result.



The final model diagnostic we performed was the check for normality of the residuals. Our initial histogram of the residuals seemed to be about normal:



Additionally, our Q-Q plot seemed to be about on a 45° line.



This led us to believe that the assumption of normality of the residuals was not violated. However, the Jarque-Bera test showed the residuals don't follow normality.

Jarque-Bera (JB):	9329.205
Prob(JB):	0.00

We choose not to worry about this result as our sample size is large and the central limit theorem guarantees that our residuals are approximately normally distributed so our inferential statistics should not be affected too badly.

V: Model selection

Quick overview of model selection strategy:

1. Best subset with adj-R2 and Mallow's Cp to give a list of candidates.
2. AIC/BIC on all candidates to find the one with the smallest value.
3. Then look into the t-test result to check significance of each predictors.

In order to select a model, we first examined the individual t tests from our initial model. This led us to the decision to remove PF, DRB, BLK and height_cm since they are not significant predictors. From there, we went with the best subset method. We chose this method because it examines all the possible models so we can confidently say our final selection is the "best" model. We believe that best subsets thoroughness outways it's high computational expense. After running our best subset function, we have at maximum 2^p models to choose from. However, some initial exploration showed most of them have a Mallow's Cp that is far away from k (where k is the number of parameters for the model). When we set the criteria that Mallow's Cp <= 200, we only get a small list of candidate models as seen below.

	number of predictors	adj_R2	Mallow's_Cp	predictors
0	8	0.807789	112.105923	C(League), C(Season), GP, ThreePA, FTA, TOV, MIN, AST
1	8	0.805637	194.135765	C(League), C(Season), GP, ThreePA, FTA, TOV, MIN, weight_kg
2	8	0.806005	180.116873	C(League), C(Season), GP, ThreePA, FTA, MIN, AST, STL
3	8	0.806245	170.956951	C(League), C(Season), GP, ThreePA, FTA, MIN, AST, weight_kg
4	8	0.806867	147.241121	C(League), C(Season), GP, ThreePA, FTA, MIN, STL, weight_kg
5	8	0.806542	159.643081	C(League), GP, ThreePA, FTA, TOV, MIN, AST, STL
6	8	0.806104	176.337174	C(League), GP, ThreePA, FTA, TOV, MIN, AST, weight_kg
7	8	0.805502	199.271497	C(League), GP, ThreePA, FTA, TOV, MIN, STL, weight_kg
8	9	0.809450	49.786098	C(League), C(Season), GP, ThreePA, FTA, TOV, MIN, AST, STL
9	9	0.809130	62.007109	C(League), C(Season), GP, ThreePA, FTA, TOV, MIN, AST, weight_kg
10	9	0.807823	111.802958	C(League), C(Season), GP, ThreePA, FTA, TOV, MIN, STL, weight_kg
11	9	0.807492	124.428061	C(League), C(Season), GP, ThreePA, FTA, MIN, AST, STL, weight_kg
12	9	0.807626	119.331848	C(League), GP, ThreePA, FTA, TOV, MIN, AST, STL, weight_kg
13	10	0.810494	11.000000	C(League), C(Season), GP, ThreePA, FTA, TOV, MIN, AST, STL, weight_kg

Since all of the models above have closed adjusted R-squared, for the same number of predictors, we choose the models with the smallest Mallow's Cp as our candidate models:

1. log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST
2. log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST+STL

3. log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST+STL+weight_kg

From here we calculated the AIC and BIC for each of these candidate models.

	model	AIC	BIC
0	log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST	-470.958219	-347.005343
1	log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST+STL	-532.733451	-401.894304
2	log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST+STL+weight_kg	-571.468002	-433.742584

This led to a clear decision for our final model:

log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST+STL+weight_kg

However, we still wanted to check the t-tests for each individual predictor in case any were not significant predictors of PTS. We found that all predictors in our final model have a p-value that is less than 0.05, except for one dummy variable of the League categorical variable. However, the other levels of the categorical variable are significant so we feel comfortable concluding the categorical variable as a whole is significant. So we conclude that all predictors are significant for the model.

VI: Final Model of Choice, Interpretation and Prediction

Our final model of choice is:

log_PTS~C(League)+C(Season)+GP+ThreePA+FTA+TOV+MIN+AST+STL+weight_kg.
When we fit this model we generate the following table:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4417	0.033	13.382	0.000	0.377	0.506
C(League)[T.Italian-Lega-Basket-Serie-A]	-0.0520	0.010	-4.973	0.000	-0.073	-0.032
C(League)[T.NBA]	-0.1145	0.009	-12.898	0.000	-0.132	-0.097
C(League)[T.Spanish-ACB]	0.0129	0.009	1.365	0.172	-0.006	0.032
C(League)[T.Turkish-BSL]	-0.0334	0.010	-3.223	0.001	-0.054	-0.013
C(Season)[T.2013 - 2014]	0.0265	0.011	2.359	0.018	0.004	0.049
C(Season)[T.2014 - 2015]	0.0535	0.011	4.754	0.000	0.031	0.075
C(Season)[T.2015 - 2016]	0.0368	0.011	3.275	0.001	0.015	0.059
C(Season)[T.2016 - 2017]	0.0634	0.011	5.668	0.000	0.041	0.085
C(Season)[T.2017 - 2018]	0.0891	0.011	8.061	0.000	0.067	0.111
C(Season)[T.2018 - 2019]	0.0989	0.011	8.878	0.000	0.077	0.121
C(Season)[T.2019 - 2020]	0.0578	0.011	5.333	0.000	0.037	0.079
GP	0.0026	0.000	19.325	0.000	0.002	0.003
ThreePA	0.0474	0.002	25.349	0.000	0.044	0.051
FTA	0.1056	0.003	38.716	0.000	0.100	0.111
TOV	0.0779	0.007	10.737	0.000	0.064	0.092
MIN	0.0452	0.001	62.008	0.000	0.044	0.047
AST	-0.0297	0.003	-10.133	0.000	-0.035	-0.024
STL	-0.0666	0.009	-7.276	0.000	-0.085	-0.049
weight_kg	0.0020	0.000	6.383	0.000	0.001	0.003

Based on the individual t test of each predictors, we conclude that all predictors in our final model are significant predictors.

The adjusted R square of this model is 0.81, meaning that the proportion of the variance for PTS that's explained by the predictors in the regression model is 81%. Although R-Squared on its own is often unhelpful, our R-Squared is close to 1, so we feel confident coming to the conclusion that our model has a good fit.

Then we do the model diagnosis of the final model. Our conclusion is that there is no multicollinearity based on the VIF factors of features. For the check of variance of residuals, we made the plot of fitted values versus residuals and found that the bandwidth changes, and the result of Breusch - Pagan Test shows that our model still has a problem of heteroscedasticity. Then we checked for the normality assumption by running the Jarque-Bera test on the residuals and found that our model still has a problem of non-normality. Non-normality is not a serious problem, because our sample size is large, the central limit theorem guarantees normality.

The issues that heteroscedasticity would cause for our model is as follows:

(1) The OLS estimate for variances (or standard errors) of the intercept and slope coefficients are generally biased.
(2) Misleading t test and anova test, incorrect prediction intervals.

For future improvements, if we want to solve the heteroscedasticity, we can consider using a weighted least square method for regression, or robust regression methods.

In regards to interpreting our model, we look to the summary table from python. Based off this table, we can make the following statements:

- Players in the Italian League score, on average, .0520 less points per game, than players in the Euroleague.
- Players in the NBA score, on average, .1145 less points per game, than players in the Euroleague.
- Players in the Turkish League score, on average, .0334 less points per game than players in the Euroleague.
- Players in the Spanish League score, on average, .0129 more points per game than players in the Euroleague. However, this p-value was greater than .05 so we conclude that there is no difference in points between players in the Euroleague and Spanish League.
- Similarly, when all the other predictors are zero, the average points scored per game by players in the Euroleague is .4417.
- Players who played in the 2013-2014 season scored, on average, 0.0265 more points per game than those in the 2012-2013 season.
- Players who played in the 2014-2015 season scored, on average, 0.0535 more points per game than those in the 2012-2013 season.
- Players who played in the 2015-2016 season scored, on average, 0.0368 more points per game than those in the 2012-2013 season.
- Players who played in the 2016-2017 season scored, on average, 0.0634 more points per game than those in the 2012-2013 season.
- Players who played in the 2017-2018 season scored, on average, 0.891 more points per game than those in the 2012-2013 season.

- Players who played in the 2018-2019 season scored, on average, 0.0989 more points per game than those in the 2012-2013 season.
- Players who played in the 2019-2020 season scored, on average, 0.0578 more points per game than those in the 2012-2013 season.
- Similarly, when all the other predictors are zero, the average points scored per game by players in the 2012-2013 season is .4417.
- For every additional game a player plays, their points per game increase by .0026.
- For every additional three pointer attempted by a player, their points per game increase by .0474.
- For every additional free throw attempted by a player, their points per game increase by .1056.
- For every additional turnover by a player, their points per game increase by .0779.
- For every additional assist a player makes, their points per game decrease by .0297.
- For every additional steal a player makes, their points per game decrease by .0666.
- For every additional kilogram a player adds to their body, their points per game increase by .0020.

Then we used our model to do some predictions on new data. Here are two data points that we made.

	League	Season	GP	ThreePA	FTA	TOV	MIN	AST	STL	weight_kg
0	NBA	2014 - 2015	69	7.88	9.99	4.56	33.33	9.13	1.11	99.0
1	Spanish-ACB	2019 - 2020	62	7.80	9.00	3.67	44.44	8.15	1.21	109.0

We chose the significance level of 5%. From the prediction summary table, for players whose league is NBA, season is 2014 - 2015, games played is 69, three-pointers attempted is 7.88, free throw attempted is 9.99, turnover is 4.56, minutes played is 33.33, assist made is 9.13, steal made is 1.11, and body weight is 99.0, the 95% confidence interval for mean response E(log_PTS | X) is [3.66, 3.75], the 95% confidence interval for individual response log_PTS | X is [3.25, 4.16].

For players whose league is Spanish-ACB, season is 2019 - 2020, games played is 62, three-pointers attempted is 7.80, free throw attempted is 9.00, turnover is 3.67, minutes played is 44.44, assist made is 8.15, steal made is 1.21, and body weight is 109.0, the 95% confidence interval for mean response E(log_PTS | X) is [4.14, 4.22], the 95% confidence interval for individual response log_PTS | X is [3.73, 4.64]. If we want to get the predicted PTS from the predicted log_PTS, we can do :

$$PTS = e^{log\_PTS}$$

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	3.703837	0.022189	3.660339	3.747335	3.246435	4.161239
1	4.183837	0.020604	4.143447	4.224226	3.726720	4.640953

It is important to remember that these prediction intervals are most likely slightly off due to our issue of heteroscedasticity, but the point predictions themselves are fine. We still include confidence intervals as we find they are interesting and if a future study could correct our heteroscedasticity problem, then they could simply re-run our notebook to get interesting intervals.

VII: Summary of Findings/Results

Our main purpose of building the model is to provide precise predictions of response variables and make inferences about model performance, such as capture the significant association between response variables and predictors as well as explain how the changes in predictors change response variables. Building the final model includes the process of making several attempts to solve the data structure problems and assumption violations. During this model diagnostics process, we learned how to detect and solve several problems. The details are as follows:

Our initial model had a problem of multicollinearity, we solved it by dropping some of the highly correlated variables.

We detected influential points using externally studentized residuals and Cook's distance, after removing those points our model performance didn't change too much. We don't know the reasons why the influential points exist. Simply deleting them might lose valuable information. So we keep them.

Our model had a problem of heteroscedasticity, we took the logarithm of the response variable and found that the bandwidth for Residuals vs Fitted Value Plot looked better. But it didn't get rid of the heteroscedasticity. Heteroscedasticity is the most serious, hardest to deal with problem. Some of the problems of heteroscedasticity include: incorrect se(beta_hat_k) in the OLS result which could lead to a misleading t-test result, the variance of y_hat_h is incorrect so we would have misleading confidence and prediction intervals (however we included them as we believe they are still interesting). However, the estimation of the coefficients and fitted values are still unbiased and reliable so we feel confident in our model.

Our model had a problem of non-normality residuals. But it is not a serious problem because the central limit theorem applies to our dataset (our n>>>30).

After running the linear regression analysis for the dataset, we conclude that League, Season, GP, ThreePA, FTA, TOV, MIN, AST, STL, weight_kg are the factors that have a significant impact on points scored per game.

We end with doing some prediction and interpretation work. We made some new data points to get the confidence interval for both mean response and individual response. This made us believe that our analysis has practical significance and can be used to solve real-world problems.

VIII:Table indicating the Work Completed by Each Member

Group Members	Ricky	Catie	Huidong
Percentage of Work	33.3	33.3	33.3
List of Work	Statement of the research problems, and a summary of methods. Regression analysis that may include but not limited to multiple linear regression model and logistic regression model. The model diagnosis that may include but not limited to: assumption validation, influential points, check for heteroscedasticity and multicollinearity.	Description of your dataset: resource, dimension, variable description, etc. The explanatory analysis may include but not limited to graphs, demographic summaries, crosstables, individual tests, etc.. And some findings at this point. Model selection Found the dataset on kaggle Helped organize meetings and wrote an initial rough draft of most of the paper. Wrote and preformed model interpretation -- specifically coefficient interpretation	Make a final model of choice , and do some prediction or interpretation work. Write a summary for your findings/results. The explanatory analysis may include but not limited to graphs, demographic summaries, crosstables, individual tests, etc.. And some findings at this point. Predictions using model

		Regression analysis that may include but not limited to multiple linear regression model and logistic regression model. Did initial model diagnosis and wrote model diagnosis Helped write the final summary Edited the paper and formatted	