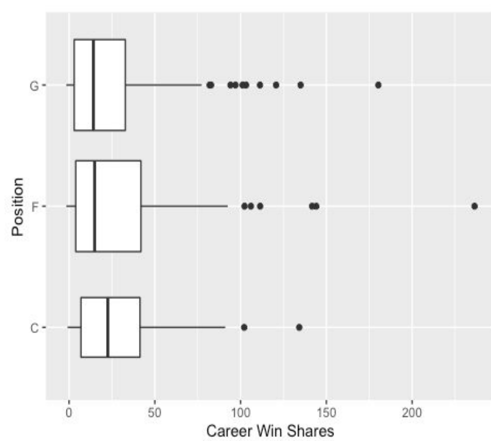
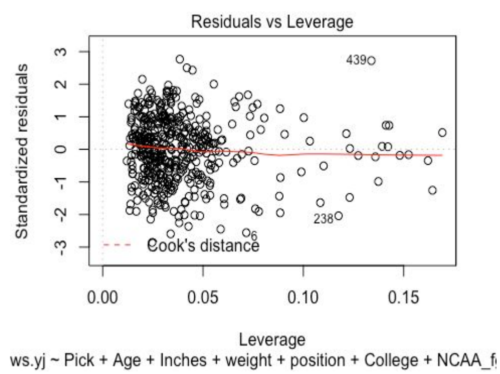
Exploratory Data Analysis Report

Micaleb Johnson, Elias Cheleuitte, Morgan Williams

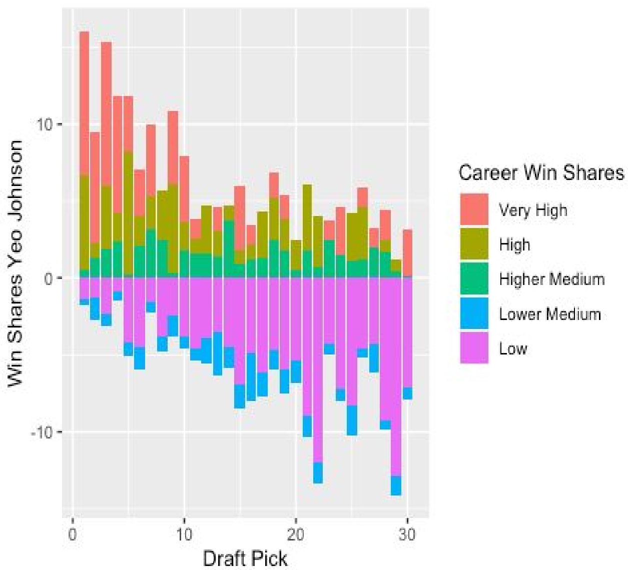
Data Cleaning and Aim Assessment:

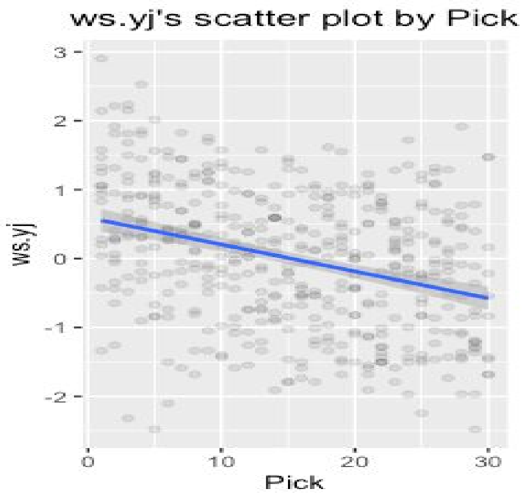
From our merged dataset, we found that certain players had a missing value for the variable 3pt percentage. Upon further investigation this was due to these players having never attempted 3 pointers, thus we decided to fill in the missing values with 0. The variable measuring player height required some adjustment because the original height format wouldn't have been usable in R programs, thereby we converted it from the format (6,8) to height in inches (80). The format for the college and position variable would have also been difficult for an R program to accurately interpret, thus we turned each variable into a factor with 3 levels. The college factor is broken down into 3 factors, 1 being high school, 2 being international, and 3 being NCAA college. Some higher outliers were present in the win shares variable, which is the metric we are using to measure the success of a player's NBA career. After running a preliminary regression model, with all players included, we didn't see any player who had an abnormally high Cook's distance or leverage. This indicated to us that these players simply had very successful careers and shouldn't be removed from our dataset.





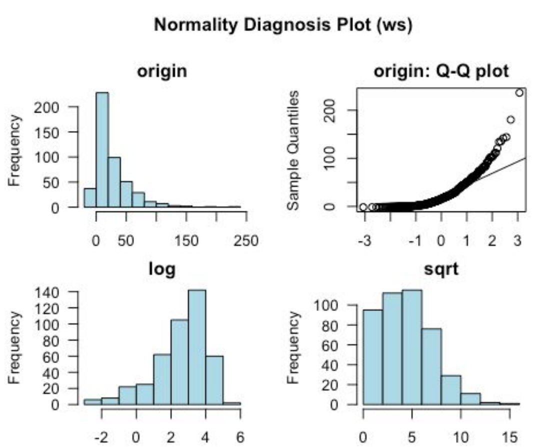
From the plots below, we see that as expected the transformed career win shares for a player decreases as draft pick increases. This thought is reinforced from our initial linear regression model predicting win shares, providing a coefficient of -.029267 for the pick variable. The plot on the lower left also shows that super star NBA players are more likely to be drafted in the early picks and players who produce very little in the NBA are more frequently drafted with higher picks. For instance, players in the 90th percentile of our data set for career win shares (indicated by the "Very High" label) are very prevalent in the top 5 picks but then greatly decrease in frequency as the pick numbers get higher. Players in the bottom quartile (labelled as "Low") of our data for win shares see a large increase in frequency of being picked as the picks increase. One interesting note is that it appears players who have decent careers (labelled as "Higher Medium") have a pick frequency that stays approximately the same throughout all the 1st round picks. From these plots, it seems that pick value will be much greater for the top 5 picks and that after that pick value will have a steep drop off and then gradually decline as each pick increases.

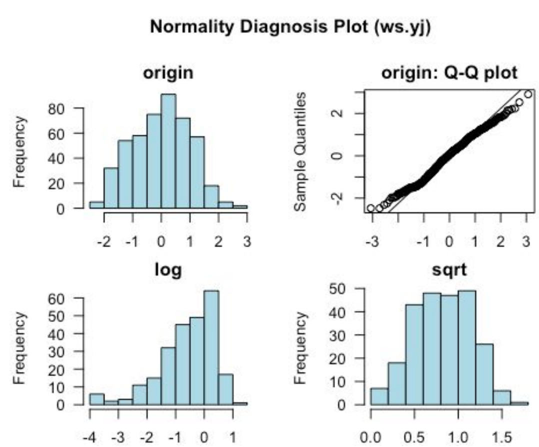




Variable Characteristics and Associations:

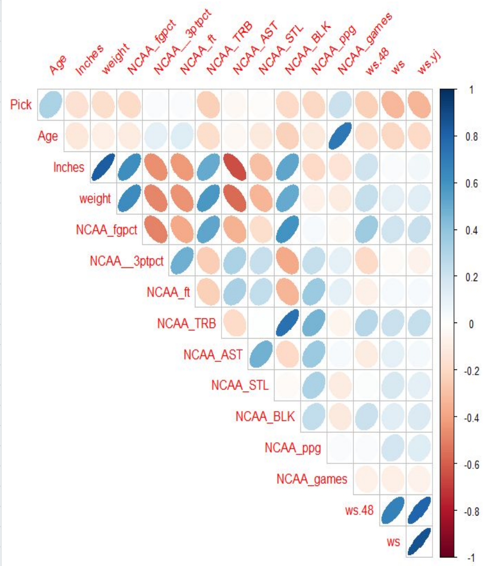
As we began working towards a linear model to predict win share totals from our statistics before being drafted,we noted that our win shares were not normally distributed. In order to normalize the distribution of win share totals as best we can, we performed a yeo-johnson transformation due to some negative win share totals.

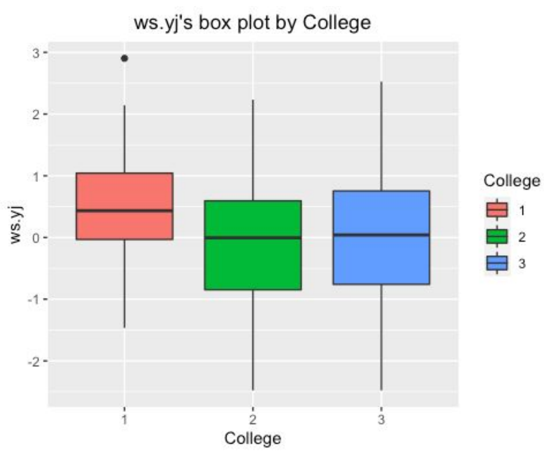


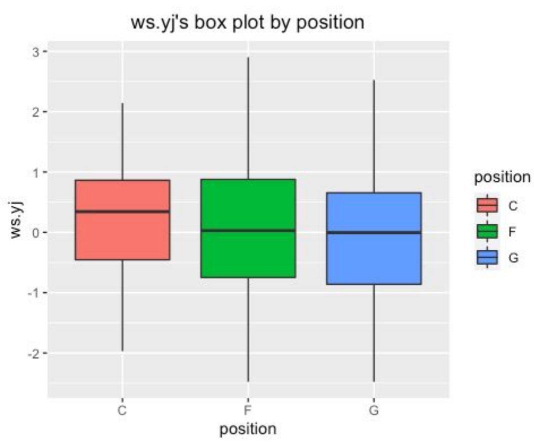


The win share totals had to be split into factors due to the different distribution of win share totals from different positions (Guard, Forward, or Center), and also the win share distributions from where they were drafted (highschool, NCAA, or international).

	variable	n	na	mean	sd	se_mean	IQR	skewness	kurtosis
1	Draft.Year	468	0	2.007547e+03	4.58466997	0.211926444	7.25000000	-0.02099616	-1.194289870
2	Age	468	0	2.042308e+01	1.39303158	0.064392908	2.00000000	0.10687493	-0.543511817
3	Inches	468	0	7.959829e+01	3.59766717	0.166302224	5.00000000	-0.15911311	-0.522309660
4	weight	468	0	2.227906e+02	26.66974156	1.232809240	36.00000000	0.17785540	-0.267978967
5	NCAA_fgpct	468	0	4.992479e-01	0.06885984	0.003183047	0.09400000	0.73467322	0.913369369
6	NCAA__3ptpct	468	0	2.901432e-01	0.14572089	0.006735950	0.10825000	-0.80573513	1.141369527
7	NCAA_ft	468	0	7.114231e-01	0.09244822	0.004273420	0.12150000	-0.66939950	0.685744392
8	NCAA_TRB	468	0	5.844637e+00	2.68775781	0.124241649	3.52500000	1.07092100	2.037044901
9	NCAA_AST	466	2	2.164485e+00	1.60539060	0.074368284	2.00000000	1.25586060	1.159316686
10	NCAA_STL	463	5	1.256587e+00	0.69880016	0.032475998	0.90000000	1.20139517	2.423981494
11	NCAA_BLK	465	3	1.098925e+00	1.24900776	0.057921348	1.10000000	2.75844717	9.924720421
12	NCAA_TOV	417	51	1.986091e+00	0.78443287	0.038413843	1.00000000	-0.10638611	0.074662542
13	NCAA_ppg	468	0	1.362274e+01	4.76738107	0.220372268	5.72500000	0.71873921	1.519185973
14	NCAA_games	468	0	7.609402e+01	38.16003333	1.763948165	66.00000000	0.31080820	-0.680918788
15	yrs	468	0	7.858974e+00	3.94303449	0.182266834	6.00000000	0.43909306	-0.515773689
16	g	468	0	4.605171e+02	289.33891593	13.374696120	420.25000000	0.47392629	-0.420784155
17	ws	468	0	2.507094e+01	29.44008196	1.360868270	32.85000000	2.29440774	8.446782057
18	ws.48	468	0	7.733761e-02	0.05949208	0.002750022	0.06450000	-1.17895836	6.388620759
19	bpm	468	0	-1.320940e+00	2.93217213	0.135539704	3.10000000	-0.97237926	6.401602053
20	vorp	468	0	6.439744e+00	12.86900404	0.594869922	8.90000000	3.99023914	26.129422154
21	ncaa_pts	468	0	1.040269e+03	627.92111044	29.025663598	895.50000000	0.94541988	1.765318356
22	tsp	429	39	5.765702e-01	0.04984457	0.002406519	0.05571517	0.92017898	11.364357914
23	ws.yr	468	0	2.513288e+00	2.32108028	0.107291974	2.85313187	1.45652125	2.916919175
24	ws.48_sqrt	437	31	2.836052e-01	0.08123545	0.003886018	0.10099567	-0.36357187	0.576479959
25	ws_log	468	0	2.627479e+00	1.31327126	0.060705986	1.94200389	-0.58409184	-0.008221672
26	ws_3rt	468	0	2.365330e+00	1.35692805	0.062724021	1.76118873	-0.48345942	0.170456900
27	ws.yj	468	0	-8.489025e-17	1.00000000	0.046225016	1.51553712	-0.06068080	-0.574266534



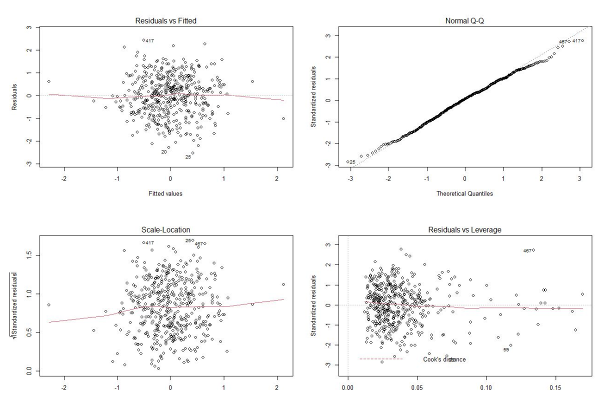




Our linear model using all the NCAA stats before they were drafted using position and college as indicator variables and our win shares transformed as the response came out to:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.722113   2.404194   2.796 0.005400 ** 
Pick         -0.029267   0.005615  -5.212 2.87e-07 ***
Age          -0.167627   0.048430  -3.461 0.000590 ***
Inches       -0.096491   0.028775  -3.353 0.000867 ***
weight        0.004722   0.003214   1.469 0.142538    
positionF    -0.185729   0.164627  -1.128 0.259856    
positionG    -0.336416   0.252970  -1.330 0.184249    
College2      0.190393   0.372664   0.511 0.609678    
College3      0.330474   0.341782   0.967 0.334115    
NCAA_fgpct    3.699696   0.941763   3.928 9.92e-05 ***
NCAA__3ptpct -0.047936   0.385975  -0.124 0.901218    
NCAA_ft       1.942799   0.615675   3.156 0.001711 ** 
NCAA_TRB      0.107783   0.033923   3.177 0.001591 ** 
NCAA_AST      0.017838   0.040422   0.441 0.659217    
NCAA_STL      0.212078   0.081980   2.587 0.010002 *  
NCAA_BLK     -0.095113   0.068778  -1.383 0.167393    
NCAA_ppg     -0.045255   0.015603  -2.900 0.003913 ** 
NCAA_games    0.003405   0.001646   2.068 0.039203 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8929 on 442 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.2268,	Adjusted R-squared:  0.197 
F-statistic: 7.625 on 17 and 442 DF,  p-value: < 2.2e-16
```



References: bestNormalize- yeo-johnson transformation, dlookr - Variable structure, associations and Normality testing, ggplot2 and Dplyr - Plotting and Data Cleaning