MKT 451 – Sports Analytics

Excel Statistics Assignment – Lahman Database

Brayden Bennett

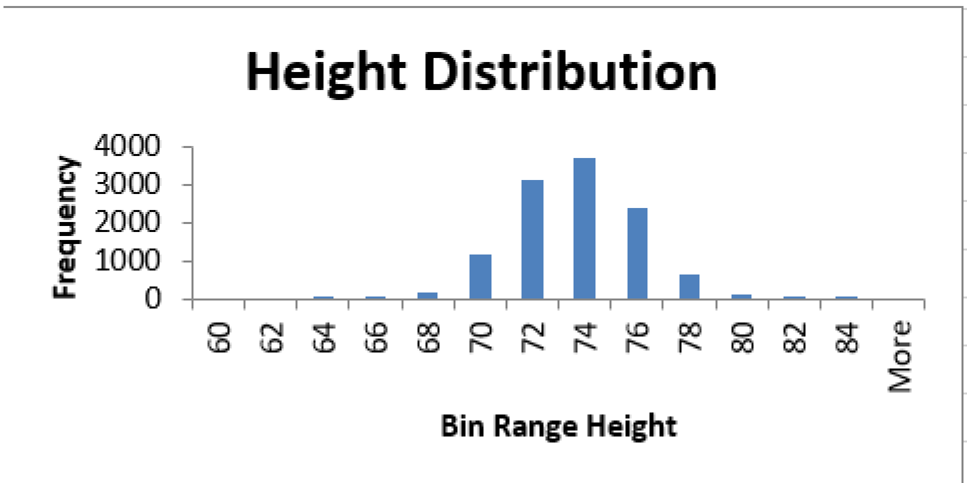
NOTE: This was a MS Excel assignment using the Lahman MLB database. The analysis included the use of ANOVA testing, t-tests, descriptive statistics, z-score analysis, regressions, some basic box plots, and more.

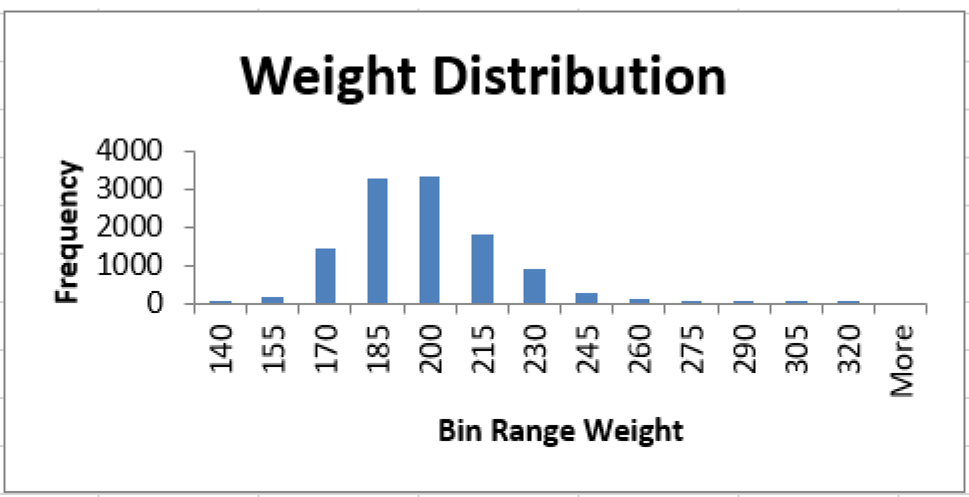
1. **Descriptive Statistics for Height and Weight for entire Master set**

weight		*height*	
Mean	192.7759	Mean	73.16875
Standard Error	0.190676	Standard Error	0.021492
Median	190	Median	73
Mode	185	Mode	72
Standard Deviation	20.2485	Standard Deviation	2.282308
Sample Variance	410.0017	Sample Variance	5.208928
Kurtosis	1.108498	Kurtosis	0.170216
Skewness	0.725439	Skewness	0.036273
Range	181	Range	20
Minimum	139	Minimum	63
Maximum	320	Maximum	83
Sum	2173934	Sum	825124
Count	11277	Count	11277

2. Height and Weight Histograms

The distributions for both height and weight are normally distributed. There are some outliers, such as the player who weighs 320 pounds and the player who is 84 inches tall, but overall, both distributions are approximately normally distributed.





3. Z – Scores

In a general sense, a z-score tells us how many standard deviations our observation is away from the mean. A positive z-score indicates our observation is above the mean, while a negative z-score indicates the observation is below the mean. In our model, using the height equation below, we find a player in the 90th percentile for height would be 76 inches tall, while the weight equation shows the 90th percentile for weight would be 220 pounds.

$$Height\ 90th\ Percentile = PERCENTILE.INC(Master!F2{:}F11278{,}0.9)$$

$$Weight\ 90th\ Percentile = PERCENTILE.INC(Master!E2{:}E11278{,}0.9)$$

If we wanted to find the percentile of a player who is 75.086 inches tall, we could find the z-score of the player at that data point, then use a z-table to find the corresponding percentile. The equation below tells us the z-score of a player who is 75.086 inches tall is 0.84. This means that the player is 0.84 standard deviations above the mean. We can then find this z-score on the z-table, and we find that this player would be in the 80th percentile as it relates to height.

$$Z - Score = STANDARDIZE(75.086, AVERAGE(Master!F2{:}F11278), STDEV.S(Master!F2{:}F11278))$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

4. Regression Predicting Weight Given Height and Position

For my regression, I chose to use the DH position as my base variable. This means that every position listed as DH was given a value of 0, and the coefficient for every other position is a measure of the distance away from the predicted weight of a DH. Since DH does not have a coefficient, the linear equation to estimate the weight of a DH is:

$$y_{weight} = 5.09(x_{height}) - 179.7$$

The coefficient for each position, as stated earlier, tells us the estimated difference from the weight of a DH. For example, catcher (C) has a coefficient of 7.34. This means that, on average, a catcher weighs an estimated 7.34 pounds more than a DH. Thus, the equation to predict the weight of a catcher would be:

$$y_{weight} = 7.34(1) + 5.09(x_{height}) - 179.7$$

****We would multiply the coefficient by 1 because position is indicated by either a 0 or 1 (0 meaning they do not play that position, 1 meaning they do)**

When looking at our model as a whole, we can look at the R^2 value, which is 0.3941 in our model. This is indicating that 39.41% of the variation of our dependent variable (weight) can be explained by the linear relationship between our independent variables (height and position). This percentage is not as high as we would like for a strong model, but it tells us that there could be variables other than height and position that can be used to predict weight. We can also look at our Significance F value to tell if we have created a good model. In our model, the Significance F value is 0, which is what we want. The Significance F value goes hand-in-hand with the p-values of individual variables, so it makes sense when we see many of our individual p-values being significant at the 0.05

alpha level. We do, however, have four positions that are not significant at the 0.05 level, and could therefore be removed from our model. These positions are P, 3B, CF, and OF. This simply means that the predicted weight of these four positions is close to the predicted weight of DH, so it gives us little benefit to include these four positions in our model. This can also be seen when looking at the coefficients of each of the four variables. They are all close to 0, showing that there is not a significant weight difference between players who play DH, P, 3B, CF, or OF.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.627784867
R Square	0.39411384
Adjusted R Square	0.393522207
Standard Error	15.76885865
Observations	11277

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	1822058.725	165642	666.146	0
Residual	11265	2801120.014	248.657		
Total	11276	4623178.739			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-179.6974358	5.495704322	-32.698	4E-224	-190.4699758	-168.9248958	-762.9933803	403.5985087
height	5.090068959	0.072603357	70.1079	0	4.947753702	5.232384216	-2.615812118	12.79595004
P	-0.491658362	1.450345164	-0.339	0.73462	-3.334588107	2.351271383	-154.4265142	153.4431975
C	7.337679383	1.52814262	4.8017	1.6E-06	4.342253042	10.33310572	-154.854342	169.5297008
1B	6.329125577	1.528963224	4.13949	3.5E-05	3.33209071	9.326160444	-155.949992	168.6082431
2B	-4.930119918	1.519298063	-3.245	0.00118	-7.908209382	-1.952030453	-166.1834092	156.3231693
3B	-1.025893337	1.589798262	-0.6453	0.51875	-4.142175502	2.090388827	-169.7618414	167.7100547
SS	-8.900480533	1.655544182	-5.3762	7.8E-08	-12.14563618	-5.655324888	-184.6144838	166.8135227
LF	4.545741221	1.614541504	2.8155	0.00488	1.380957982	7.71052446	-166.816373	175.9078555
CF	-1.53884095	1.528481181	-1.0068	0.31406	-4.53493093	1.45724903	-163.7667961	160.6891142
OF	-1.149355144	1.629612567	-0.7053	0.48064	-4.343680297	2.04497001	-174.1110623	171.812352
DH	0	0	65535	#NUM!	0	0	0	0

5. ANOVA Testing

When running an ANOVA test to compare weights at each of the positions, we find that we get a p-value of essentially 0. This tells us that at least one of our variables is significantly different than another. When looking at our averages, this makes sense, as the average weight of SS is 175 pounds, while the average weight of a 1B is 201 pounds. We can assume that this creates very little overlap between the distributions of weights and would therefore give us a low p-value. Since ANOVA testing does not tell us exactly which variables are significantly different, we could run further tests to find out why our p-value is zero. For example, we could create a box-and-whisker plot, which would give us a five-number summary and would visualize the difference in weight between positions. This would be the most useful way to proceed in further analysis, as it would show which variables are different from each other and the amount of overlap between them.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
P	5644	1108651	196.43	411.8985
C	892	175664	196.9327	285.0931
1B	886	178767	201.7686	451.7215
2B	1076	190334	176.8903	198.5163
3B	529	99347	187.8015	303.4738
SS	369	64831	175.6938	260.5065
LF	451	88139	195.4302	405.3657
CF	893	167005	187.0157	294.0917
OF	416	77933	187.3389	237.0969
DH	121	23263	192.2562	271.1921

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	599855.6	9	66650.63	186.6498	0	1.880714017
Within Groups	4023323	11267	357.0891			
Total	4623179	11276				

6. T-Test

When we run a Two-Sample T-Test to compare the weights between SS and 2B, we discover that the two-tailed p-value is 0.21, which is not significant at the 0.05 alpha level. This means that we fail to reject our null hypothesis, which was that the difference between our means was 0. This makes sense, as our test also shows the means and variances between the weights at our two positions are very similar to each other. With relatively basic baseball knowledge, we realize that both of these positions are defensive-minded, where quick feet and agility are needed to get to a ball in the hole or turn a double play. We would assume that the weights between positions are similar. It is much more difficult to "hide" a 250-pound power hitter at shortstop or second base than it is in left field, for example. If we ran another test comparing the weight between shortstop and left field, we could assume that our results would differ, and the p-value would most likely be much lower.

t-Test: Two-Sample Assuming Unequal Variances

	weight of 2B	*weight of SS*
Mean	176.8903346	175.6937669
Variance	198.5163344	260.50651
Observations	1076	369
Hypothesized Mean Difference	0	
df	572	
t Stat	1.268021226	
P(T<=t) one-tail	0.102653126	
t Critical one-tail	1.647521905	
P(T<=t) two-tail	0.205306251	
t Critical two-tail	1.964119952	

7. Comparing Weights Grouped by End Date of Career

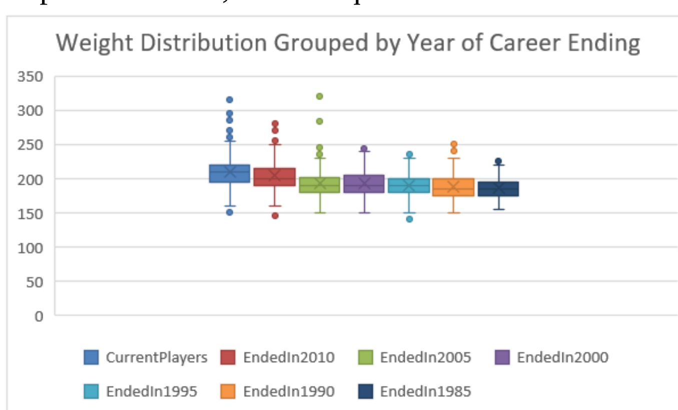
If asked to compare weights grouped by the end date of a player's career, there are many different analysis tools that we could use. For example, we could simply create descriptive statistics for each of the groups, which would look like this:

CurrentPlayers		*CareerEnded2010*		*CareerEnded2005*		*CareerEnded2000*		*CareerEnded1995*		*CareerEnded1990*		*CareerEnded1985*	
Mean	209.4185	Mean	204.4408	Mean	193.4375	Mean	192.7882	Mean	190.5722	Mean	187.5967	Mean	187.0132
Standard Error	0.582072	Standard Error	1.481535	Standard Error	1.442307	Standard Error	1.289544	Standard Error	1.33879	Standard Error	1.330701	Standard Error	1.251141
Median	210	Median	200	Median	190	Median	190	Median	190	Median	185	Median	185
Mode	220	Mode	200	Mode	200	Mode	190	Mode	190	Mode	180	Mode	190
Standard Deviation	21.1397	Standard Deviation	21.52053	Standard Deviation	20.80124	Standard Deviation	18.37318	Standard Deviation	18.30767	Standard Deviation	17.90276	Standard Deviation	15.4251
Sample Variance	446.8869	Sample Variance	463.1334	Sample Variance	432.6917	Sample Variance	337.5737	Sample Variance	335.1708	Sample Variance	320.5087	Sample Variance	237.9336
Kurtosis	0.73145	Kurtosis	1.531336	Kurtosis	7.358249	Kurtosis	0.323439	Kurtosis	0.0232	Kurtosis	0.557583	Kurtosis	-0.10137
Skewness	0.310378	Skewness	0.718617	Skewness	1.691729	Skewness	0.468627	Skewness	-0.03499	Skewness	0.464408	Skewness	0.402172
Range	165	Range	135	Range	170	Range	100	Range	100	Range	100	Range	75
Minimum	150	Minimum	145	Minimum	150	Minimum	150	Minimum	140	Minimum	150	Minimum	155
Maximum	315	Maximum	280	Maximum	320	Maximum	250	Maximum	240	Maximum	250	Maximum	230
Sum	276223	Sum	43137	Sum	40235	Sum	39136	Sum	35637	Sum	33955	Sum	28426
Count	1319	Count	211	Count	208	Count	203	Count	187	Count	181	Count	152

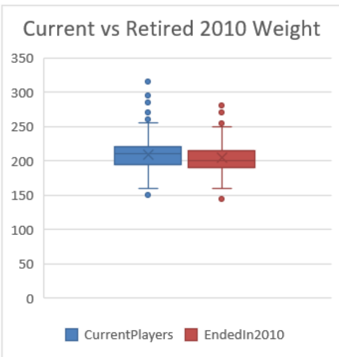
But, as we can see, this is difficult to read and even more difficult to interpret the between-group differences. If we wanted to go one step further to simplify this, we could create a five-number summary for each of the groups, which eliminates many of the unnecessary statistics, such as kurtosis and skewness. The five-number summary for each group would look like this:

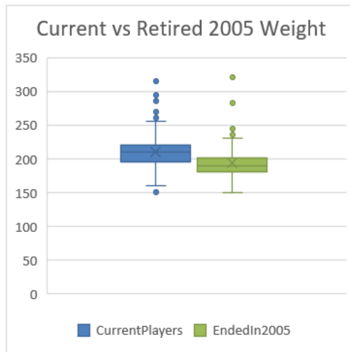
	CurrentPlayers	EndedIn2010	EndedIn2005	EndedIn2000	EndedIn1995	EndedIn1990	EndedIn1985
MIN	150	145	150	150	140	150	155
Q1	195	190	180	180	180	175	175
MEDIAN	210	200	190	190	190	185	185
Q3	220	215	200.25	205	200	200	195
MAX	315	280	320	250	240	250	230

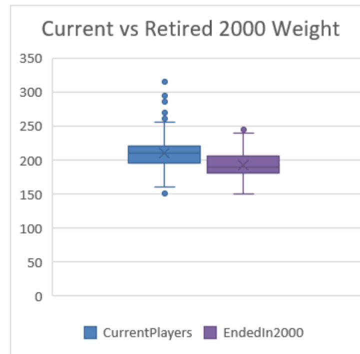
While this is marginally better than the descriptive statistics, it still requires a detailed analysis. We want our data to be easily interpreted and understood, which is where a box plot can come in handy. If we create a box plot with each of the seven groups, we get this visualization:

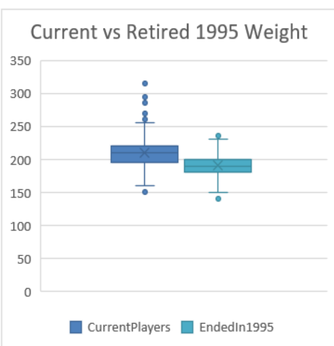


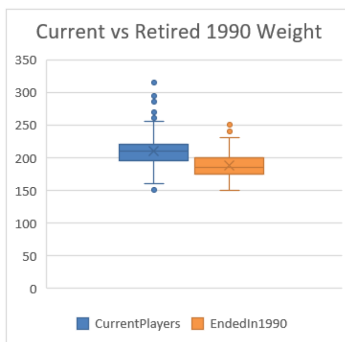
This is far more beneficial than descriptive statistics or a five-number summary, but we can take box plots one step further and break them down by individual groups, as well. We can compare the weights of current players to each of the other groups individually, which highlights key differences that may be overlooked when looking at the chart above.

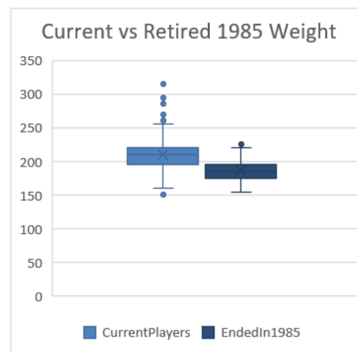












When looking at each graph individually, we see that the median weight of current players is the highest it has been in the last 30 years. There is, however, one thing that stands out. In the graph with players whose careers ended in 2005, we see a few outliers on the upper part of the box plot. As steroid use ran rampant throughout the 90’s and early 2000’s, it is possible that these points are the results of the steroid era. The graph shows that while steroid use around this time was definitely a factor, the majority of players were not using steroids. From 1995 to 2005, we see that Quartile 1 was around

180 pounds, while Quartile 3 was around 200 pounds. Using statistical knowledge, we know that around 50% of our data falls within this interval. Many people regard the Steroid Era as a tarnished time in baseball history but is important to note that while a few “bad apples” may have seen benefits of steroid use, the bulk of players were not using steroids.