# An Analysis of Reported Heights and Weights of Current NFL Players

*Albert Ding*

*November 28, 2018*

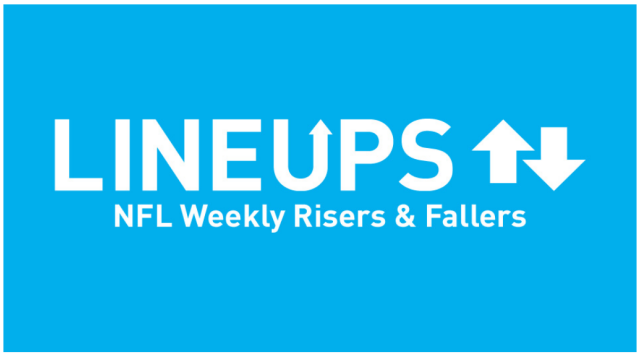
## Summary

This project examines the height and weight of NFL players for potential inaccuracy due to the self-reported nature of these figures. I establish that the distribution does not follow the Benford distribution before testing both attributes for normality based on QQ plot visualizations and the Shapiro-Wilk test. After establishing that NFL players' heights and weights were not normally distributed, I segmented the population into different groups by position and examined whether they were normal or not.

For most intra-positions heights and weights, the distributions looked close to normal distribution. For offensive line weights which did not appear normal, I then ran chi-squared tests on each quantile of the distribution to see which quantiles deviated the most in order to generate suspects the same way that Benford package would have. Lastly, I visualized and explored some of the other relationships in the player dataset.

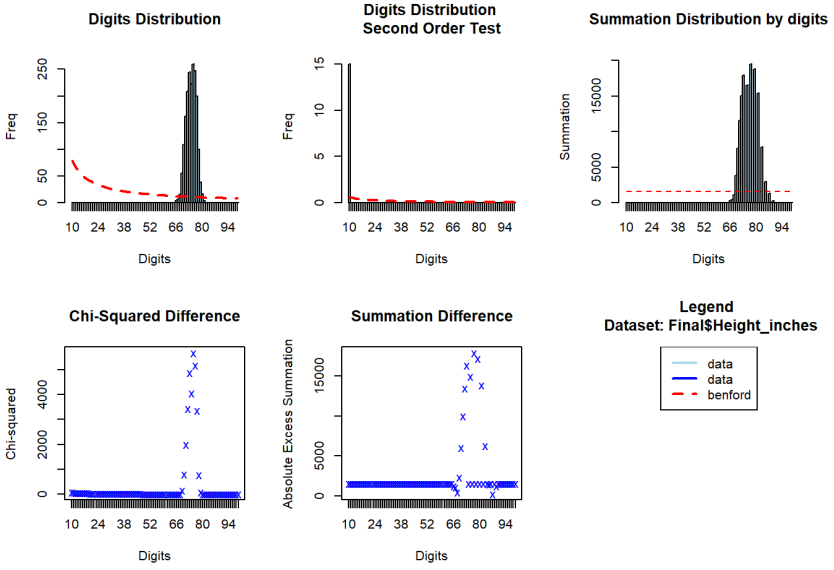
## Data Scraping and Cleaning

I scraped the data from www.lineups.com which provides analytics for fantasy sports. I chose this website over others as their data seemed complete and reliable but also easy to scrape without having too parse too much HTML.



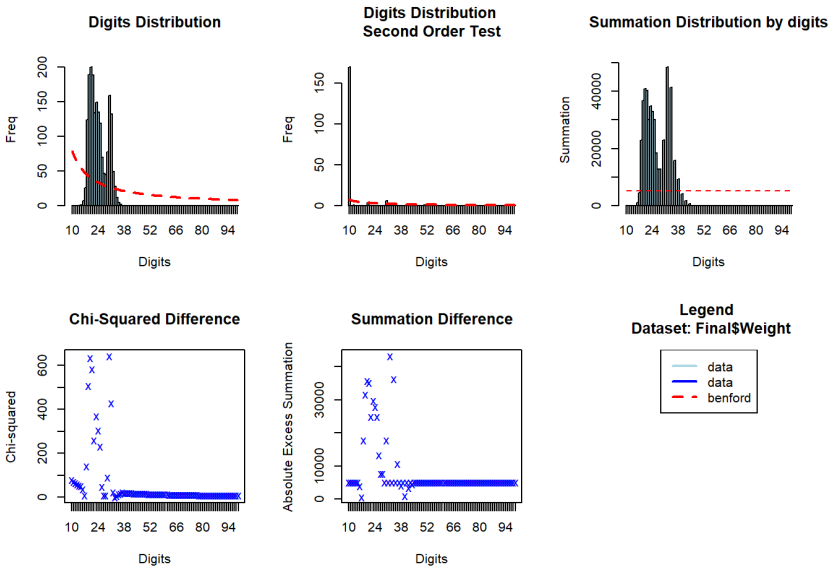

## Benford analysis on Distribution of Height in Inches

We can see that the distribution of heights does not follow the Benford Distribution. All the heights are between 66 inches and 81 inches. The plots generated from Benford are meaningless in this context, but we will apply the same method of getting suspects by focusing on portions of the distribution with the largest chi-squared test statistics later on.



# Benford Analysis on Weight in Pounds

We can see that weight doesn't follow the benford analysis either. All the weights are between 149 and 362 pounds and do not follow a Benford distribution.

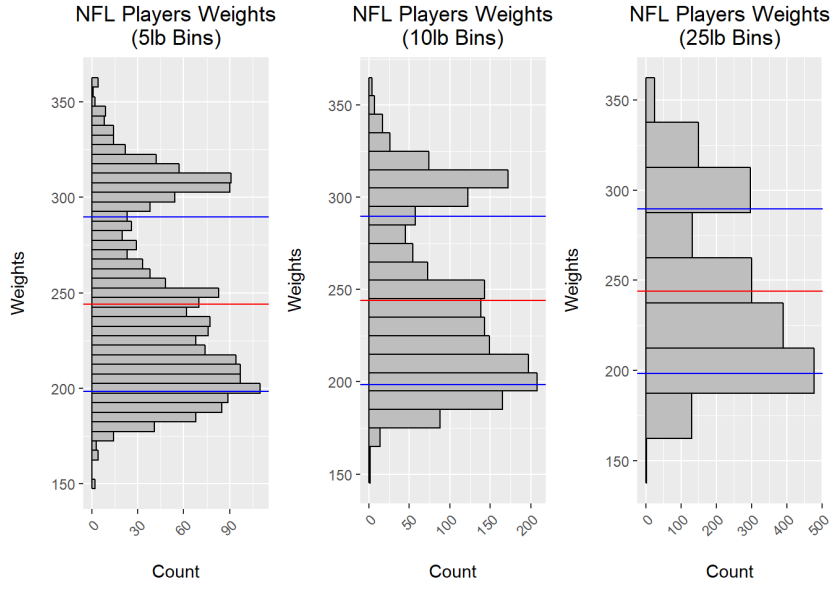



# Basic Visualizations of Distributions

## Weight Distribution

Looking at histograms of weights. Weights range from 149 to 362 lbs. Mean weight is 244 with a standard deviation of 46 lbs marked by the red and blue lines below respectively. Lining up different bin sizes for comparative purposes. This looks like a multi-modal distribution.

| | |
|---|---|
| Max_weight | 362 |
| Min_weight | 149 |
| Mean_weight_rounded | 244 |
| SD_weight_rounded | 46 |




We can already see that the distribution is not normal but let's see where values lie with respect to one and two standard deviations from the mean.

```
## [1] 0.98
```

```
## [1] 0.58
```

98% lies within two standard deviation and 59% lies within one standard deviation. This seems to suggest that the weights are less heavily concentrated within one sd of the mean than expected if the distribution was normal and more heavily concentrated within two sd than expected if the distribution were normal.

### Interpretation of Shapiro-Wilk Test and Q-Q Plots for Weight

Now, let's look at the QQ plot and the Shapiro-Wilk test of normality. The Shapiro-Wilk tests the null hypothesis that a sample x1, …, xn came from a normally distributed population.

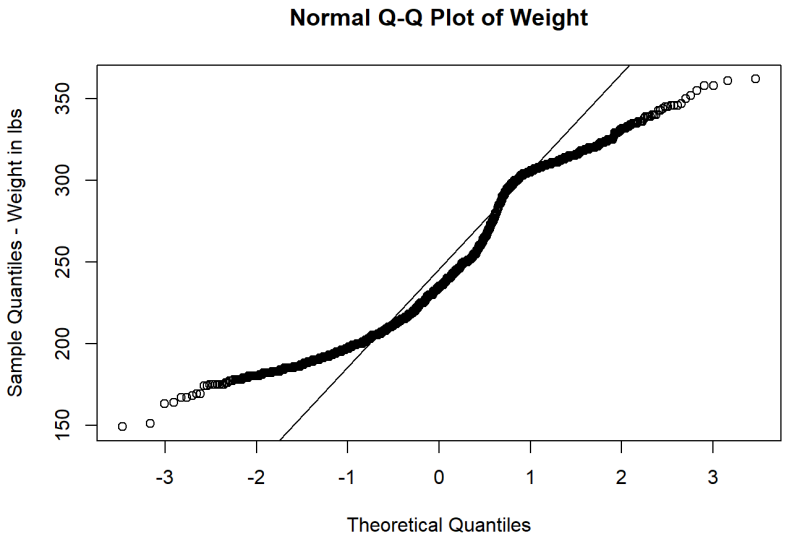
The test statistic is:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where

$x_{(i)}$ : The ith order statistic, or ith-smallest number in the sample. $\bar{x} = (x_1 + \ldots + x_1)/n$ is the sample mean. The coefficients ai are given by: $(a_1 + \ldots + a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$ where $m = (m_1, \ldots, m_n)^T$ is a vector made of the expected values of the order statistics of i.i.d random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

We can see that the p-value is close to 0 for weight. The QQ plot appears to agree as the line deviates significantly from the normal line. We can reject the null hypothesis that the samples of weight comes from a population which has a normal distribution. This aligns with what we saw in the histograms visually.

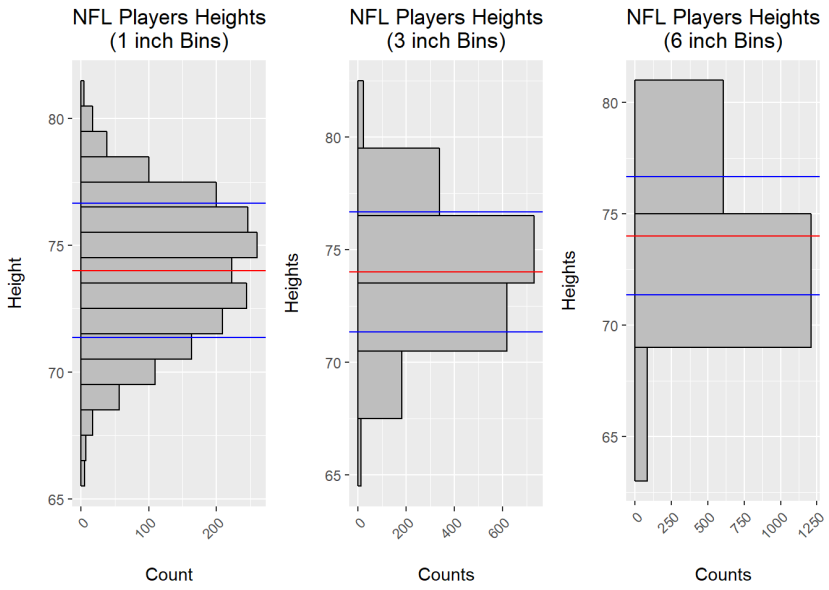



```
##
##  Shapiro-Wilk normality test
##
## data:  Final$Weight
## W = 0.93497, p-value < 2.2e-16
```

## Height Distribution

Let's look at heights now. Heights range from 5 feet 6 inches (66 inches) to 6 feet 9 inches (81 inches). Mean weight is 6 feet 2 inches (74 inches) with a standard deviation of three inches. Lining up different bin sizes for comparative purposes. This definitely looks closer to a bell curve shape than the weight distributions and doesn't exhibit the multiple peaks either. We do note however that there are only 16 values that the height takes in this distribution (66 - 81 inches in full inch increments) because height is reported to the nearest inch.

Max_height_ 81
Min_height_ 66
Mean_height_74
SD_height_ 3



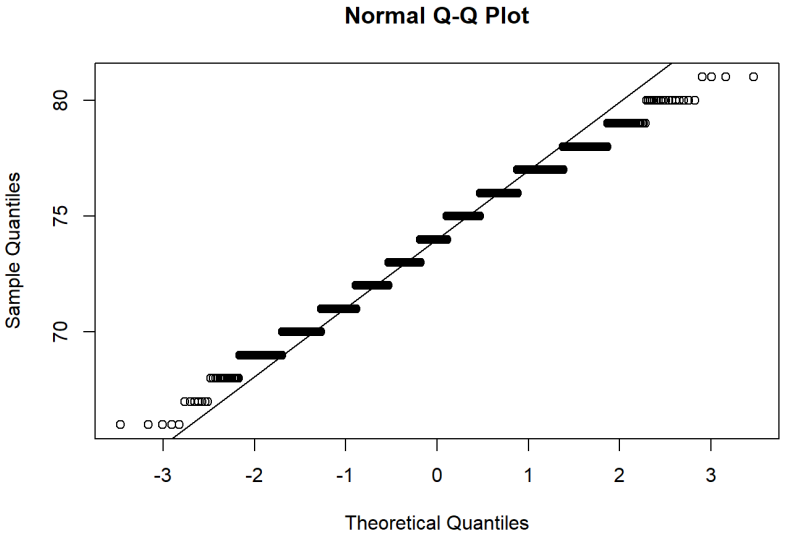

Let's see if the distribution obeys the rules of thumb for 68% within one standard deviation and 95% within two standard deviation. We see that 97% lies within two standard deviation and 62% lies within one standard deviation. That seems to be not too far off the normal distribution.

```
## [1] 0.97
```

```
## [1] 0.62
```

## Interpretation of Shapiro-Wilk Test and Q-Q Plots for Height

Let's look at the QQ plot - there's a weird pattern occuring that is causing the deviations to all occur in a straight line. Looking at the Shapiro-Wilk test for normality it looks like the null hypothesis that the samples are being drawn from a normal distribution is also being rejected.

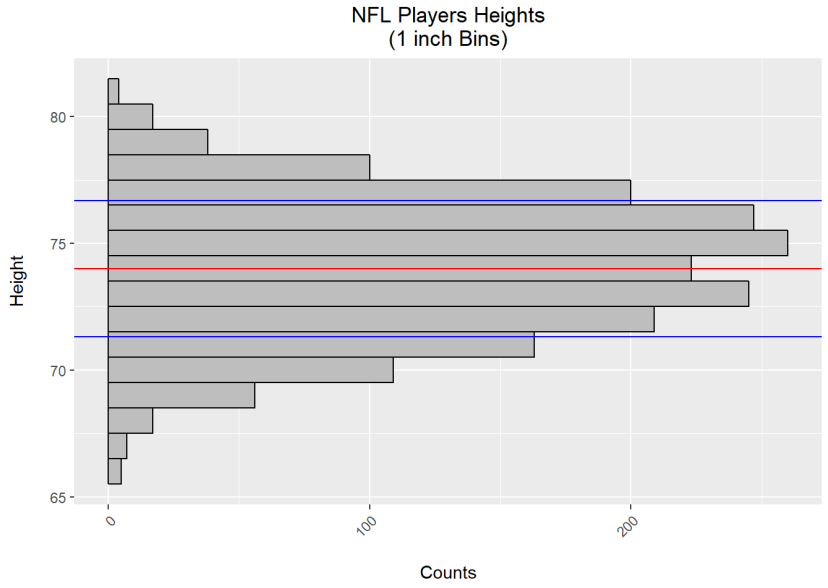


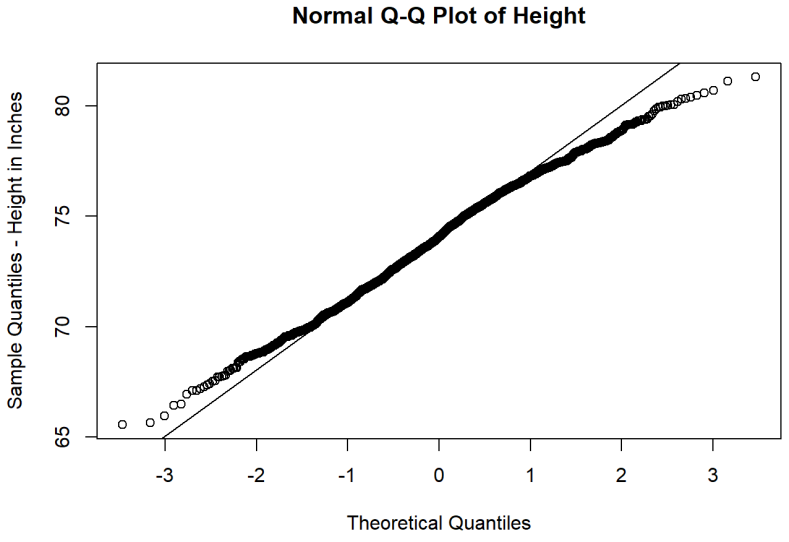

```
##
##  Shapiro-Wilk normality test
##
## data:  Final$Height_inches
## W = 0.9818, p-value = 7.898e-15
```

## Adding Decimals to Height Based on Random Draws from Uniform Distribution

We suspect that this has to do with the discreteness of the heights because there are only 16 values in the height distribution. We add some random variation from the uniform distribution for +/- 0.5 inches. We hypothesize that this might be in line with what happens in the NFL since heights might be rounded slightly up or down.

The histogram looks like a bell curve and the QQ plot looks mostly normal but with light tails. The p-value is still quite small for the Shapiro-Wilk test and we reject the normal hypothesis that this distribtuon could have been drawn from a normal distribution.

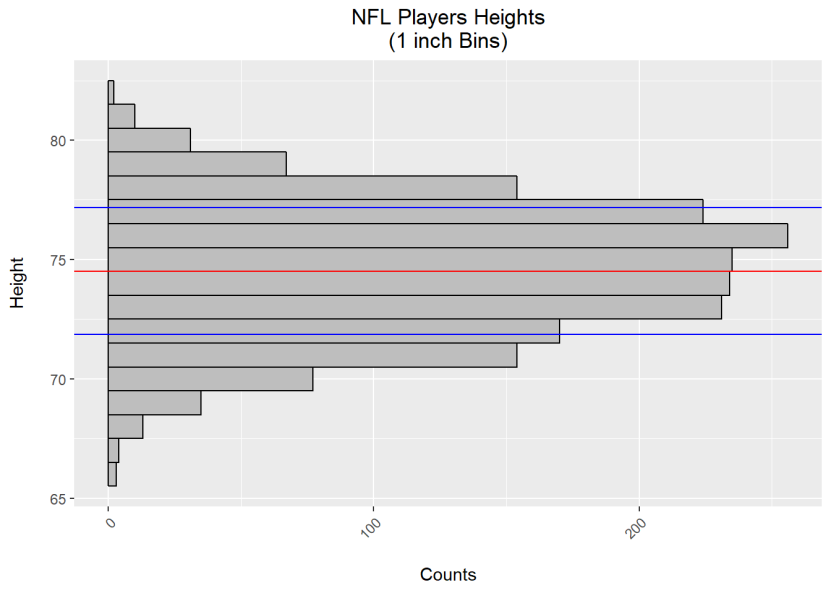


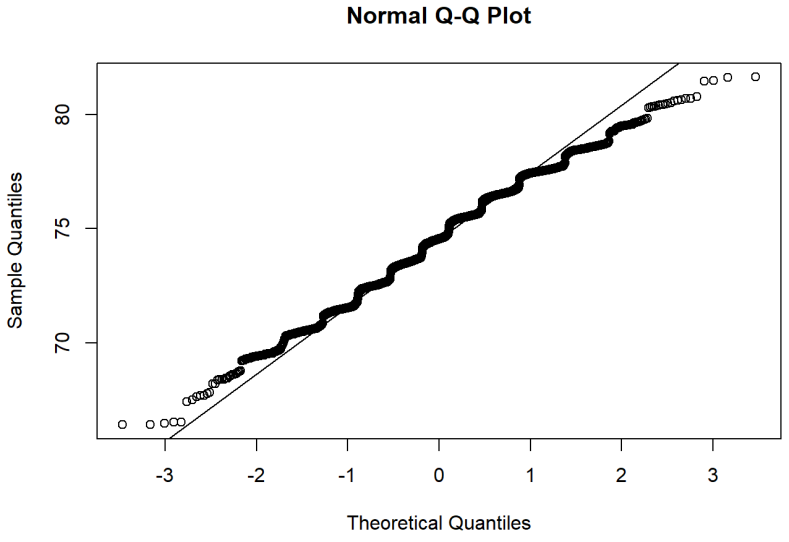




```
##
##  Shapiro-Wilk normality test
##
## data:  Final$Height_inches_uniform
## W = 0.9938, p-value = 3.786e-07
```

## Adding Decimals to Height Based on Random Draws from Normal Distribution

How about if we draw decimal endings for the height from the normal distribution? Still looks like there's weird tails but for both random draws for decimals, there appears to be fat tails as we can see from the deviation from the QQ normal plot.




```
##
##  Shapiro-Wilk normality test
##
## data:  Final$Height_inches_normal
## W = 0.99096, p-value = 1.78e-09
```

Let's do a quick check to see what percentage of values lie within one and two standard deviations. We still get the same values 97% within two standard deviations and 62% within one.

## Conclusions from Basic Visualizations of Height and Weights

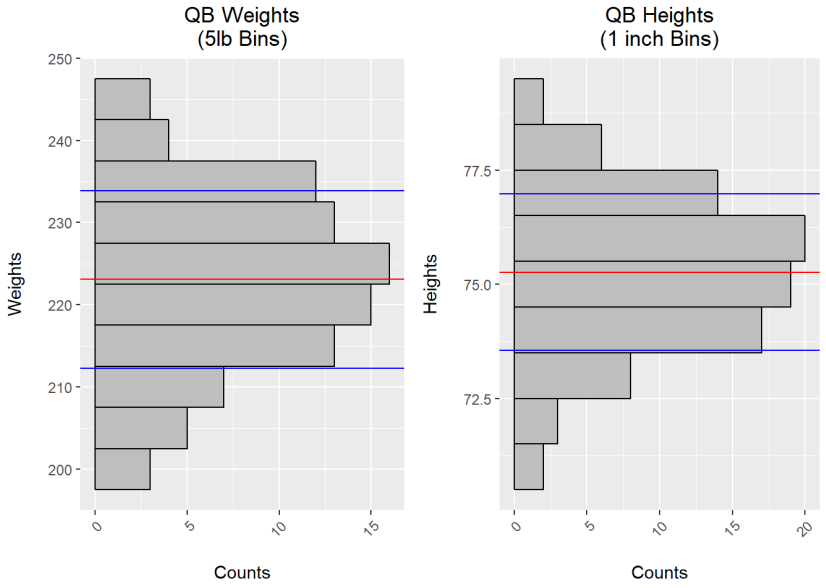
We know that weights are definitely not normally distributed and heights appear a bit off as well even when correcting for the discreteness caused by rounding of reporting heights. We hypothesize that it may be that each position is drawn from a different distribution as quarterbacks tend to be a certain height and size - six feet four and two hundred ten pounds might be a protoypical size while running backs might be expected to be five feet ten and two hundred twenty poinds. Let's take a look at the analysis by position. I focus on positions on offense for this analysis just to narrow the scope slightly and avoid needless repetition.

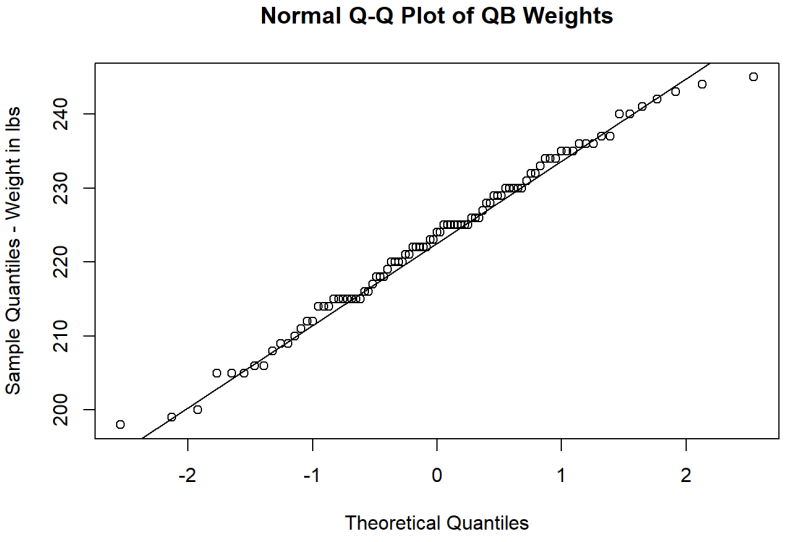
# By Position Visualizations

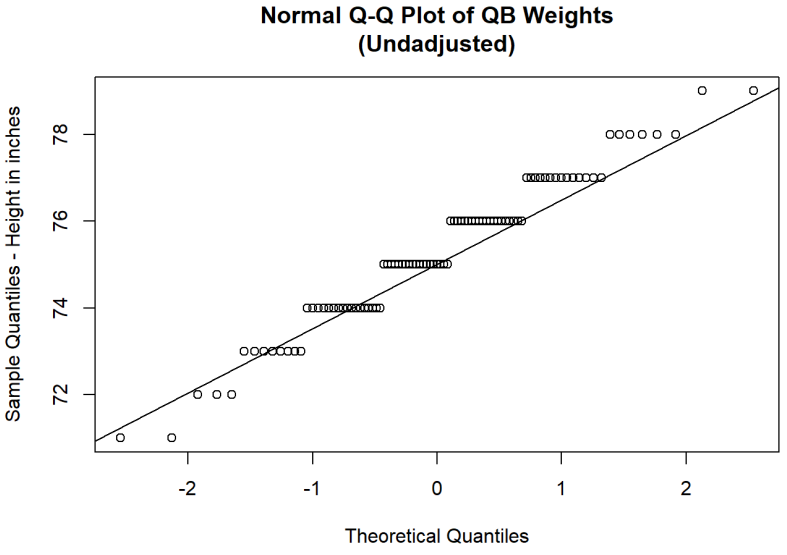
## Quarterback Visualizations

We'll start with the QB - these weights are far more normally distributed. We fail to reject the null hypothesis for quarterbacks.

For the heights, we see something unusual. The distribution looks similar to the weights but the bucketing of the categories looks discrete which alters the shapiro-wilks calculation. Testing for the typical normality values about 95% are within two standard deviations and 62% within two standard deviations.



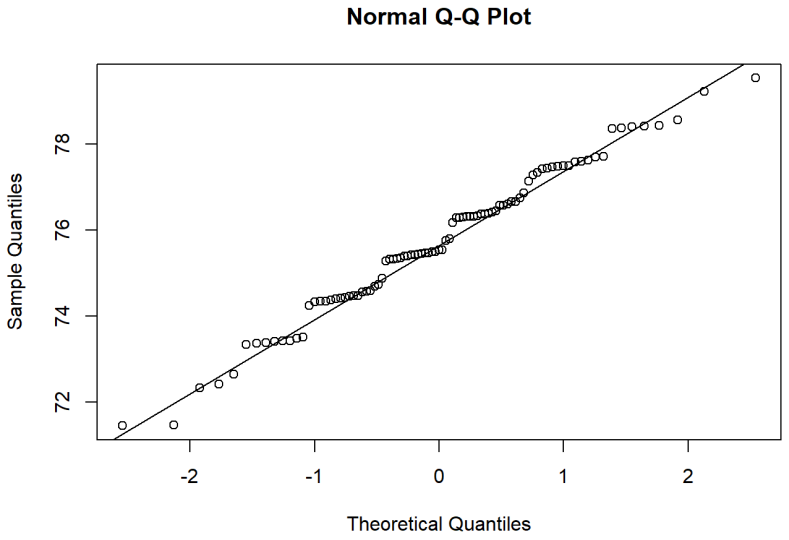

```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Weight
## W = 0.98809, p-value = 0.5831
```

```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Height_inches
## W = 0.96711, p-value = 0.0211
```

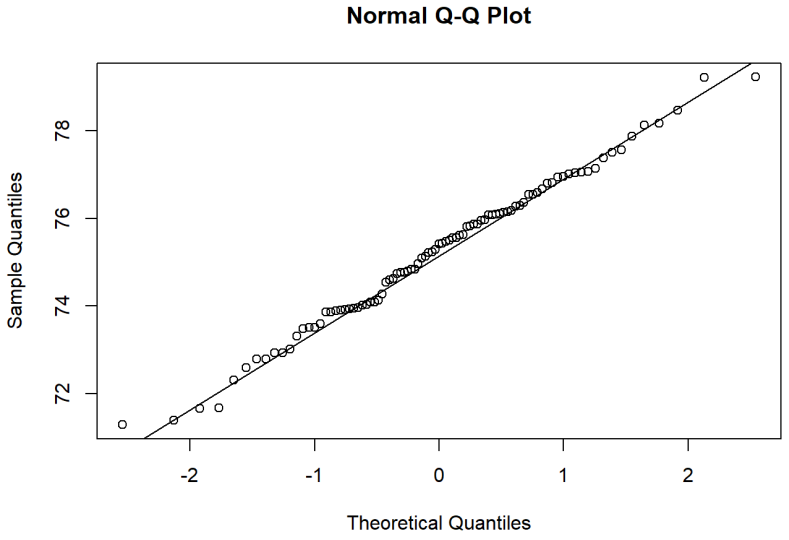
```
## [1] 0.956044
```

```
## [1] 0.6153846
```

Because the heights are constrained to such few whole numbers, the height becomes similar to a set of ordinal categories. We intuit that this may have resulted in such a low p-value in the test of normality and we add decimal endings to the height drawn from the uniform distribution and re-run the QQ plot and Shapiro-Wilk test of normality. We do the same thing for decimal endings from the the normal distribution.




```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Height_inches_Adjusted
## W = 0.98511, p-value = 0.3889
```

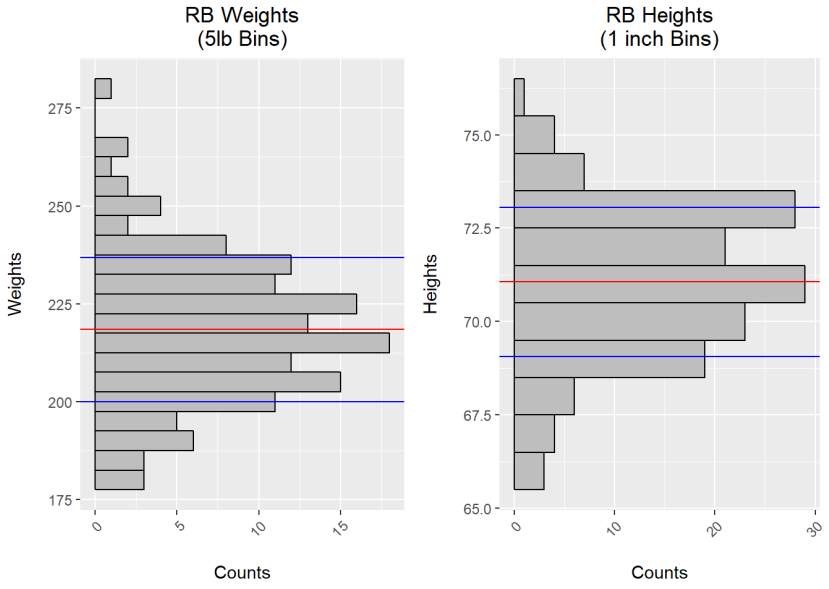


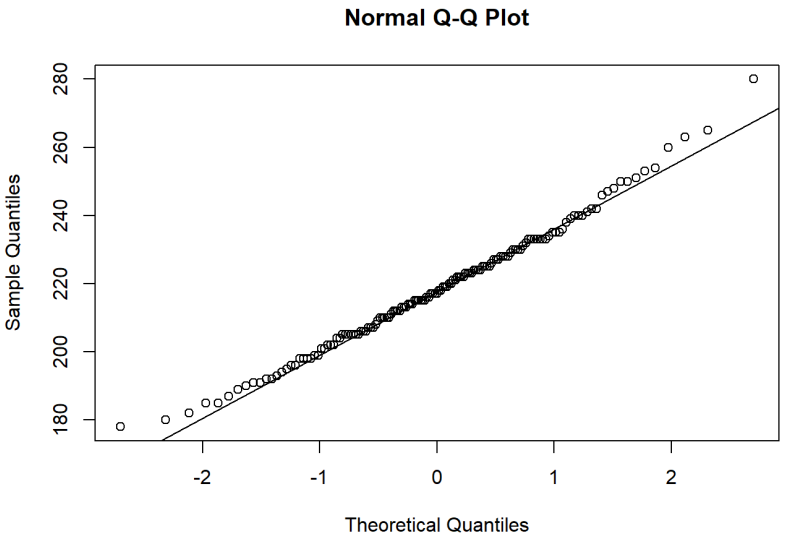

```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Height_inches_uniform
## W = 0.99083, p-value = 0.7844
```

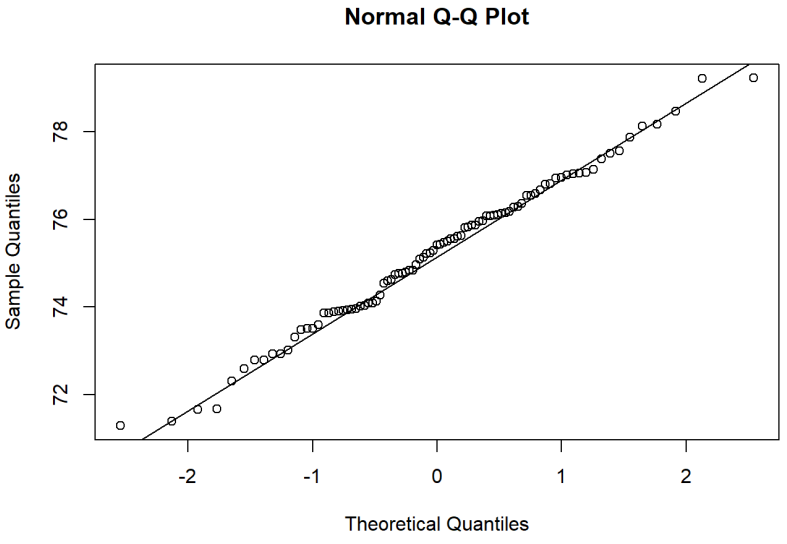
As we suspected, this fixes the normality deviation. Does this inhibit our ability to catch outliers or false data? Perhaps - we'll come back to revisit this issue.

## Running Back Visualizations

Next, let's look at running backs, inclusive of both fullbacks and running backs - these weights are far more normally distributed. We fail to reject the null hypothesis for running backs as well. We address the discreteness issue of running backs and see that the heights also appear normally distributed based on the p-value, QQ plot and histogram distribution.



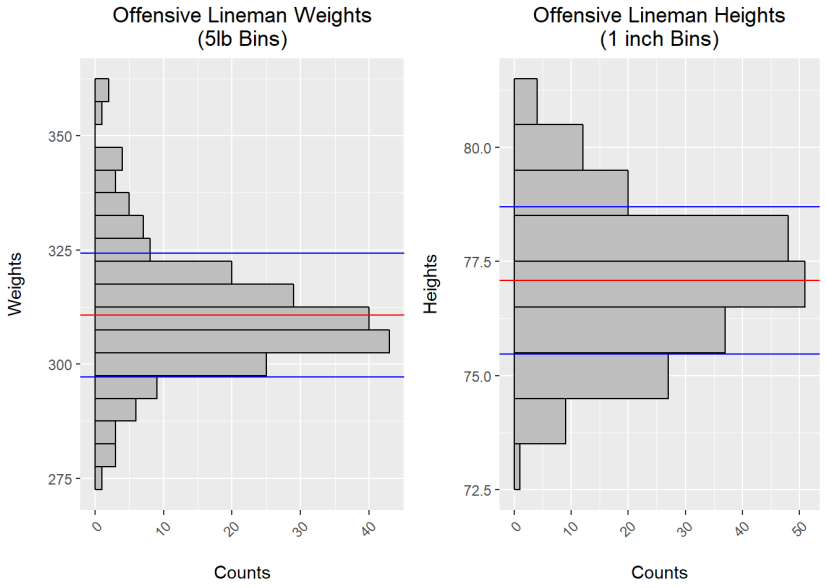



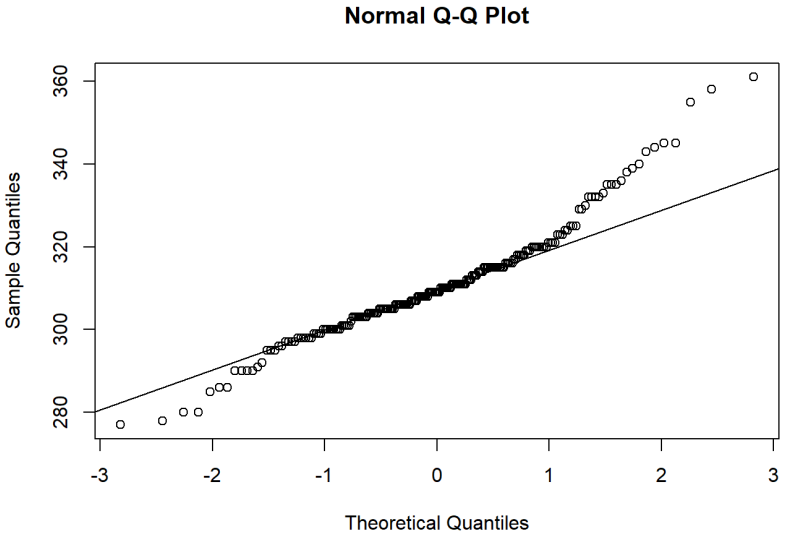
```
##
##  Shapiro-Wilk normality test
##
## data:  RB_Final$Weight
## W = 0.99022, p-value = 0.4099
```

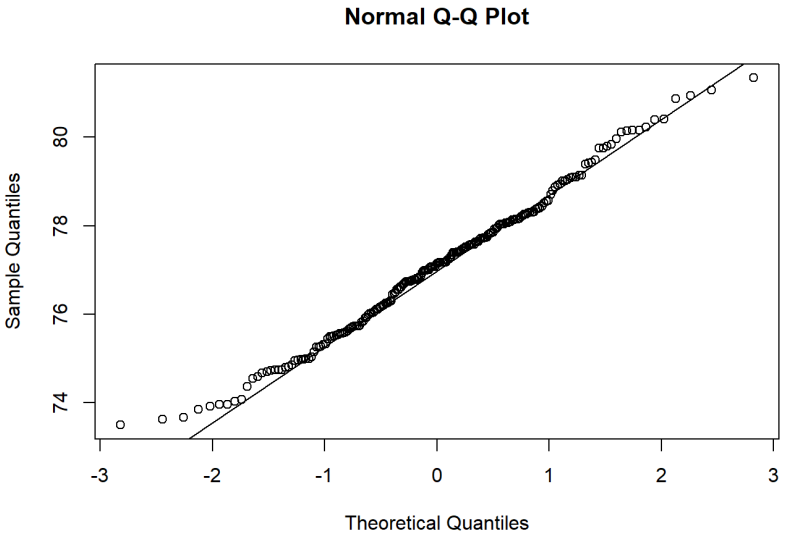
```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Height_inches_uniform
## W = 0.99083, p-value = 0.7844
```

## Offensive lineman Visualizations

Next, let's look at offensive linemen - these weights are not normally distributed. We reject the null hypothesis for offensive lineman and note from the QQ plot that the distribution exhibits heavy tails. This makes sense given the wider variation of playstyles among offensive lineman than running backs or quarterbacks. There are certain offensive lineman who might be pass block specialists who are lighter and left tackles will generally be much heavier. Heights however are normally distributed. This is an interesting place to explore further.



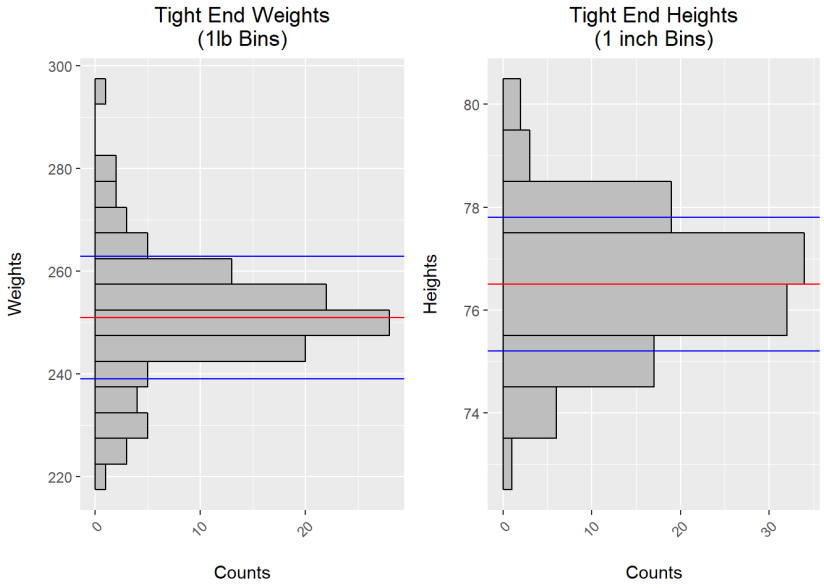



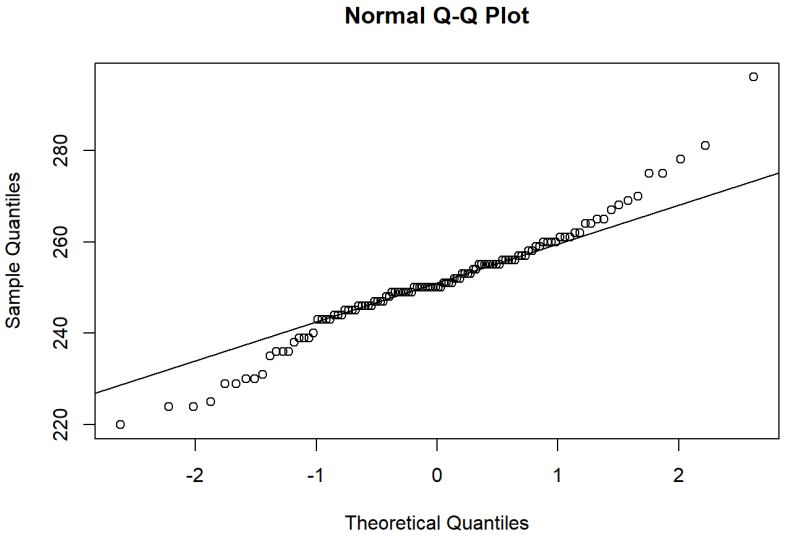

```
##
##  Shapiro-Wilk normality test
##
## data:  OL_Final$Weight
## W = 0.95225, p-value = 1.956e-06
```

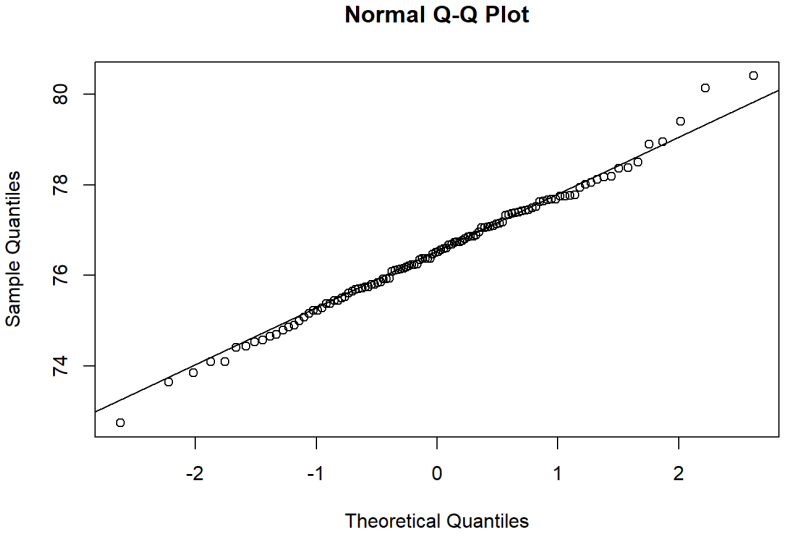
```
##
##  Shapiro-Wilk normality test
##
## data:  OL_Final$Height_inches_uniform
## W = 0.99144, p-value = 0.2569
```

## Tight End Visualizations

There's significant variation for tight ends, which does make sense, because they are stratified into different roles - blocking, receiving and hybrid. The receiving tight ends are generally a bit shorter and lighter while the blocking tight ends are generally taller and heavier. Interestingly, we rejected the null hypothesis for the weight distribution being drawn from normal due to fat tails but not the height. Tight end weight is a candidate for further examination.



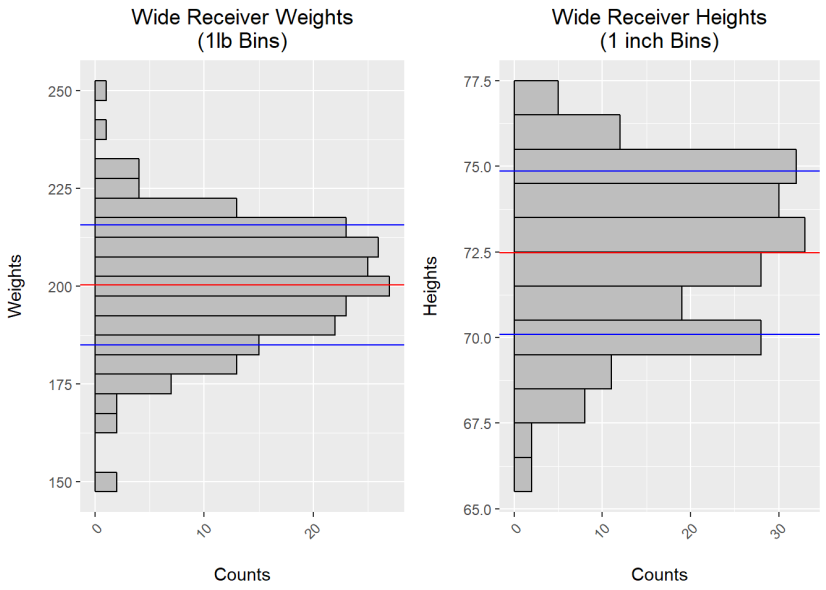



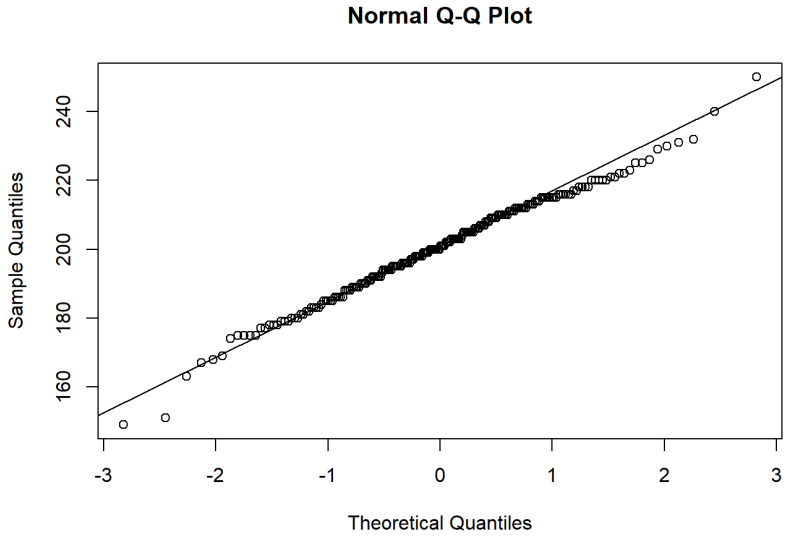

```
##
##  Shapiro-Wilk normality test
##
## data:  TE_Final$Weight
## W = 0.96545, p-value = 0.004833
```

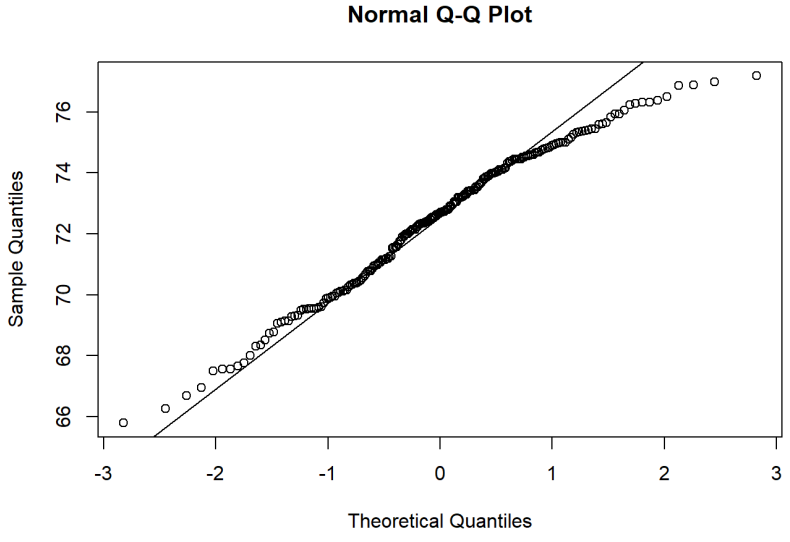
```
##
##  Shapiro-Wilk normality test
##
## data:  TE_Final$Height_inches_uniform
## W = 0.99422, p-value = 0.9204
```

## Wide Receiver Visualizations

Wide receivers have weights that appear somewhat normally distributed but not heights. Weide reciever heights is another candidate for further examination.

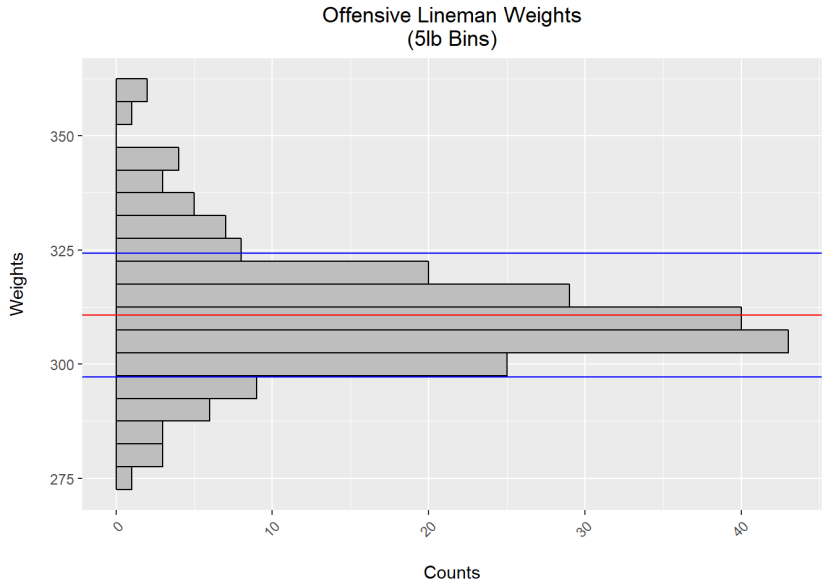



# Illustrative Investigation Of Suspects by Chi-Squared Values

This might be a good place to investigate a bit deeper on deviations from the expected quantiles of normal distribution. We choose to explore extreme deviations of weight for offensive lineman. Let's take a look at the descriptive statistics.




In my histogram, I have bucketed the weights are in five pound increments and I have 85 different pounds in the weight range so 17 buckets total. Just looking at it visually we see a slight red skew and hypothesize that the values on the upper extreme might be prime suspects.

I'm going to use each bucket to calculate what the expected density of their midpoint and multiply that by the area to get how many samples are expected in that portion of the distribution. This will help us find the abnormal deviations above and below the normal density values. I convert these to chi squared values.

The expected densities are calculated using the cumulative distribution function. For standard normal, this is:

$$F(X) = \int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

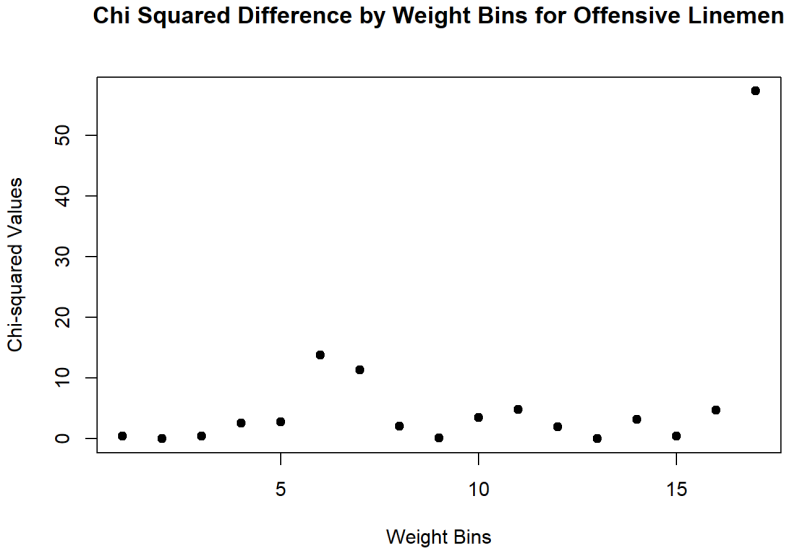
This has no closed form solution and is computed numerically. We can subtract adjacent CDFs to get the expected area in a certain range and then multiply by the number of players to get how many players should be in that height or weight range if it follows the normal distribution.

Once we get the expected, we can generate chi squared values based on the following test statistic:

$$\chi^2 = \frac{(Observed - Expected)^2}{Expected}$$

We then select the bins with the largest chi squared test statistic values - this is how the benford analysis package selects suspects.

Let's take a look at plot of the chi-squared deviations for each range of weights. The buckets are numbered in chronological order in five pound increments so for example bucket one is 277-281 lbs. BUcket two is 282-286 lbs and so on and so forth. We see that bucket 17 has by far the largest chi squared value and investiage this further.




This would be the equivalent of the top value of the get suspects function in the Benford analysis. Rather than relying on pre-written function, we've rolled our sleeves up and found the suspects ourselves. We see that the expected value at this range of a normal distribution is far less than one player, so it is suspicious that we hvae two players here.

| | Name | Pos | Team_name | Height_inches | Weight | Age | Exp | Rating | Depth | Drafted | Draft_round | Draft_pick | College | Height_inches_uniform | Hei |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

| | Name | Pos | Team_name | Height_inches | Weight | Age | Exp | Rating | Depth | Drafted | Draft_round | Draft_pick | College | Height_inches_uniform | Heig |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 129 | Marcus Cannon | OT | new-england-patriots | 77 | 358 | 30 | 8 | 80 | 1 | 2011 | 5 | 138 | TCU | 77.05385 | |
| 171 | Zach Banner | OT | pittsburgh-steelers | 80 | 361 | 24 | 2 | 69 | 1 | 2017 | 4 | 137 | Southern Cal | 80.14967 | |

The two players are Marcus Cannon of the New England Patriots and Zach Banner of the Pittsburg Steelers - shown below.

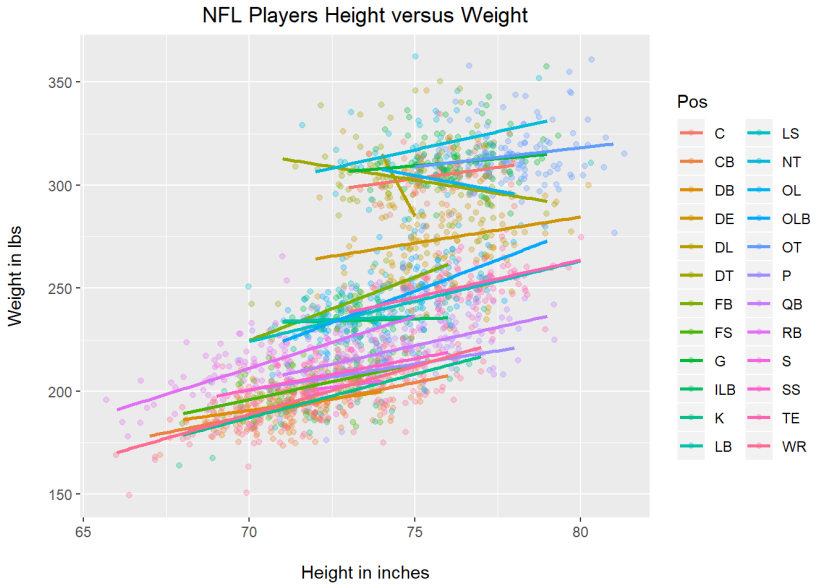

Doing some quick Google searches of both, we see that their listed weights differ depending on the website listing and that both have struggled to keep their weight under control. In fact, they both have contractive incentives paying them bonuses to come to camp under a certain weight. Based on their listed weights on lineups.com, it doesn't seem like they have had success recently.

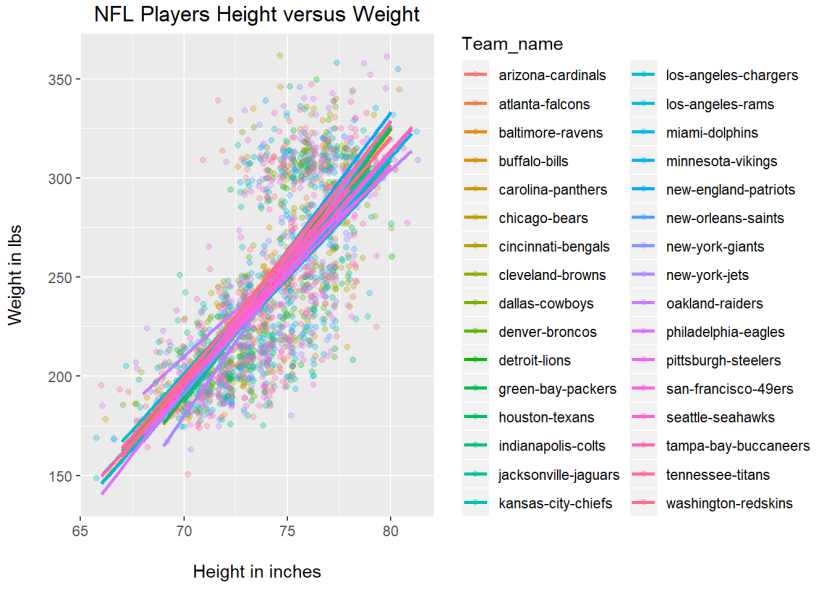
We believe our Chi-squared method has achieved some level of success given that we were able to find a couple players that had some unusual deviations in their weight.

# Visualization of bivariate relationships

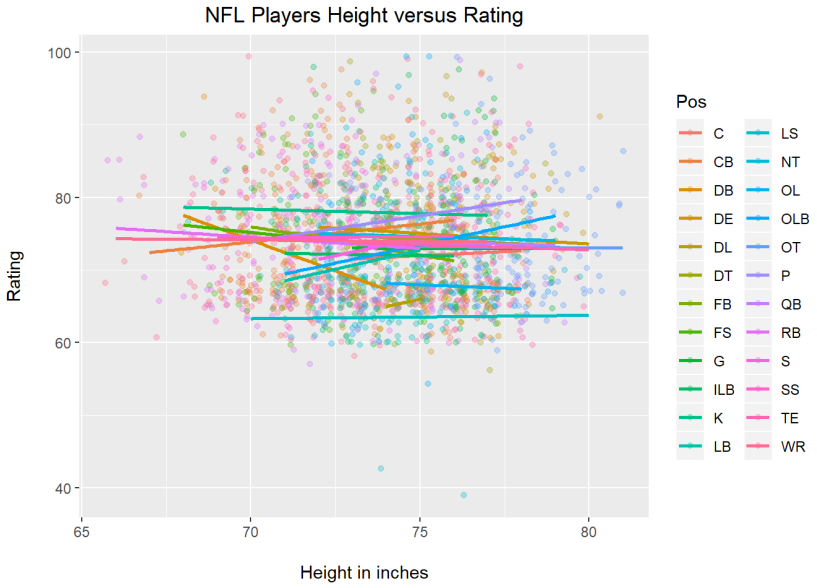
After such tedious examinations of distributions, let's have some fun with visualizations.

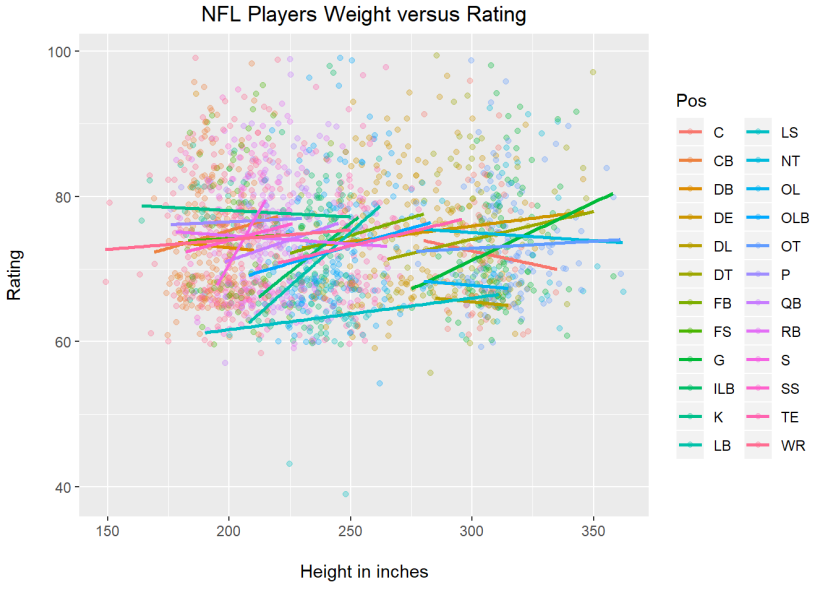
If we plot a linear regression through the data we see that the relationships between height and weight vary significantly by position but not by team. What this suggests to us is that the relationship between height and weight across teams are far more consistent than across positions. This intuitively makes sense as well. Teams have the same composition of players by position which averages out to a similar relationship for height and weight while the relationship between height and weight will be drastically different for each position. For example, some running backs might be shorter and heavier due to their play style but among quarterbacks, ones who are heavier and shorter stature will be considered too heavy or out of shape as there's no reason to carry that additional weight.



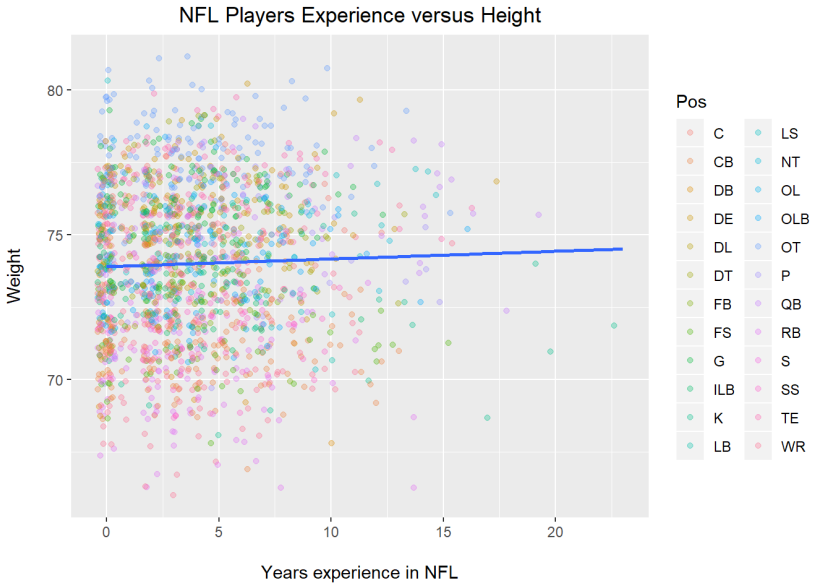



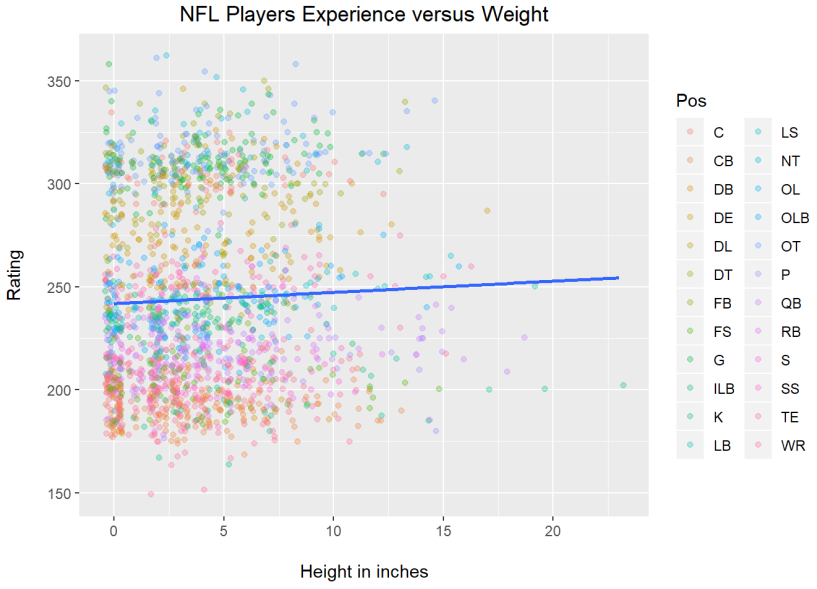

Let's take a look at the relationship between rating and height and rating and weight.



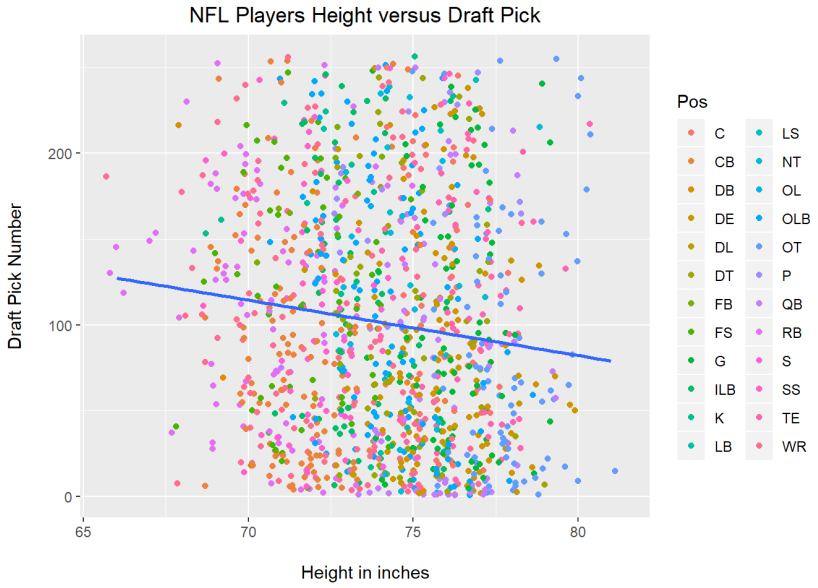



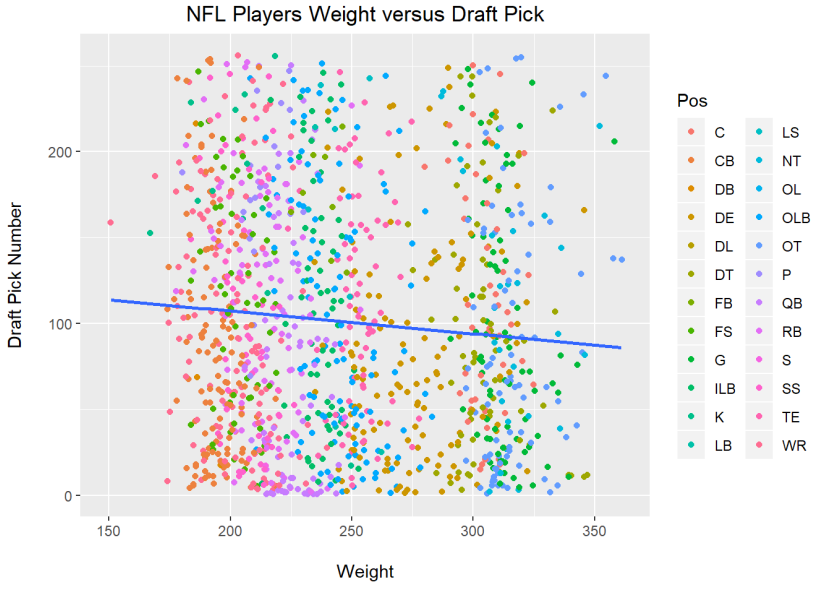

Is there a relationship between height and experience? In other words have players been getting shorter or taller over time? How about for weight? There's not really a clear trend.



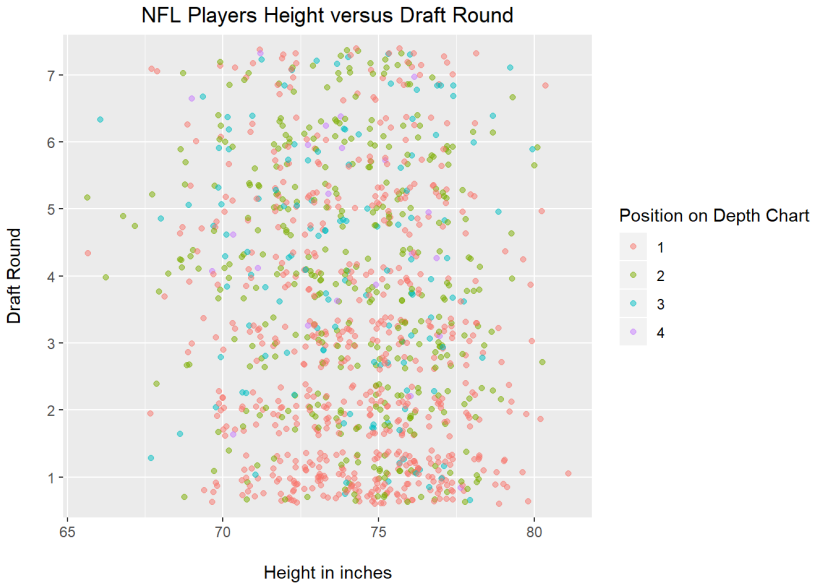



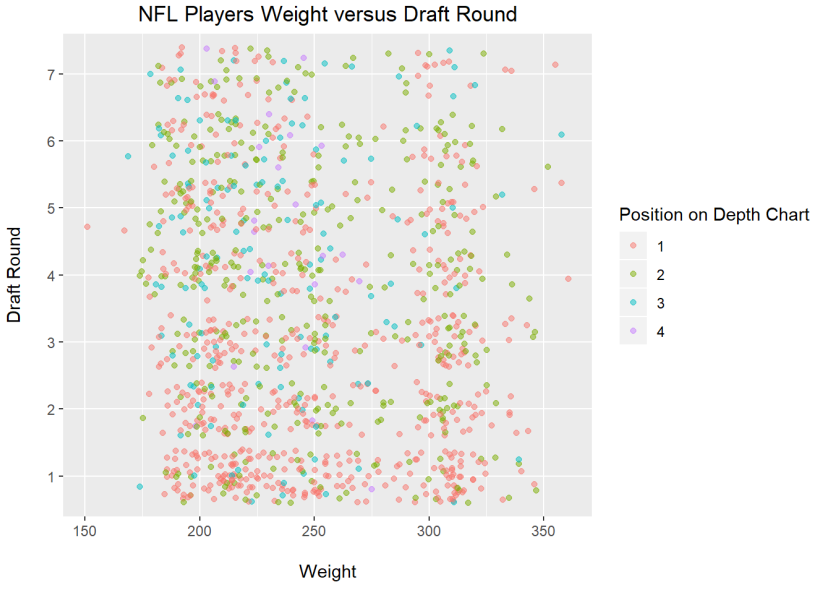

How about the relationship between draft pick number and height? We expect that taller and heavier players may have certain physical attributes that may allow them to be on average drafted earlier in the draft. The relationship is not that clear cut. Though if we allow R to calculate a linear regression through the data, it does suggest that taller and heavier players are on average drafted slightly earlier.

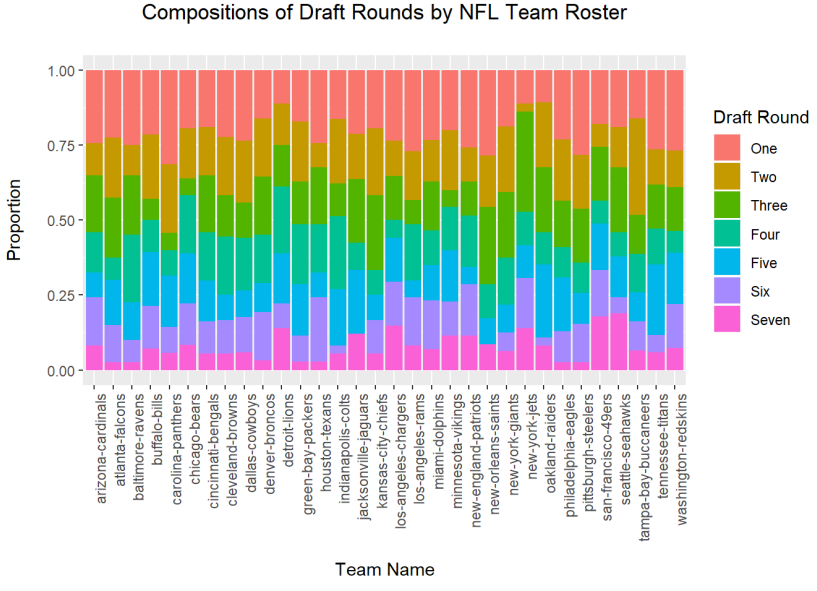


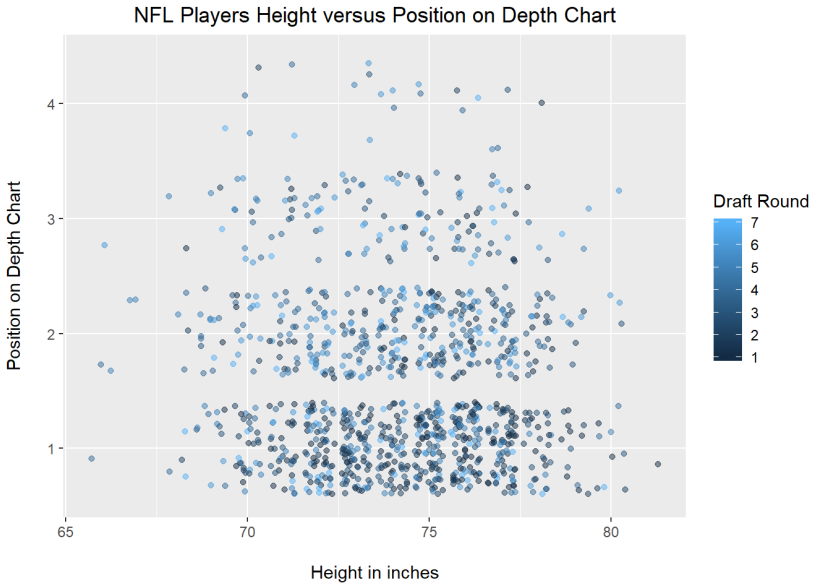

Draft round versus height and weight below with fill showing position on depth chart. We can see that height and weight have less of an effect on where you're drafted but there are definitely minimum heights and weights for getting drafted in general. There does seem to be a correlation between being drafted earlier and having a starting position. We explore this further in our next plots.

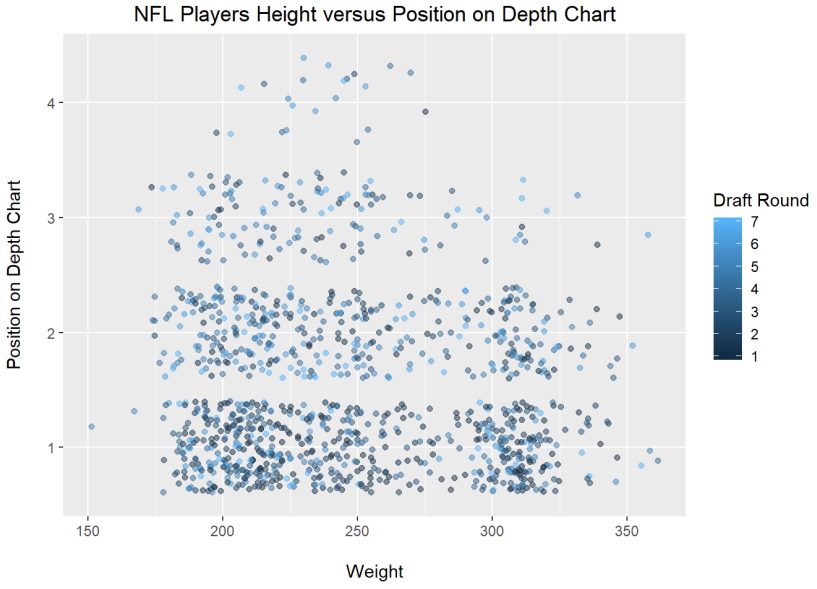



Position on depth chart versus height and weight below with fill showing round drafted. This makes it look like that heavier and taller players are slightly higher on the depth chart but that those higher on the depth chart are drafted earlier and those drafted earlier are higher on the depth chart.


NFL Players Height versus Position on Depth Chart
Position on Depth Chart
4
3
2
1
65
70
75
80
Height in inches
Draft Round
7
6
5
4
3
2
1


NFL Players Height versus Position on Depth Chart
Position on Depth Chart
4
3
2
1
150
200
250
300
350
Weight
Draft Round
7
6
5
4
3
2
1