An Analysis of Reported Heights and Weights of Current NFL Players

Albert Ding

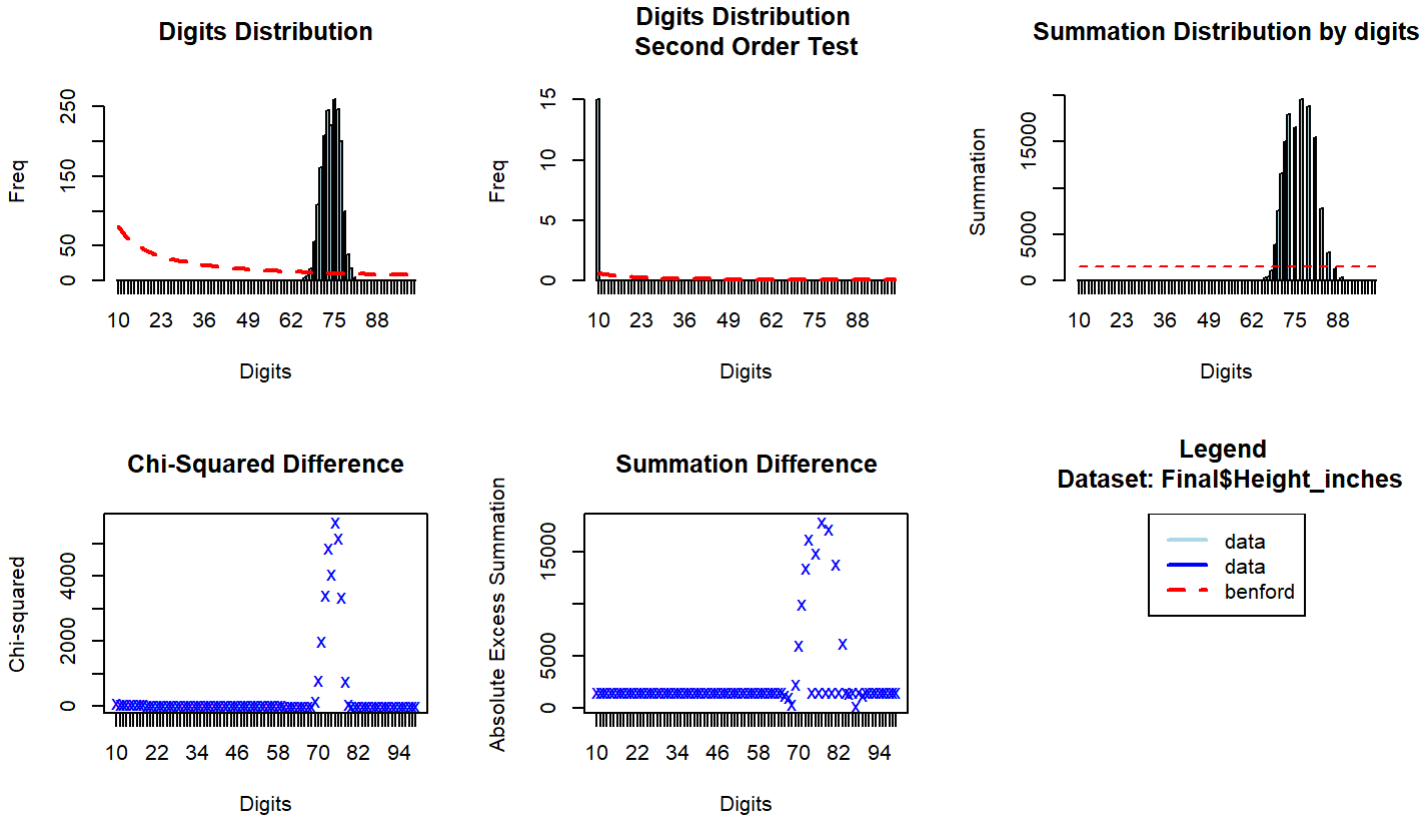
December 2018

Summary

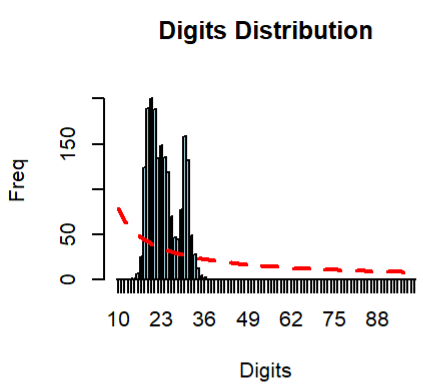
- Project examines heights and weights of current NFL players for inaccuracies arising from self-reported nature of figures
- Neither player heights, nor player weights follow Benford distribution, but they should both be approximately normal if their disributions are similar to those of the overall population of adult males
- Population further segmented into groups by player position and examined for normality
- For offensive linemen weight distribution, I reproduced the Benford getSuspects function by running Chi-squared tests on different quantiles to see which ones deviated the most from the expected normal densities
- After significant analytical investigation, I visualize and explore other relationships in the dataset

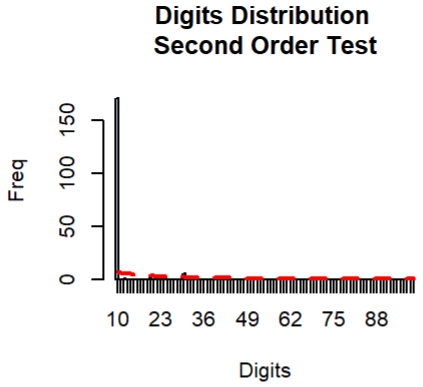
Data Scraped from Sports Analytics Website Lineups.com

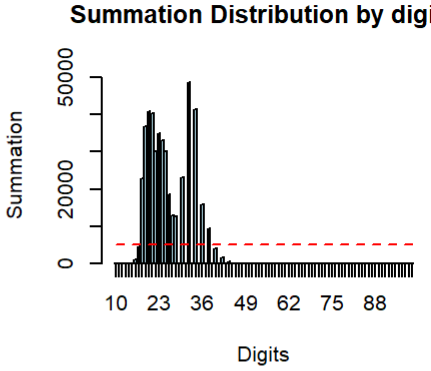
Player Heights Don't Follow Benford Distribution and Range from 66 to 81 Inches

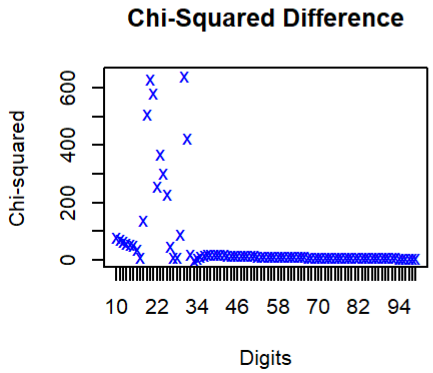


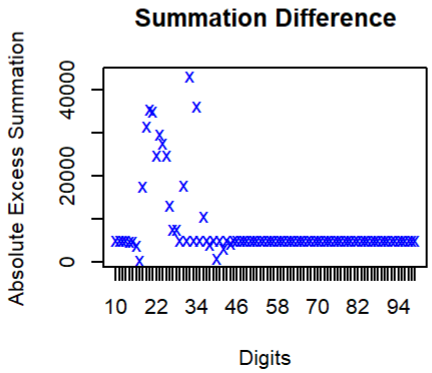
Player Weights Don't Follow Benford Either and Range from 149 to 362 lbs

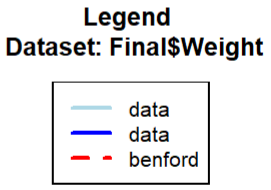




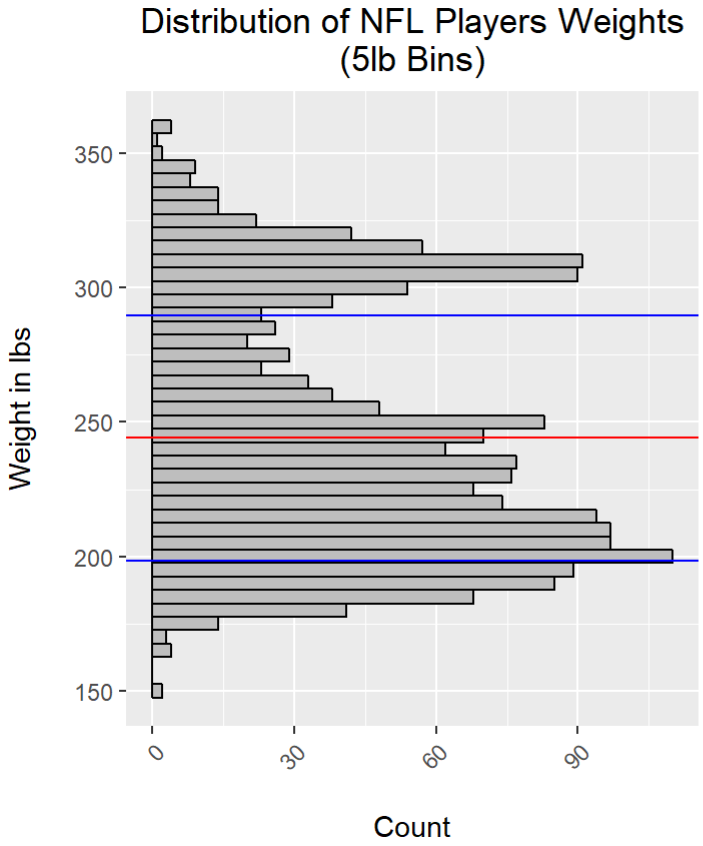


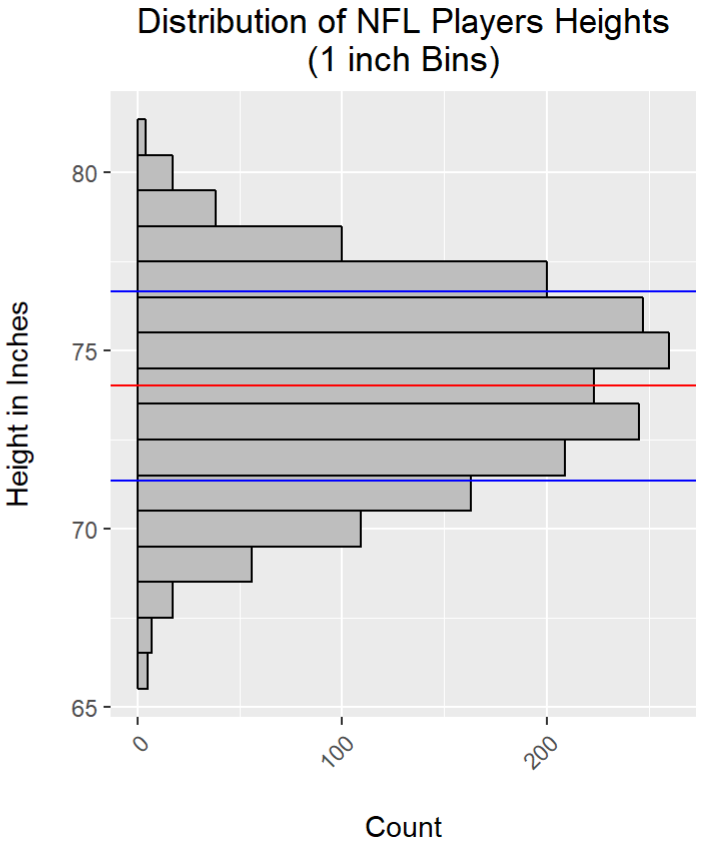




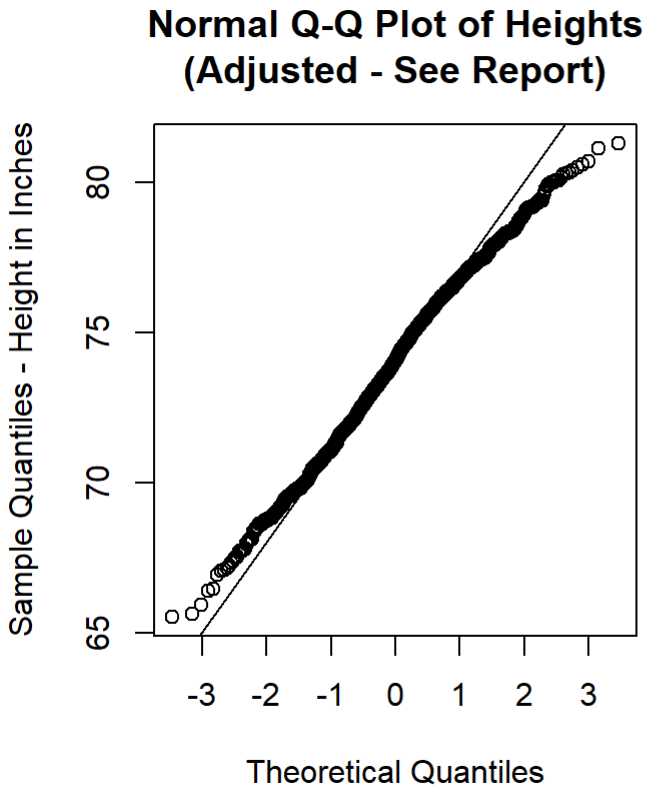


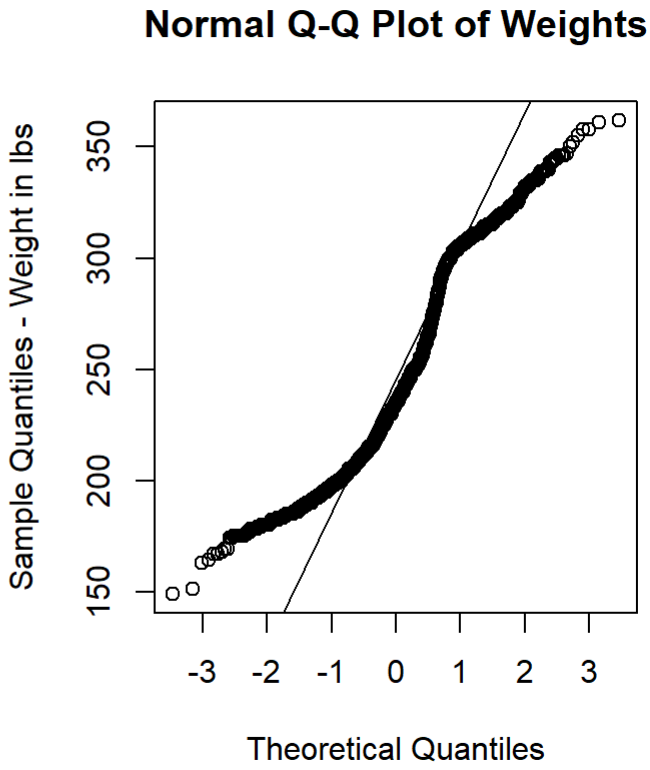
Weights Appear Multi-Modal While Heights Have Bell Shape



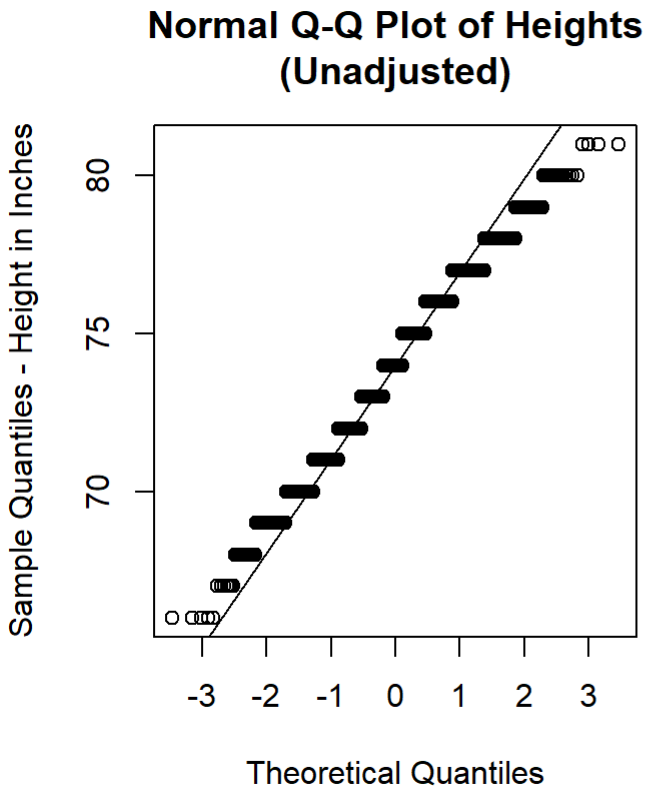


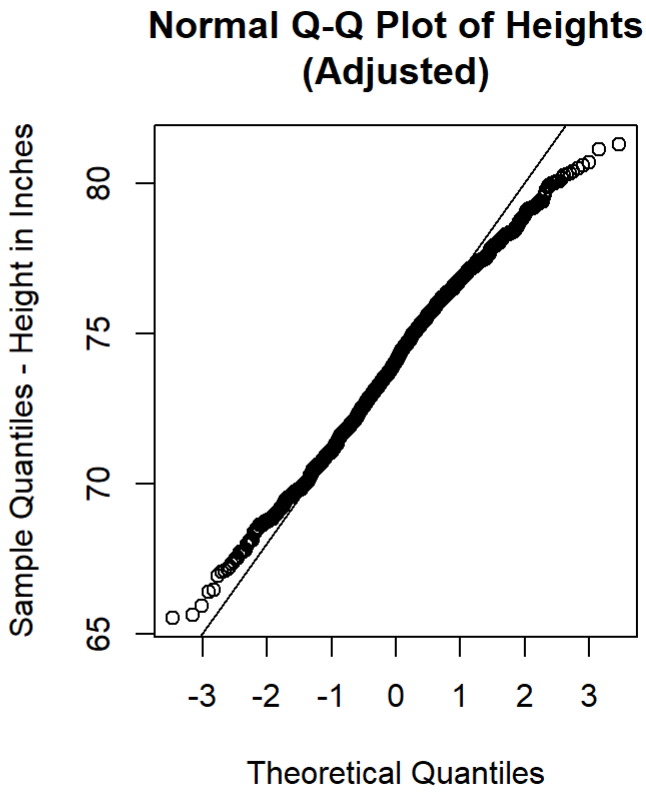
Q-Q Plots Indicate Neither Are Drawn from Normally Distributed Population





Player Heights Were Adjusted With Randomly Drawn Decimal Endings To Address Discreteness





Shapiro-Wilk Test of Normality Agrees They're Not Normally Distributed

- Shapiro-Wilk tests null hypothesis that a sample came from a normally distributed population. p-value is less than 0.05 so we can reject the null hypothesis at 95% LOC that height and weight were drawn from normally distributed population (See report for details)

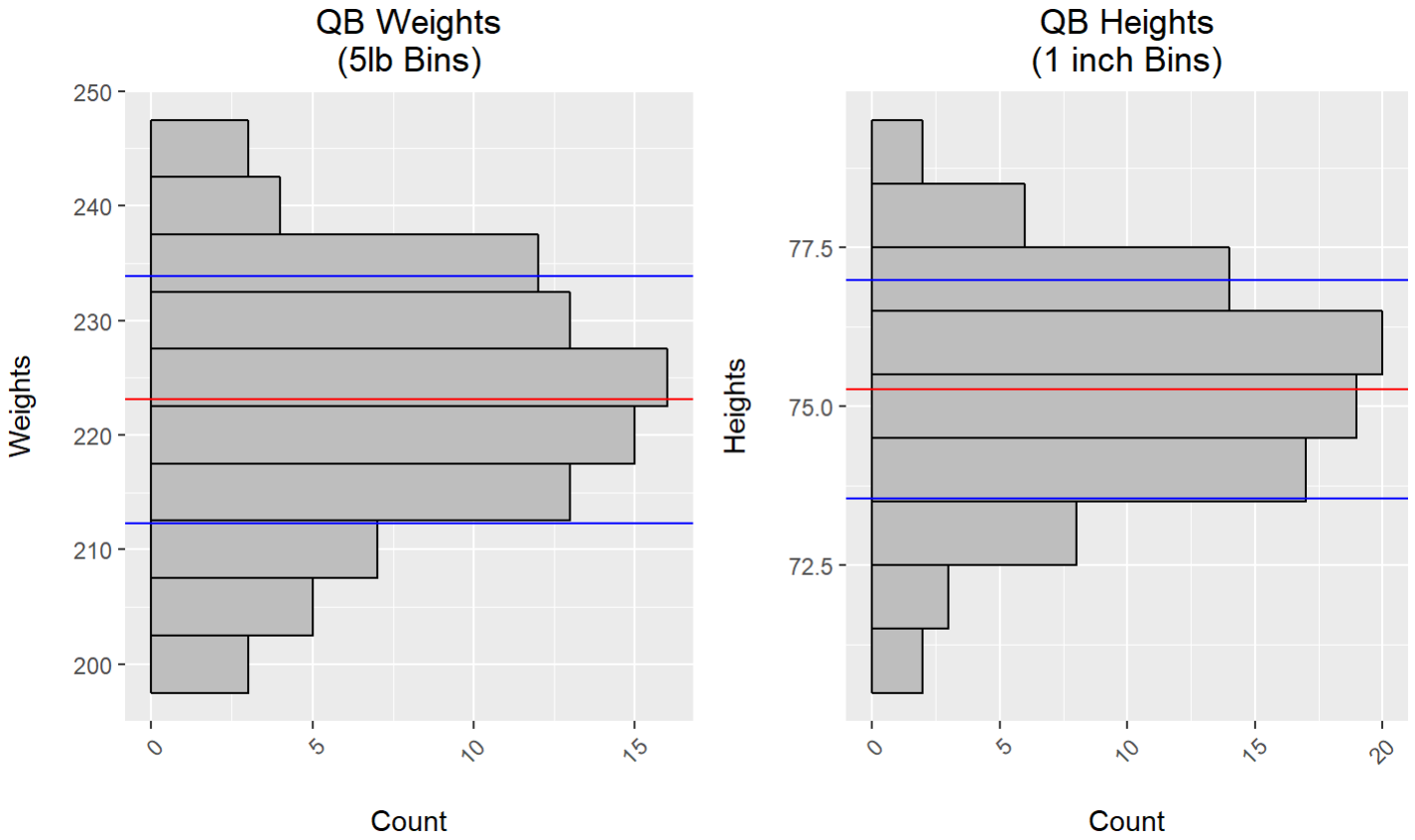
```
##
##  Shapiro-Wilk normality test
##
## data:  Final$Weight
## W = 0.93497, p-value < 2.2e-16
```

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

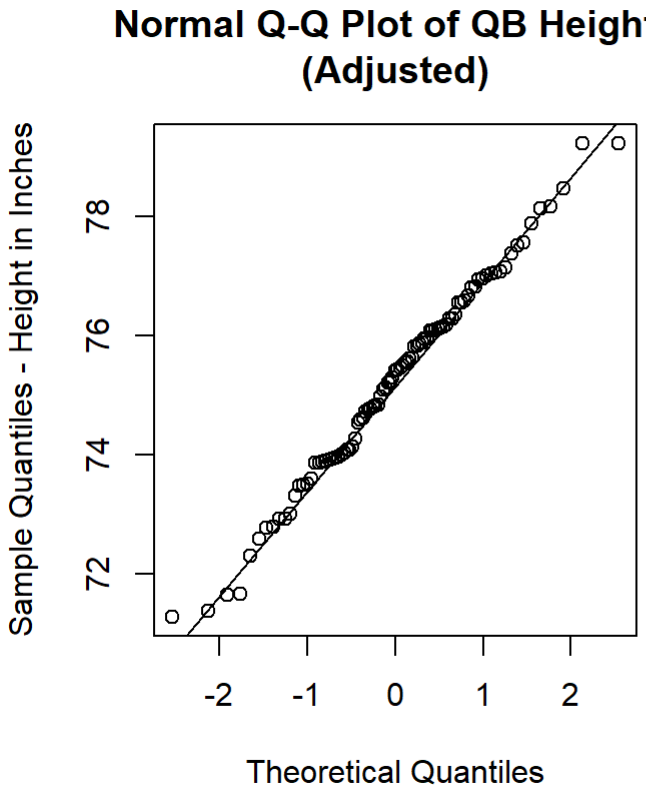
Conclusions from Basic Visualizations of Height and Weights

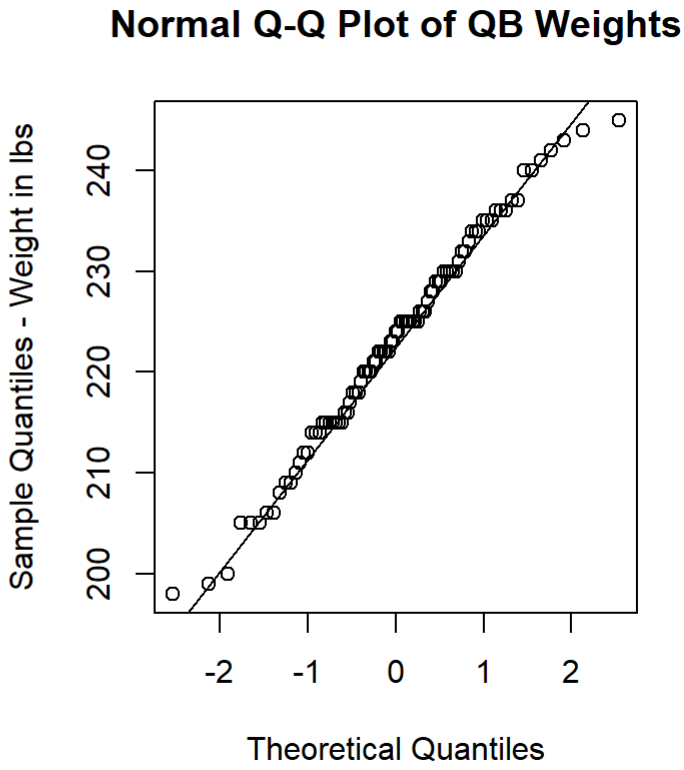
- Player weights were not normally distributed and heights appear off as well
- Height distribution has elements of discreteness caused by rounding to the nearest inch in reported height values
- There were only 16 different values for height: 66 inches to 81 inches inclusive
- I addressed height discreteness by adding decimal endings from uniform distribution centered at zero and plus / minus half an inch to mimic rounding in real life
- Next, I'll analyze the distribution of heights and weights by position

Quarterback Height and Weight Distributions Look Bell Shaped



QQ Plots for Quarterback Suggest Normal Distribution





Shapiro-Wilk Tests Agrees That Quarterback Distributions May Be Normal

- Both have p-values far greater than 0.05. Notably, if the heights are not adjusted, the p-value becomes small enough to reject the null hypothesis.

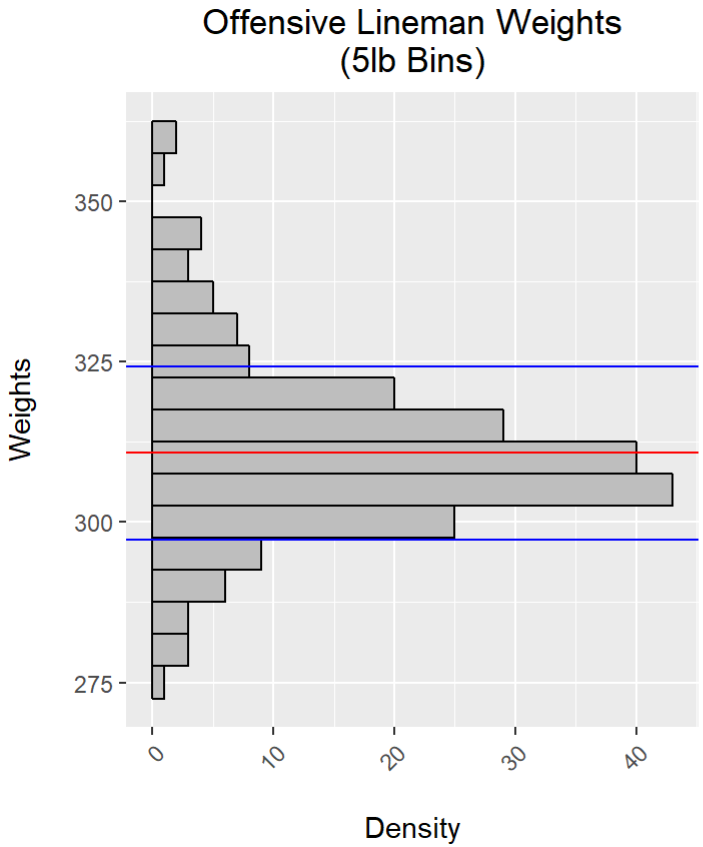
```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Weight
## W = 0.98809, p-value = 0.5831
```

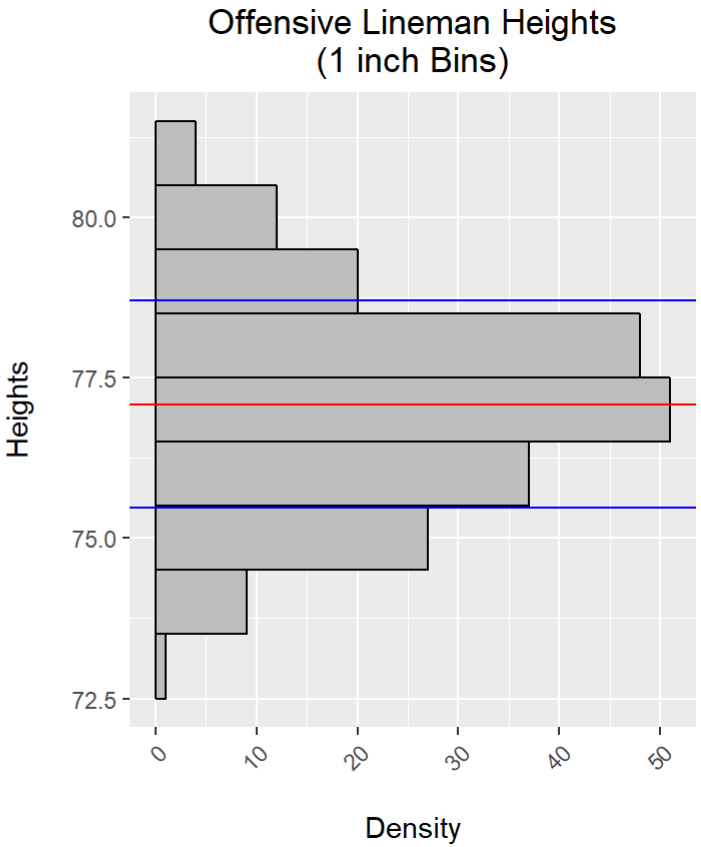
```
##
##  Shapiro-Wilk normality test
##
## data:  QB_Final$Height_inches_uniform
## W = 0.99083, p-value = 0.7844
```

In Full Report and Shiny App, We Also Examine Distributions, QQ-Plots, and Tests of Normality for Other Positions

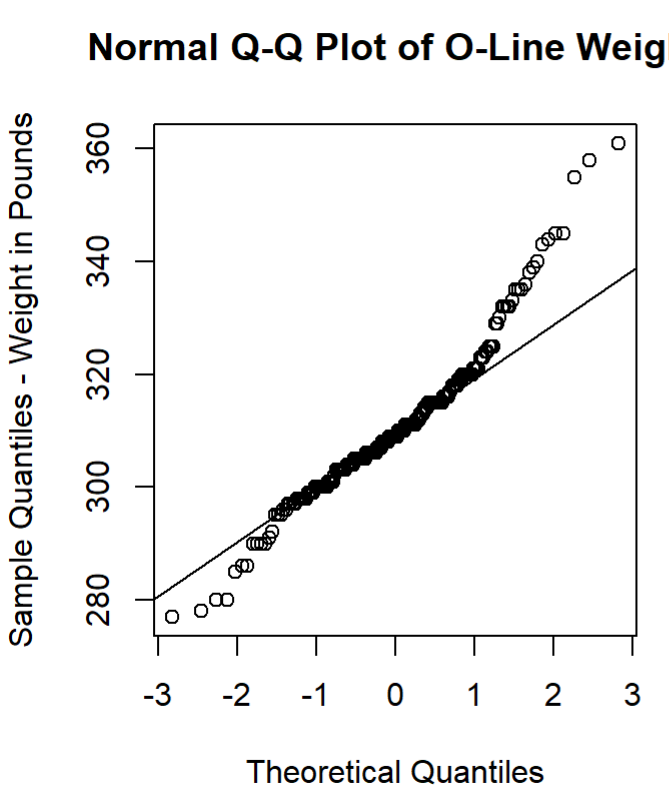
- Omitted from report for sake of compactness and brevity
- We'll look at offensive linemen next for which we apply further testing on to retrieve suspects

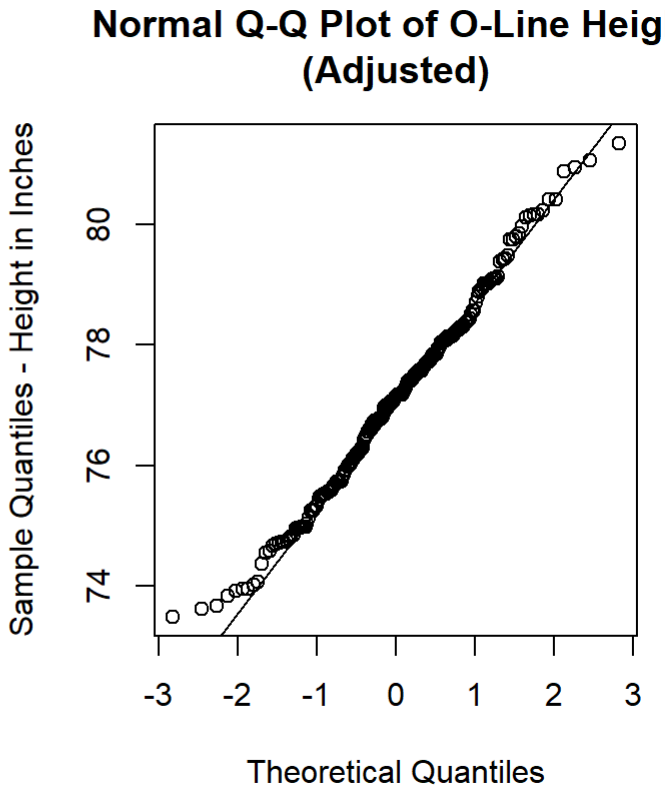
Offensive Line: Distributions Look Bell Shaped But Skewed For Weight





Q-Q Plots Indicate Height Close to Normal But Weight is Not





Shapiro-Wilk Tests Agrees Offensive Line Height May be Normal But Not Weight

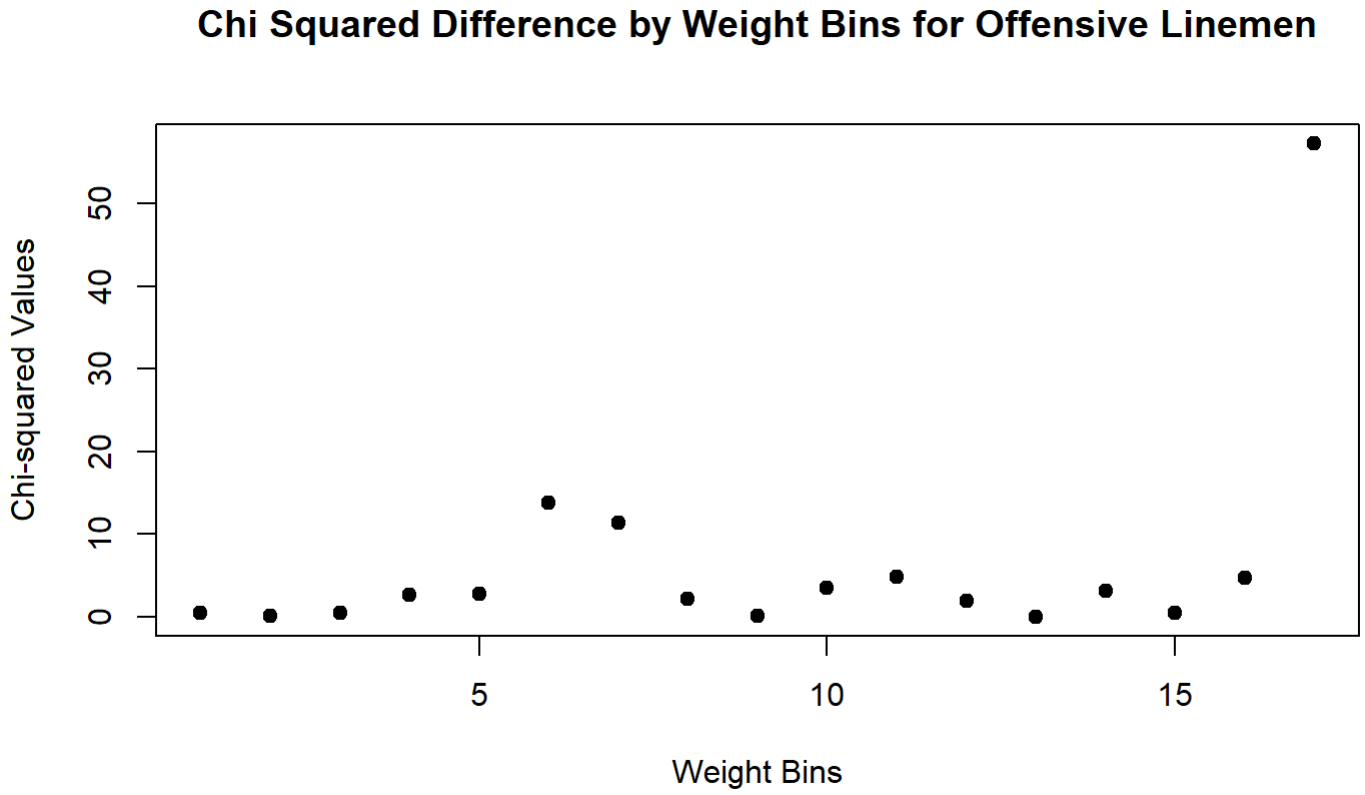
```
##
##  Shapiro-Wilk normality test
##
## data:  OL_Final$Weight
## W = 0.95225, p-value = 1.956e-06
```

```
##
##  Shapiro-Wilk normality test
##
## data:  OL_Final$Height_inches_uniform
## W = 0.99144, p-value = 0.257
```

Using The Chi-Squared Test Statistic, We'll Find Weight Bins With The Largest Deviations

- In histogram, weights bucketed in five pound increments with 85 different values in that range so 17 buckets total
- Using each bucket, we can calculate expected area under the curve from pnorm function and multiply by total sample size to get how many samples expected in that quantile
- From there, we can calculate observed minus expected and generate chi-squared values for each bin
- We then select the bins with the largest chi squared test statistic values - this is how the benford analysis package selects suspects

Biggest Suspects Are in the Heaviest Quantile - See Below

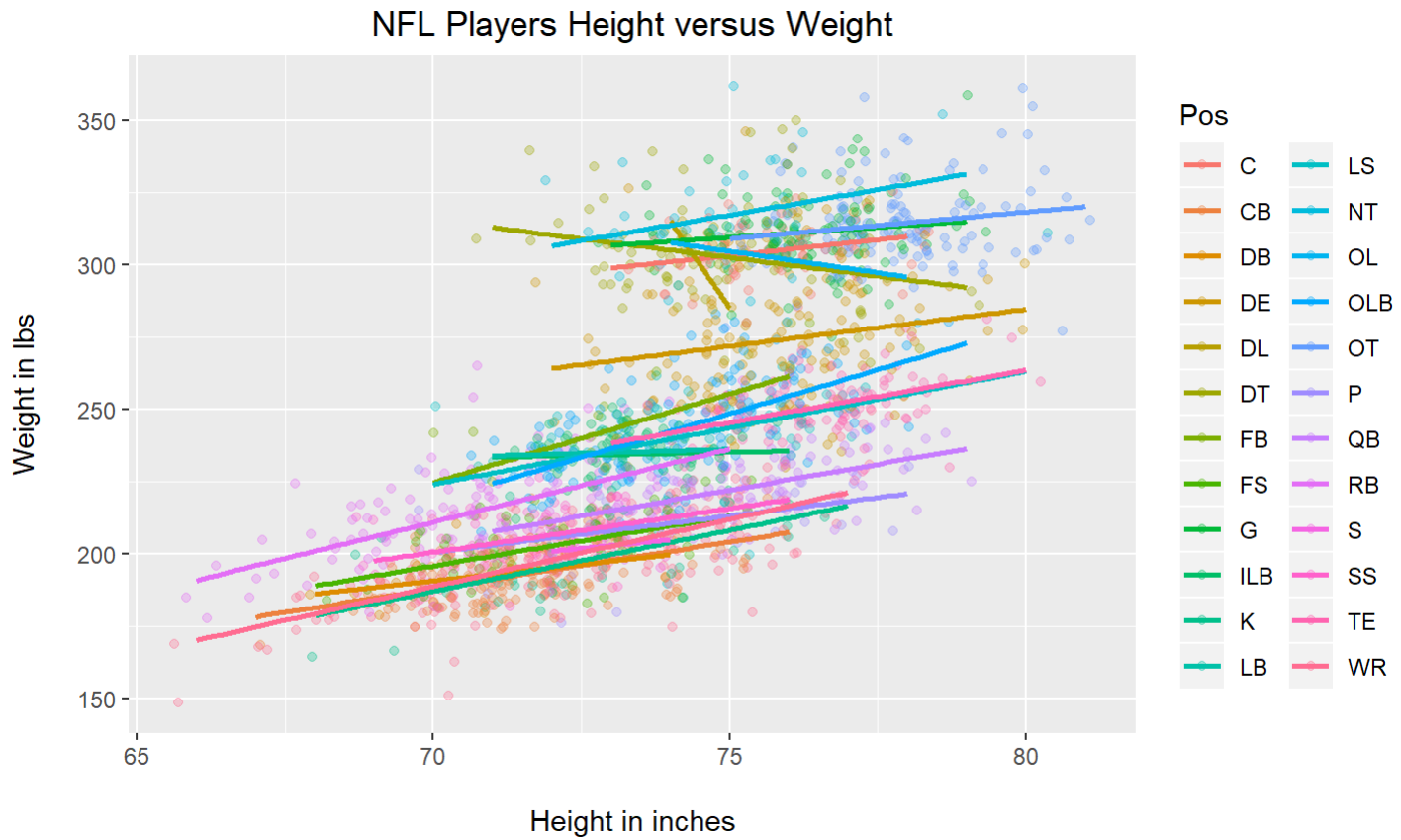


These players are Marcus Cannon of the New England Patriots and Zach Banner of the Pittsburgh Steelers

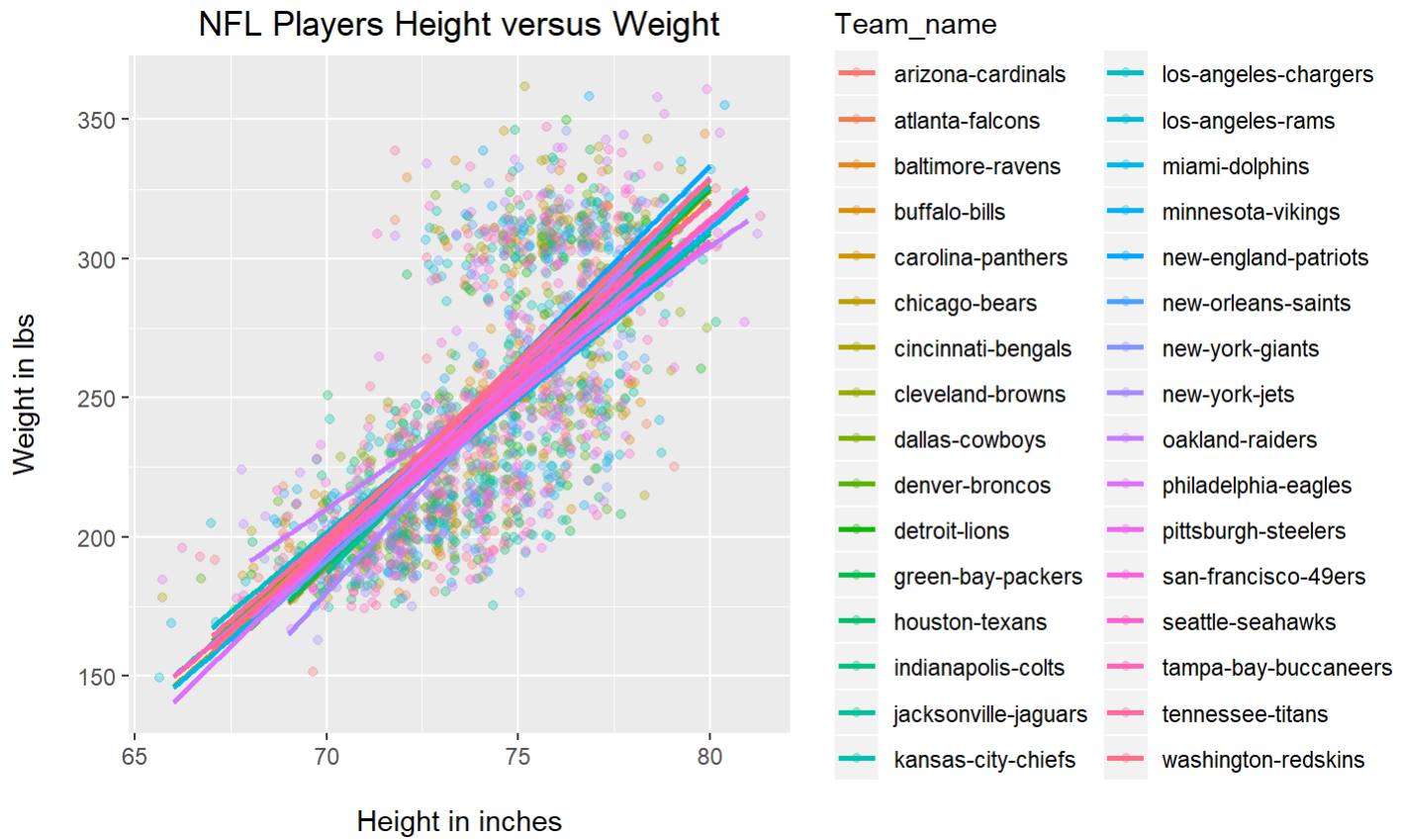
Both players Struggled With Weight in the Past and Have Weight-Related Incentives in Contracts

- Both players are listed at varying weights depending on the website and time period and have struggled with being too heavy in the past
- Both also have performance incentives in their contract for making a certain weight during training camp
- Our method of getting suspects via the largest Chi-Squared values appears somewhat successful in finding unusual observations in this test case

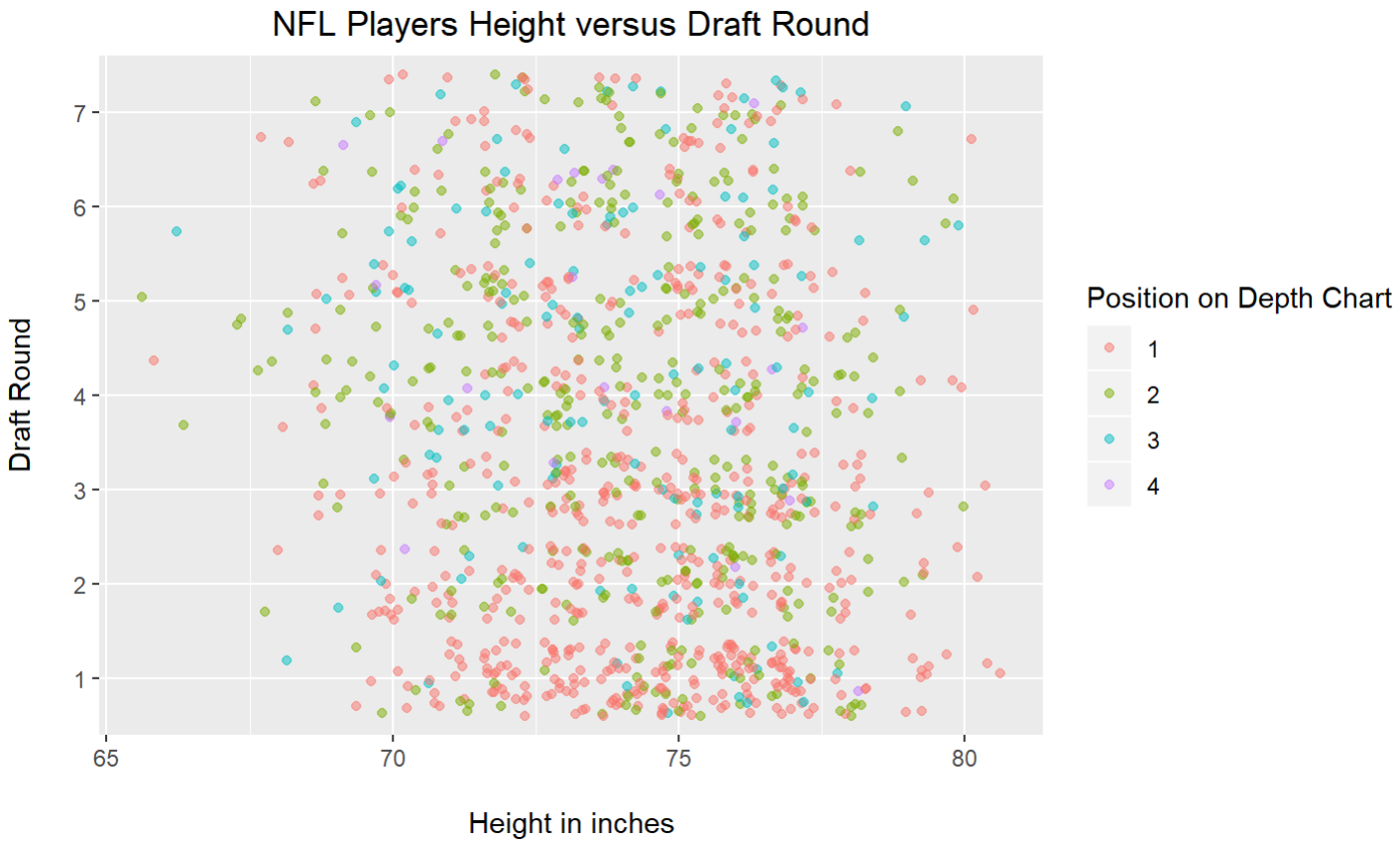
Some Other Selected Relationships Visualized (More in Report/Shiny) - 1



Some Other Selected Relationships Visualized (More in Report/Shiny) - 2



Some Other Selected Relationships Visualized (More in Report/Shiny) - 3



Some Other Selected Relationships Visualized (More in Report/Shiny) - 4