ReadMe

Madhu Suhaas Reddy Nachannagari

November 2024

1. Summary:

- Loaded in Kaggle dataset regarding NBA draft combine data since 2000 with nearly 1500 players, hoping to analyze key differences and trends across various different variables and measures mostly based on position and height for every player who attended the combine in the period.
- Performed hypothesis testing to answer 3 research questions that reveal meaningful findings regarding NBA combine data and player performance.
- Analyses and visualizations were implemented solely using R code to demonstrate effectiveness with technical testing and data handling.

2. Introduction:

This project takes Kaggle NBA Draft Combine data and hopes to analyze many measures of athletic prowess that players have been demonstrating for the last 20 years. We hope to compare different positions and different physiques, hoping to make meaningful connections and comparisons between the state of their athlete and their performance in a given exercise during the draft combine .

3. Data Exploration:

3.1 Data Preparation

The first step was to access the Kaggle folders through the API, which I hoped to do without downloading on my computer as the set was quite large.

We had to remove some non-accessible values and aggregate certain columns to make for easier data analysis. For example, we have combination positions such as C-PF and PF-C, which are important in certain discussions but they further divide up the dataset and make whatever conclusions we may find about each position to be weaker by decreasing the sample set. By considering every player as primarily one of 5 positions (PG, SG, SF, PF, and C) - if they are a combination position whichever one comes first (like PG from PG-SG) - we can make more general and stronger conclusions from our data analysis.

3.1 Data Visualization and Analysis

After accessing the dataset with 47 variables of data, we choose 8 variables of interest:

- **Wingspan**: The length from fingertip to fingertip with arms fully extended, measured in inches. It gives an idea of a player's reach and defensive abilities, important for defense and rebounding.

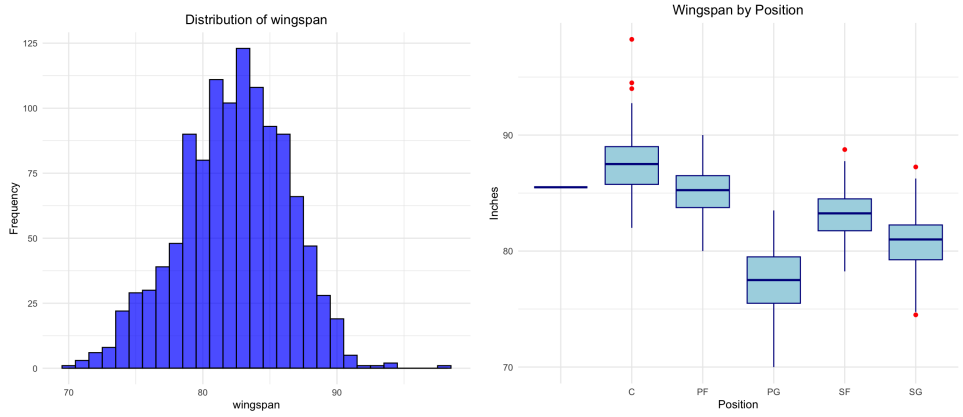

Figure 1 and 2: Histogram of wingspan distribution (Left) and Boxplots of wingspan by position (Right)

As we can see, there is a somewhat normal distribution of wingspans of players that have attended the combine over the past 20 years, with the most frequent ones being somewhere between 80-90 inches. The boxplot shows us that there are differing means and the “big men positions” (C, PF, SF) seem to have the highest means and extremes in comparison to the others.

- **Max_vertical_leap**: The highest a player can jump from a standing position, measured in inches. It shows explosive power and athleticism.

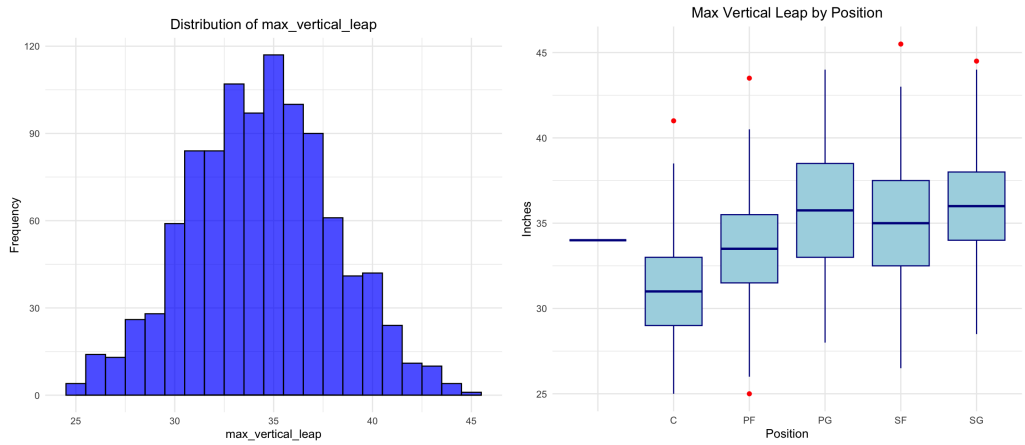

Figure 3 and 4: Histogram of max vertical leap distribution (Left) and Boxplots of verticals by position (Right)

There seems to be a normal distribution of values, making it suited for hypothesis testing based off the histogram. Most leaps seem to center around 35 inches. The boxplot shows us that the ball handling positions (SG, PG) seem to have higher means of vertical leaps, however SF has the single highest outlier of the bunch.

- **Lane_agility_time**: The time it takes to complete a lateral agility test, measured in seconds. It reflects a player’s quickness moving side to side.

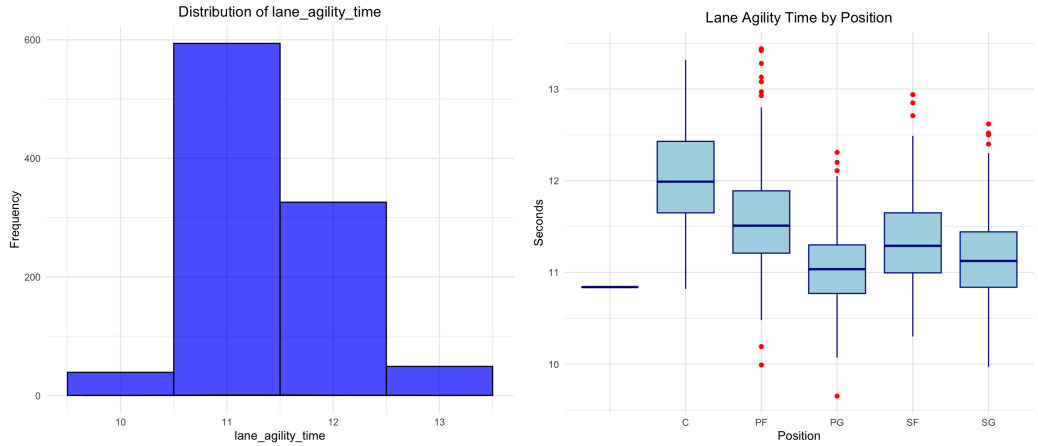

Figure 5 and 6: Histogram of lane agility times distribution (Left) and Boxplots of lane agility times by position (Right)

There seems to be a very small range of values that all players fall within, between 10 and 13 with a few outliers. Big men positions (C, PF) tend to take longer than the shorter, quicker positions like PG and SG, as can be seen with their IQR of C being clearly above the IQRs of the PG and SG.

- **Position**: This is the player's role on the court (e.g., Point Guard, Shooting Guard, Small Forward, Power Forward, Center). Different positions have unique responsibilities and physical characteristics.

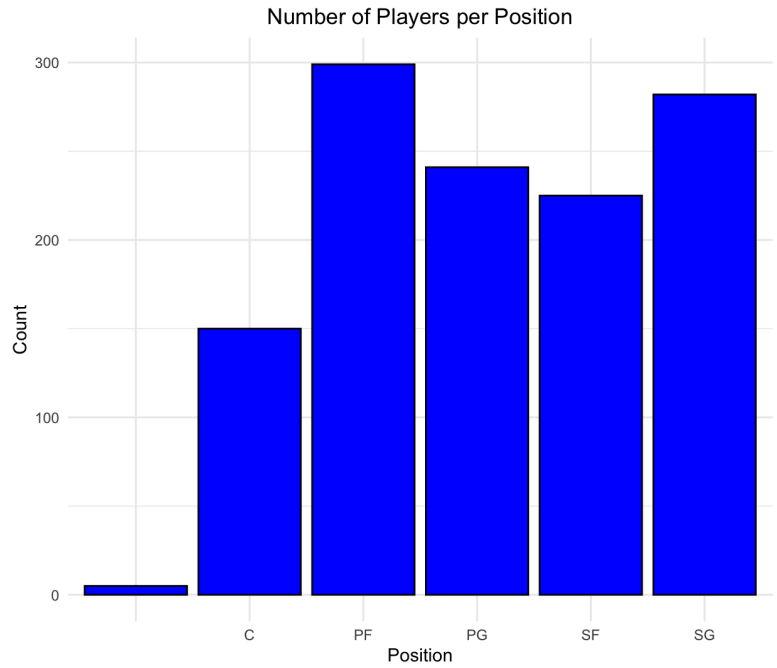

Figure 7: Counts of Players that attend the draft combine for each person.

The counts show us that centers are most likely underrepresented in the group statistics, such as the histograms, thus our analysis by positions is a fair and intuitive way to analyze the data in the combine.

- **Season**: The year the player participated in the NBA Combine. It helps track changes in player performance and physical stats over time.

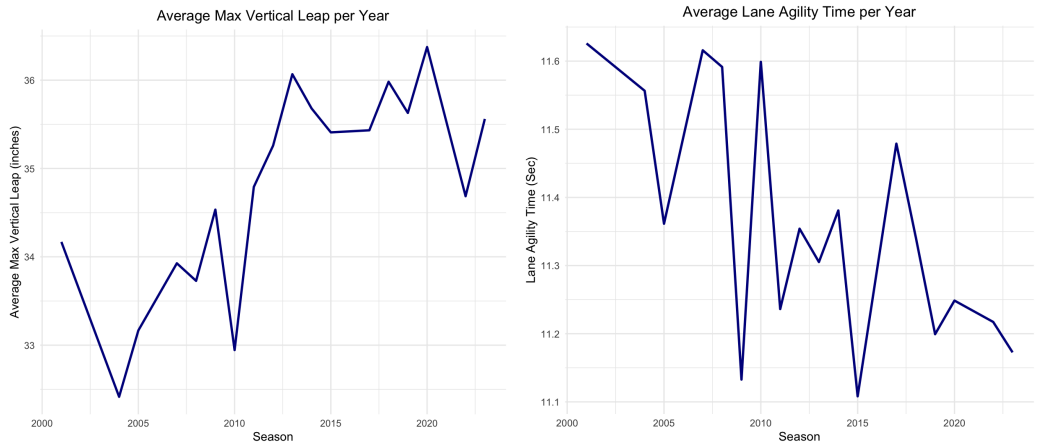

Figure 8 and 9: Line plots of average max vertical leap per year (left) and average lane agility time per year (right)

Seasons help us visualize general trends. As can be seen above with variables we've already explored, we can see the progression of max vertical leaps each year rising and dipping until it increased to where it is today, while lane agility time decreases over the years (quicker people).

- **Standing_reach**: How high a player can reach while standing flat-footed, measured in inches. It's a more wholistic measure than height which does not account for arm, shoulder, and neck length.

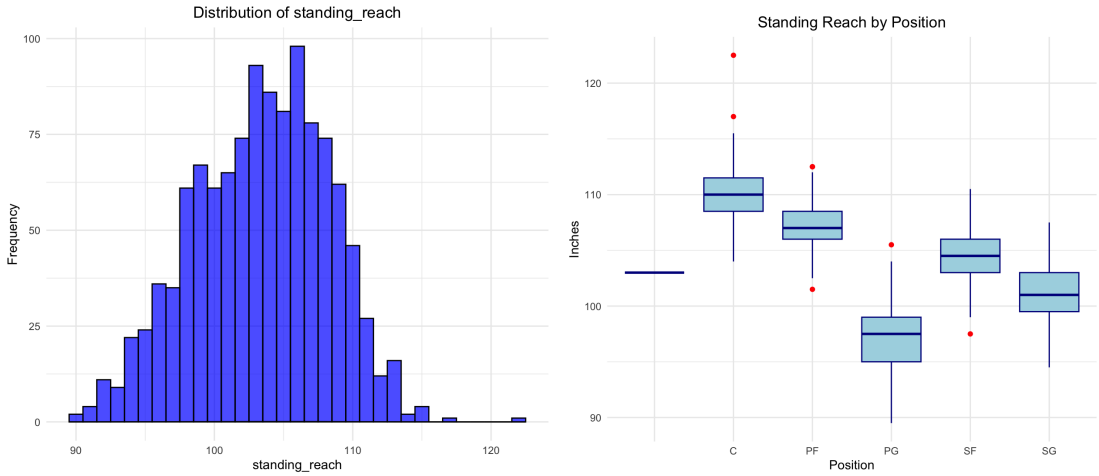

Figure 10 and 11: Histogram of standing reach distribution (Left) and Boxplots of standing reach by position (Right)

The histogram is again normal and indicates that the mean standing reach is somewhere around 105 inches, while the boxplots here show something similar to wingspans, however, each position has a smaller range this time, indicating that standing reach is a balancing variable that accounts for everything. To have an outlier in standing reach is to be truly out of the norm physically for your position in many different ways, not just one specific physical aspect.

- **Bench_press**: The number of 185-pound repetitions a player can do. It measures upper body strength.

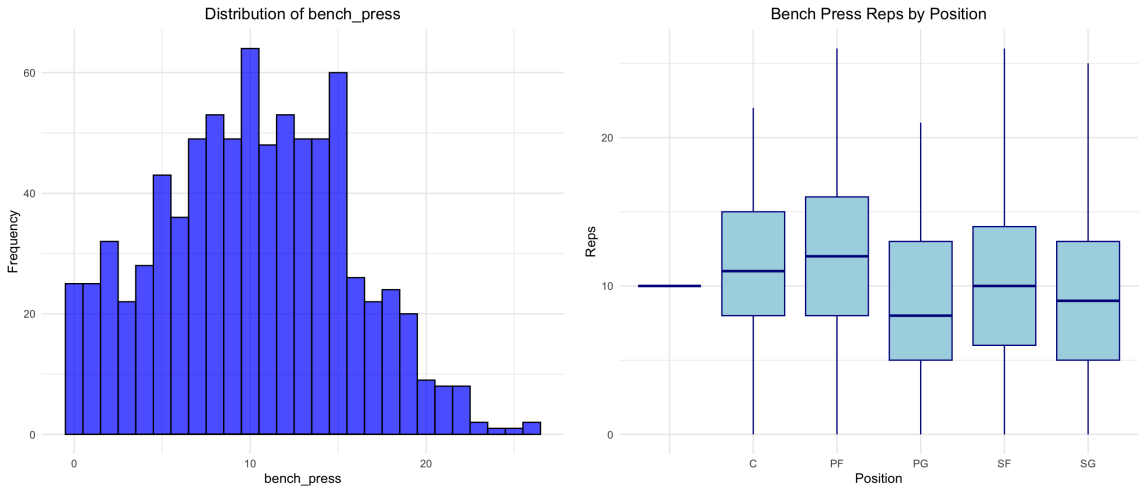

Figure 12 and 13: Histogram of bench press distribution (Left) and Boxplots of bench press reps by position (Right)

The histogram indicates a normal distribution and an average rep range of nearly 13 for each player, but the boxplot itself seems to have many outliers and the range for each and every position is relatively large. The IQRs of all the positions overlap a lot, also indicating that there is no significant difference, but this can further be tested with a null hypothesis and a possible hypothesis test.

From the variables and data above, just a surface level understanding of their functions as well as the general trends we can possibly see with the naked eye, we hope to confirm some assumptions as well as discern certain differences in datasets and their sample statistics. More specifically:

3.2a Research Question 1

Do players in different positions differ SIGNIFICANTLY in their max vertical leaps?

We ask this question because as we saw in the visual, there SEEMS to be differences between the positions, but there seems to be much overlap between adjacent box plots for positions. This question aims to definitively confirm and identify any significant difference. Depending on what we find, we can interpret the following:

- **Significant Difference**: If guards jump higher, it confirms the traditional view that smaller players develop explosiveness to offset their size disadvantage. This would emphasize vertical leap as a key trait for guards.
- **No Significant Difference**: If there's no difference, it challenges this view, suggesting that explosiveness is now valued across all positions. In this case, vertical leap may not be the key differentiator for productivity if the larger players can match it and negate the advantage, other traits (e.g., skill, strength) might be more important in compensating for Guards' innate lesser stature and size.

3.2b Test for Research Question 1 (**ANOVA**)

We choose an ANOVA test because we are choosing the means of 5 different groups, making T-tests irrelevant as they only test for two. Some of the assumptions we make are:

- **Normal Distribution:** We know this because of the Central Limit Theorem, there's much more than 30 data points - 1,500 actually - and according to the histograms from the last section there is also a visually apparent normal distribution.
- **Independent Data Points:** Each row in the dataset is an independent player.

We operate on the following hypotheses with an alpha value of 0.05 used to infer the results:

- **Null Hypothesis** (H_0)**:** $\mu_{PG}=\mu_{SG}=\mu_{SF}=\mu_{PF}=\mu_{C}$ (Mean max vertical leaps are equal across all positions:)
- **Alternative Hypothesis** (H_1)**:** *At least one position's mean max vertical leap differs*

```
> # Anova testing implementation
> max_vertical_anova_df = select_stats[!is.na(select_stats$position) & select_stats$position != "", ]
> anova_result <- aov(max_vertical_leap ~ position, data = max_vertical_anova_df)
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)
position      4   2608   651.9   60.39 <2e-16 ***
Residuals  1011  10914    10.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14: Anova Test Results

The p-value (< 2e-16) is far below 0.05, so we reject H_0. This indicates that at least one position has a significantly different mean max vertical leap.

```
> tukey_result <- TukeyHSD(anova_result)
> tukey_result
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = max_vertical_leap ~ position, data = max_vertical_anova_df)

$position
            diff         lwr       upr     p adj
PF-C   2.4613479  1.46206000 3.4606358 0.0000000
PG-C   4.7328057  3.69901012 5.7666014 0.0000000
SF-C   4.1295018  3.08609452 5.1729091 0.0000000
SG-C   4.9178089  3.91392149 5.9216964 0.0000000
PG-PF  2.2714579  1.42736985 3.1155459 0.0000000
SF-PF  1.6681539  0.81232104 2.5239868 0.0000012
SG-PF  2.4564610  1.64928005 3.2636420 0.0000000
SF-PG -0.6033039 -1.49918733 0.2925794 0.3509217
SG-PG  0.1850032 -0.66452508 1.0345315 0.9758280
SG-SF  0.7883071 -0.07289182 1.6495061 0.0910808
```

Figure 15: Tukey's HSD results

We then use Tukey's HSD to identify the positions that significantly differ. We can see that since that p-adj is less than 0.05, PF-C, PG-C, SF-C, SG-C, PG-PF, SF-PF, SG-PF are all positions whose mean max verticals differ significantly, and positive differences in favor of the smaller positions indicate they are more explosive.

Based off the interpretations we set up earlier, these significant differences mean that maximum vertical is a valuable trait that helps smaller positions compensate for their lack of stature when going up against less explosive, bigger players.

3.3a Research Question 2

Is there a significant relationship between wingspan and standing reach?

We ask this question because intuitively these are two variables that will relate with one another as a result of wingspan often directly contributing to a higher standing reach. However, how closely they are related could be different, as standing reach isn't necessarily just height plus half the wingspan, wingspan also takes into account shoulder and chest width, as well as hand size - things that often are considered in standing reach.. Standing reach on the other hand is the combination of your body up til your shoulders, and one arm raised from that point, thus it ignores significant parts of wingspan, and that absence of factors may or may not make the relationship weaker.

- **Close Relationship**: We can use wingspan to predict standing reach even if we are not able to calculate it exactly, and use them interchangeably as strong indicators of lateral reach and vertical reach.
- **Weak Relationship**: We have to think of them as independent and assume that there are other more important factors in standing reach such as other body proportions, and one cannot be used to infer the same thing as the other nor can they predict one another.

3.3b Test for Research Question 2 (**Correlation Analysis and Linear Regression**)

We choose a Correlation test because we are trying to find how closely two variables are related to one another in terms of predicting one another successfully. If we can derive a linear regression from it, then they are closely related and can be used to infer conclusions from/for one another. We operate under the following assumptions:

- **Normal Distribution:** We know this because of the Central Limit Theorem, there's much more than 30 data points - 1,500 actually - and according to the histograms from the last section there is also a visually apparent normal distribution.
- **Independent Data Points:** Each row in the dataset is an independent player.
- **Linear Relationship:** Both increase in a proportional manner, a higher wingspan intuitively means higher standing reach.

We operate on the following hypotheses with an alpha value of 0.05 used to infer the results:

- **Null Hypothesis (H_0):** *There is no significant correlation between wingspan and standing reach*.
- **Alternative Hypothesis (H_1):** *There is a significant correlation between wingspan and standing reach*.

```
> correlation

        Pearson's product-moment correlation

data:  reach_stats$wingspan and reach_stats$standing_reach
t = 71.787, df = 1149, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8931463 0.9142726
sample estimates:
      cor
0.9042614
```

Figure 16: Correlation Analysis results

The p-value (< 2e-16) is far below 0.05, so we reject H_0. This suggests that there is a significant positive correlation between wingspan and standing reach among basketball players. The correlation coefficient of 0.904 indicates a strong positive linear relationship, meaning that as wingspan increases, standing reach also should increase.

```
Call:
lm(formula = standing_reach ~ wingspan, data = reach_stats)

Residuals:
   Min     1Q Median     3Q    Max
-8.068 -1.287  0.096  1.373  8.041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.09451    1.27598   9.479   <2e-16 ***
wingspan     1.10948    0.01546  71.787   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.084 on 1149 degrees of freedom
Multiple R-squared:  0.8177,    Adjusted R-squared:  0.8175
F-statistic:  5153 on 1 and 1149 DF,  p-value: < 2.2e-16
```

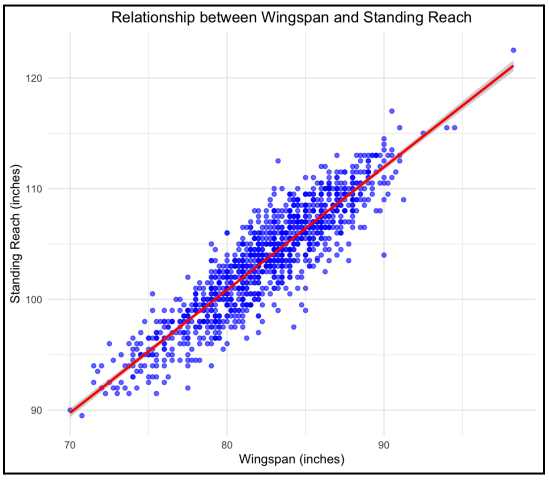

Figure 17 and 18: Linear Regression Model results and graph

Since the p-value for both the intercept and the slope of the regression model is less than 0.05, we reject the null hypothesis. This indicates that there is a significant linear relationship between wingspan and standing reach. The positive slope (1.10948) indicates that as wingspan increases, standing reach also increases.

Using our interpretations from earlier, with both the correlation analysis and the linear regression model indicating that there is a strong positive relationship between Wingspan and Standing reach, we can conclude that the other factors are indeed negligible and they are good indicators of reach in general in place of one another, and wingspan can be used to predict standing reach.

3.4a Research Question 3

Do players with a higher wingspan have significantly higher bench press reps than players with a smaller wingspan?

We ask this question because bench press is a great indicator of strength for an athlete, but does the length of their arms either hamper or increase their ability to get in more reps at the draft combine? While longer arms mean more required range of motion for the average person and is thus traditionally considered a burden to someone bench pressing. For NBA athletes however, a longer wingspan could mean the opposite for various reasons such as being a sign of an elite athlete who can lift heavy despite the longer arms in comparison to a smaller wingspan who is usually not a good NBA player. Or not, this test aims to answer that question. We operate under the following assumptions

- **Significant Difference**: Higher wingspan means better on the bench press test, on average, compared to players with a smaller wingspan. This difference is unlikely to be due to random chance, and wingspan may be a contributing factor to strength in a player and thus more valuable than we initially thought.
- **No Significant Difference**: Wingspan does not have a significant impact on bench press performance. Any observed differences in bench press reps could be due to random variability, and wingspan may not be a useful predictor for strength in this case.

3.4b Test for Research Question 3 (**2-Sample T-Test**)

We choose a **2-sample t-test** because we are comparing the means of 2 different groups of high wingspan players and low wingspan players, and a t-test is appropriate when comparing the means of two independent samples.

- **Normal Distribution:** We know this because of the Central Limit Theorem, there's much more than 30 data points - 1,500 actually - and according to the histograms from the last section there is also a visually apparent normal distribution.
- **Independent Data Points:** Each row in the dataset is an independent player.

We operate on the following hypotheses with an alpha value of 0.05 used to infer the results:

- **Null Hypothesis** (H_0)**:** $\mu_L = \mu_H$ (Mean bench presses are equal in both wingspan groups)
- **Alternative Hypothesis** (H_1)**:** $\mu_L \neq \mu_H$ (Mean bench presses are different in both wingspan groups)

```
        Welch Two Sample t-test

data:  high_wingspan_group and low_wingspan_group
t = 3.0642, df = 804.82, p-value = 0.002255
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4154862 1.8965867
sample estimates:
mean of x mean of y
10.746835  9.590799
```

Figure 19: 2-Sample T-Test Results

The p-value 0.00255 is below 0.05, so we reject H_0. There is sufficient evidence that there is a difference in the bench presses of players with high wingspans compared to those with low wingspans.

```
> print(t_test_result$conf.int)
[1] 0.4154862 1.8965867
attr(,"conf.level")
[1] 0.95
```

Figure 20: Confidence Interval

We are 95% confident that the true difference in the mean values between the high wingspan group and low wingspan group is between 0.4155 and 1.8966. Since the interval does not equal 0, this is a meaningful interval and since it is a positive difference between high wingspan and low wingspan groups that we are comparing - when coupled with the result of the t-test - we can conclude that there is a higher mean bench press for higher wingspan groups than lower wingspan groups.

4. Conclusion:

In conclusion, this analysis offers valuable insights into how various physical traits relate to player performance at the NBA Draft Combine. By examining the factors highlighted throughout this report, we were able to glean meaningful conclusions about player performance and player attributes. Using statistical methods to analyze each of the research questions, we present the numerous utilities of the various types of methods one can use to analyze data, and moving forward we can integrate even more statistics and even more methods to enhance predictions of player success and gain a deeper understanding of how physical traits vary across different positions.