* A note specifying which dataset you analyzed
* A statement of the question(s) you posed
* A description of what you did to investigate those questions
* Documentation of any data wrangling you did
* Summary statistics and plots communicating your final results

**A note specifying which dataset you analyzed**

The required datasets are obtained from stats.nba.com. JSON files were downloaded, and the converted into CSVs. The dataset features full relevant data from the 2017-2018 NBA regular season.

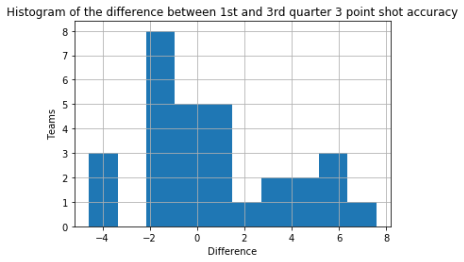
Based on the data we’ll be able to extrapolate to the whole league, and due to the everchanging nature of the game, data from the latest season will give us the best possible insight into the future.

This is an observational study, not an experiment, so we won’t be able to establish any causation, only correlation. But there is still a lot of value in that.

**A statement of the question(s) you posed**

1. Does the team 3 point shot accuracy differ in the 1st and the 3rd quarters?
2. Is there a relationship between position and age?
3. In games where the lead is small at the half (within 5 points), does being ahead correlate with winning?
4. Is the origin makeup of NBA players different across positions?
5. Does an increase in wingspan relate to increase in the number of deflections and steals?

**1. Does the team 3 point shot accuracy differ in the 1st and the 3rd quarters?**



Note that this graph represents the difference “1st quarter 3p percentage – 3rd quarter 3p percentage”. Negative values mean increase of the 3p shot accuracy in the 3rd quarter, and positive values mean decrease in the 3p shot accuracy in the 3rd quarter.

Descriptive statistics:

|  |  |
| --- | --- |
| count | 30.000000 |
| mean | 0.674881 |
| std | 3.145046 |
| min | -4.581340 |
| 25% | -1.381035 |
| 50% | -0.091312 |
| 75% | 2.487729 |
| max | 7.577884 |

Paired C->Q t-test

t-statistic=1.1753325804178147, p-value=0.24942155727039717

Not significant to reject the null hypothesis.

Considering their drastic improvement from 1st quarter to 3rd Minnesota may consider incorporate the 3-point shooting more in the warmup. Doesn’t apply to Golden State, because they already shoot above average in the first quarter, but go superhuman in the 3rd (the infamous “3rd quarter Warriors” can be seen here quantified).

**2. In games where the lead is small at the half (within 5 points), does being ahead correlate with winning?**

Two-way table:

|  |  |  |  |
| --- | --- | --- | --- |
| 1st half win / WL | L | W | All |
| False | 121 | 94 | 215 |
| True | 94 | 137 | 231 |
| All | 215 | 231 | 446 |

Conditional percentages:

|  |  |  |
| --- | --- | --- |
| 1st half win / WL | L | W |
| False | 0.562791 | 0.437209 |
| True | 0.406926 | 0.593074 |
| All | 0.482063 | 0.517937 |

C->C Chi-square test for independence

χ-statistic = 10.834980000149809, p-value = 0.0009960039387090216

At 5 it does significantly

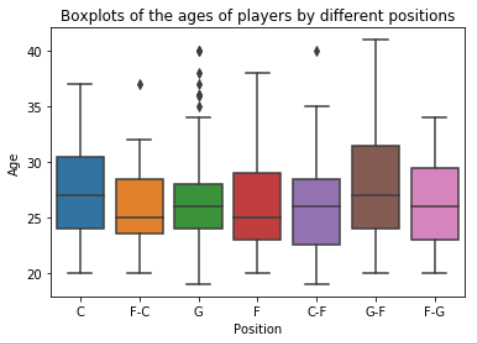
Note that at 4 doesn’t (don’t have evidence).

Exactly 4 points: χ-statistic = 0.768171114599685, p-value = 0.38078342795092013

Within 4 points: χ-statistic = 2.1520447236647535, p-value = 0.14238016058673728

Surprising that at 4 (relatively big lead) we don’t have it. Maybe it’s the complacency of winning at half-time when the game is close.

**3. Is there a relationship between position and age?**



Descriptive statistics:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PLAYER\_POSITION | count | mean | std | min | 25% | 50% | 75% | max |
| C | 63.0 | 27.111111 | 4.399821 | 20.0 | 24.0 | 27.0 | 30.5 | 37.0 |
| C-F | 23.0 | 26.217391 | 5.071816 | 19.0 | 22.5 | 26.0 | 28.5 | 40.0 |
| F | 150.0 | 26.426667 | 4.045623 | 20.0 | 23.0 | 25.0 | 29.0 | 38.0 |
| F-C | 31.0 | 25.935484 | 3.881026 | 20.0 | 23.5 | 25.0 | 28.5 | 37.0 |
| F-G | 23.0 | 26.782609 | 4.155572 | 20.0 | 23.0 | 26.0 | 29.5 | 34.0 |
| G | 205.0 | 26.321951 | 4.006582 | 19.0 | 24.0 | 26.0 | 28.0 | 40.0 |
| G-F | 39.0 | 27.717949 | 5.155216 | 20.0 | 24.0 | 27.0 | 31.5 | 41.0 |

Multiple C->Q ANOVA

There are outliers in the positions “C-F” and “F-C” - Dirk Nowitzki and Nick Collison, and the sample sizes may not necessarily be enough to guarantee normal distribution of the …

Running ANOVA on the full data:

statistic=0.955603760244476, p-value=0.45477740597358385

Running ANOVA on the data without the outliers in the "C-F' and 'F-C" positions:

statistic=1.311318715208154, p-value=0.2501211510855592

The outliers did not influence the results.

Not statistically significant. Can’t reject the null hypothesis.

**4. Is the origin makeup of NBA players different across positions?**

Remove data entry outlier Walter Lemon Jr. He's not the only "pure" PG, other pure PGs were classified as guards, so we’ll consider this data entry error.

Two-way table:

|  |  |  |  |
| --- | --- | --- | --- |
| Origin /  PLAYER\_POSITION | International | USA | All |
| C | 30 | 33 | 63 |
| C-F | 8 | 15 | 23 |
| F | 27 123 | 150 |  |
| F-C | 10 | 21 | 31 |
| F-G | 4 | 19 | 23 |
| G | 28 | 177 | 205 |
| G-F | 7 | 32 | 39 |
| All | 114 | 420 | 534 |

Conditional percentages:

|  |  |  |
| --- | --- | --- |
| Origin /  PLAYER\_POSITION | International | USA |
| C | 0.476190 | 0.523810 |
| C-F | 0.347826 | 0.652174 |
| F | 0.180000 | 0.820000 |
| F-C | 0.322581 | 0.677419 |
| F-G | 0.173913 | 0.826087 |
| G | 0.136585 | 0.863415 |
| G-F | 0.179487 | 0.820513 |
| All | 0.213483 | 0.786517 |

C->C Chi-square test for independence

statistic= 39.26852697080573, p-value=6.340368194661366e-07

p-value practically zero, statistically significant

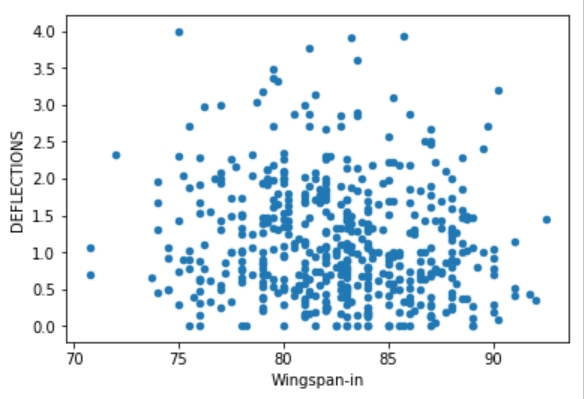
shows that international basketball more likely to be centers or big men, and the trend in US basketball towards guards (“smallball”)

**5. Does an increase in wingspan relate to increase in the number of deflections and steals?**

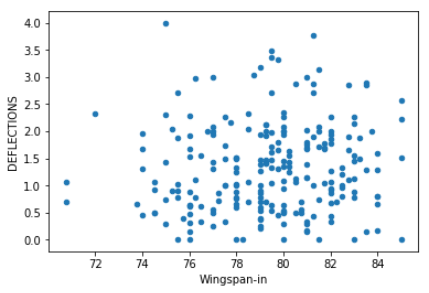
We’ll choose only players who had played more than 15 games.

The graphs below show the relationship between height in inches and the average number of deflections recorded by NBA players during regular season.

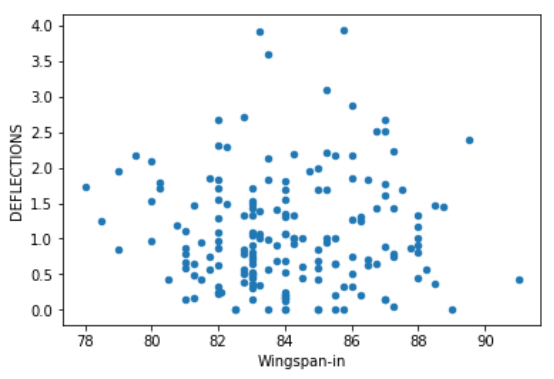
Scatterplot for all the players in the NBA:



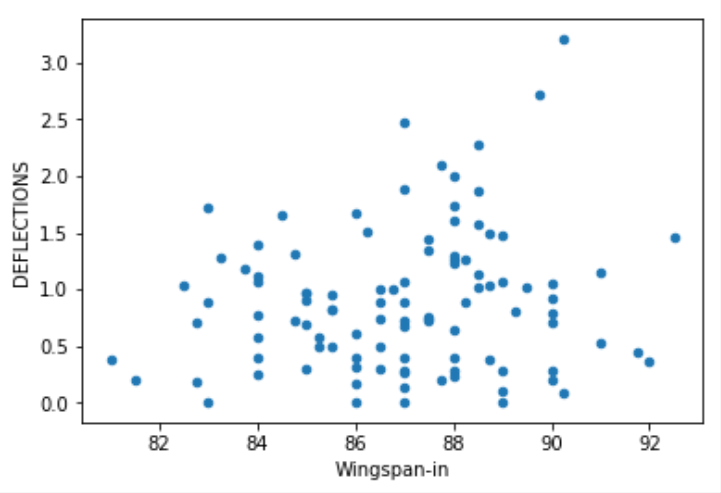
For guards:



For forwards:



For centers:

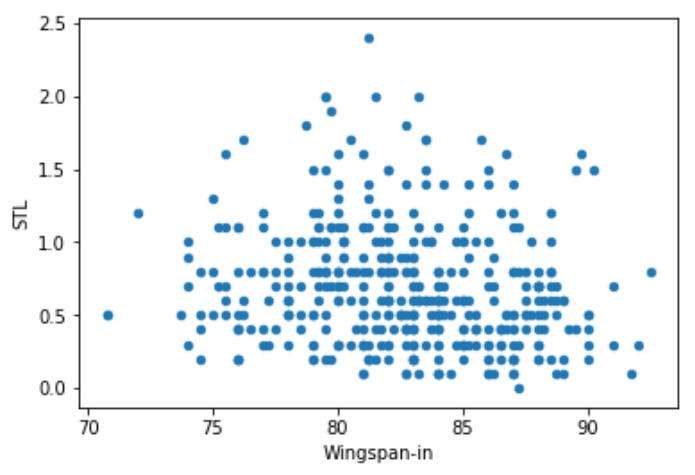


Q->Q regression t-test for the slope.

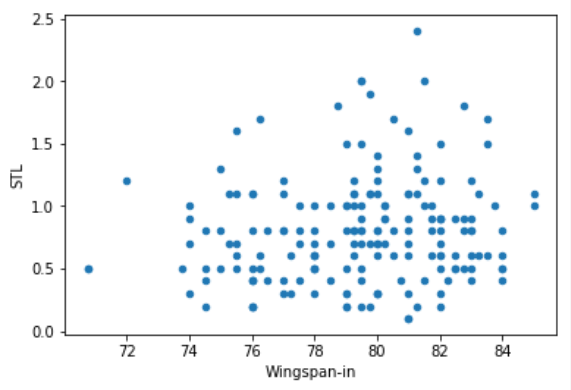
The closest we have is in the case of centers, with correlation coefficient r-value=0.14536953391373944, and test p-value=0.14288568244203492.

The graphs below show the relationship between height in inches and the average number of steals recorded by NBA players players during regular season.

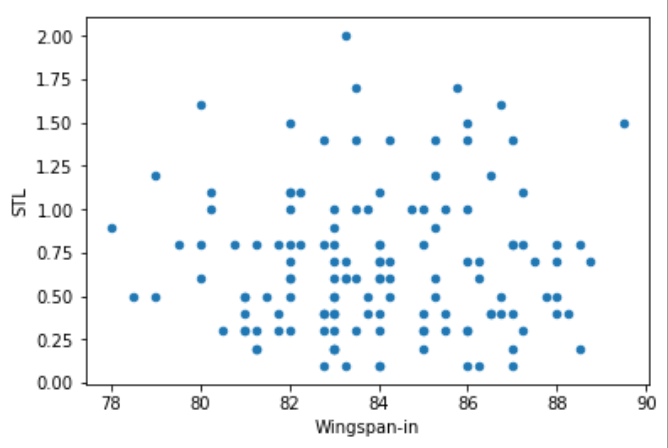
Scatterplot for all the players in the NBA:



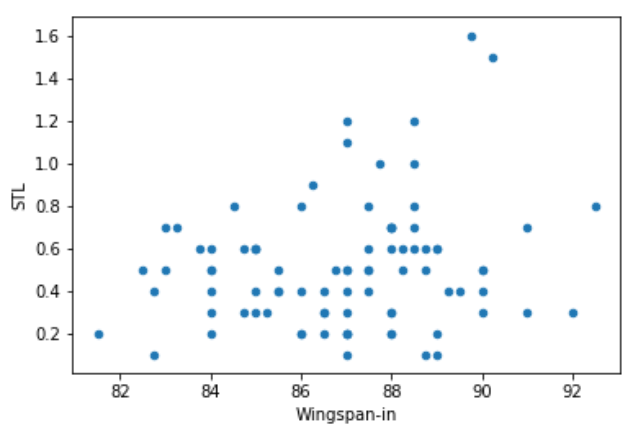
For guards:



For forwards:



For centers:



Q->Q regression t-test for the slope.

None of the relationships are significant at 0.05 significance level with the selected minimum game cutoff of 15, although for guards is close at p-value=0.05779801365934409. But with bigger cutoffs for games played (i.e. 20, 30, 40) the p-value is even bigger, so the relationship doesn’t uphold).

For centers correlation coefficient r-value=0.19791927433929551, and the test p-value=0.06612281263777153. Increasing the cutoff for minimum games played to 25 we get r-value=0.2165473177952725, and p-value=0.046525524153957186.

At the cutoff of 30 games minimum, we get r-value=0.25191047346980966, p-value=0.02242571458638422. It further upholds at the cutoff level of 40 games.

For our purposes, we can consider this statistically significant, and reject the null hypothesis. We can conclude that for centers there is, if weak, a positive linear relationship between wingspan in inches and number of steals recorded.

Documentation of any data wrangling you did

Summary statistics and plots communicating your final results

Extra questions for inquiry.

1 C- > Q. Two-sample t-test

Do international run more?

2 Multiple C->Q ANOVA

Rookie-sophomore-veteran? Contested shots

Do different positions run more?

3. C->C Chi-square test for independence

4. Q->Q regression t-test for the slope

Pace true shooting

Draft number height? Not linear