## horizontal line



Baseball Data Analysis

December 14, 2022

**─**

Selam Mekonnen

STAT 301 Final Project

# Introduction

Data analysis is an essential aspect of statistics that is applicable in real-world situations. A lot of fields use data analytics to deal with different types of huge data and the sports industry is no different. Baseball is one of the biggest sports in the world and the 3rd most played sport in the US, and thereby it applies data analytics to deal with large amounts of data.

In this paper, we will be analyzing Roanoke College’s baseball data and discussing some of the aspects of the analysis and the data itself. We have two datasets that we will be using to answer two separate problems. For the first problem, we will be conducting a hypothesis test and for the second problem, we will be creating contingency tables and running a Chi-square test.

# Analysis

## Problem 1

The first analysis focuses on two baseball data variables i.e., Horizontal break & Exit speed. Horizontal break is the difference between where the pitch would have crossed the front of home plate side-wise if it was thrown perfectly straight and where it actually crossed. The second variable is Exit speed which is the speed of the baseball as it comes off the bat, right after a batter makes the contact. For the first variable, we will specifically try to look at whether or not the exit speed is different when horizontal break is less than two and when horizontal break is greater than two. So, the variables of interest are as follows;

X1 = Horizontal Break < 2

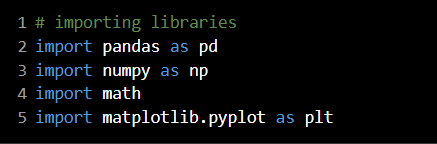
X2 = Horizontal Break > 2

Y = Exit Speed

To organize and analyze the data, we are using Python software. Python is a programming language that can be used for different purposes including data analytics. It is a very widely used software that has easy features to organize data so we chose to use it over other software like Excel or R Studio.

# Data

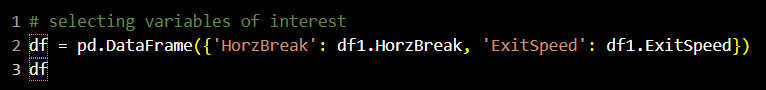
The first step we did for the analysis in Python was importing the libraries we will be using:

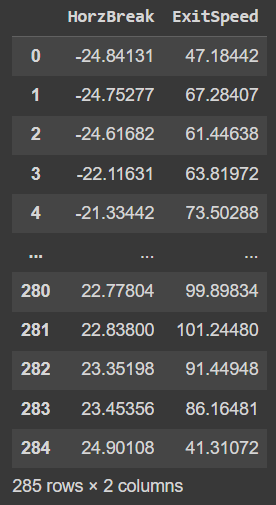


We then uploaded the first CSV file and viewed it:

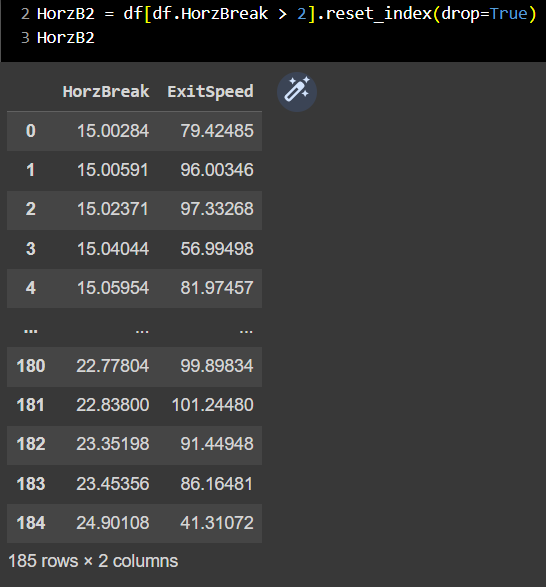
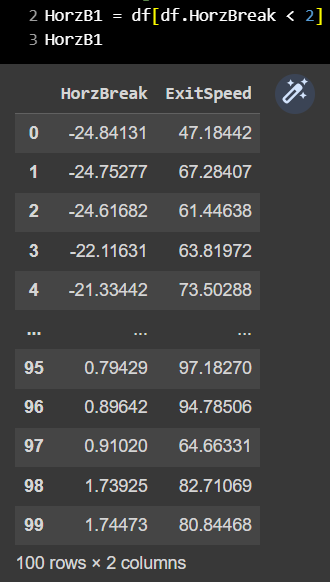


After looking at the data and making sure there aren’t any missing values or duplicates, we then selected the variables we are interested in analyzing and ended up with the following organized dataset.

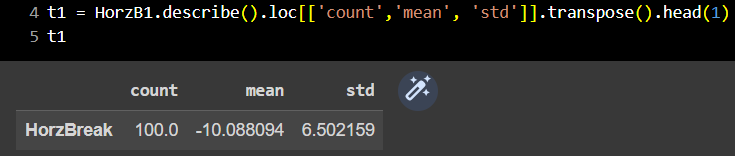


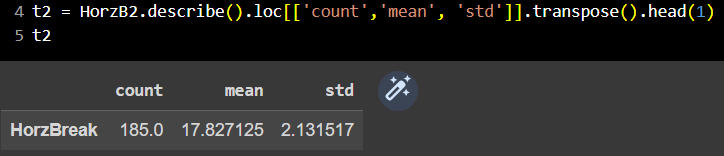


Following, we split the dataset by filtering it using the first variable (I.e., horizontal break) in order to get the X1 & X2 variables we stated earlier.

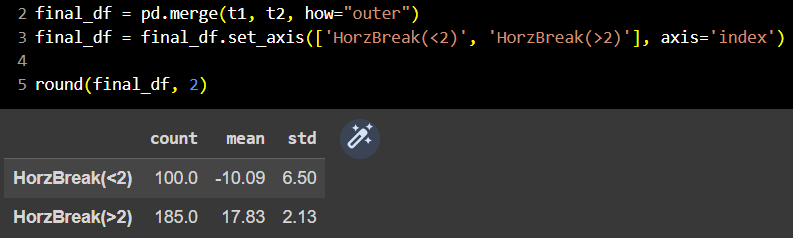


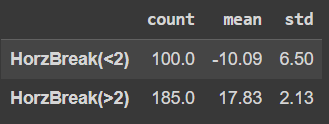
We then computed the mean and standard deviation for both tables (I.e., horizontal break < 2 & horizontal break > 2)





Finally, we merged the two tables containing the descriptive stats together and ended up with the following final table:





# Expectations

Before we performed our analysis, we tried to discuss what we should expect to see from the analysis. We stated that we would expect to see different exit speeds from the pitches that had a horizontal break of less than 2 versus the ones that had a horizontal break greater than 2. This is because the batter will have a slightly different point of contact or swing depending on the pitch that is coming in at them. Pitchers are trained to throw different pitches based on the player up to bat, how the game is going, who is in fielding positions, the number of runners on base, and many other aspects. As a batter, there is no way to correctly predict the type of pitch you will get 100% of the time. That being said, their position may not always be right for what they are hitting against, therefore creating a difference in exit speed and a possible skew in the data.

# Hypothesis Test

The hypothesis test is focused on whether or not the exit speed is different for horizontal breaks less than two and horizontal breaks greater than two. Therefore, our null hypothesis is that the exit speed means are not different for horizontal breaks that are less than two and greater than two. And our alternative hypothesis is that the exit speed means are different for horizontal breaks that are less than two and greater than two.

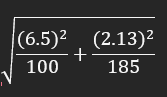
**Ho:**  **(µ1 = µ2)**

**Ha: (µ1 ≠ µ2)**

We have a large sample size for our data set (I.e., n > 30) and we are comparing two means thus, we are conducting a normal distribution test. To conduct the test, we will need the test statistic and the mean.

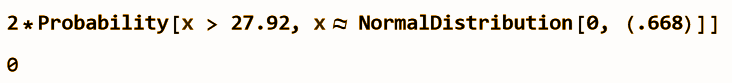
Therefore. we calculate the mean and the test statistic by using the table we analyzed earlier (where, **µ1**= -10.09,  **µ2**= 17.83, **Sd1** = 6.5, & **Sd2** = 2.13 ).

**Mean = µ2 – µ1**

= 17.83 – (-10.09) = **27.92**

**Test-Statistic = = .668**

We then used Mathematica to conduct the normal distribution test. (Since it is two-sided we multiplied the result by 2)



# Interpretation

With a 99% confidence level, we can reject the null hypothesis that the mean of the exit speed for horizontal breaks less than two is the same as the mean for horizontal breaks greater than two. The p-value of zero gives us the confidence that there is no possibility for the null hypothesis to be correct, therefore, we can say that there is significant evidence that the exit speed means are different for horizontal breaks less than two and greater than two.

The result we got made us hesitate at first because p-values are rarely exactly zero, however, the fact that it was zero suggests that there is no possibility for the null hypothesis to be correct. Another explanation is that it is actually not exactly zero, but it is very close to zero so Mathematica rounded it off.

# How can this information be useful to a baseball player/coach?

If we were reporting this information to a baseball team, we would be able to tell them that the horizontal break of their pitch would affect their exit speed. Pitchers and pitching coaches can use this information to determine what kind of pitches they want to practice. It would be beneficial to practice the pitch that results in the exit speed they prefer, which in most cases, would be the lower speed because it is easier to field a ball that is not moving as fast.

On the other side of things, coaches can use this information to improve their teams at bats. As a batter, you want to have a higher exit speed. Again, it is easier to field a ball that is moving at a slower speed so having a higher exit speed would give you a higher chance of getting on base. As a batting coach, you would want your players to practice the pitches that are causing a lower exit speed to improve the bats where they are getting these types of pitches.

# Limitation

The sample size in this example is small, having more data would be able to give us more accurate results. Our data set has 285 rows of data and each pitcher has a maximum pitch count for the day which is most likely somewhere around 100 pitches so there is a chance this data is only showing three different pitchers. This could be a conflict because each pitcher has different levels of difficulty when it comes to batting against them, regardless of what the horizontal break is. This data is also only looking at the Roanoke College Baseball team and their opponent. By looking at only a select number of teams, there is no way to know if this analysis is as accurate as it could be.

Another limitation is that we are also only looking at horizontal break and whether it is less than or greater than 2. Limiting this category to these 2 groups can show us which one is better but there are lots of values in each group. We could break this up into more detailed groups with fewer values in each one.

Another approach that could be taken is looking at more variables that could affect exit speed. When reporting our data to a team, we might tell them the result of our horizontal break analysis will help them. However, reporting an analysis with multiple variables that all relate to possible effects on exit speed would help them more because they could see what they want their team to work on all around not just in one spot.

## Problem 2

Our second analysis is on two baseball data variables from a different dataset. The variables are Tagged pitch type & Play result. Tagged pitch type tells us what kind of pitch the pitcher threw for this play. In this second problem, we are looking at whether the pitch was tagged as a fastball or other. And the second variable, Play result, is what resulted from the play. The results we are looking at are categorized into single and extra. In this case, extra means that the result of the play was either a double, triple, or homerun.

The variables we will look at are as follows;

X = Tagged Pitch Type

* Fastball/ Other

Y = Play Result

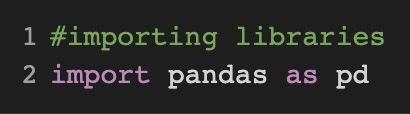
* Single/ Extra

# Expectation

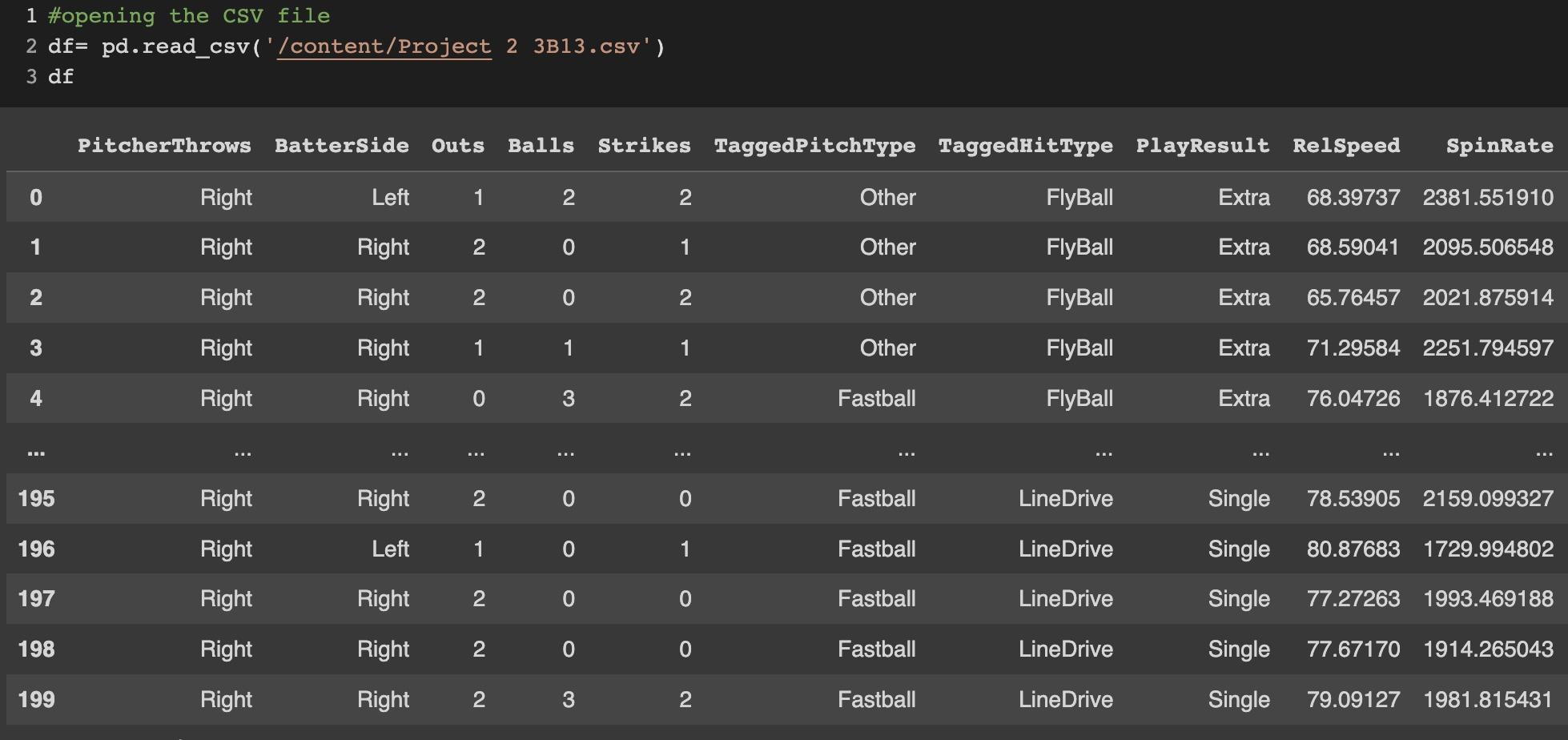
We would expect that the variables tagged pitch type and play result would be independent of each other. We believe this because it is not necessarily the type of the pitch that results in the play but the values of the pitch itself. The type of pitch does not determine what kind of ball the batter is getting. Each pitcher has a different fastball, curveball, slider, etc. Additionally, we are only looking at the difference between a fastball and everything else and because of this, there is not enough data to state there will be a difference in play results.

# Data

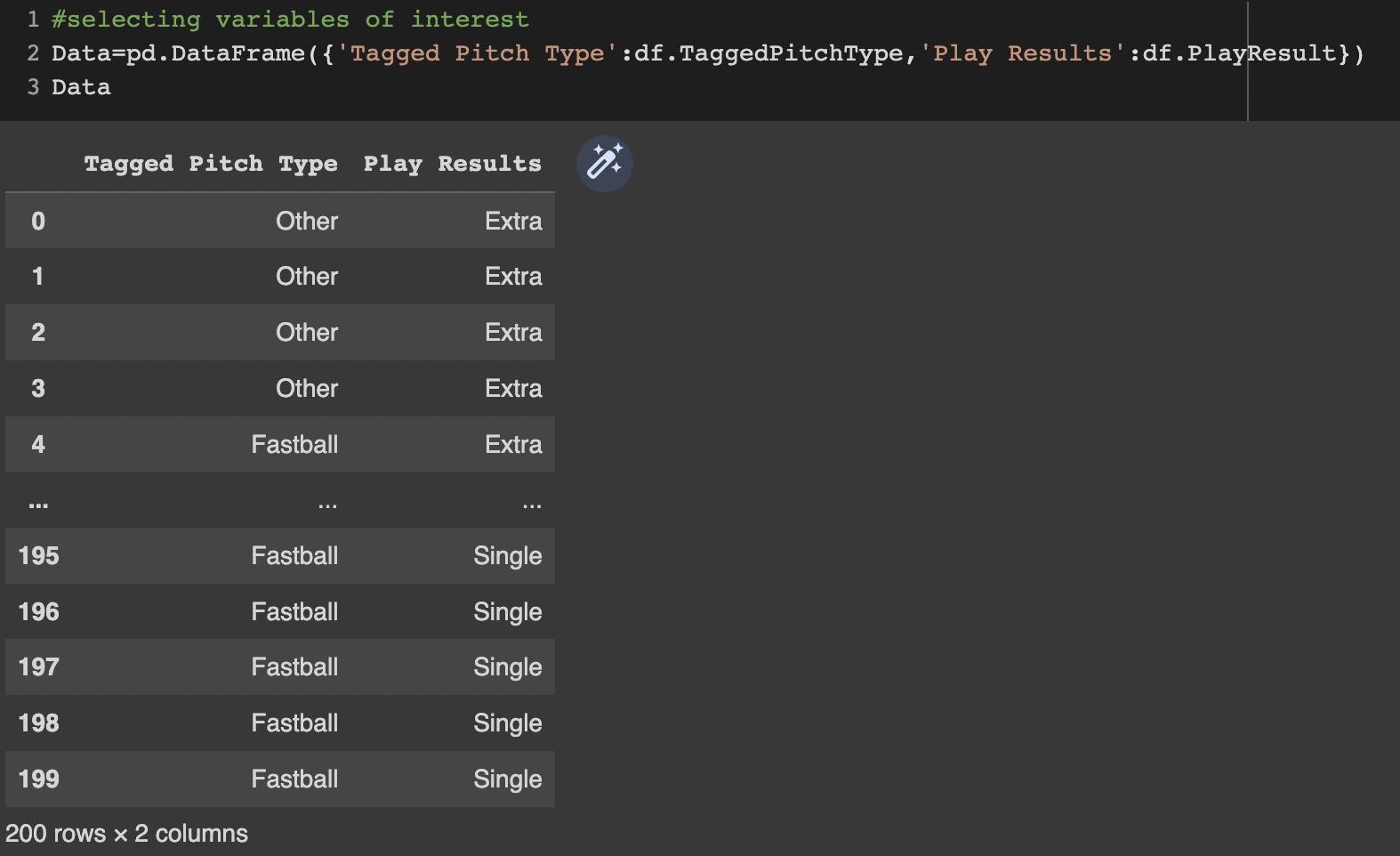
For this dataset, we similarly used Python to analyze the dataset and find the values for the actual contingency table. We first imported the libraries we will be using:



We then uploaded the second CSV file and named it df and viewed it:

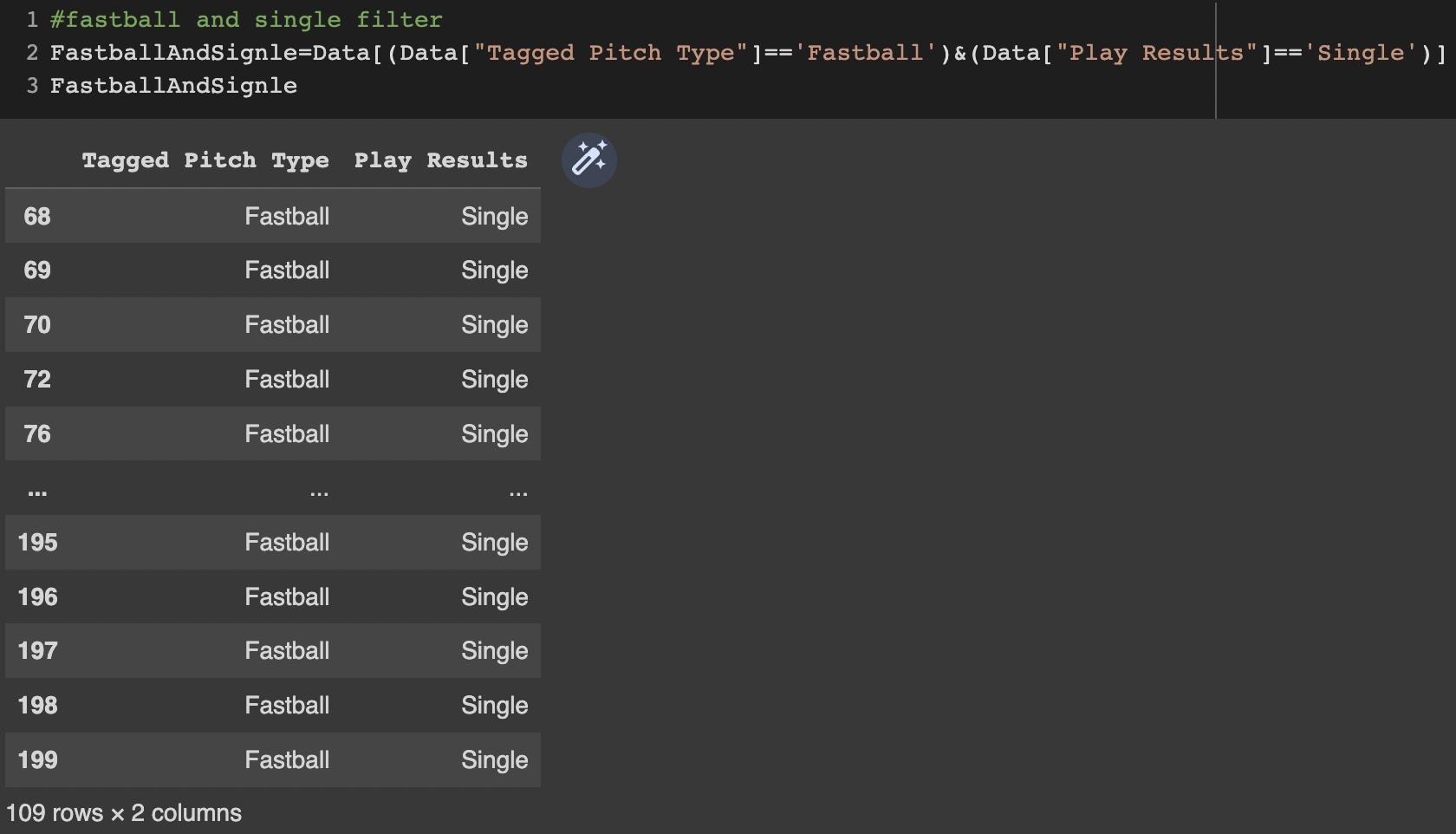


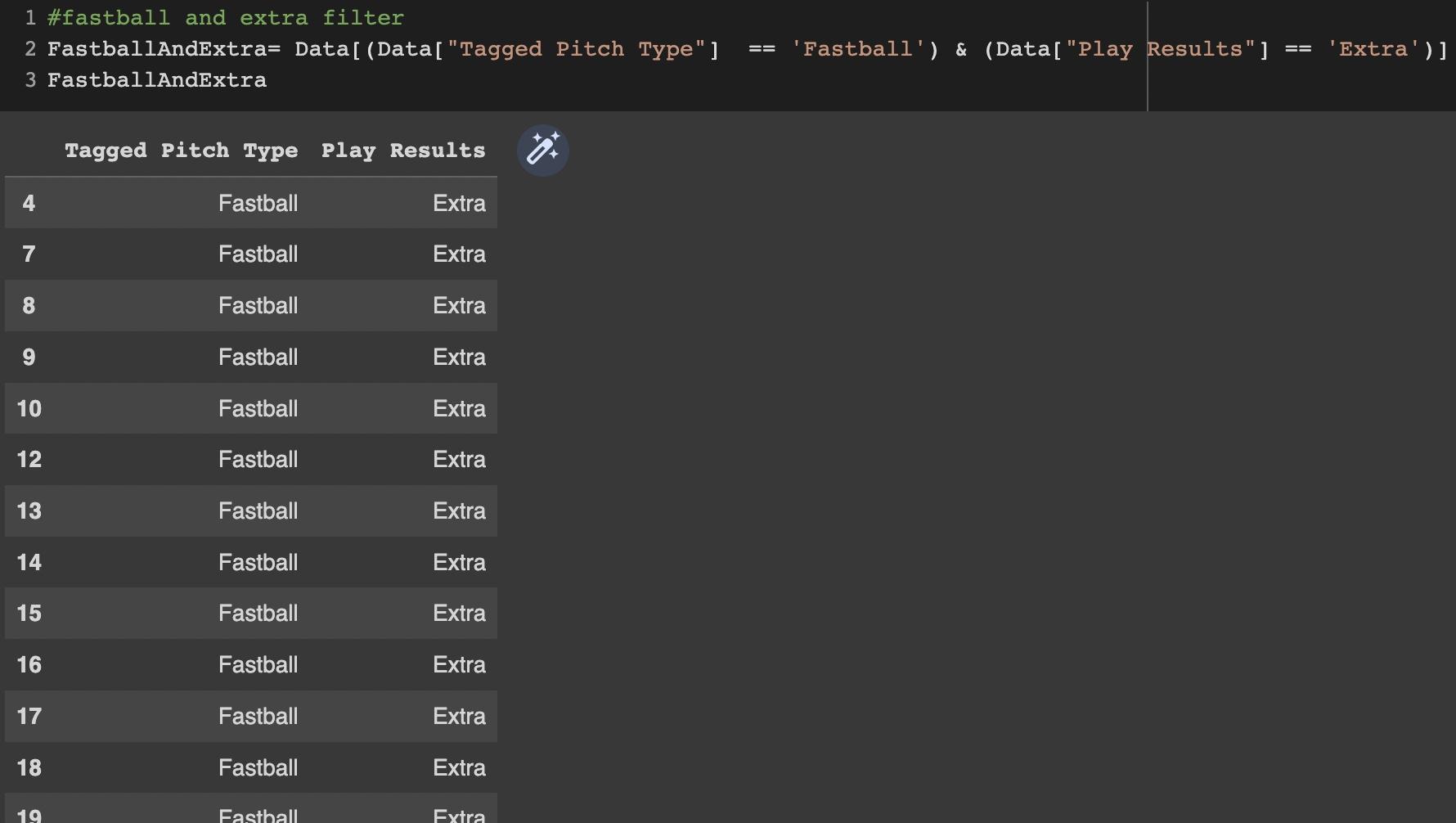
After making sure there isn't any issue with the dataset, we made a new dataset that consisted of tagged pitch type and play results. We called this data set Data:



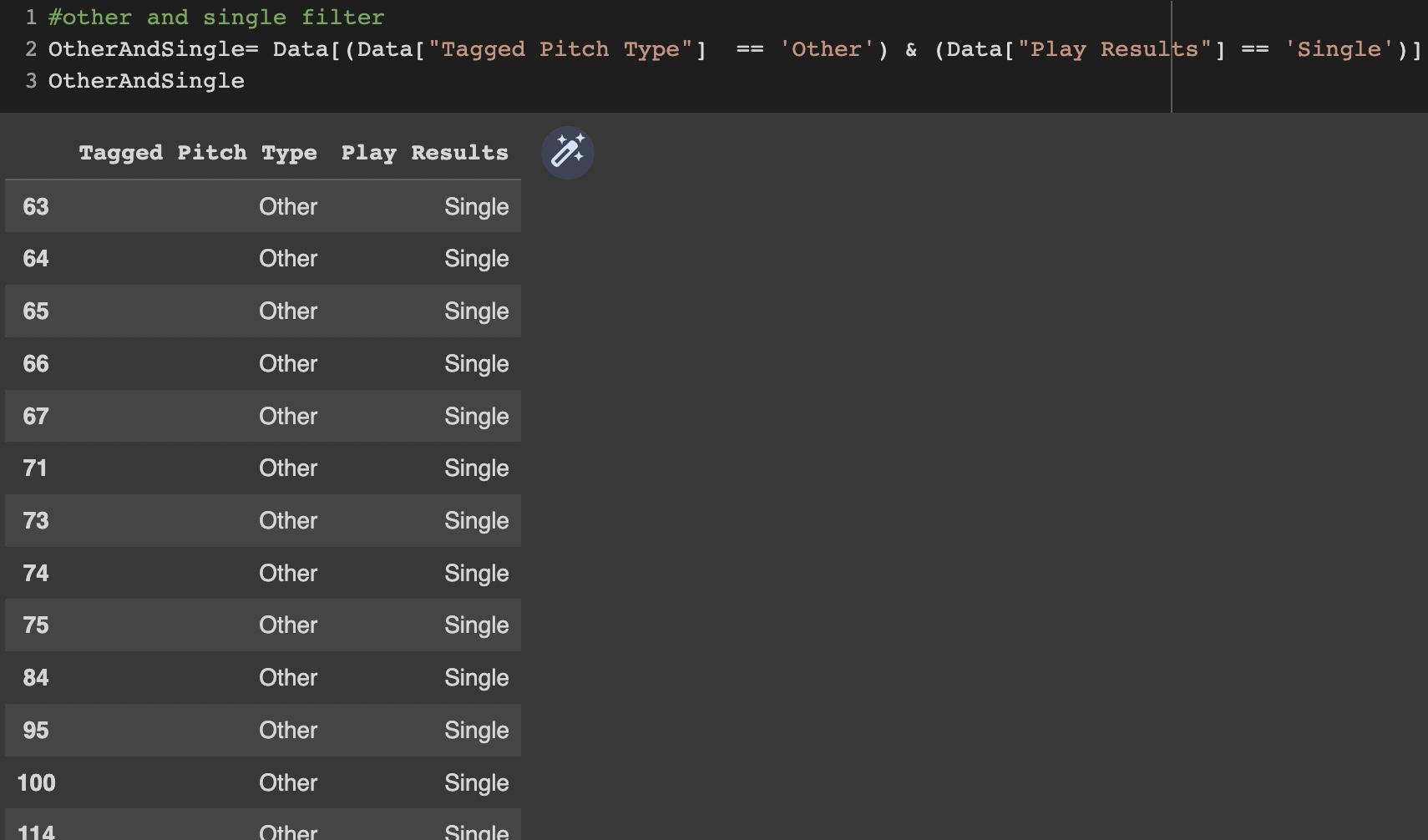
The next step was finding the actual values for the contingency table. We did this by filtering out the two variables we were looking for in each section. This was done four times for the values of fastball and single, fastball and extra, other and single, and other and extra:

Fastball and Single:

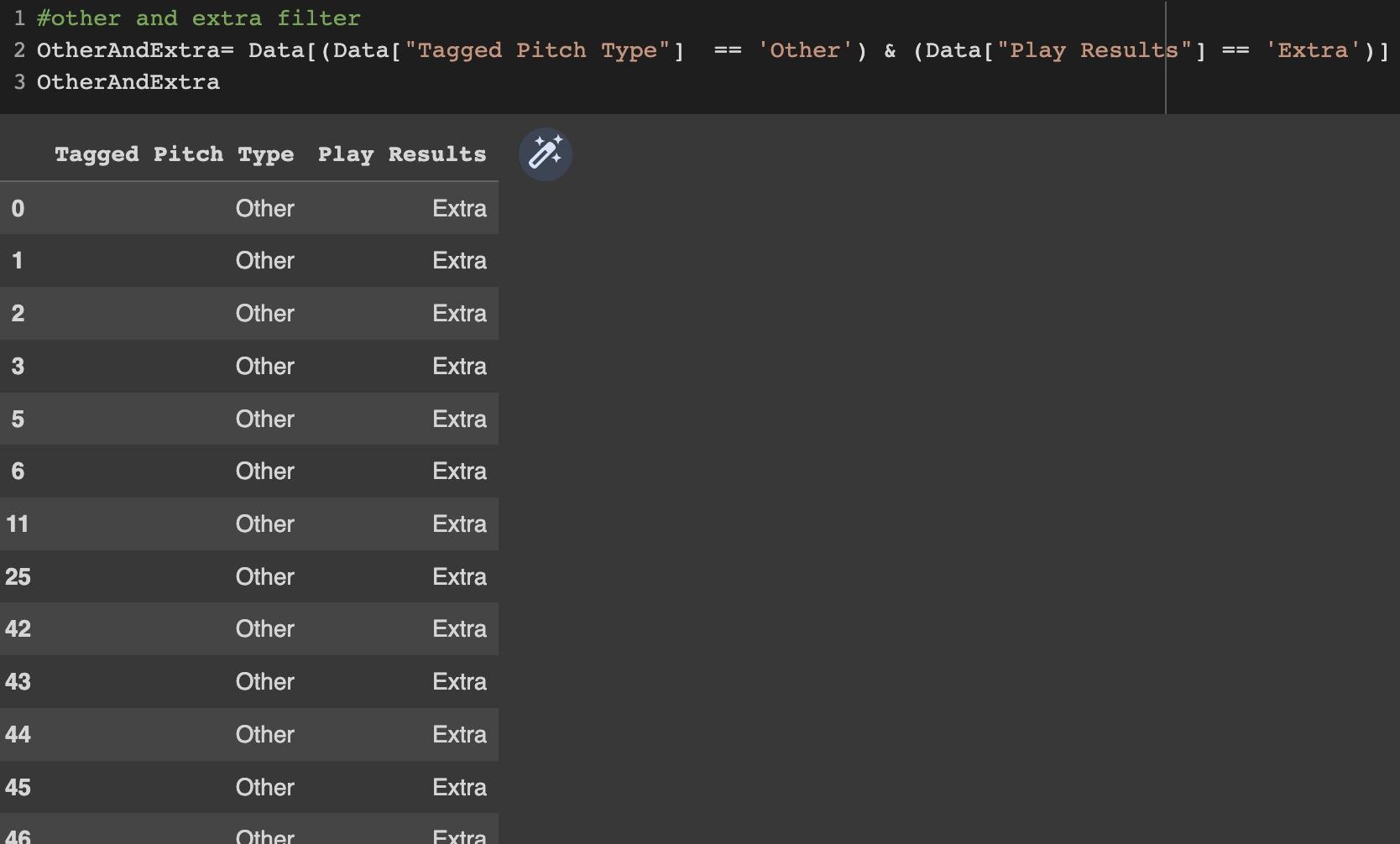


Fastball and Extra:

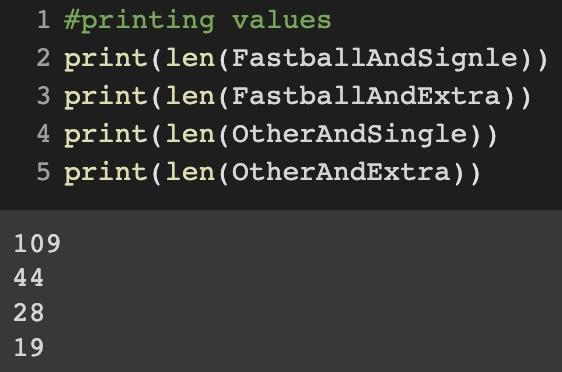
Other and Single:



Other and Extra:



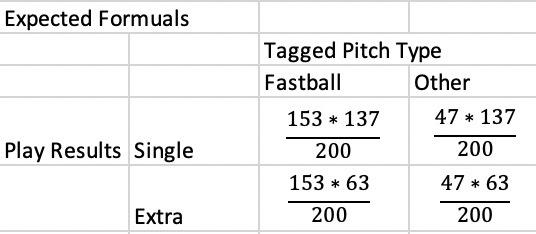
We then found the length of each filter to determine how many plays from the original dataset had the parameters we were looking for:



This is how we ended up with the contingency table we did with the actual values that which is displayed in the next section (i.e., *Table 1*).

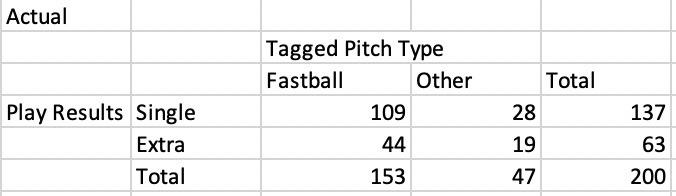
Using the values from table 1, we were also able to compute the expected values and create a second contingency table (i.e., *Table 2*).

The equation to find the expected values is the probability of the first parameter times the probability of the second parameter. This simplifies to the total of the first parameter times the total of the second parameter divided by the overall total. The equations are displayed below:

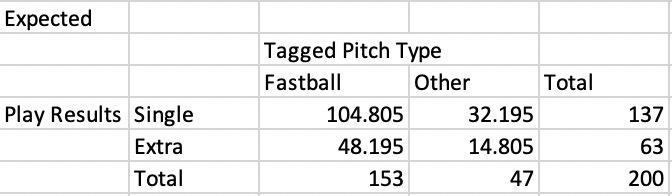


# Contingency Tables

## **Table 1:**



## **Table 2:**



# Hypothesis Test

The hypothesis test for this problem focuses on whether or not the two variables of tagged pitch type and play results are independent or not. Our null hypothesis is that the two variables are independent, and our alternative hypothesis is that the two variables are not independent.

In order to find the p-value, we conducted a Chi-square test using Mathematica. Since we need the test statistic to calculate the p-value, we used the equation to compute the test statistic and then calculated the degrees of freedom. The degrees of freedom is computed by multiplying the degrees of freedom of the tagged pitch type and the degrees of freedom of the play results.

**Test-statistic:**

**df**: (2-1)\*(2-1) = 1

We then used Mathematica to conduct the Chi-square test:

# Interpretation

With an alpha level of .05 and a p-value of 0.132, we fail to reject the null hypothesis and we conclude that the two variables are independent. This means that the type of pitch and the play results are not related based on the parameters we looked at. We would have no way of predicting what the play result would be based on the tagged pitch type.

# How can this information be useful to a baseball player/coach?

If we were reporting this information to a baseball team, we would be able to tell them that the type of pitch and the result of the play are not directly related. Coaches can use this information to determine that it is not just important for their pitchers to perfect just one type of pitch but that they should practice them all if they are concerned about play results.

This information would be more helpful to players/coaches if there were more variables involved. This would be helpful because we could look at what affects play results the most and the team could focus on that. This problem only tells them that pitch type does not affect this.

# Limitations

Similar to problem 1, we have a small sample size which most likely consists of only a couple of pitchers from one or two teams. This could be skewing the data because baseball players hit differently depending on the pitcher. Only having data from a couple of teams could also be giving us inaccurate results due to the fact that it's most likely the same players hitting every time with a couple of exceptions so it represents the sample of baseball players from these teams rather than the population of baseball players overall. Additionally, we only looked at if the pitch was a fastball or not. This could be affecting our analysis if the other pitch types cause different results from each other.

# Conclusion

Overall, we can conclude that horizontal break does have an effect on exit speed and that the type of pitch is not directly related to the result of the play. If we were to do something similar to this project in the future, with the intent of sharing our findings with the Roanoke College baseball team, we would add more variables and make our hypothesis more detailed by not limiting the variables as much. This would allow players and coaches to have more information on how they can improve their games.

# Citation

Caffo, B., Wackerly, D., Mendenhall, W., Scheaffer, R., & Jones, G. (2002). *Mathematical Statistics with applications* (7th ed.). Duxbury.