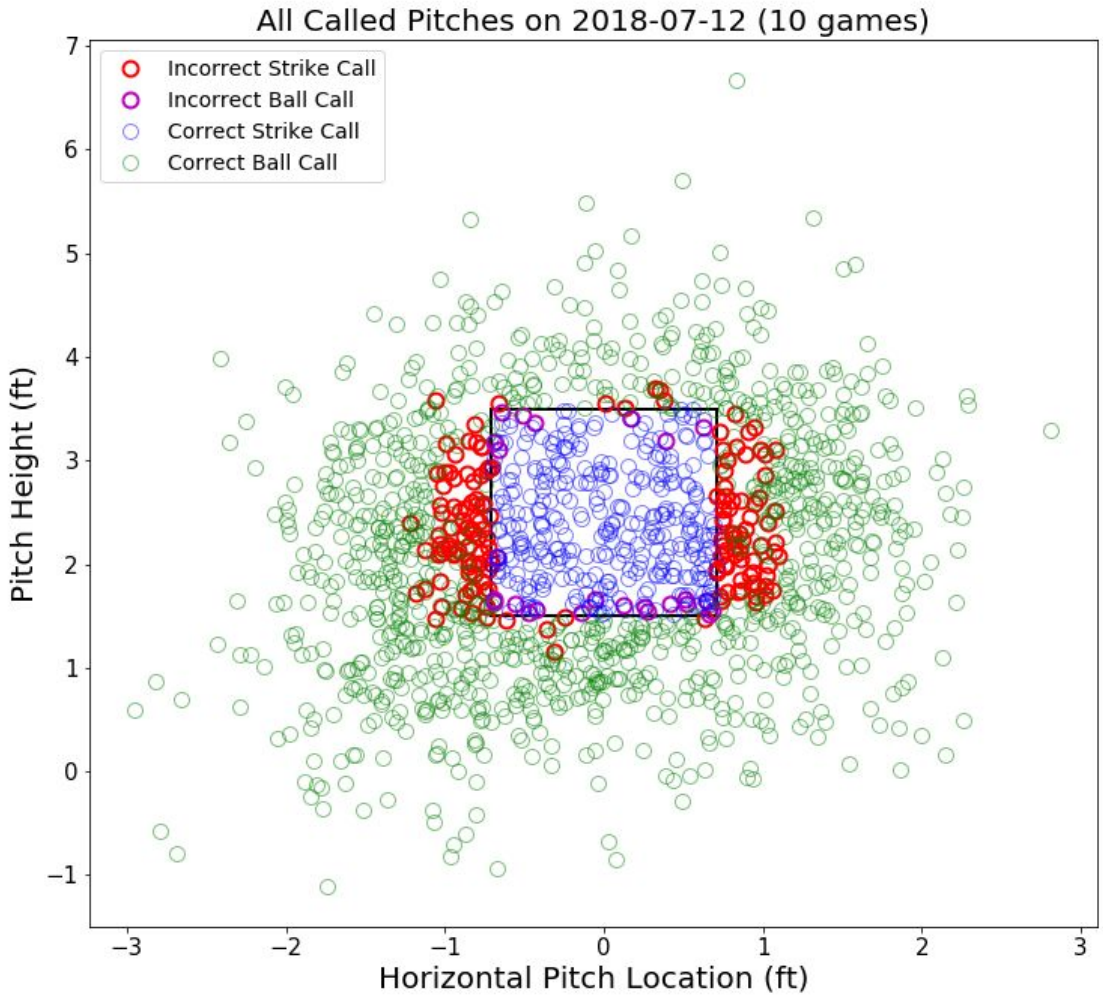
Zachary Kopstein
2.980 - Team MLB
Individual Technical Deliverable #1

Problem Statement: Choose a set of 10 games and visualize balls called, strikes called, and mistakes made by the umpire.

Analysis/Visualization 1:

As someone completely unfamiliar with SQL and not very comfortable in Python, I first set out to recreate the pitch location vs pitch height visualization from our project overview document, since I knew what that should look like. My visualization had one key difference, which was to graphically differentiate between the two types of umpire error - strikes that should have been called balls, and balls that should have been called strikes.

Since there is not yet the ability to sort the data by individual games, I chose to look at all games from a single date that happened to have 10 games played on it - July 12, 2018. I used four SQL queries to separate the 4 different potential umpire calls - correctly called strike, correctly called ball, incorrectly called strike, and incorrectly called ball. I did this using the parameters PlateLocationHeight, PlateLocationSide, and PitchCall. I compared the horizontal location to the width of home plate (which is 17 inches wide), and compared the vertical location to an "average" vertical strike zone: 1.5 to 3.5 feet off the ground. After determining which calls were accurate and which weren't, I plotted them using matplotlib. Full code can be found in the appendix at the end of the document. Here's the resulting graphic:



Note that currently there are a few circles which appear to be mislabeled (i.e. red circles, or "incorrect strike calls" that are touching the strike zone). This is due to the fact that the marker size (more representative of a ball, but not precisely accurate) was not taken into account during the analysis. This means that only the center of the data point was determined to be inside or outside of the strike zone, which is something that can and should be modified going forward.

This plot gives us a few key insights into the nature of umpiring and how an automatic strike zone might affect the game. First, the plot seems to indicate that many more balls are incorrectly called strikes by umpires than strikes that are incorrectly called balls. This appears to show that umpires have a larger zone than the "true" strike zone. In addition, this plot indicates that umpires are worse

at judging pitches side to side than they are at judging pitches vertically. There are lots of balls a few inches inside or outside of the strike zone that are called strikes, while there are relatively few missed calls above or below the strike zone.

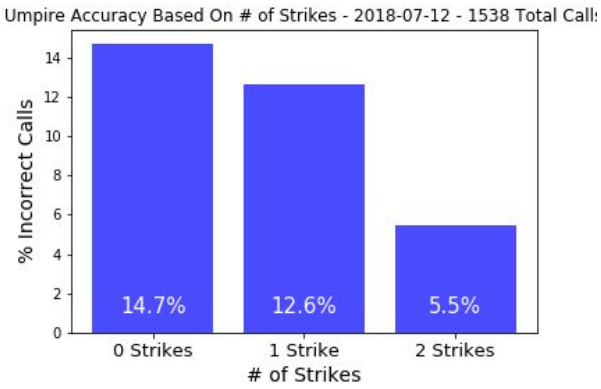
Suspicious Data/Limitations:

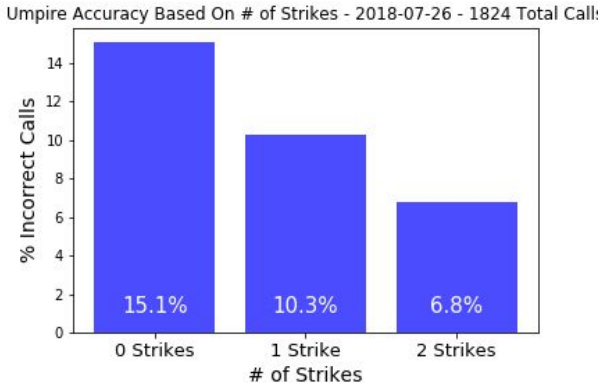
Based on my personal experience watching MLB games, this data does not appear to be suspicious. I'd expect many missed calls from side to side, as umpires tend to reward pitchers who throw it exactly where the catcher sets up, even if that's an inch or two off of the plate. Pitches an inch or two high or low tend to be rewarded less. One drawback of my analysis technique is that the vertical strike zone doesn't account for individual height differences. For instance a 6'6" batter and 5'10" batter have the same strike zone for this analysis, even though their knees and chest are at different heights. This could lead to some vertical errors in my pitch classification and mischaracterization of whether a call is correct or not.

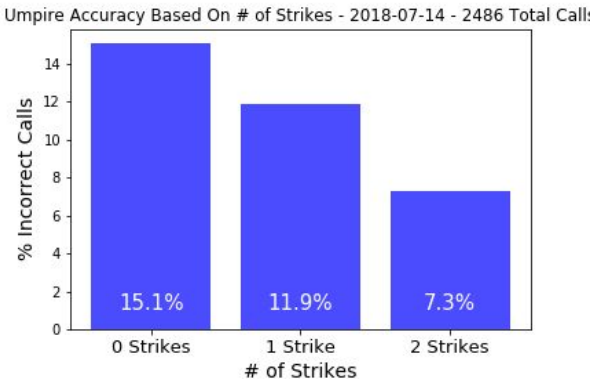
Analysis/Visualization 2:

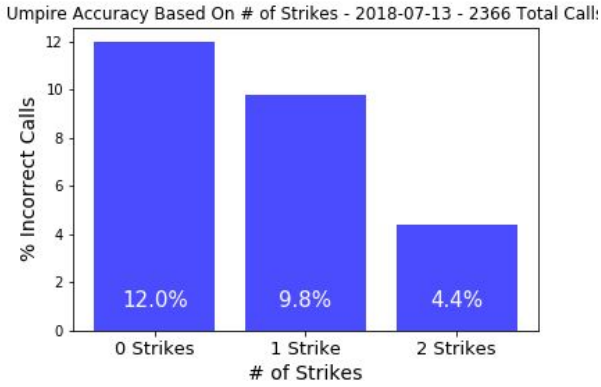
I did a second analysis because I was interested in seeing how umpire accuracy changes with the number of strikes there are in the count. Do umpires tend to "lock in" more when there are two strikes?

For this analysis, I repeated the four queries (for all games in one day) from the previous analysis for each of three scenarios: counts with zero strikes, counts with one strike, and counts with two strikes. I found the percentage of missed calls for each of these buckets by adding both umpire error types together and dividing by the total number of calls made that day. Then, I plotted the results using bar graphs in matplotlib to more easily visualize how those percentages differ based on the number of strikes. I analyzed 10 different days (roughly 150 games), and the resulting data were quite consistent from day to day. Representative bar graphs of 4 different days are shown below:









The results from this analysis were fairly surprising to me. This data seems to suggest that umpires are more than two times more likely to make an error when there are no strikes than they are when there are two strikes. This indicates that maybe umpires might indeed "lock in" and become more accurate with two strikes, when their call means the difference between a strikeout or the at bat continuing.

Suspicious Data/Limitations:

Although it's possible I messed up my code, this data seems to be accurate and consistent across all dates, and does make some logical sense as well. Even though I acknowledged earlier that the vertical strike zone assumptions made in this analysis may lead to some errors, those same errors would be consistent across the three different strike buckets used in this secondary analysis, and thus the results should still be significant. Results from more thorough analyses like this one could be used in future MLB umpire trainings, as being aware of biases based on count could lead to more accurate pitch calling.

Appendix A— Visualizing balls/strikes/missed calls on 7/12/2018

```
##### SQL QUERIES: #####

import matplotlib.pyplot as plt
%matplotlib inline

sql1 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12)
    OR (PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5))
    """

sql2 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12
    AND PlateLocationSide <= 8.5/12) AND (PlateLocationHeight BETWEEN 1.5 AND
3.5))
    """

sql3 = """

    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5)))
    """

sql4 = """
```

```
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5)))
    """
missed_location_strike_df = bq_client.query(sql1).to_dataframe()
missed_location_ball_df = bq_client.query(sql2).to_dataframe()
all_other_strikes_df = bq_client.query(sql3).to_dataframe()
all_other_balls_df = bq_client.query(sql4).to_dataframe()


##### PLOTTING: #####

fig = plt.figure(figsize=(12,24))

ax1 = fig.add_subplot(212)
plt.plot('PlateLocationSide', 'PlateLocationHeight', 'ro', label = 'Incorrect
Strike Call', markersize = 10, markeredgewidth = 2, mfc='none',
data=missed_location_strike_df)

ax1 = fig.add_subplot(212)
plt.plot('PlateLocationSide', 'PlateLocationHeight', 'mo', label = 'Incorrect
Ball Call', markersize = 10, markeredgewidth = 2, mfc='none',
data=missed_location_ball_df)

ax1 = fig.add_subplot(212)
plt.plot('PlateLocationSide', 'PlateLocationHeight', 'bo', label = 'Correct
Strike Call', markersize = 10, alpha = 0.5, mfc='none',
data=all_other_strikes_df)

ax1 = fig.add_subplot(212)
plt.plot('PlateLocationSide', 'PlateLocationHeight', 'go', label = 'Correct
Ball Call',markersize = 10, alpha = 0.5, mfc='none', data=all_other_balls_df)
```

```
plt.legend(loc='upper left', fontsize = 14)
plt.xlabel('Horizontal Pitch Location (ft)', fontsize = 20)
plt.ylabel('Pitch Height (ft)', fontsize = 20)
plt.title('All Called Pitches on 2018-07-12 (10 games)', fontsize = 20)
ax1.tick_params(axis='both', which='major', labelsize=15)

rectangle = plt.Rectangle((-8.5/12,1.5), 8.5/6, 2, fc='none',ec="black",lw='2')
plt.gca().add_patch(rectangle)

from google.colab import files
plt.savefig("2980.png")
files.download("2980.png")
```

Appendix B— Visualizing missed call percentage based on number of strikes:

```
import matplotlib.pyplot as plt
%matplotlib inline


#### SQL QUERIES ####

#### 0 strike queries: ####
sql01 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12)
    OR (PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)) AND (Strikes =
0)
    """
sql02 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12
    AND PlateLocationSide <= 8.5/12) AND (PlateLocationHeight BETWEEN 1.5 AND
3.5)) AND (Strikes = 0)
    """
sql03 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5))) AND (Strikes = 0)
    """
```

```
sql04 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5))) AND (Strikes = 0)
    """

#### 1 strike queries: ####
sql11 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12)
    OR (PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)) AND (Strikes =
1)
    """
sql12 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12
    AND PlateLocationSide <= 8.5/12) AND (PlateLocationHeight BETWEEN 1.5 AND
3.5)) AND (Strikes = 1)
    """
sql13 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled"))
```

```
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5))) AND (Strikes = 1)
    """
sql14 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5))) AND (Strikes = 1)
    """
#### 2 strike queries: ####

sql21 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12)
    OR (PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)) AND (Strikes =
2)
    """
sql22 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12
    AND PlateLocationSide <= 8.5/12) AND (PlateLocationHeight BETWEEN 1.5 AND
3.5)) AND (Strikes = 2)
    """
```

```
sql23 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "StrikeCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5))) AND (Strikes = 2)
    """
sql24 = """
    SELECT PlateLocationHeight, PlateLocationSide, Inning, PitchCall, Balls,
Strikes
    FROM pitches.mlb_pitches_3years
    WHERE Date ="2018-07-12" AND ((PitchCall like "BallCalled"))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "StrikeCalled")) AND
((PlateLocationSide <= -8.5/12 OR PlateLocationSide >= 8.5/12) OR
(PlateLocationHeight < 1.5 OR PlateLocationHeight > 3.5)))
    AND NOT (Date ="2018-07-12" AND ((PitchCall like "BallCalled")) AND
((PlateLocationSide >= -8.5/12 AND PlateLocationSide <= 8.5/12) AND
(PlateLocationHeight BETWEEN 1.5 AND 3.5))) AND (Strikes = 2)
    """

missed_strike_zero_strikes_df = bq_client.query(sql01).to_dataframe()
missed_strike_one_strike_df = bq_client.query(sql11).to_dataframe()
missed_strike_two_strikes_df = bq_client.query(sql21).to_dataframe()

num_missed_strikes_zero_strikes = len(missed_strike_zero_strikes_df)
num_missed_strikes_one_strike = len(missed_strike_one_strike_df)
num_missed_strikes_two_strikes = len(missed_strike_two_strikes_df)

#######
missed_ball_zero_strikes_df = bq_client.query(sql02).to_dataframe()
missed_ball_one_strike_df = bq_client.query(sql12).to_dataframe()
missed_ball_two_strikes_df = bq_client.query(sql22).to_dataframe()

num_missed_balls_zero_strikes = len(missed_ball_zero_strikes_df)
```

```
num_missed_balls_one_strike = len(missed_ball_one_strike_df)
num_missed_balls_two_strikes = len(missed_ball_two_strikes_df)

#######
all_other_strikes_zero_strikes_df = bq_client.query(sql03).to_dataframe()
all_other_strikes_one_strike_df = bq_client.query(sql13).to_dataframe()
all_other_strikes_two_strikes_df = bq_client.query(sql23).to_dataframe()

all_other_balls_zero_strikes_df = bq_client.query(sql04).to_dataframe()
all_other_balls_one_strike_df = bq_client.query(sql14).to_dataframe()
all_other_balls_two_strikes_df = bq_client.query(sql24).to_dataframe()

#######
num_correct_calls_zero_strikes =
len(all_other_strikes_zero_strikes_df)+len(all_other_balls_zero_strikes_df)
num_correct_calls_one_strike =
len(all_other_strikes_one_strike_df)+len(all_other_balls_one_strike_df)
num_correct_calls_two_strikes =
len(all_other_strikes_two_strikes_df)+len(all_other_balls_two_strikes_df)

#######
num_missed_calls_zero_strikes = num_missed_strikes_zero_strikes +
num_missed_balls_zero_strikes
num_missed_calls_one_strike = num_missed_strikes_one_strike +
num_missed_balls_one_strike
num_missed_calls_two_strikes = num_missed_strikes_two_strikes +
num_missed_balls_two_strikes

######

total_calls_zero_strikes = num_correct_calls_zero_strikes +
num_missed_calls_zero_strikes
total_calls_one_strike = num_correct_calls_one_strike +
num_missed_calls_one_strike
total_calls_two_strikes = num_correct_calls_two_strikes +
num_missed_calls_two_strikes

total_calls = total_calls_one_strike + total_calls_two_strikes +
total_calls_zero_strikes
```

```
labels = ('0 Strikes', '1 Strike', '2 Strikes')
percentages = [100*num_missed_calls_zero_strikes/total_calls_zero_strikes,
100*num_missed_calls_one_strike/total_calls_one_strike,
100*num_missed_calls_two_strikes/total_calls_two_strikes]
percentages = list(np.around(np.array(percentages),1))

#### PLOTTING BAR CHART: ####

plt.bar(labels, percentages, align='center', alpha=0.7, color = 'blue')

plt.xticks(labels, fontsize = 13)
plt.xlabel('# of Strikes', fontsize = 14)
plt.ylabel('% Incorrect Calls', fontsize = 14)
plt.title('Umpire Accuracy Based On # of Strikes - 2018-07-12 - ' +
str(total_calls) + ' Total Calls')

plt.text(0, 1, str(percentages[0]) + '%', horizontalalignment='center',
fontsize = 15, color = 'white')
plt.text(1, 1, str(percentages[1]) + '%', horizontalalignment='center',
fontsize = 15, color = 'white')
plt.text(2, 1, str(percentages[2]) + '%', horizontalalignment='center',
fontsize = 15, color = 'white')


from google.colab import files
plt.savefig("2980.png")
files.download("2980.png")

plt.show()
```