Professional Sports Teams and Crime Rates

INFOW18 Summer 2020
By Steven Burstyn, Isabel Zheng, and Jack Barrera

Context

Professional sports teams provide an opportunity for cities to come together in support of a common entity. Fans across the United States take pride in the success of their hometown teams. However, what happens when a team fails to perform to the expectations of their fanbase, and exactly to what extent does team disappointment impact fan behavior? Our project investigates this question, by looking how a city's sports team's performance affects the crime rates in the city to find out exactly how invested fans are in their teams.

Questions

- In less successful years for a sports team, do local crime rates increase?
- What metric of a team's success best correlates with crime rates?
- Do more serious crimes have a significantly increased association with team performance in comparison to all crimes?
- Do year over year trends and team momentum have an impact?

Dataset Choice

We've decided to focus on a single city's crime data, and only one of their teams in order to minimize variation, emphasizing that our analysis can be extrapolated to other teams and cities. When considering candidates, we knew we needed a city with fairly high population and crime rates to provide a large pool data year over year. We also needed a team that has had both success and failure for the period of data being analyzed. Our city of choice was Chicago and our team of choice was the Chicago Bulls. Chicago has a high population, high overall crime volume and accessible data, and the Bulls are one of the most successful franchises in sports history, but have had varying results recently.

Source Data

Our data consists of two main data sets, Chicago crime data from 2001 to present and the Chicago Bulls historical performance data.

Our source for the crime data was the Chicago Data Portal:
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

The size of the CSV dataset is 7165841 rows and 22 columns, with variables categorized as:

ID - Unique identifier for the record
Case Number - The Chicago Police Department RD Number
Date - Date when the incident occurred
Block - The partially redacted address where the incident occurred
IUCR - The Illinois Uniform Crime Reporting code.
Primary Type - The primary description of the IUCR code
Description - The secondary description of the IUCR code
Location Description - Description of the location where the incident occurred
Arrest - Indicates whether an arrest was made.
Domestic - Indicates whether the incident was domestic-related
Beat - Indicates the beat where the incident occurred.
District - Indicates the police district where the incident occurred
Ward - The ward (City Council district) where the incident occurred
Community Area - Indicates the community area where the incident occurred
FBI Code - Indicates the crime classification as outlined by the FBI
X Coordinate - The x coordinate of the location where the incident occurred
Y Coordinate - The y coordinate of the location where the incident occurred
Year - Year the incident occurred
Updated On - Date and time the record was last updated
Latitude - The latitude of the location where the incident occurred
Longitude - The longitude of the location where the incident occurred
Location - The location where the incident occurred in a format that allows for creation of maps

Our source for the team performance data was Basketball Reference: https://www.basketball-reference.com/teams/CHI/

The size of the CSV dataset was 54 rows and 19 columns. The rows each represent a season and the variables are as follows:

Season
League
Team
Wins
Losses
Win/Loss Percentage

Finish in Division
Simple Rating System - takes into account point differential and strength of schedule
Pace - Estimate of possessions per 48 minutes
Rel Pace - Team's possessions per 48 minutes relative to the league
Offensive Rating - Estimate of team points scored per 100 possessions
Relative Offensive Rating - Team's Offensive Rating relative to the rest of the league
Defensive Rating - Estimate of points allowed per 100 possessions
Relative Defensive Rating - Team's Defensive Rating relative to the rest of the league
Playoff Results
Coach
Top Win Share - Highest win share player for the season

Assumptions

Our crime data is categorized by year, but the NBA season begins in October and ends in June, partially spanning two years. We've made the assumption that the year for crimes corresponds with statistics from the year in which the NBA season ends, due to the fact that the emotional impact of the season on the fanbase will likely continue throughout the rest of the year.

Later in our analysis, we will look into how arrests are affected by the Chicago Bulls performance, and we assume that arrests mean that it's a more serious crime. The data is categorized, but with 36 categories, we decided to use arrests as a binary evaluation of a crime being more serious for simplification.

Preliminary Crime Data Exploration and Cleaning

Our first task was to explore and validate the crime data, a relatively large dataset, and isolate the variable most relevant to our questions and goals. We first grouped the data by year, and took a count of total data points for each corresponding year.

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
485779	486762	475960	469395	453731	448136	437039	427095	392762	370393	351870
2012	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**		
336124	307271	275517	264422	269380	268621	268115	259640	117829		

Statistical Data for Crime Counts by Year

count	mean	std	min	25%	50%	75%	max
19.00	370,948.00	87,547.87	259,640.00	272,448.50	370,393.00	450,933.50	486,762.00

It's important to note here that the total number of crimes decreases each year from 2003 on and in our time period, the total number of crimes decreased by 75%, an extremely significant change. Also, 2020 has a much lower value in comparison to the rest of the years, likely because the year is not complete. For that reason, it is not included in our statistical analysis, and won't be included in our later visualizations.

We then looked at the number of crimes that resulted in eventual arrests, citing our assumption from earlier where an arrest is indicative of a more serious crime.

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
141910	141555	141577	144690	140902	135394	131862	109973	110784	100501	96256

2012	2013	2014	2015	2016	2017	2018	2019	2020
90610	86489	79578	69970	52936	52546	53678	55355	19543

Statistical Data for Arrest Counts by Year

count	mean	std	min	25%	50%	75%	max
19.00	101,924.53	34,564.79	52,546.00	74,774.00	100,501.00	138,148.00	144,690.00

Now that the variables are isolated and we've calculated the corresponding statistical data, we can merge the data frames and calculate the proportion of crimes that resulted in an arrest.

Year	Total Crimes	Total Arrests	% of Crimes Resulting in Arrests
2001	485779	141910	0.29
2002	486762	141555	0.29
2003	475960	141577	0.30
2004	469395	144690	0.31
2005	453731	140902	0.31
2006	448136	135394	0.30
2007	437039	131862	0.30
2008	427095	109973	0.26
2009	392762	110784	0.28
2010	370393	100501	0.27
2011	351870	96256	0.27
2012	336124	90610	0.27
2013	307271	86489	0.28
2014	275517	79578	0.29
2015	264422	69970	0.26
2016	269380	52936	0.20
2017	268621	52546	0.20
2018	268115	53678	0.20
2019	259640	55355	0.21
2020	117829	19543	0.17

Statistical Data for Arrest Proportion

count	mean	std	min	25%	50%	75%	max
19.00	0.27	0.04	0.20	0.26	0.28	0.29	0.31

Our crime data has been sorted, and we've isolated the variables necessary for our primary analysis. Our intent at this point is to return back to this dataset with our initial findings from the crime and arrest data, and hopefully investigate other variables later.

Preliminary NBA Data Analysis and Cleaning

Our next task is to clean up and validate the Chicago Bulls historical performance dataset, which is much smaller in comparison. Our metric of choice for the team success is the Win Loss percentage, but we also included the SRS as a secondary statistics, which is a unique rating system that considers team performance relative to the rest of the league in the corresponding year.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
W/L%	0.18	0.26	0.37	0.28	0.57	0.50	0.60	0.40	0.50	0.50	0.76	0.76	0.55
SRS	-9.09	-8.52	-5.31	-6.69	0.65	0.51	4.52	-3.19	-0.16	-1.64	6.53	7.43	-0.02

2014	2015	2016	2017	2018	2019	2020
0.58	0.61	0.51	0.50	0.33	0.27	0.34
1.20	2.54	-1.46	0.03	-6.84	-8.32	-3.85

Statistical Data for W/L% and SRS

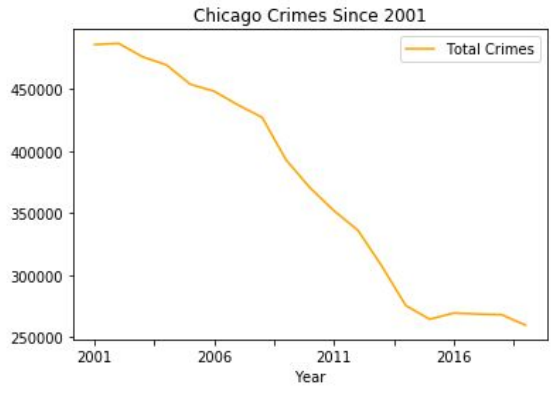
	count	mean	std	min	25%	50%	75%	max
W/L%	20.00	0.47	0.16	0.18	0.34	0.50	0.58	0.76
SRS	20.00	-1.58	4.87	-9.09	-5.66	-0.81	0.79	7.43

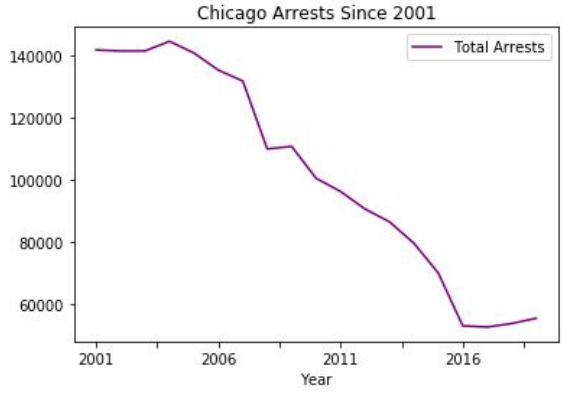
Now we combine both the crime data frame and the Chicago Bulls data frames to get our final data frame for analysis.

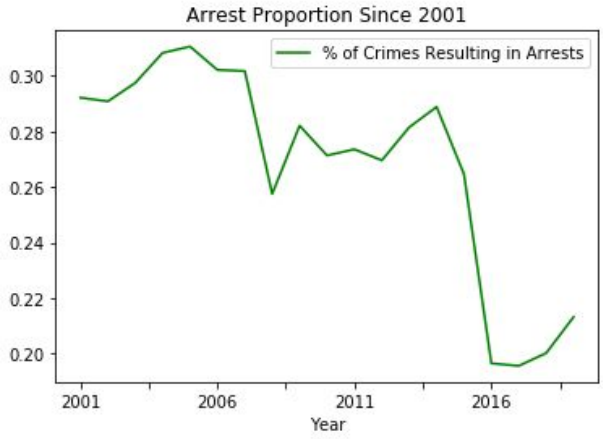
Year	Total Crimes	Total Arrests	% of Crimes Resulting in Arrests	W/L%	SRS
2001	485779	141910	0.29	0.18	-9.09
2002	486762	141555	0.29	0.26	-8.52
2003	475960	141577	0.30	0.37	-5.31
2004	469395	144690	0.31	0.28	-6.69
2005	453731	140902	0.31	0.57	0.65
2006	448136	135394	0.30	0.50	0.51
2007	437039	131862	0.30	0.60	4.52
2008	427095	109973	0.26	0.40	-3.19
2009	392762	110784	0.28	0.50	-0.16
2010	370393	100501	0.27	0.50	-1.64
2011	351870	96256	0.27	0.76	6.53
2012	336124	90610	0.27	0.76	7.43
2013	307271	86489	0.28	0.55	-0.02
2014	275517	79578	0.29	0.58	1.20
2015	264422	69970	0.26	0.61	2.54
2016	269380	52936	0.20	0.51	-1.46
2017	268621	52546	0.20	0.50	0.03
2018	268115	53678	0.20	0.33	-6.84
2019	259640	55355	0.21	0.27	-8.32
2020	117829	19543	0.17	0.34	-3.85

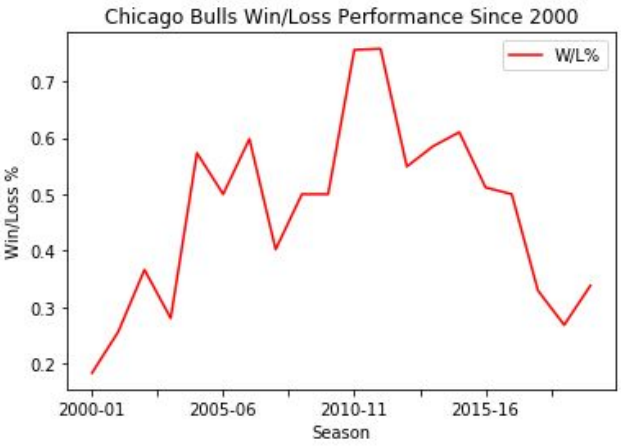
Analysis

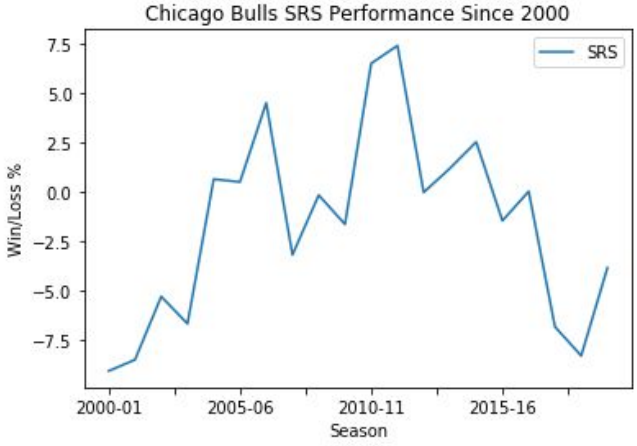
For our analysis, we created plots mapping the changes in values through the years as a starting point for investigating relationships and correlations.











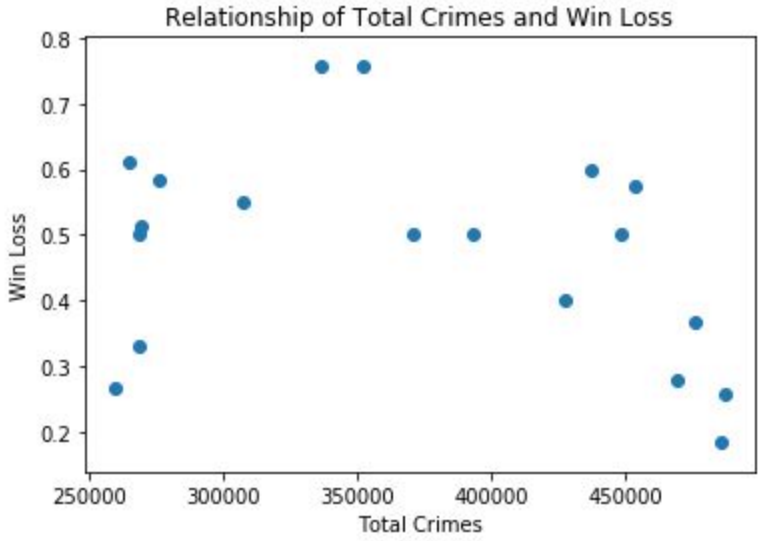
The crime graphs present the trend of significant decreasing crime and arrest rates in the period of our analysis and the smaller relative decrease in the proportion of crimes that result in arrest. The Bulls graphs show the high variance of the team's success in the period, giving a broad range of results that allows for analysis of highly successful years, and not as successful years.

Now we use the correlation function in pandas to calculate the pairwise correlations between the variables in our data frame, electing for the standard Pearson method.

	Total Crimes	Total Arrests	% of Crimes Resulting in Arrests	W/L%	SRS
Total Crimes	1.00	0.98	0.82	-0.18	-0.15
Total Arrests	0.98	1.00	0.90	-0.11	-0.08
% of Crimes Resulting in Arrests	0.82	0.90	1.00	0.17	0.17
W/L%	-0.18	-0.11	0.17	1.00	0.98
SRS	-0.15	-0.08	0.17	0.98	1.00

These numbers at face value represent the strength of association between the two variables. The most important data point relative to our analysis is the negative correlation between Win Loss percentage and total crimes and total arrests. It is a small correlation value, however it statistically states, that as one variable increases, the other decreases.

We also created a scatter plot of our data points to visualize the correlation and trends.



Overall, there is just insufficient data to come to any conclusions based on this visualization, but there may be a small pattern beginning to form.

While we sought out to see if team performance would affect crime rates, and while the statistical data does support our investigation, it would be bold to say that our claim has been proven true. There are many confounding factors to make a concrete claim, and we will investigate these further.

Confounding Factors

The biggest red flag during our data analysis was the consistent decreasing value of total crimes in Chicago. While it's not certain the reasoning for this, it is surely some extent of institutional change for such a significant decrease that requires additional research, and this does harshly skew our analysis. Also, on the Chicago Bulls data side, our hypothesis implies that social behavior is affected by the team's performance. However, there is a psychological element to this response. Individuals respond more

harshly unrelated to win/loss percentage, such as recent years performance compared to the current year, management decisions, and playoff performance.

It is important to note that in the years of approximately 2000-2001, the Chicago Bulls were undergoing a rebuilding process, causing them to have some all-time low scores in their history. Our data depicts a version of the Chicago Bulls different from the Chicago Bulls of the 1900s, one with constant change. This rebuilding process and lack of stability may also contribute as a confounding factor to the large fluctuations in scores.

The attack on 9/11 may also be an important factor to keep in mind. With an ominous atmosphere and looming presence of fear, basketball was likely the last thing on everyone's minds. This famous attack likely contributed to the low performance of the Chicago Bulls in the early 2000s. The aftermath of the attack including PTSD and paranoia may have affected individuals directly related to the Chicago Bulls. The attack has inspired countless other attacks such as the anthrax letters which may be a reason for the increase in crime in 2001-2003. With increased security and policing, this could also explain the increase in the arrest proportion in that time period.

While our initial plan was to investigate further into the different types of crimes and other indicators of team success, further analysis would also be affected by these confounding factors. The total crime change is very significant and it would be difficult to make any strong conclusions because of its implied impact. Instead there are different plans that would serve our original goals better which will be explained in the conclusion.

Conclusion

Our idea was to take a creative approach by looking into seemingly unrelated variables to see if we could find an interesting correlation. However, in taking on this task, it was difficult to navigate the numerous factors that could be skewing our data, due to the minimal quantitative relationship between the Chicago Bulls performance and crime rates in the city. The final result is a small amount of evidence supporting our original claim, and a lot more data analysis required to come to a conclusion. We've identified an action plan that could build upon our findings, and hopefully find more concrete evidence.

First, we would need to minimize the effects of the main confounding factor, which is the consistent decrease in total crimes in the period of our analysis. This could be done in two ways, the first being doing research into the institutional and societal reasons that

have caused such a significant drop. From there, we could implement this additional variable into our analysis, and use it as a reference point, and even possibly a predictor. Another method would be to investigate different variables that are relative to the number of total crimes instead of using the total number as one of our variables. Some ideas we have are to use a value pegged against the overall change that would represent crime rates, or even certain indicators of violence as proportions of the total amount of crimes.

Second, we would need substantially more historical years of crime data, which would help in multiple ways. If we were able to go beyond 2001, we could investigate the timeline of the decreasing total crimes. By identifying when the decrease started or other significant historical changes, we can investigate the systematic cause of the changes, and adjust our analysis accordingly. Also, with more total data, we have more evidence and our scatter plot visualizations will become more insightful, as patterns will be easier to see. It would be also interesting to analyze the years where Michael Jordan was on the Bulls, since that was one of the most successful and exciting periods in professional sports history, and something the city was historically very proud of. Another approach would be to incorporate different teams and cities as additional data as well. Ultimately, the options to expand this analysis are extensive, but for a successful result, it’s important to take a qualitative perspective.