**NBA Attendance**

**Exploratory Analysis**

Elijah Sartin, [esartin@bellarmine.edu](mailto:esartin@bellarmine.edu)

1. **INTRODUCTION**

This data set contains statistics determined prior to NBA games taking place. It was collected from <https://www.basketball-reference.com/> using Python and the BeautifulSoup package. This data was selected due to a desire to interact with sports in some capacity throughout this project as well as the NBA being the only major sport in season. Focusing on attendance allows the eventual statistical modeling to be meaningful as opposed to the guesswork of predicting performance related factors.

1. **DATA SET DESCRIPTION**

This data contains 5457 samples with 10 columns with various data types. A complete listing in shown in **Table 1**. Due to the nature of data collection there is no missing data. There were a few instances of the attendance value not being present due to COVID restrictions, these were eliminated from the data in MySQL prior to analyzing it.

**Table 1: Data Types and Missing Data**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Missing Data (%)* |
| Game Time | Interval / timedelta64[ns] | 0% |
| Game Date | Interval / datetime64[ns] | 0% |
| Arena | Nominal / object | 0% |
| Home Team | Nominal / object | 0% |
| Home Wins | Ratio / int64 | 0% |
| Home Losses | Ratio / int64 | 0% |
| Away Team | Nominal / object | 0% |
| Away Wins | Ratio / int64 | 0% |
| Away Losses | Ratio / int64 | 0% |
| Attendance | Ratio / int64 | 0% |

1. **Data Set Summary Statistics**

This section provides a brief statistical summary of the variables in the data set.

**Table 2: Summary Statistics for XXX (name of dataset)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Variable Name* | *Count* | *Mean* | *Standard Deviation* | *Min* | *25th* | *50th* | *75th* | *Max* |
| Game Time | 5457 | 19:47:51 | 1:43:52 | 12:00:00 | 19:00:00 | 20:00:00 | 21:00:00 | 22:30:00 |
| Home Wins | 5457 | 18.735 | 13.009 | 0 | 8 | 17 | 27 | 64 |
| Home Losses | 5457 | 18.744 | 13.197 | 0 | 8 | 17 | 27 | 65 |
| Away Wins | 5457 | 18.765 | 13.131 | 0 | 8 | 17 | 27 | 65 |
| Away Losses | 5457 | 18.723 | 12.984 | 0 | 8 | 17 | 27 | 64 |
| Attendance | 5457 | 17738.213 | 2082.070 | 6960 | 16449 | 18071 | 19432 | 22983 |

Proportions for Home Team (n=5457)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Atlanta Hawks | *181* | *3.317* |
| Boston Celtics | *181* | *3.317* |
| Brooklyn Nets | *180* | *3.299* |
| Charlotte Hornets | *177* | *3.244* |
| Chicago Bulls | *181* | *3.317* |
| Cleveland Cavaliers | *185* | *3.390* |
| Dallas Mavericks | *184* | *3.372* |
| Denver Nuggets | *188* | *3.445* |
| Detroit Pistons | *183* | *3.353* |
| Golden State Warriors | *187* | *3.427* |
| Houston Rockets | *182* | *3.335* |
| Indiana Pacers | *183* | *3.353* |
| Los Angeles Clippers | *183* | *3.353* |
| Los Angeles Lakers | *179* | *3.280* |
| Memphis Grizzlies | *180* | *3.299* |
| Miami Heat | *181* | *3.316* |
| Milwaukee Bucks | *184* | *3.372* |
| Minnesota Timberwolves | *184* | *3.372* |
| New Orleans Pelicans | *181* | *3.317* |
| New York Knicks | *182* | *3.335* |
| Oklahoma City Thunder | *183* | *3.353* |
| Orlando Magic | *182* | *3.335* |
| Philadelphia 76ers | *184* | *3.372* |
| Phoenix Suns | *184* | *3.372* |
| Portland Trail Blazers | *183* | *3.353* |
| Sacramento Kings | *180* | *3.299* |
| San Antonio Spurs | *183* | *3.353* |
| Toronto Raptors | *175* | *3.207* |
| Utah Jazz | *181* | *3.317* |
| Washington Wizards | *176* | *3.225* |

Proportions for Away Team (n=5457)

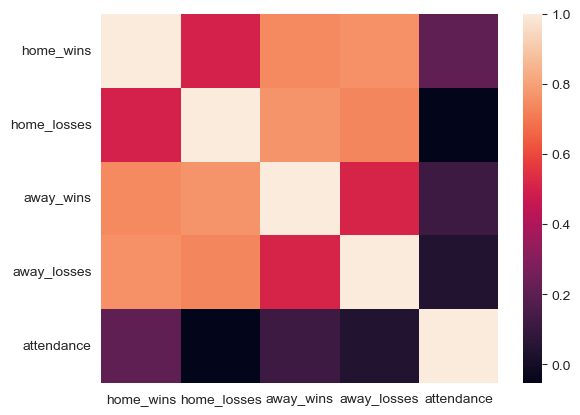
|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| Atlanta Hawks | *182* | *3.335* |
| Boston Celtics | *185* | *3.390* |
| Brooklyn Nets | *185* | *3.390* |
| Charlotte Hornets | *185* | *3.390* |
| Chicago Bulls | *179* | *3.280* |
| Cleveland Cavaliers | *178* | *3.262* |
| Dallas Mavericks | *181* | *3.317* |
| Denver Nuggets | *181* | *3.317* |
| Detroit Pistons | *184* | *3.372* |
| Golden State Warriors | *180* | *3.299* |
| Houston Rockets | *185* | *3.390* |
| Indiana Pacers | *184* | *3.372* |
| Los Angeles Clippers | *186* | *3.408* |
| Los Angeles Lakers | *182* | *3.335* |
| Memphis Grizzlies | *181* | *3.317* |
| Miami Heat | *181* | *3.317* |
| Milwaukee Bucks | *184* | *3.372* |
| Minnesota Timberwolves | *179* | *3.280* |
| New Orleans Pelicans | *179* | *3.280* |
| New York Knicks | *180* | *3.299* |
| Oklahoma City Thunder | *180* | *3.299* |
| Orlando Magic | *183* | *3.353* |
| Philadelphia 76ers | *183* | *3.353* |
| Phoenix Suns | *177* | *3.244* |
| Portland Trail Blazers | *184* | *3.372* |
| Sacramento Kings | *178* | *3.262* |
| San Antonio Spurs | *182* | *3.335* |
| Toronto Raptors | *182* | *3.335* |
| Utah Jazz | *184* | *3.372* |
| Washington Wizards | *183* | *3.353* |

Proportions for Arena (n=5457)

|  |  |  |
| --- | --- | --- |
| *Category* | *Frequency* | *Proportion (%)* |
| AT&T Center | *182* | *3.335* |
| AccorHotels Arena | *2* | *0.037* |
| Air Canada Centre | *41* | *0.751* |
| American Airlines Center | *184* | *3.372* |
| AmericanAirlines Arena | *114* | *2.089* |
| Amway Center | *180* | *3.299* |
| BMO Harris Bradley Center | *40* | *0.733* |
| Ball Arena | *68* | *1.246* |
| Bankers Life Fieldhouse | *116* | *2.126* |
| Barclays Center | *178* | *3.262* |
| Capital One Arena | *175* | *3.207* |
| Chase Center | *100* | *1.833* |
| Chesapeake Energy Arena | *117* | *2.144* |
| Crypto.com Arena | *132* | *2.419* |
| FTX Arena | *67* | *1.228* |
| FedEx Forum | *180* | *3.299* |
| Fiserv Forum | *143* | *2.620* |
| Footprint Center | *27* | *0.495* |
| Gainbridge Fieldhouse | *67* | *1.228* |
| Golden 1 Center | *180* | *3.299* |
| Little Caesars Arena | *181* | *3.317* |
| Madison Square Garden (IV) | *182* | *3.335* |
| Mexico City Arena | *7* | *0.129* |
| Moda Center | *183* | *3.353* |
| Oracle Arena | *87* | *1.594* |
| Paycom Center | *66* | *1.209* |
| Pepsi Center | *120* | *2.199* |
| Philips Arena | *41* | *0.751* |
| Phoenix Suns Arena | *40* | *0.733* |
| Quicken Loans Arena | *82* | *1.503* |
| Rocket Mortgage Fieldhouse | *103* | *1.887* |
| STAPLES Center | *230* | *4.215* |
| Scotiabank Arena | *134* | *2.456* |
| Smoothie King Center | *181* | *3.317* |
| Spectrum Center | *176* | *3.225* |
| State Farm Arena | *140* | *2.566* |
| TD Garden | *181* | *3.317* |
| Talking Stick Resort Arena | *116* | *2.126* |
| Target Center | *184* | *3.372* |
| The O2 Arena | *2* | *0.037* |
| Toyota Center | *182* | *3.335* |
| UW-Milwaukee Panther Arena | *1* | *0.018* |
| United Center | *181* | *3.317* |
| Vivint Arena | *25* | *0.458* |
| Vivint Smart Home Arena | *156* | *2.859* |
| Wells Fargo Center | *183* | *3.353* |

Table 4: Correlation Table/Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Home Wins | Home Losses | Away Wins | Away Losses | Attendance |
| Home Wins | 1 | 0.499762 | 0.741714 | 0.759406 | 0.208892 |
| Home Losses | 0.499762 | 1 | 0.765475 | 0.735253 | -0.053710 |
| Away Wins | 0.741714 | 0.765475 | 1 | 0.507283 | 0.114854 |
| Away Losses | 0.759406 | 0.735253 | 0.507283 | 1 | 0.041689 |
| Attendance | 0.208892 | -0.053710 | 0.114854 | 0.041689 | 1 |



1. **DATA SET GRAPHICAL EXPLORATION**

This data set has proven difficult to create meaningful visualizations with. Due to the nature of the statistics collected all the significant figures revolve around attendance. Figures 1-4 show the distribution of the home and away teams’ number of wins and losses going into each game. All 4 of the graphs show similar distributions that are bimodal with right skew. This shows that most games occur with records below 30 wins or losses, which makes sense due to the NBA season only having 82 games. Figure 5 is rather important as it shows a histogram of attendance which has left skew and is bimodal, with one mode around 18,000 and the other slightly under 20,000. There are a few values south of 10,000 that appear to be outliers of sorts. Figures 5-9 suffer from the same issue as Figures 1-4 in that they show a culmination of a team’s number of wins/losses going into home and road games. It gives a general idea of how successful each team is but is not as accurate of a depiction as result of game statistics would be. However, end of game statistics cannot have any effect on attendance so are not included. The numbers are much higher than total NBA games played due to the calculation being a sum of records going into each game. It is still possible to see that teams such as the Boston Celtics and Milwaukee Bucks have been rather successful since 2017 and the Atlanta Hawks have been terrible. Figure 10 shows the average attendance for each team when playing at home, teams such as the Chicago Bulls and Philadelphia 76ers show very high attendance numbers. There is a fair amount of disparity for average attendance amongst home teams. Figure 11 shows the average attendance based on the away team and has much less disparity than the home graph. The Boston Celtics remain high but are joined by the Lakers at the top. The difference between these two graphs seems to emphasize the impact of the home team on attendance being much higher. Figure 12 shows average attendance based on arena and the variance between each can most likely be attributed to the max capacity of each venue, of course in tandem with the teams playing. Figures 13-16 show the average attendance across 4 different seasons. There are not any major trends across the graphs except for slight increases in attendance at the beginning and end of each season. Figure 17 shows average attendance based on day of the week. Saturday appears to have the largest positive impact with Monday and Wednesday negatively impacting attendance. The other 4 days of the week have around the same average attendance.

*Chart, histogram

Description automatically generated***Figure 1: Histogram of Home Team Wins on Record (single plot)**

**Chart, histogram

Description automatically generated**

**Figure 2: Histogram of Home Team Losses on Record (single plot)**

**Chart, line chart, histogram

Description automatically generated**

**Figure 3: Histogram of Away Team Wins on Record (single plot)**

**Chart, histogram

Description automatically generated**

**Figure 4: Histogram of Away Team Losses on Record (single plot)**

**Chart, histogram

Description automatically generated**

**Figure 5: Histogram of Attendance (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 6: Bar Plot of Home Wins by Team (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 7: Bar Plot of Home Losses by Team (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 8: Bar Plot of Away Wins by Team (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 9: Bar Plot of Away Losses by Team (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 10: Bar Plot of Average Attendance by Home Team (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 11: Bar Plot of Average Attendance by Away Team (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 12: Bar Plot of Average Attendance by Arena (single plot)**

**Chart

Description automatically generated**

**Figure 13: Line Plot of Average Attendance During 2017/2018 Season (single plot)**

**Chart, line chart

Description automatically generated**

**Figure 14: Line Plot of Average Attendance During 2018/2019 Season (single plot)**

**Chart

Description automatically generated**

**Figure 15: Line Plot of Average Attendance During 2019/2020 Season (single plot)**

**Graphical user interface, chart

Description automatically generated**

**Figure 16: Line Plot of Average Attendance During 2021/2022 Season (single plot)**

**Chart, bar chart

Description automatically generated**

**Figure 17: Bar Plot of Average Attendance by Day of Week (single plot)**

1. **SUMMARY OF FINDINGS**

Through the exploration of the data a few concerns were brought to the surface. The heatmap shows little correlation between any of the other variables and attendance. Visualizing the records in a meaningful manner proved difficult so only time will tell if it has any sort of impact on attendance. Game time was not able to be graphed and is the only statistic that will have to simply be dropped as accounting for time zones will only further complicate the variable despite its impact on attendance most likely being miniscule. It is hard to gauge what kind of effect record will have on attendance based on the distributions all being rather similar, it will be a run and see situation. Attendance has its ebbs and flows throughout each season, dates such as holidays may account for the highs and lows present throughout different months. The impact of home team on attendance cannot be understated and should play a huge role in the prediction of attendance along with the max capacity of the venue being played at. Day of the week also shows similar promises with each day, despite being close in value, having distinct levels of attendance compared to the other days. Games occurring on Saturday versus those on Monday should show vastly different results and have significance in the model.

Screenshot of Dashboard:

Chart

Description automatically generated with low confidence