NBA MARKEt size Analysis

# *By Michael Kim*

# introduction

My project involves analyzing American city media market sizes with NBA records and the amount of all stars. I was originally interested in seeing if there’s a correlation between both media market sizes and the number of all stars in these cities.

As a basketball fan, I was interested in proposing this topic because I was curious to see how much of a factor a team’s media market plays in the drawing talented All-Star players. When talking to supporters of smaller market teams, they always complain their window of opportunity closes once their star players have an opportunity to join a bigger market team. While a player will always get the biggest contract staying with his original team, the earning potential entering a bigger market has the potential to increase dramatically with larger fan base and opportunity for off-the-court earnings (real estate, venture capital, endorsements, etc).

Furthermore, this assignment was a great way for me to combine my outside of school interests with an academic topic. In the nature of professional sports, it’s clear that talent combined with solid team management and coaching will propel a team to better win-loss records and winning championships. However, I wanted to explore a factor that could possibly influence a reason why teams are able to get so much talent.

# data sources

I downloaded two datasets for my project: NBA Market Sizes and Ticket Sales and NBA All Stars past 16 years.

**NBA All Star Data**

**Description:** I used the Data.world website which a csv containing every single all star, their team, height, weight, college, selection type, NBA Draft status, and nationality from 2000-2016. Since I was only interested in the amount of all stars each team had over the past 17 years, I extracted team and count which I will talk about in my Data Manipulation section.

**Size:** 440 records/61kb

**Location:** https://data.world/gmoney/nba-all-stars-2000-2016

**Format:** CSV file  
**Access Method:** Data.world has complied the data of all the NBA All Stars from 2000-2106 and I downloaded the data to a csv file.

**NBA Market Size and Ticket Data**

**Description:** An article from SB Nation detailed a dataset containing every NBA team’s win percentage, home win percentage, market size rank, and median remaining home ticket price.

**Size:** 30 rows/2 KB

**Location:** <https://www.slcdunk.com/research-statistics-analytics/2016/3/21/11278814/nba-tickets-los-angeles-lakers-golden-state-warriors-san-antonio-spurs-utah-jazz-market-size>

**Format:** CSV File   
**Access Method:** SB Nation had a table containing the data in a similar format to an excel spreadsheet so I copied and pasted the data into a file without any issues.

# Data Manipulation

**Step 1: Read the NBA All Star roster data**

Using the allstars function, I create a dictionary with the key being the NBA team and the value being the number of all stars each team has produced using the NBA All Star roster data. The allstars function takes a file as its mandatory input (NBA All Star Data.csv and returns a dictionary that maps the NBA team to its amount of all stars over the past 17 years.

**Step 2: Data Cleaning**

In addition, the allstars function also accounts for necessary cleaning of the dataset. As a lot has changed during the NBA landscape, I had to delete the New Jersey Nets, Charlotte Bobcats, and the New Orleans Hornets as keys since they are no longer teams in the NBA. In addition, there was also an empty string that was a key so I had to delete that as well. However, I added the total All Stars the Nets, Bobcats, and Hornets received to the Brooklyn Nets, Charlotte Hornets, and the New Orleans Pelicans respectively since I figured media markets would be in support for the teams that recently moved to their city. Since Brooklyn is so close to New Jersey, I figured the Nets would’ve retained their same media market.

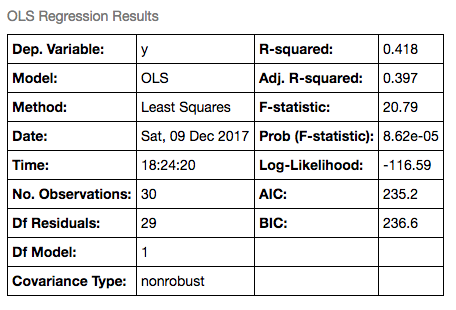
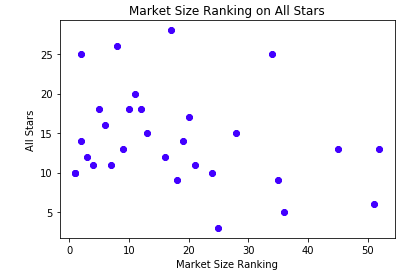
**Step 3: Read the NBA Market Sizes Data**

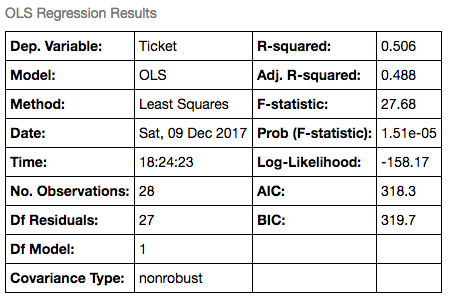
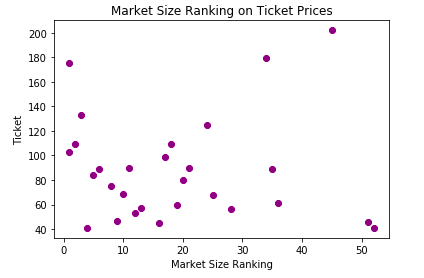
Using the the mastercsv function I created, I used the dictionary created in the previous step as the input and write the masterfile.csv output which is an extension of the NBAData.csv (Nba Market Size and Ticket Data). In this function, I’m adding the All Star Column by making All Star row the values from the allstars dictionary made in the previous step. Using the CSV module for reading and writing CSVs, the output file contained all the data I was looking to analyze for NBA Markets.

**Step 4: Scatterplots with Matplolib and Pandas**

The last few blocks of code were written to visualize my findings using matplotlib, pandas, and the statsmodels api for correlation results. The next section of my write-up will show the graphs that I was able to make along with the regression outputs.

# Visualizations





# Visualization Analysis

For my first graph, I explored the correlation for my initial question: Does the market size of city affect the amount of all stars that want to play for the team? By looking the graph, there seems to be a slightly inverse relationship: as the market size ranking is higher (smaller market) the number of all stars is lower. However, by looking at the correlation coefficient (R2), there doesn’t seem to be much of a relationship between market size and all stars. More or less, this is what I expected because middle market teams like the San Antonio Spurs and the Cleveland Cavaliers have been able to generate all star selections through their team play and ability to attract talent due to continued success.

For my second graph, I analyzed a correlation between market size ranking and median ticket prices. Although this graph was constructed in from the source I got the data from (SB Nation), I wanted to remove the Lakers and Warriors who were notable outliers with median ticket prices being $630 and $362, respectively (higher than any other team by a mile!). The Lakers are arguably the most popular team in NBA history and the Warriors are the arguably the most popular team in the NBA right now so their demand for tickets is high, explaining the outrageously high prices. When running the regression, the R2 with the Lakers and Warriors included was 11% and by removing the outliers, the R2 ended up being 50.6%, a much higher value but clearly, media market size still doesn’t explain all the variability of ticket prices.

While working on this project, I was able to come up with many takeaways. First, if I had more time, I definitely would’ve incorporated more variables for a multiple linear regression in order to build a more representative model. Based on my results, it’s pretty clear that the dependent variables I made have more than one factor that influences their variability. Secondly, I found myself extremely engaged in this project and discovered how powerful and diverse of a tool Python truly is. Although I found myself initially struggling to figure out which methods from the class we had learned to incorporate in the project, looking at previous homework assignments to see what manipulations would be appropriate helped me tremendously in the process. I’m excited to engage in future analysis projects that combine more of my interests while discovering Python’s full potential.

**Resources**:

In addition to the datasets, the link below was extremely helpful in aiding me with the code I had to write in order to produce visualizations and regression outputs.

<https://www.datarobot.com/blog/multiple-regression-using-statsmodels/>