Predicting MLB Market Attendance/Valuation with US Income and Housing Data

Team 116

Joshua Mayanja, Yutai Liu, Warren Spann, Anthony Palmeri, Jonathan Martin

Master of Science in Analytics, Georgia Institute of Technology

MGT 6203: Data Analytics for Business

Professor Bien

July 6, 2023

Primary Research Question

Can MLB attendance be predicted based on household valuation and income data?

Problem Overview

Starting an MLB franchise is a very expensive process. The MLB commissioner mentioned that the fees for the next expansion franchise "will be in the $2.2 billion range. With the stakes being very high for both the MLB's brand and potential owners, basing this decision on data would be an obvious choice. The purpose of our analysis is to figure out if housing data and income correlate with MLB attendance. We would then return to the MLB with the factors that highly correlate with increased attendance. Additionally, as the MLB looks to expand into new cities, we can provide recommendations on the most profitable expansion cities.

Data Cleaning/Transformation

The data for this project was collected from three unique sources. The first of these is a csv file found on Kaggle (1) with data from the US Census between the years of 2011 and 2021. The dataset contains data points for each zip code across that ten year span on 110 features based around household income and home value data. Before this data set could be used, its data points had to be converted from zip codes into MLB markets.

To transform this data, zip codes for each market were gathered from Zillow (2) (with a market defined as the metropolitan area for the city in which the MLB franchise is located). Once a market had its zip codes defined, the data from each zip code in the market was summed by year, leaving 10 data points for each market (one for each year). Repeating this process for multiple markets yielded two new, cleaned datasets: one for the twenty five current MLB

markets, and one for twenty five potential new markets (Albuquerque, Austin, Birmingham, Buffalo, Charlotte, Columbus, Hartford, Honolulu, Indianapolis, Jacksonville, Las Vegas, Louisville, Memphis, Nashville, New Orleans, Oklahoma City, Omaha, Orlando, Portland, Raleigh-Durham, Sacramento, Salt Lake City, San Antonio, San Juan, Virginia Beach). Note that twenty five current markets are being used as opposed to thirty, as there was no Census data for Toronto as it is outside of the United States, and because four markets have two teams, so the data for both teams in each of those markets was combined.

Once these two new data sets were created, two new features were added to each: one to represent if a market had a domed stadium, and one to represent the number of teams a market contained. These were added to serve as dummy variables to see if they had any significance in attendance or valuation prediction at all. Note that all new markets have data points for both domed and open stadiums so both scenarios can be tested, therefore each new market has twenty data points.

The final two datasets being used for this project are of yearly attendance numbers, and of franchise valuation for each MLB team between 2011-2021. Both of these datasets had to be scraped from the internet, and that process was very similar for both. Attendance data was collected from baseball-reference.com (3), and the valuation data from Wikipedia (4) and Forbes (5). For each market, yearly attendance and valuation numbers were pulled and saved in both json and yml format. Once all of this data was scraped, two new features were added to the cleaned current markets data set, one for attendance and one for valuation, as attendance will be a dependent variable, and valuation potentially could be as well.

At the moment, all data has been successfully transformed and cleaned as described above. The final step that needs to be taken before model building can begin is dividing the

current market data into a training and test set, and figuring out how much data should be allocated for each.

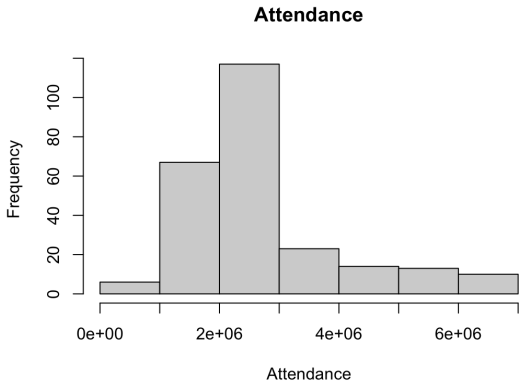
Initial Exploratory Data Analysis

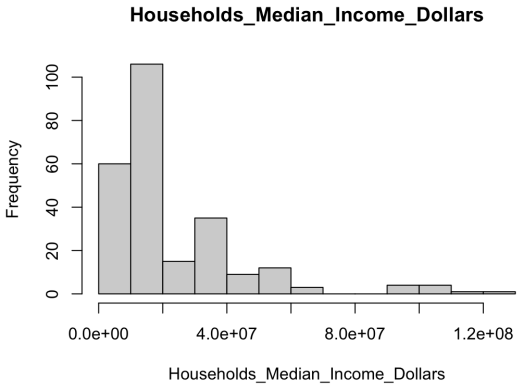
After compiling the data into two dataframes, we've begun exploring the variables we have and their relationships with attendance. For our initial EDA, we'll be analyzing current markets in the MLB. The first thing to consider was the 2020 MLB season, which was canceled due to COVID. Because attendance across that year is 0, 2020 has been removed from our EDA. Next for the analysis, we've selected a couple of the attributes to analyze. In the future, we will need to conduct variable selection to find the best variables to predict on, however for the time being, we have selected the following variables:

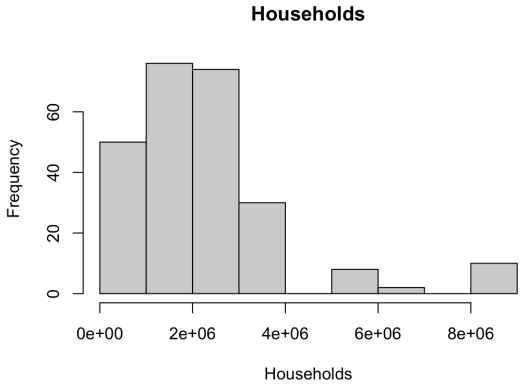
- Attendance
- Number of Families
- Number of Married Couple Families
- Number of Nonfamily Households
- Median Household Income

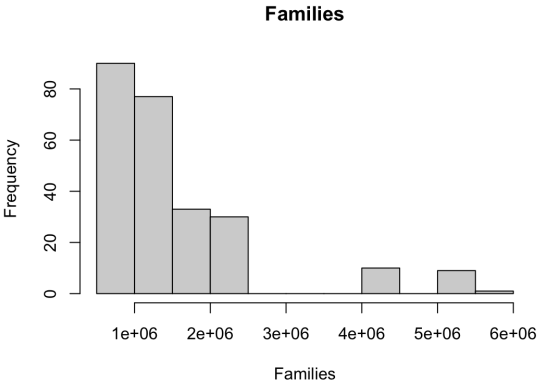
Data Distribution

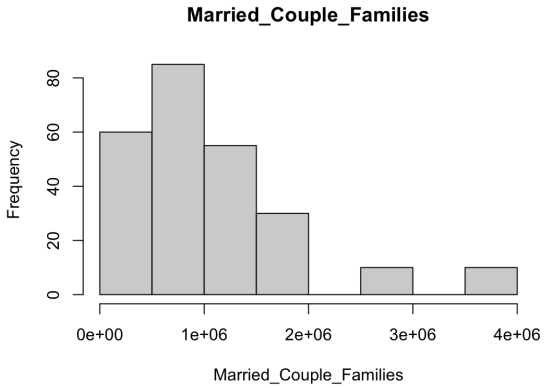
As seen below, all of our attributes have a right skew and depending on the model we use, we'll likely need to normalize the data. The spread of the data also appears to be wide, likely due to larger market teams when compared to smaller cities and markets. Several transformations will be considered to shrink the spread and attempt to normalize, including boxcox and performing log transformation. For this analysis, we've used a log transformation.

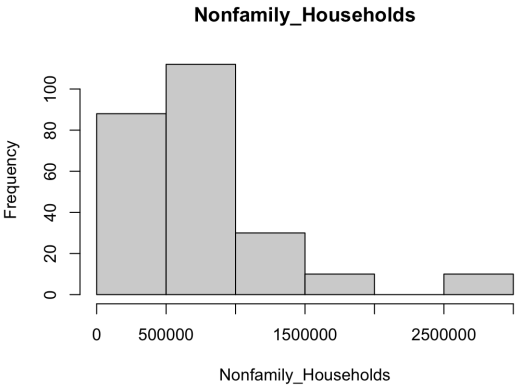

Attendance
Frequency
0
20
40
60
80
100
0e+00
2e+06
4e+06
6e+06
Attendance


Households_Median_Income_Dollars
Frequency
0
20
40
60
80
100
0.0e+00
4.0e+07
8.0e+07
1.2e+08
Households_Median_Income_Dollars

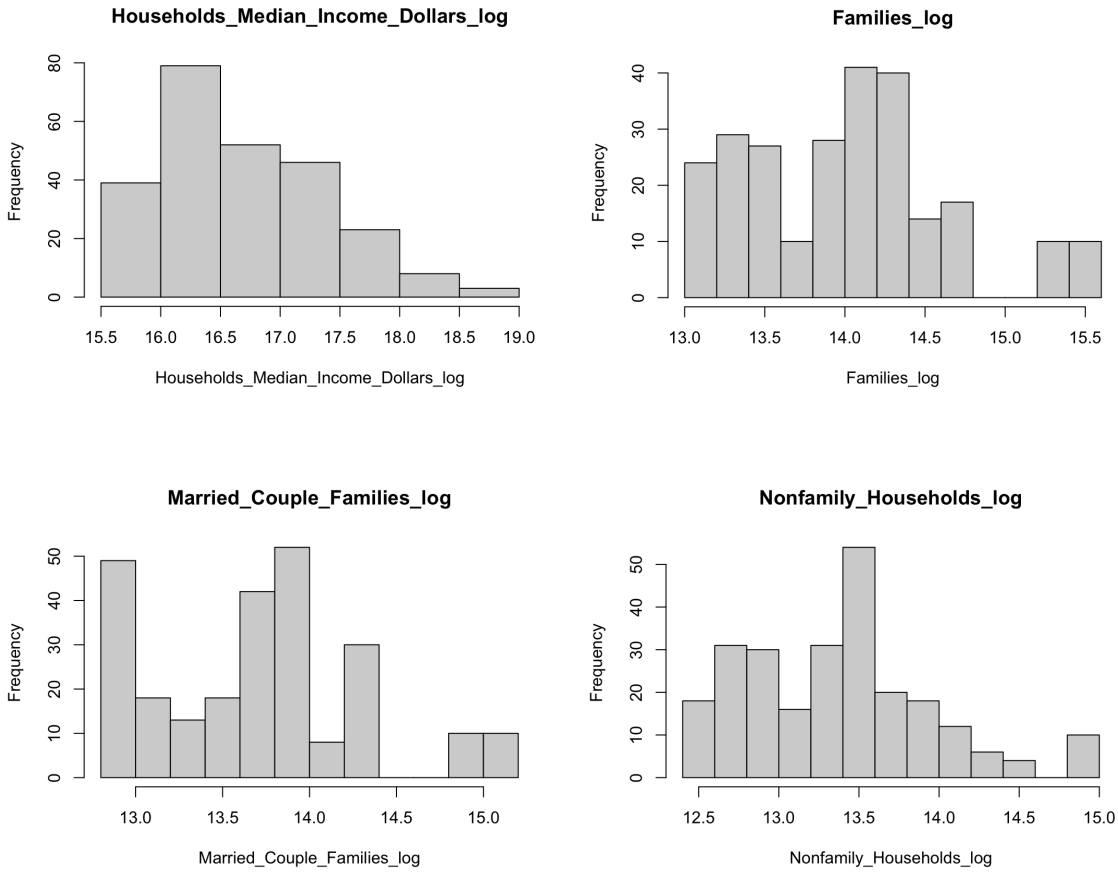

Households
Frequency
0
20
40
60
0e+00
2e+06
4e+06
6e+06
8e+06
Households


Families
Frequency
0
20
40
60
80
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
Families


Married_Couple_Families
Frequency
0
20
40
60
80
0e+00
1e+06
2e+06
3e+06
4e+06
Married_Couple_Families


Nonfamily_Households
Frequency
0
20
40
60
80
100
0
500000
1500000
2500000
Nonfamily_Households

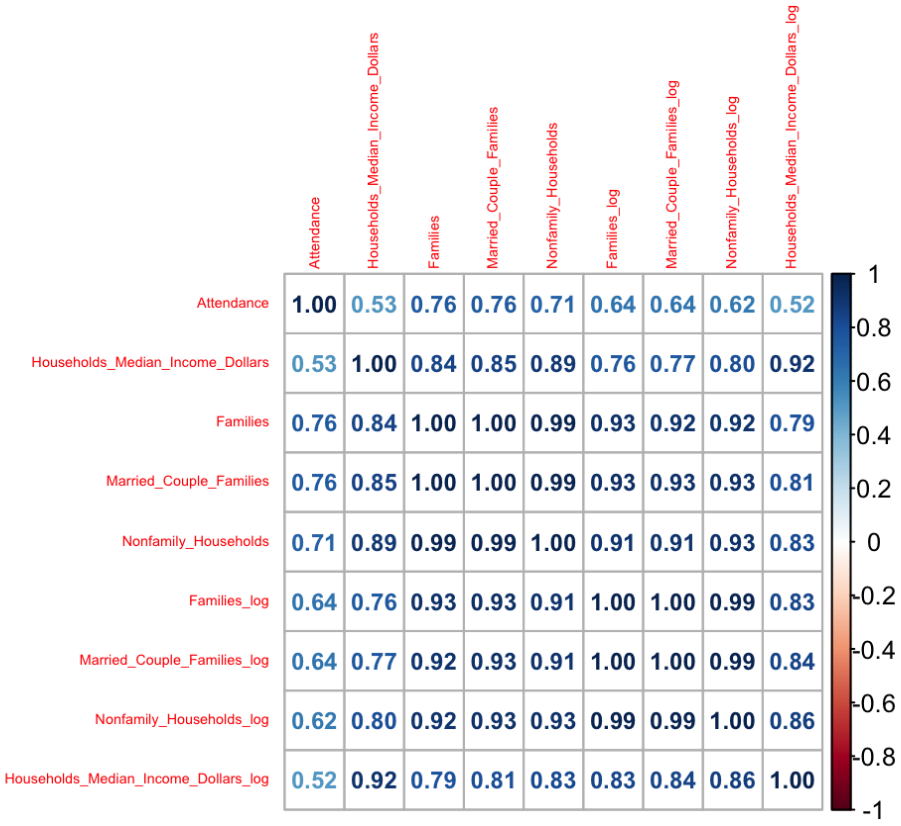
Log Transformations

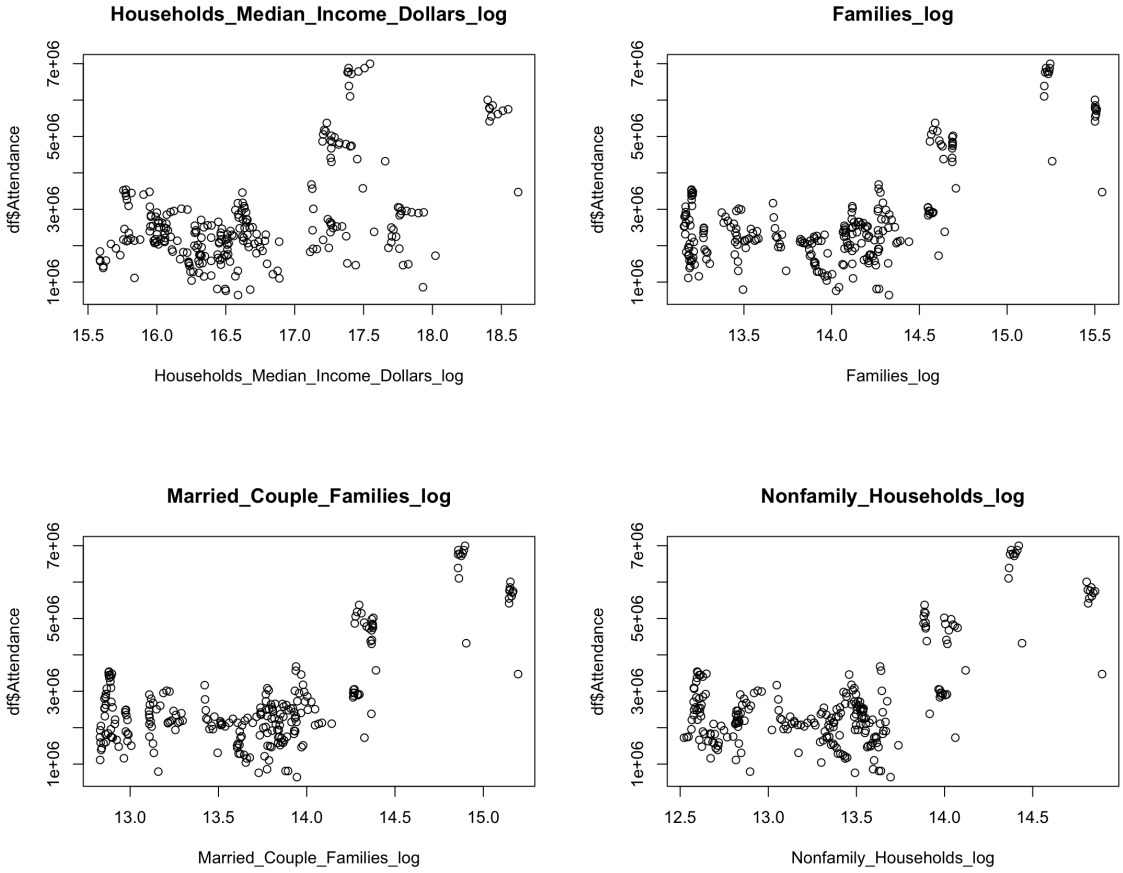


After performing a log transformation, many of the variables more closely resemble a normal distribution. Additional transformations will be explored in future analysis to find what will work best for prediction. Depending on the models we also select, standardization of the data might be necessary.

Attribute Relationships

Upon further analysis of our data, the selected attributes all have a mild positive correlation with attendance. Our initial expectation was as household income increased, we'd expect projected attendance to increase as well. However, it appears that attendance actually shares a stronger relationship with the number of households in an area. In addition to this, the independent variables selected for analysis appear to be correlated with each other. This would introduce multicollinearity into our model so this is an issue to consider. Potentially, looking at the total number of households or the best predictor between families, married couple families, and nonfamily households could offer better results. The scatter plots also appear to have several outliers with clusters suggesting that larger cities often have higher attendance when compared to smaller cities. This conclusion makes sense when considering large market teams vs. small market teams and a team's ability to pay players (and ultimately perform better).





Methods

With the data cleaned and transformed, and some initial EDA performed, the next step for this project will be to generate a model to predict attendance. The initial hypothesis is that one of four markets (Nashville, Charlotte, Salt Lake City, or Las Vegas) will be predicted to have the highest attendance, as these are the four cities which have been explored for expansion most by the MLB at the current moment. However, twenty one other markets will also be tested in this project, so it would not be surprising if any of these four cities were not predicted as the top new market.

The model that will be used to predict attendance will be either a linear regression model,

XGBoost model, or a neural network. All three model types will be built and tested, and the strongest model is the one that will be used to predict attendance. Some variable selection will have to occur before this model is created, as it is not practical to use all 110 features as variables. Only the most significant variables need to be kept.

In addition to the attendance prediction model, a clustering model will also be created to group markets together. This will allow for the creation of a visual that will show how attendance in new markets is predicted to behave in comparison to existing markets. The clustering model will either be a K-Means model, K-Nearest Neighbor model, or hierarchical based model. Once again, all three model types will be created, and the strongest one will be used.

Literature Survey

The theory and assumptions for this project derived from two literatures. Firstly, in Lee's study (6), the sequential test method was used to identify the common factors in MLB attendance with data from 1904 to 2012. Both team-specific factors (performance) and economic factors (per capita GDP) were tested using a sequential test method and was found that per capita GDP was among the significant factors in different models. And it was concluded that the combination of time trend and per capita GDP is likely to be a common factor influencing attendance at MLB. Thus this study provides evidence and base aligning with the assumption of this project in using housing and income data for predicting MLB attendance and expansion.

However, in Hong, Modelleo and Coates' study(7), it was suggested that income was often found not to be a significant determinant of attendance, possibly because of the lack of variation over the league's season or because it is not a good measure of the purchasing power and economic circumstances of the fans. Instead, the coincident indexes have an advantage over

income per capita because they vary by month within the season and because they combine information from several indicators so they can better reflect the current economic conditions. Using 4,696 games during the 2008 and 2009 seasons, the study investigated the effect of the economic crisis on attendance in MLB. The empirical evidence from the study indicates the recent economic crisis contributed to a decline in MLB attendance over the period 2008 through 2009; deteriorating economic circumstances explain a decline of about 6% compared to the reported decline of 6.77%. Nevertheless, this study also pointed out that there were limitations to this study such as the lack of weather data and playoff uncertainty.

Hence, in these two articles, income related variables were studied and different conclusions were drawn. Our project will use these as evidence and base to further examine how the housing and income data influence MLB attendance and use it for making MLB expansion decisions.

References

(1) Gendron, Clay. “Us Household Income by ZIP Code 2021-2011.” *Kaggle*, 14 June 2023, www.kaggle.com/datasets/claygendron/us-household-income-by-zip-code-2021-2011.

(2) “Browse Zillow by State/Province.” *Zillow*, www.zillow.com/browse/homes/. Accessed 3 July 2023.

(3) “List of All the Major League Baseball Teams.” *Baseball Reference*, www.baseball-reference.com/teams/. Accessed 2 July 2023.

(4) “Forbes List of the Most Valuable MLB Clubs.” *Wikipedia*, 2 July 2023, en.wikipedia.org/wiki/Forbes_list_of_the_most_valuable_MLB_clubs.

(5) Badenhausen, Kurt. “Full List: Baseball’s Most Valuable Teams.” *Forbes*, 23 Mar. 2011, www.forbes.com/2011/03/22/mets-yankees-phillies-dodgers-baseball-valuations_slide.html?sh=c503cef22937.

(6) Young H, Lee. (2018) Common Factors in Major League Baseball Game Attendance. Journal of Sports Economics.

(7) Sung Il Hong, Michael Mondello, and Dennis Coates. (2011) An Examination of the Effects of the Recent Economic Crisis on Major League Baseball (MLB) Attendance Demand. North American Association of Sports Economists.