Predicting MLB Market Attendance 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 20, 2023

Abstract

This analysis takes housing and income data from the US census and determines if it is possible to use it for predicting yearly attendance for an MLB market. The end goal was to provide recommendations for top expansion markets for the league as they begin to explore expansion into new markets A multiple linear regression model, an extreme gradient boosted model (xgboost), and a deep learning model were created to explore this relationship. Two models provided very similar predictions and the xgboost model was selected as the final model to use for predictions, as it had a low margin of error and accounted for a lot of variance. It was found that attendance could accurately be predicted from this data, with the top potential new markets being San Antonio, Oklahoma City, Charlotte, Portland (OR), and Indianapolis.

Introduction

Over the past few years, Major League Baseball has been exploring expansion, with the idea of adding a couple more teams into brand new markets. Starting a franchise in the MLB is a very expensive process with a total cost in the range of $2.2 billion. With so much money on the line, the stakes are very high for the MLB, potential owners, and the economy of the markets. A failing franchise would lose billions for the league and the ownership group, and the losses could devastate the market's economy, which could have long term and far reaching rippling effects. It is vital that any new franchises the MLB chooses to instate are placed in an environment where they will succeed, so basing the location of the new markets on data is an obvious choice. The US census provides housing and income data on every zip code in the country. Grouping these zip codes into markets, this analysis will explore this primary research question:

Is it possible for MLB attendance to be predicted based on household valuation and income data?

Attendance is a good estimator of franchise health as higher attendance leads to a higher demand for tickets, and a higher demand for tickets leads to more tickets sold at a higher price, which leads to higher revenue and a more profitable franchise. This analysis will determine if there is any correlation between attendance and this census data, and if there is, it will then return to the MLB the factors that are most significant in predicting attendance as well as attendance predictions for each market. It is important that these predictions are strong, as there is so much on the line for so many parties in this process.

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 (TX), Birmingham

(AL), Buffalo, Charlotte, Columbus (OH), Hartford, Honolulu, Indianapolis, Jacksonville, Las Vegas, Louisville, Memphis, Nashville, New Orleans, Oklahoma City, Omaha, Orlando, Portland (OR), Raleigh-Durham (NC), Sacramento, Salt Lake City, San Antonio, San Juan (PR), 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 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 yaml 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 is the dependent variable (note: valuation data was collected with the goal of creating a model to predict valuations for the new markets, but that proved impossible with the housing and income data alone, as the best model created was far from strong enough to use, so the attention was turned strictly to attendance).

Once the data was cleaned and formatted as described above, the current market data needed to be split into a training and test set. Since there were only 276 data points, the data was too small to include a validation set, and more data needed to be allocated to the training set to ensure that it had enough data to generate a strong model. An 80/20 split was determined sufficient to meet these demands, with 80% of the data being allocated to the training set, and 20% to the test set. Since each data point represents a specific point in time (a year) the data could not be assigned randomly, as the distribution of years across each could be skewed. A rotation was used to split the data instead. For each market, there were 10 years of data. For the first market, 2011 and 2016 were assigned to the test set, and the rest in the training set. The next market had 2012 and 2017 go to the test set. This pattern repeated for the remaining eight markets, with the year wrapping back to 2011 once one of the pointers passed 2021. The result was an even distribution of years across both sets, with each year being represented 20 times in the training set, and 5 times in the test set.

Exploratory Data Analysis

After compiling the two dataframes, one for current markets and one for new market cities, we began exploring the relationships between the variables. The first thing noted was the year 2020 contained no attendance due to COVID, so the housing data for this year was excluded in the modeling. Another consideration was certain markets contained two teams in their market (New York, San Francisco, and Chicago). Some of our models removed these observations for performance as well as the fact that the MLB is looking to add teams to non-existing markets.

The distribution of our attributes was explored next. All of the variables from the housing dataset showed a right skew, therefore, depending on the model used, the data will need to be normalized to match model assumptions. The spread of these values was also wide, likely due to

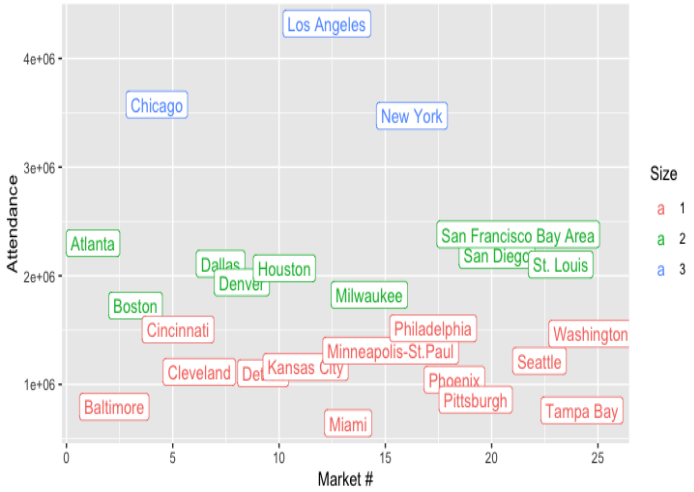
larger market teams being compared to smaller market teams. For this reason, several transformation techniques were attempted in the modeling process including Boxcox and logarithmic transformation to reduce the spread of the data.

Next, we analyzed the correlations between the attributes. Certain models including regression are highly impacted from multicollinearity and require there to be a linear relationship between the predictor variable and the independent variables. The correlation tests depicted an issue we faced during modeling, multicollinearity, with almost all of the attributes having a strong correlation with one another. Depending on the model selected, this issue will need to be addressed, potentially with dimension reducing techniques such as PCA, which can help eliminate correlations between variables.

Lastly, we tested for missing values in the data, as well as outliers and the cause of the outliers. Luckily, the housing dataset and MLB attendance datasets were cleaned thoroughly and did not contain any missing values. A Z-score test in combination with interquartile range was also performed to assess for outliers. It was discovered that there were approximately 20 outliers per column with the maximum returning a Z-score of 6.010637 found in the households with more than $200,000 or more income, nearly 6 standard deviations from the mean. While there appear to be a fair bit of outliers in the data, this can be likely attributed to larger cities compared to the average city and do not represent measurement errors or poor sampling.

Attendance Clustering

To better understand the attendance behavior between different markets, we used the 2021 attendance data of the 25 current markets to form cluster markets. Our project used K-means clustering to form 3 clusters (it is defined that 3 is the best number using the elbow method). The 3 cluster markets (Large, Medium and Small) have cluster means of 1114513, 2071340, 3789160 respectively. We expect the 25 potential new markets in the same clusters (Large, Medium and Small) would have similar attendance behavior. And based on the predicted attendance data from XGBoost model, the new markets are clustered into two groups (Small and Medium). We will compare the attendance behavior between current markets and new markets based on the cluster group it is in.



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. Three models were created to predict attendance: a linear regression model, an extreme gradient boosted model, and a deep learning model. Additionally, prediction results were clustered with the current market attendances to compare behaviors.

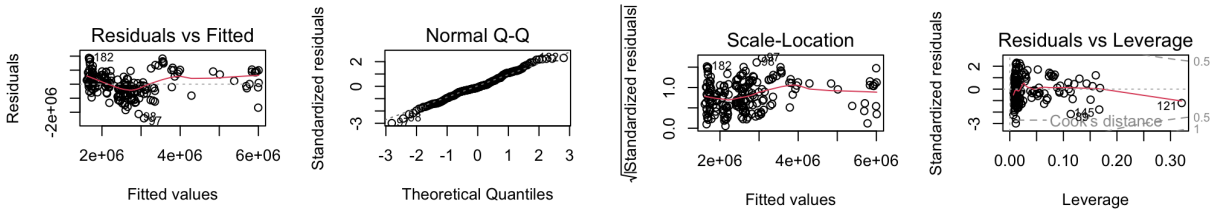
Multiple Linear Regression Model with Principal Component Analysis

A multiple linear regression model was implemented as the first model. Based on the initial EDA conducted, the model was tested and trained on the training and testing data, and final predictions were made using the new market data.

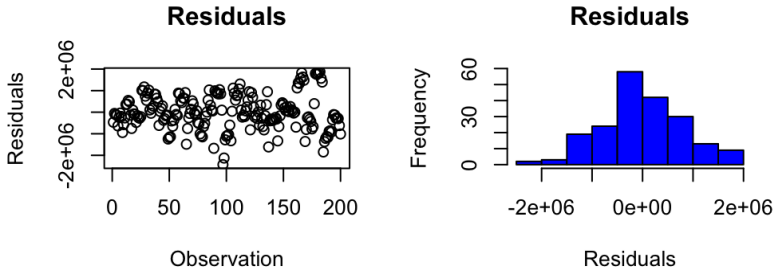
It was necessary to first identify which features would be significant for prediction. It was noted that following the assumptions made in linear regression would be a challenge when developing this model. High dimensionality and multicollinearity of the data were the main challenges encountered. Furthermore, the data showed skewness in nearly all attributes and certain extraneous columns, such as the error columns, which would need to be removed. In order to obtain the best results, a number of different models were attempted, including lasso and ridge regression, logarithmic transformations, Boxcox transformations, and finally principal component analysis. Lasso and ridge regression were initially tested, but they failed to select the most important variables because of the multicollinearity issue. Similar results were obtained when exploring Boxcox and logarithmic transformations. The Residuals vs. Fitted plot also revealed patterns in the residuals, and the Q-Q plot displayed curved tails, which would break the assumptions for regression.

The final method explored was principal component analysis, which involved selecting only variables that showed a moderate correlation with the predictor variable. After normalizing the data and creating 26 principal components, the first 6 components were found to account for a cumulative variance coverage of 0.99790. These components were selected for the model after testing their statistical significance. The regression analysis identified principal components PC1, PC2, PC3, PC4, and PC6 as statistically significant, resulting in the highest adjusted R^2 value (0.6192). Therefore, they were used in the final model.

The model also returned promising results in the form of plots when analyzing the residuals and fitted values. The residuals vs. fitted plot showed a reasonably linear fit with a slight peak and trough, making it the best-fit method out of the methods tested. The Q-Q plot also exhibited a relatively linear fit with slight tails at the end of the most extreme quartiles, outperforming other implemented methods.



After selecting the method for regression, testing was implemented on the testing data set, which comprised of 50 observations transformed using the earlier PCA. The Mean Absolute Error (MAE), measuring the absolute magnitude of the model's errors, was found to be 678,492.7 or 8,376.45 per home game. Additionally, the Root Mean Squared Error (RMSE), representing the deviation of the residuals from the actual values, was 831,613.2 or 10,266.83 per home game. Our R^2 on the test data returned 0.63, which covers an acceptable portion of the explained variance of attendance. The residuals also demonstrated little correlation with the observations and exhibited a normal distribution; both important assumptions to meet. Overall, the regression performed relatively mediocre in projecting attendance in the current MLB market. Missing 10,266.83 per home game could account for nearly $308,004.90 (roughly $30 a ticket per game) in missing or over-projected revenue without accounting for market size.

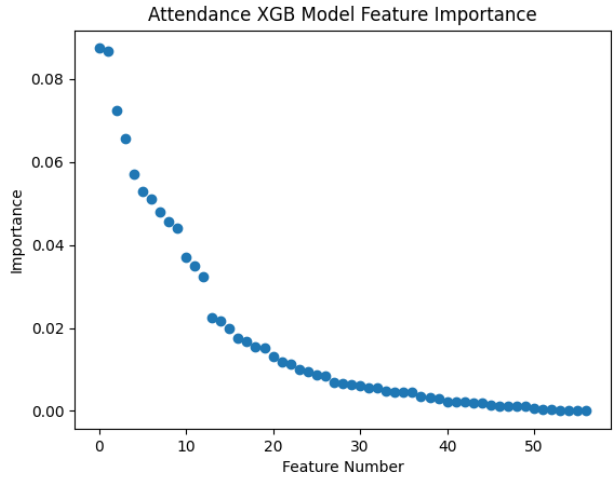


After assessing the performance of the model, the final projections were calculated using the new market data compiled in our data cleaning. The final predictions made on the new market data projected Portland, Sacramento, Orlando, Las Vegas, and Charlotte as the top 5 cities with San Juan projecting the least attendance for total attendance. Our initial hypothesis did include Charlotte and Las Vegas as the model predicted, however, the other 2 cities were not projected in the top 5. This could be attributed to California and Oregon having a larger population and average income leading to attendance in Portland and Sacramento being projected higher than expected (see figure 1 in appendix for full regression predictions).

Extreme Gradient Boosted Model

The second model tested was an Extreme Gradient Boosted (xgboost) model. This algorithm uses ensemble learning with many regression trees, with each tree being gradient boosted by the residuals of the tree built before it, minimizing the error at an “extreme” level (hence the name for the algorithm). Since the algorithm is based on regression decision trees, the data did not need to be scaled or normalized before model creation could begin, as decision trees are not very sensitive to the scale of the data’s features.

While the scale of the features in this dataset did not have to be adjusted, feature selection did have to occur, as the data included 110 features which could lead to overfitting if all features were used. To find the most significant features to use in the model, an xgboost model was created and fit to the entire dataset (both training and test sets). The xgboost python library was used for this, and the model created with this library tracks the most significant features it found. These features were plotted in decreasing order of importance on the figure below.



	Feature	Importance
1	Nonfamily Households Nonfamily Income in the P...	0.087442
2	Households $50,000 to $74,999	0.086789
3	Families $75,000 to $99,999	0.072436
4	Married-Couple Families $25,000 to $34,999	0.065592
5	Households Household Income in the Past 12 Months	0.057059
6	Married-Couple Families Less Than $10,000	0.052882
7	Households $35,000 to $49,999	0.051040
8	Families $100,000 to $149,999	0.047944
9	Families Family Income in the Past 12 Months	0.045706
10	Married-Couple Families $75,000 to $99,999	0.044016
11	Households $200,000 or More	0.037053
12	Nonfamily Households	0.034997
13	Married-Couple Families	0.032341
14	Dome	0.022506
15	Nonfamily Households $50,000 to $74,999	0.021793

The curve of the graph has a point where features start losing significance at a very low rate, so the elbow method can be used to analyze it. 0.02 seemed to be the point at which importance seems to stop decreasing at a high rate, so that was set as the elbow on the graph. There are 15 features with importance of 0.02 or higher, so these became the 15 most significant features, and all were used when creating the final mode. See them listed in the table above with their importance level[1].

Now that features were selected, a new xgboost model could be created and trained on the training set, using only these 15 features. The xgboost algorithm has many hyperparameters that can greatly impact model performance. Through trial and error, it was found that four of them significantly impacted this model's performance: gamma (a regularization parameter), learning rate (the step size for each iteration between trees), max depth (the maximum depth per tree), and num estimators (the number of regression trees used by the model). Grid search k-fold cross validation (with k=5) was used to find the best combination, with gamma being tested between 0-5 inclusive, learning rate being between 0.1-1 inclusive, max depth between 3-10 inclusive, and num estimators between 100-1000 inclusive. The results yielded the best parameters to be gamma = 0, learning rate = 0.1, max depth = 3, and num estimators = 400.

A final xgboost model was then created using the best parameters and most significant features. This model was trained on the training set, and then tested against the test set. The results showed that the model was indeed very solid, as it had high R-squared (0.8316) and Adjusted R-squared (0.7345) values, meaning that the model is accounting for most of the variability in the data. In addition, the model had a root mean squared error of 268,577.7372, meaning that on average a market's attendance prediction is off by this amount. However note that these predictions are for yearly attendance, so dividing the RMSE by 81 scales it down to a per game basis (as each team plays 81 home games). When scaled down, the RMSE is 3,315.7745, so on average a market's per game attendance is off by just over three thousand people. While this number could be lower, it is important to note that the average MLB stadium holds tens of thousands of people, and the average MLB attendance per game over this time span was 27,250. 3,316 people would fill only about 1-3 sections in stadiums this big, and would not be very noticeable if missing in their crowds.

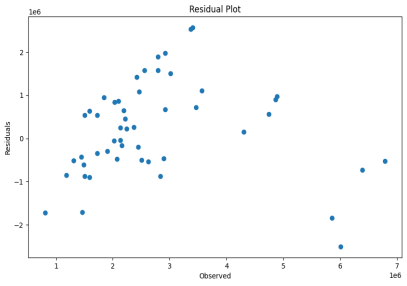
With a low RMSE and high R-squared values, this model seems to be capturing the most

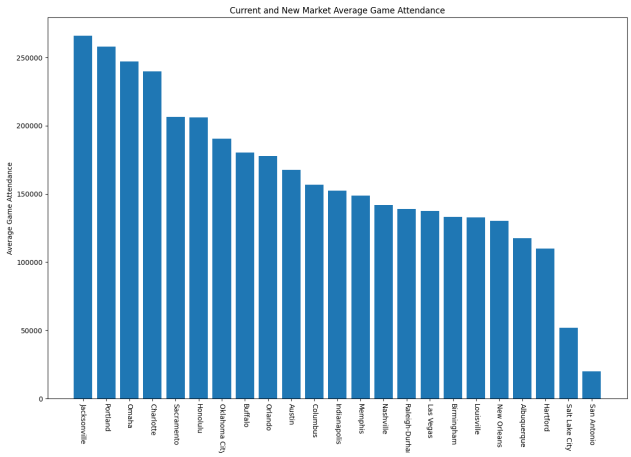
[1] Note that feature 1 in the table is: Nonfamily Households Nonfamily Income in the Past 12 Months

important aspects of MLB markets, and should give strong predictions for the new markets being tested.

Deep Learning Model

As an auxiliary model we trained and tested a deep learning model with the tensorflow package in python. Error and historical columns were removed as well as data in 2020. A regression model was created with two hidden layers. To compile the model, the Adam optimization algorithm was used due to its efficiency and robustness. For the loss function, mean squared error was used due to it being a regression problem. The model was then trained with a batch size of 32 and for 50 epochs. This returned a model with a Root Mean Squared Error of 1,185,855 (annual) and an R-squared of 0.2554. Below is the residual plot and there does seem to be some correlation. Additionally, the model did seem to overestimate attendance on a per game basis as seen in the attendance plot. Due to the low R-squared and high RMSE score, the other two models seem to be better fits.





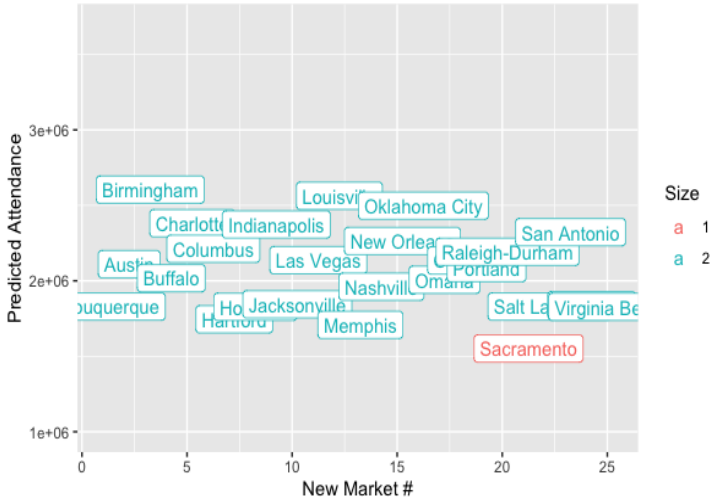
Results

Based on its high R-squared values, and its relatively low RMSE, the xgboost model will be used as the final model to predict attendance for the new markets. However, it is important to note that two models (the multiple linear regression model and the xgboost model) provided very similar predictions in terms of how well each potential new market performed against the rest of the league. This provides extra assurance that our predictions and recommendations will be accurate, as we have multiple models confirming similar results. The decision to use the xgboost model for final attendance predictions was made solely on the fact that it had the lowest margin of error, and accounted for more variability.

When predictions are made using this model, the results are returned on a yearly total scale. For easier interpretation, the results will be scaled down to an average per game basis. Again, note that the predictions were made on a yearly basis over the course of 10 years, and the best stadium prediction (either dome or open) was the only prediction kept for each market.

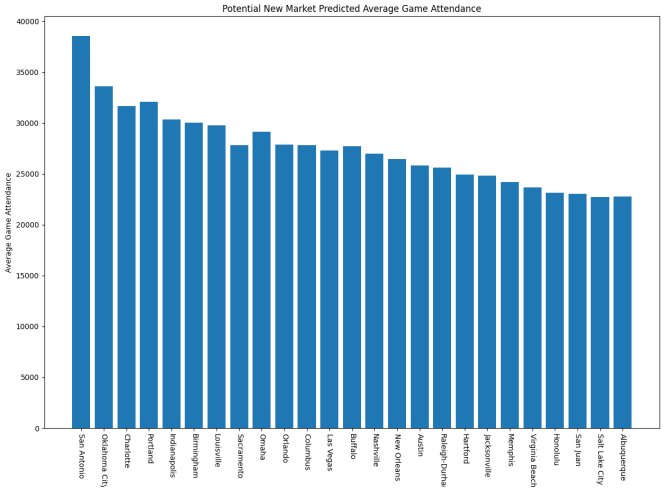
A first look at the predicted attendance for new markets took the 2021 predicted attendance and attempted to place each one of the three clusters (small, medium, large) of

existing markets created when initially exploring the data.



Note from the plot above that the new markets are clustered into only two groups (small and medium), with Sacramento being the only small market. All of the other markets are medium, which is promising as that provides access to more potential fans, which would provide an opportunity to build a bigger fan base, build a bigger stadium and sell more tickets.

With an idea of how the markets will behave, observe the yearly attendance predictions of each new market below:

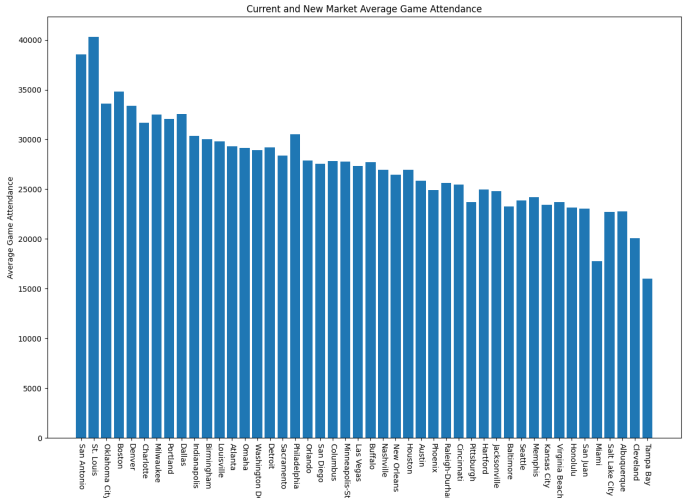


	Market	Dome	Place Score	Average Game Attendance
1	San Antonio	0	258	38569.058642
2	Oklahoma City	0	467	33599.898889
3	Charlotte	0	751	31682.262716
4	Portland	0	974	32084.454074
5	Indianapolis	0	1196	30347.221605
6	Birmingham	0	1213	30037.002716
7	Louisville	0	1329	29788.208025
8	Sacramento	1	1546	27821.792469
9	Omaha	1	1597	29155.437901
10	Orlando	1	2039	27863.531975
11	Columbus	0	2134	27826.336790
12	Las Vegas	0	2277	27304.976049
13	Buffalo	0	2343	27726.860247
14	Nashville	1	2429	26970.788519
15	New Orleans	1	2552	26439.141235
16	Austin	1	2923	25823.160864
17	Raleigh-Durham	1	2972	25599.929506
18	Hartford	1	3252	24947.110494
19	Jacksonville	1	3278	24804.652716
20	Memphis	1	3565	24174.643580
21	Virginia Beach	1	3806	23680.214691
22	Honolulu	1	4092	23145.567901
23	San Juan	1	4136	23024.434691
24	Salt Lake City	1	4228	22716.680494
25	Albuquerque	1	4366	22768.498889

From these results, it is interesting to note that only one of the four markets predicted to be the best in the initial hypothesis (Charlotte) made it into the top 5 performing markets. Note

that these figures are ordered by a metric created called place score. Place score is the sum of the position of the yearly attendance numbers for each position (for example, if a market was first in attendance for all 10 years evaluated, its place score would be 10). A lower place score means a team had higher attendance numbers over these 10 years, as that is an indicator of their yearly attendance numbers placing higher on a more consistent basis.

The predicted top five performing cities based on place score are San Antonio, Oklahoma City, Charlotte, Portland, and Indianapolis, so the model believes that these markets will consistently draw large crowds to games. As stated earlier, it is interesting to note that Charlotte was the only city of the four on the initial hypothesis that was predicted to perform consistently well. Las Vegas and Nashville were predicted to draw average crowds, and Salt Lake City was predicted to be a poor performer as well. This also holds true when these markets are compared to the rest of the league.



	Market	Dome	Place Score	Average Game Attendance
1	San Antonio	0	465	38569.0
2	St. Louis	0	497	40329.0
3	Oklahoma City	0	913	33600.0
4	Boston	0	1081	34786.0
5	Denver	0	1202	33364.0
6	Charlotte	0	1330	31682.0
7	Milwaukee	1	1557	32487.0
8	Portland	0	1563	32084.0
9	Dallas	1	1727	32549.0
10	Indianapolis	0	1904	30347.0
11	Birmingham	0	1965	30037.0
12	Louisville	0	2099	29788.0
13	Atlanta	0	2420	29301.0
14	Omaha	1	2447	29155.0
15	Washington DC	0	2510	28919.0
16	Detroit	0	2685	29226.0
17	Sacramento	0	2760	28361.0
18	Philadelphia	0	2990	30538.0
19	Orlando	1	3018	27864.0
20	San Diego	0	3102	27575.0
21	Columbus	0	3145	27826.0
22	Minneapolis-St.Paul	0	3283	27754.0
23	Las Vegas	0	3322	27305.0
24	Buffalo	0	3381	27727.0
25	Nashville	1	3517	26971.0

Note that the graph and the table[2] are sorted by a recalculated place score, which now includes all new markets and all current markets that only have one team. The model predicts San Antonio and Oklahoma City would consistently be two of the best performing markets in the league, keeping pace with two of the league's biggest markets in Boston and St. Louis. Charlotte, Portland, and Indianapolis are all also predicted to consistently draw strong crowds by major league standards as they are all predicted to be in the top 10 for single team markets. The crowds drawn in Las Vegas and Nashville are average on this scale, and Salt Lake City is down with the worst performers in the league.

Next Steps

To conclude, we believe that our research question has been confirmed: *can MLB attendance be predicted based on household valuations and income*? Despite the considerable success of predicting using income data, there are a few steps that we would like to take in order to improve our analysis before finalizing our recommendations. A random sampling of residents

[2] Note that the table only includes the Top 25 markets by place score, while the graph includes all markets

in the community could be used to collect data on factors outside of housing, including the level of baseball interest in the community. Infrastructure is an important consideration as well. Does the city have stadium plans or the capacity to handle a professional sports team? Does the city already have another professional sports team and how would this impact our analysis? In the case of the Atlanta Braves, who recently relocated from downtown Atlanta to Cobb County and Truist Park in the Battery, how do government programs such as SPLOST impact the stadium and tourism? The last thing we considered is overall approximate distance between current MLB markets. In a separate analysis, the population for the closest distance between each current team was analyzed; the results showed the Braves having the largest total population at nearly 30 million people compared to Houston who was the next closest at 20 million people (see figure 2 in appendix). This could suggest an opportunity to expand in the south.

Conclusion

The results of this study have shown that it is possible to predict attendance for an MLB market from housing and income data, and that the MLB may want to reconsider which markets it wants to place a new franchise in. Out of the four markets they are exploring now, there is strong indication from the two models of three generated in this analysis that only Charlotte would have consistently high attendance numbers. Nashville and Las Vegas were predicted to draw average crowds, and Salt Lake City would draw some of the smallest crowds in the league.

The model used for the final predictions was the xgboost model, as it had the lowest margin of error, and accounted for the most variability. However, the predictions for this model were still off by about 3,316 people per game, which does still translate into a few million dollars missed annually. While this would be the equivalent of a franchise paying one extra major league player (not an overly expensive cost for the team), these predictions could be better. Income and housing data does capture a lot of variability as it describes who lives in a market and how they are living, but as stated earlier there are other factors that could help the model account for more variance. Lowering the RMSE per game of the model down into the teens or hundreds would be the goal, as this would only translate to thousands of dollars missed annually. Including features such as market infrastructure, and market saturation could make this a possibility.

All things considered, the models created using the census and housing income data are very strong and can provide some very helpful insights to the MLB at the current stage of their expansion process. Based on the findings of the analysis, it is recommended that the MLB explore San Antonio and Oklahoma City as its primary expansion targets. Charlotte, Portland (OR), and Indianapolis are the three next best markets that should be explored if either of San Antonio or Oklahoma City are not a fit.

Appendix

Figure 1, Regression Final Results:

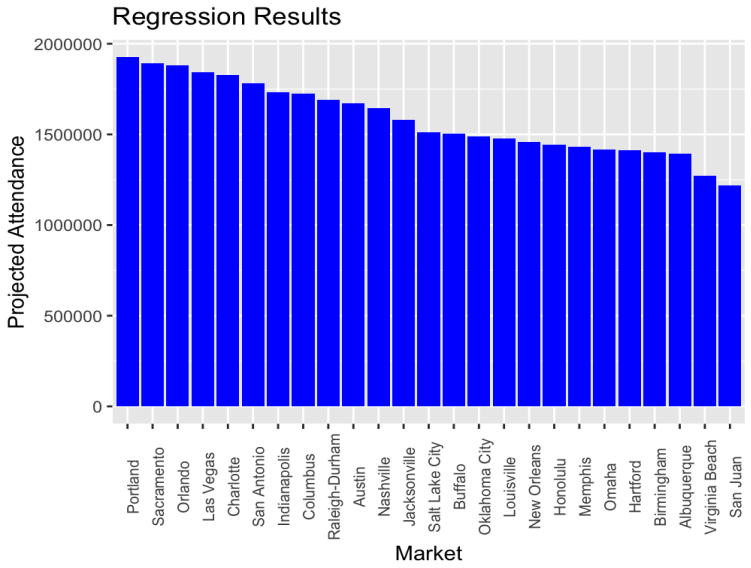
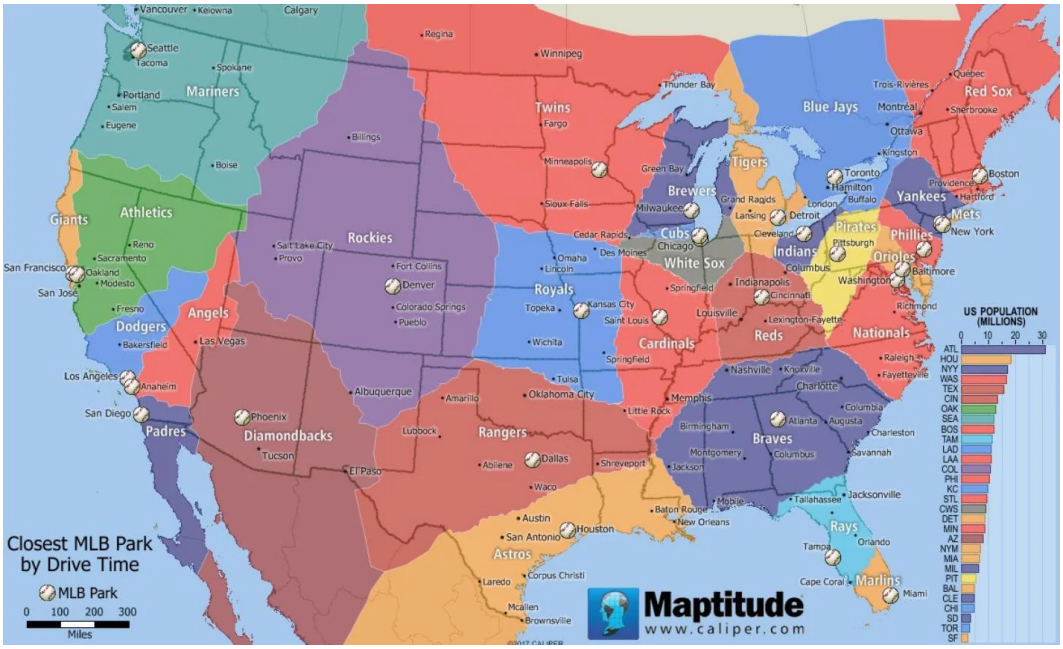


Figure 2, MLB Team Distance:



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.