MGT 6203 Group Project Proposal

TEAM INFORMATION (1 point)

Team #: 116

Team Members:

1. Jonathan Martin; 903858534
    - My name is Jonathan Martin and I currently work as a Business Intelligence Consultant in the legal field, mostly helping law firms and corporate law departments make sense of their spend.
2. Anthony Palmeri; 903737678
    - My name is Anthony Palmeri and I currently live in Atlanta, GA. I have a background in Finance but currently work at a tech startup as a Data Analyst. I have worked on data visualization projects based around diversity/inclusion and stock prices.
3. Yutai Liu; yliu648
    - I work in a consulting company as a project analyst with a background in civil engineering. My most recent analytical project was my graduate thesis on using non-destructive technology to detect the deterioration of concrete.
4. Joshua McLennan Mayanja; 903858147; jmayanja3
    - I currently work in the software industry as a Solutions Consultant. My background is in math/computer science, and the sports industry. I've done prior modeling/visualization projects based on sports, fraud/criminal investigations, and generated language.
5. Warren Spann 903833672 wspann3
    - I work in the software industry as an Enterprise Data Analyst. I've done work in accounting, marketing, sales, operations, ect. reporting on different metrics and KPI's.

OBJECTIVE/PROBLEM (5 points)

Project Title: Analyzing Housing and Income Data for MLB Expansion

Background Information on chosen project topic:

With the MLB looking to expand into additional cities, we thought about how we could analyze what the best expansion city would be. Our ideas were to base the analysis around housing and income data in the relevant city.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

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.

State your Primary Research Question (RQ):

Can MLB attendance be predicted based on housing and income data?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. How does an expansion team impact the housing prices in the community?
2. How does team valuation correlate with housing prices in the community?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

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.

https://www.espn.com/mlb/story/_/id/31346969/expansion-fees-major-league-baseball-teams-rise-22-billion-range

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

https://www.kaggle.com/datasets/claygendron/us-household-income-by-zip-code-2021-2011 – Household Income Data

https://www.baseball-reference.com/teams/TBD/1998-schedule-scores.shtml – MLB Attendance Numbers

https://en.wikipedia.org/wiki/Forbes_list_of_the_most_valuable_MLB_clubs – MLB Team Valuations per Year 2012-2021

Data Description (describe each of your data sources, include screenshots of a few rows of data):

Household Income Data – This dataset is composed of statistics gathered by the US census for US household income. This data was compiled in 2021 and ranges from the years 2011 through 2021 per zipcode. Some of the columns include total houses in a given income range, median and mean income per zipcode, as well as metrics on family, single, and nonfamily incomes.

# ZIP	Geography	Geographi...	Households	Household...	Household...	#
00601	860Z200US00601	ZCTA5 00601	5397.0	264.0	33.2	4.

MLB Attendance Numbers – This source contains general statistics per game for all MLB teams. The dataset is mainly composed of data provided by the major league teams and composed by Retrosheet. The main table we can plan to use contains dates for each game during the regular season as well as the recorded attendance.

Gm#	Date		Tm		Opp	W/L	R	RA	Inn	W-L	Rank	GB	Win	Loss	Save	Time	D/N	Attendance	cLI	Streak	Orig. Scheduled
1	Tuesday, Mar 31	boxscore	TBD		DET	L	6	11		0-1	4	0.5	Thompson	Alvarez		3:01	N	45,369	.96	–	

MLB Team Valuations per Year 2012-2021 – This dataset includes each team's valuations (per US million) by year. The years included in this dataset range from 2012 to 2021 per MLB team according to Forbes' historic valuation.

Team	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012
New York Yankees	6,000	5,300	5,000	4,600	4,000	3,700	3,400	3,200	2,500	2,300

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

For our general analysis, we'll likely use attendance as a dependent variable as this is the variable we'll be looking to predict. We can use team valuation, household mean/median incomes, total families, total nonfamily households to predict attendance per year, as well as make recommendations for cities viable for MLB expansion.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Our approach is to take housing income data as well as home value data to try to predict attendance for an MLB market. We will be using a linear regression model to do this, and potentially a few clustering models to group together similar markets. At the moment, we do not have any transformations to our data that need to occur. We will use years of data from 20 markets to train our model, and the remaining 10 markets to test. Once the model is trained, we will remove all statistically insignificant predictors to optimize the model.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

At the moment, the MLB is exploring expansion. Four cities they are very interested in are Charlotte, Las Vegas, Nashville, and Salt Lake City. Our hypothesis is that our model will predict one of these four cities to be the best city for expansion (they will have the highest predicted attendance/highest revenue). However, we will be testing many other cities around the country that do not have franchises as well, so a city outside of these four might prove itself worthy of relocation attention.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Our analysis will help the MLB make a decision worth billions of dollars. If our model is strong, then we will provide the MLB with a strong recommendation on where to place a new franchise. Placing this new team in the right market is critical for the MLB, as it would bring in billions of dollars in revenue to the league, and boost a city with an influx of jobs, tourism dollars, and a new way to tie the community together. Placing the franchise in the wrong market would open the league to billions of dollars in losses, and could leave a community ravaged. It's important that we provide the league with a strong recommendation so that they can make the decision that is best for all parties.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

- project proposal - June 21, 2023;
- project proposal video - July 2, 2023
- progress report - July 9, 2023
- project final report - July 20, 2023
- final project video presentation - July 23, 2023
- final project report with R code and slides - July 23, 2023