2019

ETL PROJECT

TEAM 1

Nahielys La Fontaine

Melissa Morgan

Sabir Samlani

Julia Thompson

Rocio Zambrano

ETL (Extract, Transform & Load) Project

MLB Team Profitability & Performance

PROJECT OVERVIEW

Task:

Our team was tasked with a hypothetical, typical work assignment at our hypothetical job – to take a bunch of data from multiple sources and migrate that data to a production database. However, our team was able to select the subject and data sources.

For this project, our team knew it was essential to properly prepare and format the data to successfully load it to a central host, or database storage system. This process of extracting data from multiple source systems, transforming it to suit business needs, and loading it into a destination database is commonly called ETL (extract, transform, and load).

While it is usually defined with three distinct steps, our team found that these steps are not as clean cut and the ETL process is much broader and requires a variety of actions – sometimes revisiting each of the three steps multiple times. But regardless of the steps, the purpose of the ETL process is always the same – it allows an organization to analyze and report on data more easily and effectively.

Subject:

Major League Baseball (MLB) is comprised of 30 teams located throughout the United States (except for the Toronto Blue Jays in Canada) and its regular season schedule consists of 162 games played by each team. The MLB has the highest season attendance of any sports league in the world, generating a total revenue of almost $10 billion league-wide and an average of nearly $330 million per team (as of season-end 2018).

But if we take a closer look at the numbers, is this average truly representative of all teams in the entire league? Are there outliers skewing the data? More importantly (or at least what piqued our curiosity the most) is the question, do teams with the highest payroll and average salary per player outperform the teams with the lowest payrolls?

Project Objective:

WHEN IT COMES TO THE MLB – CAN MONEY BUY WORLD CHAMPIONSHIPS?

For this project, we began to explore this question and other profit- and performance-related questions by creating a clean, relational database of tables/collections setup to easily perform analysis and create accompanying visualizations.

PROJECT SUMMARY

Extract:

The main objective of this phase is to find sources and retrieve all necessary data to move to the next phase. Our team began the extraction phase of the project by searching for relatively clean and reliable data associated with MLB team performance and profitability for the recently clinched 2019 season.

As a group, while searching for and ultimately selecting our data and sources, we put emphasis on seeking a correlation between a team's monetary data and its overall performance (aka wins and losses) for the 2019 season

Transform:

Our first steps in cleaning up the datasets involved removing unnecessary and blank columns. Most of our data cleaning and transformation was ensuring that every column in each table was assigned the correct data type (integer, varying character, numerical, etc), as well as assigning primary keys.

Load:

The load portion of our project happened in three separate phases but ultimately, we created a database with tables/collections ready to perform queries to suit a desired criterion.

The following pages provide more details of our process and methodology for this ETL project.

Data Gathering

Sources:

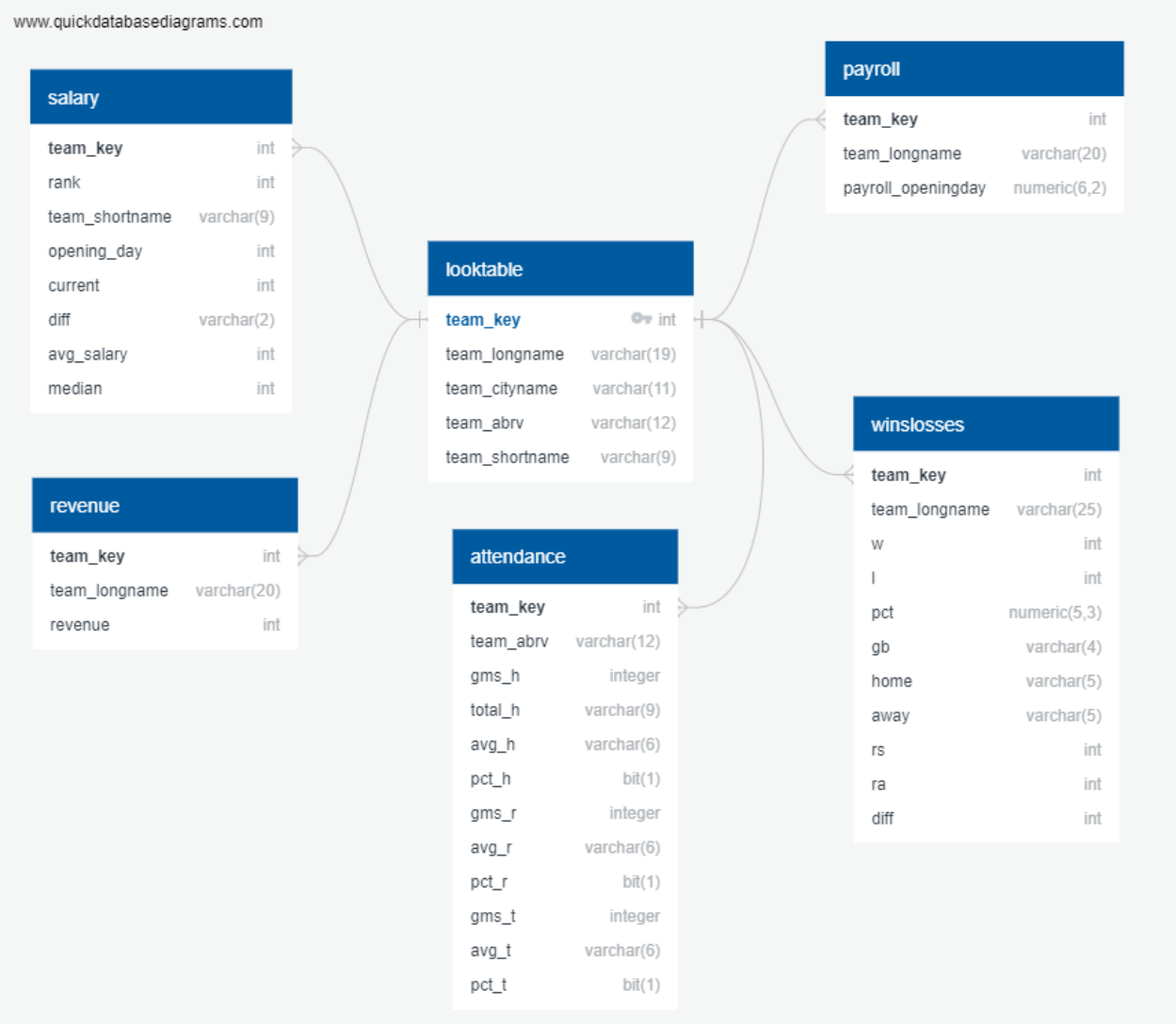
Data	Source	File
Attendance by MLB Team	ESPN	MLB_2019_Attendance.csv
Revenue per MLB Team	Statista – Forbes	MLB_Revenue.csv
2019 MLB Team Payroll	Statista – CBS Sports	MLB_Team_Payroll_2019.csv
Salary Info by MLB Team	USA Today	Teams Salary.csv
2019 MLB Team Wins/Losses	Fox Sports	MLB_2019_Wins_Losses.csv

INITIAL DATABASE

- Created a looktable to assign a key and columns for the multiple variations of a team's name.
- Using SQL, created a schema to create each table.
  Link to our Initial Schema
- Populated each table with data by importing the corresponding csv file.
- Assigned primary keys to each table via pgAdmin.
- Created a 'mlb_db' database in pgAdmin 4.
  schemas_mlb_db\schema_mlb_db_initial.sql
- Created the following tables within 'mlb_db'
  - attendance
  - looktable
  - payroll
  - revenue
  - salary
  - winslosses

Initial Entity Relationship Diagram (ERD)

ERD created from our five original datasets (plus the looktable we created)



After reviewing the initial database and ERD, we decided that including two more datasets would help to better the story – stadium capacity per team and average ticket price per team.

Additional Data Sources

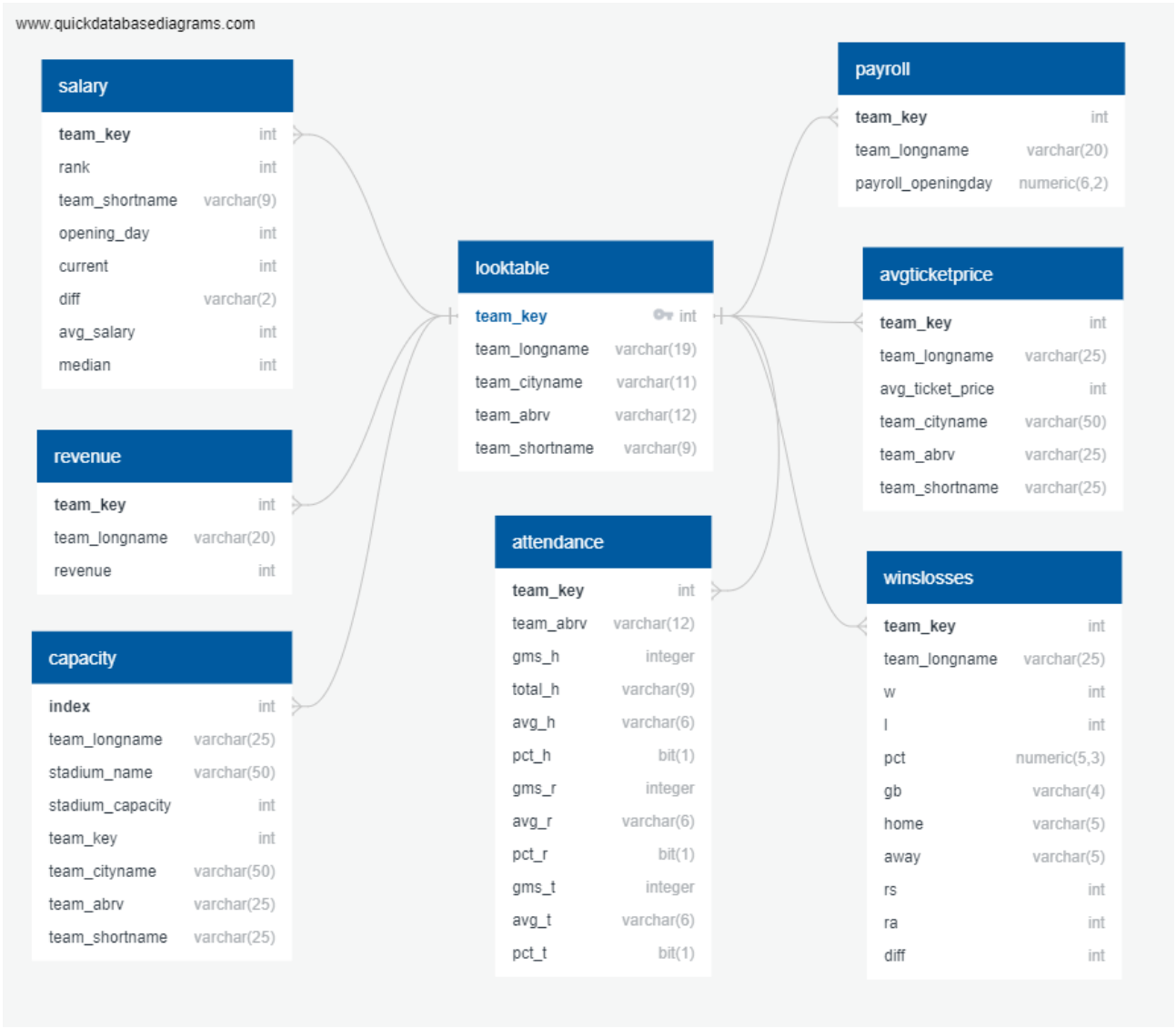
Data	Source	File
Average Price per Ticket	ESPN	MLB_Avg_Ticket_Price.csv
MLB Ballpark Capacity by Team	Statista – Forbes	MLB_Capacity.csv

Jupyter Notebook

For this second round of ETL, we took a different approach and used Jupyter notebook.

- Read in the two new csv files.
- Created a connection to our postgresql 'mlb_db'
- Pulled in the looktable to get keys
- Created a filtered dataframe pulling specific columns from each new set of data
- Merged each dataframe with the looktable dataframe
- Reviewed the datatypes for each newly merged table to confirm accuracy
- Pushed (appended) each new table to the database
- Confirmed successfully updated the database

ERD (Version_2)



Final Steps

We next concluded that one more round of cleaning and merging tables would result in the most streamlined and useful database that could easily be utilized for analysis on our subject.

MERGE TABLES

Capacity & Attendance

Link to Table Merge - Capacity & Attendance

Revenue & Ticket Prices

Link to Table Merge - Revenue & Ticket Price

Payroll & Salaries

Link to Table Merge - Salary & Payroll

Deleted the tables(collections) containing duplicated data sets

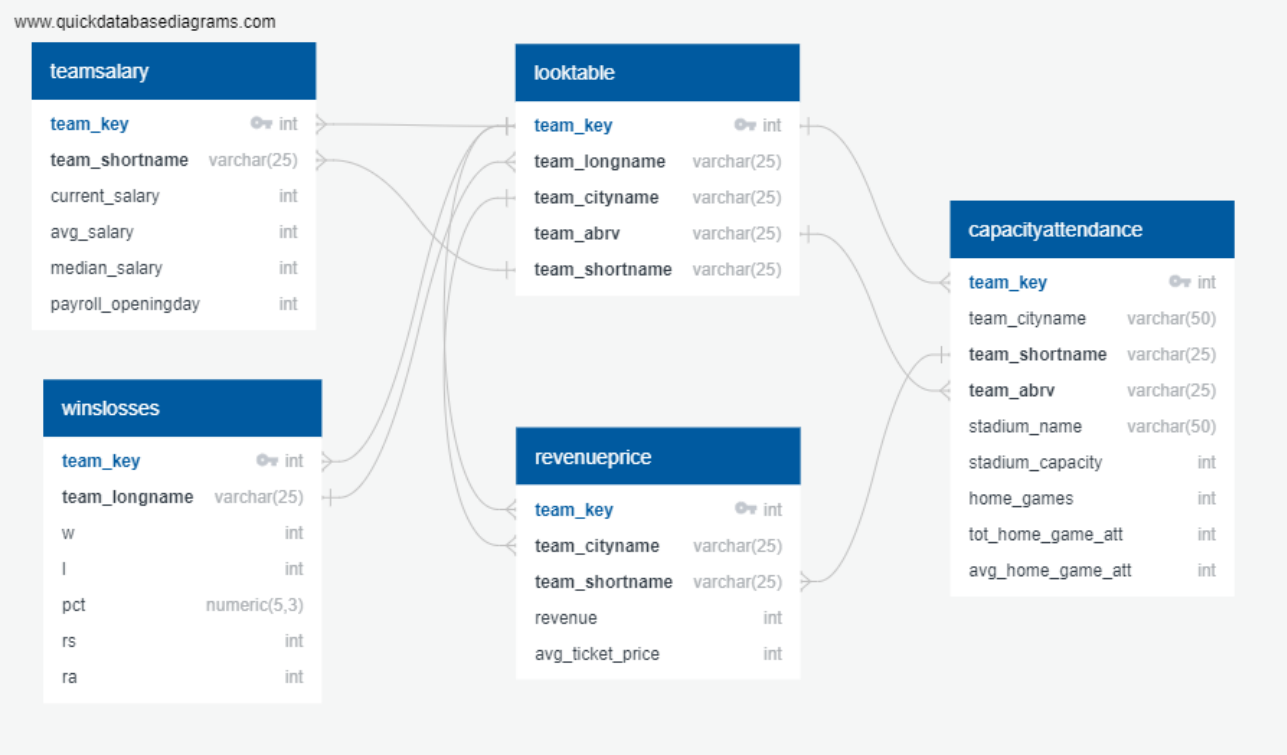
> attendance
> avgticketprice
> capacity
> payroll
> revenue
> salary

The end result:

Confirmed our cleaned tables had been added to the database by checking engine for current tables and querying each new table

FINAL DATABASE

Final ERD



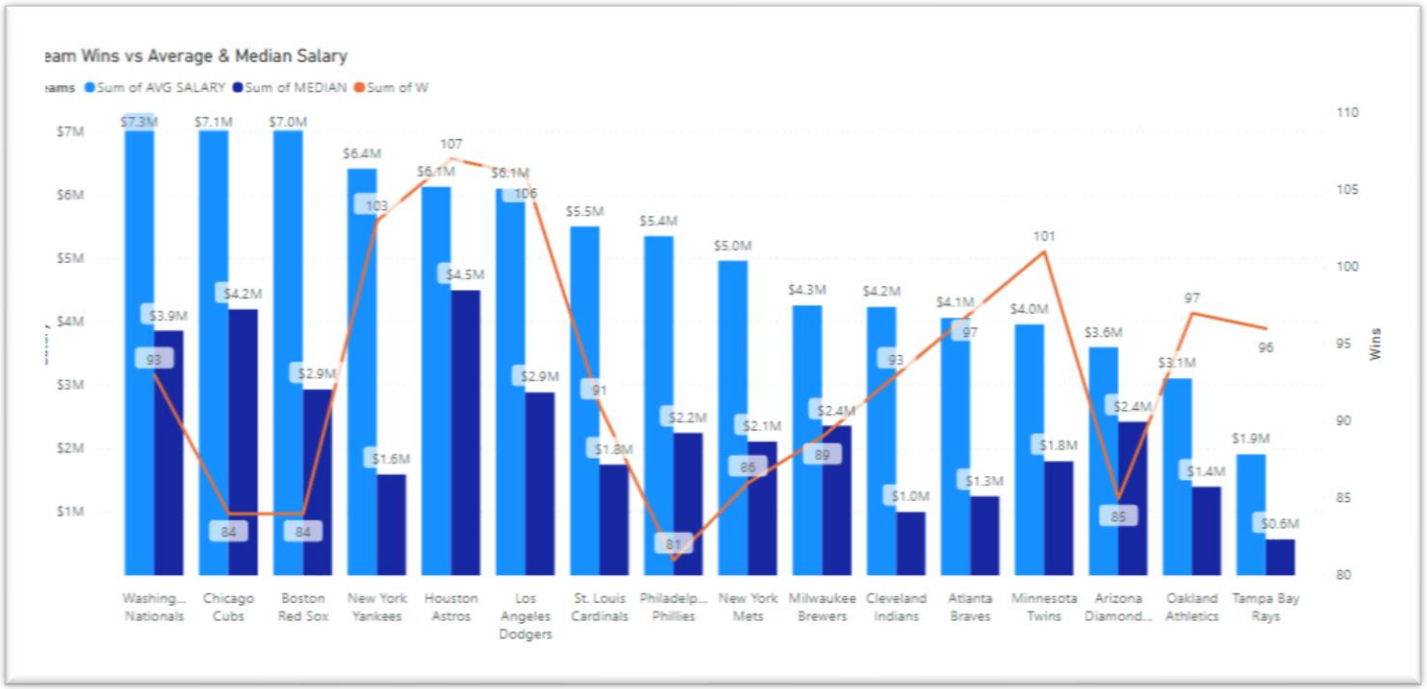
Link to Final Schema

While all three schemas will create a viable 'mlb_db', we believe our final product resulted in the most logical and useful database.

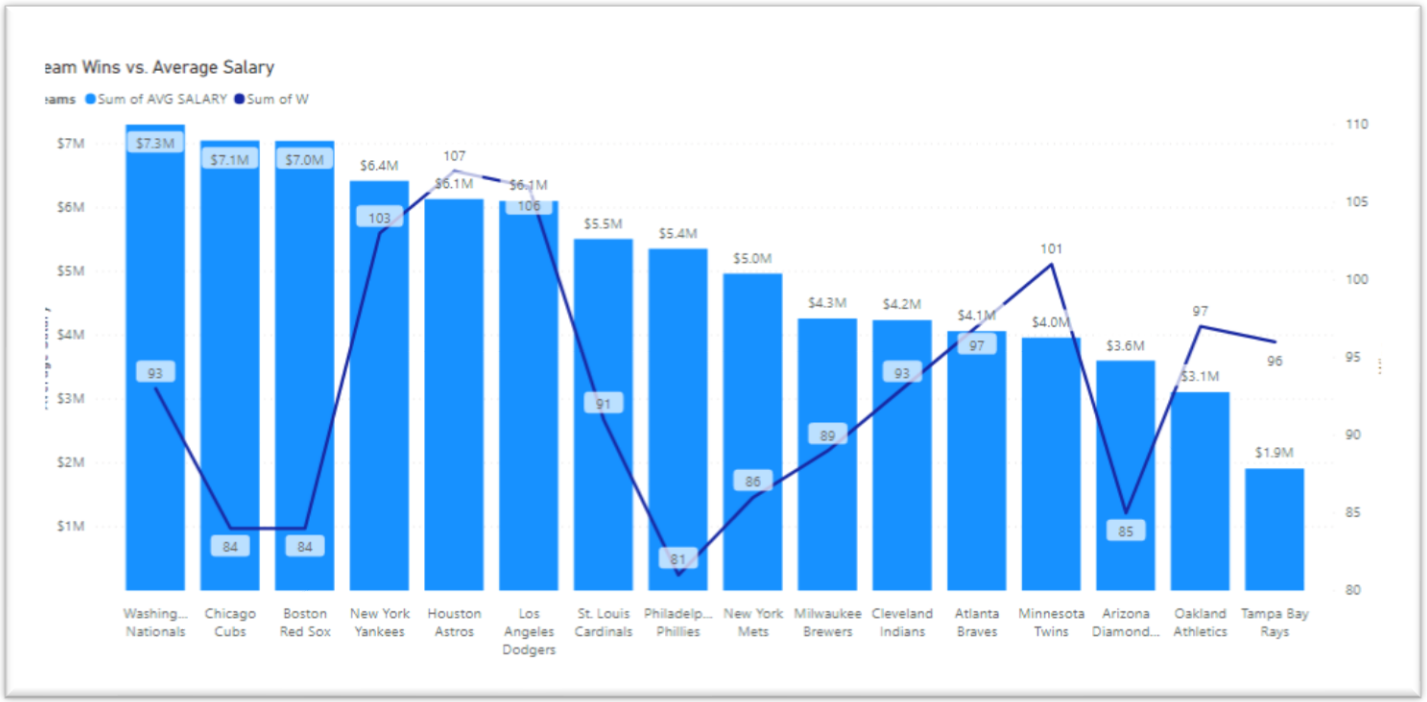
VISUALIZATIONS

The execution of our ETL project resulted in a clean, relational database completely ready for data analysis and visualizations. The following are a few examples of visualizations that are easily executed using the 'mlb_db' database we created.

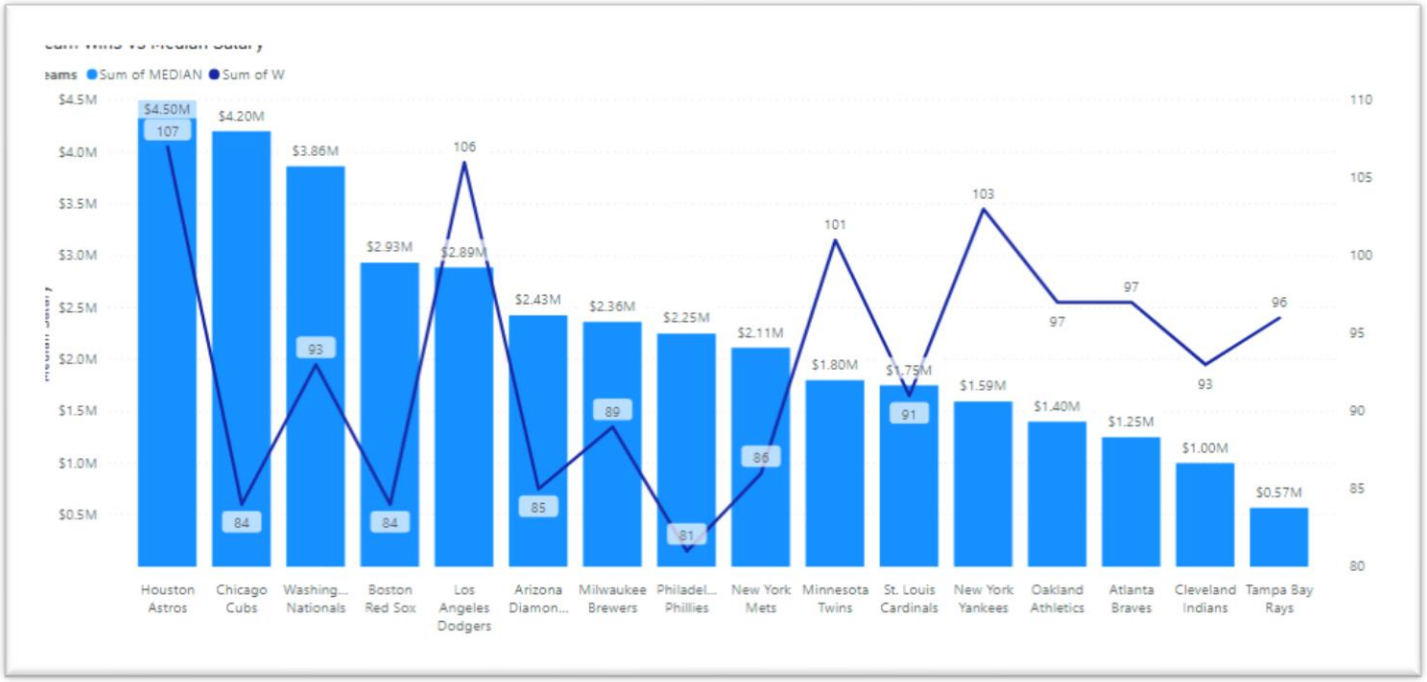
Average & Median Salary vs. Wins



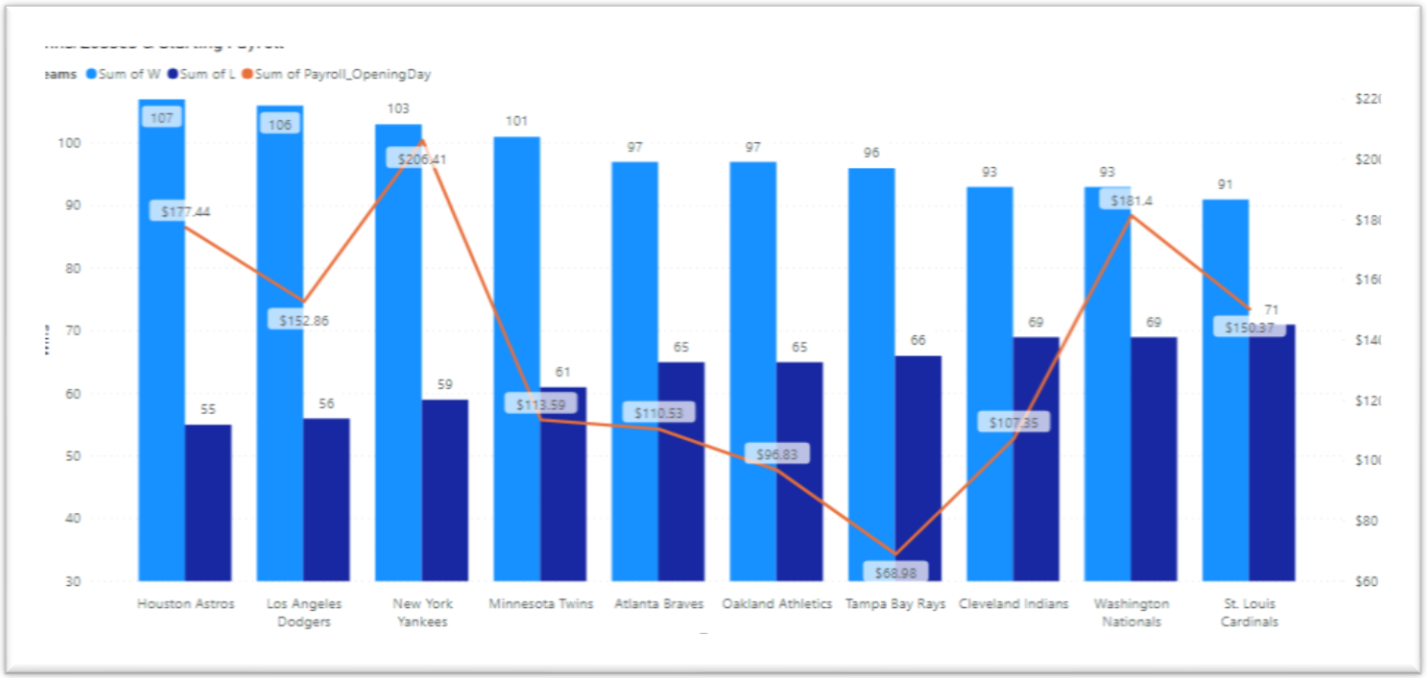
Average Salary vs. Wins



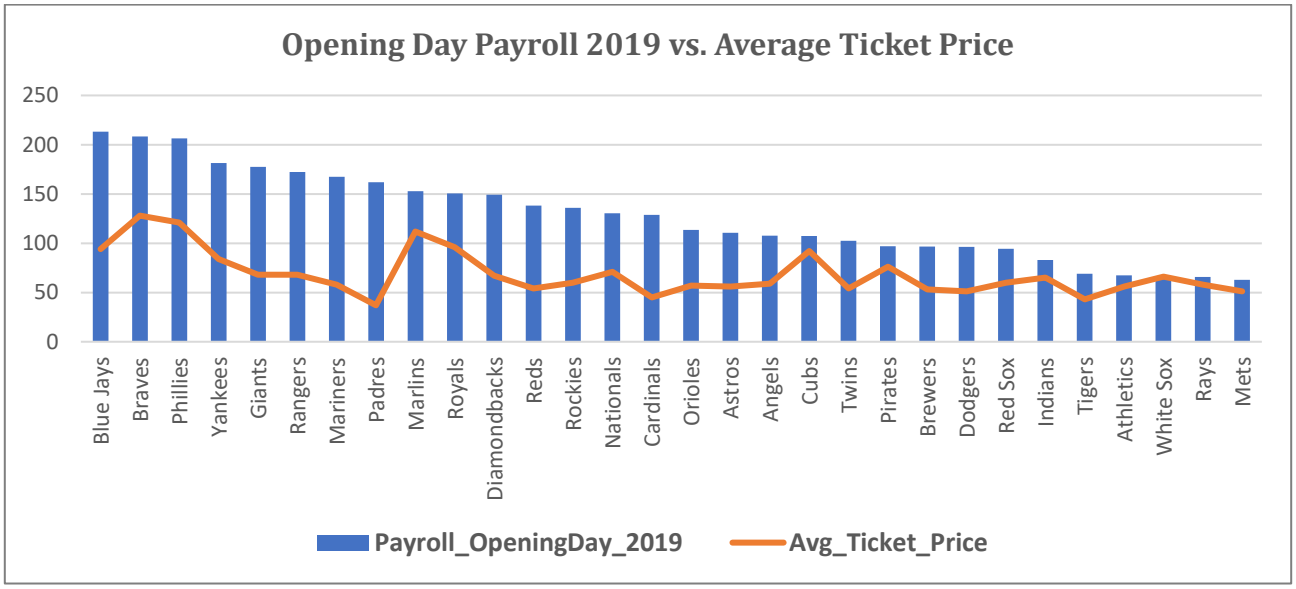
Median Salary vs. Wins

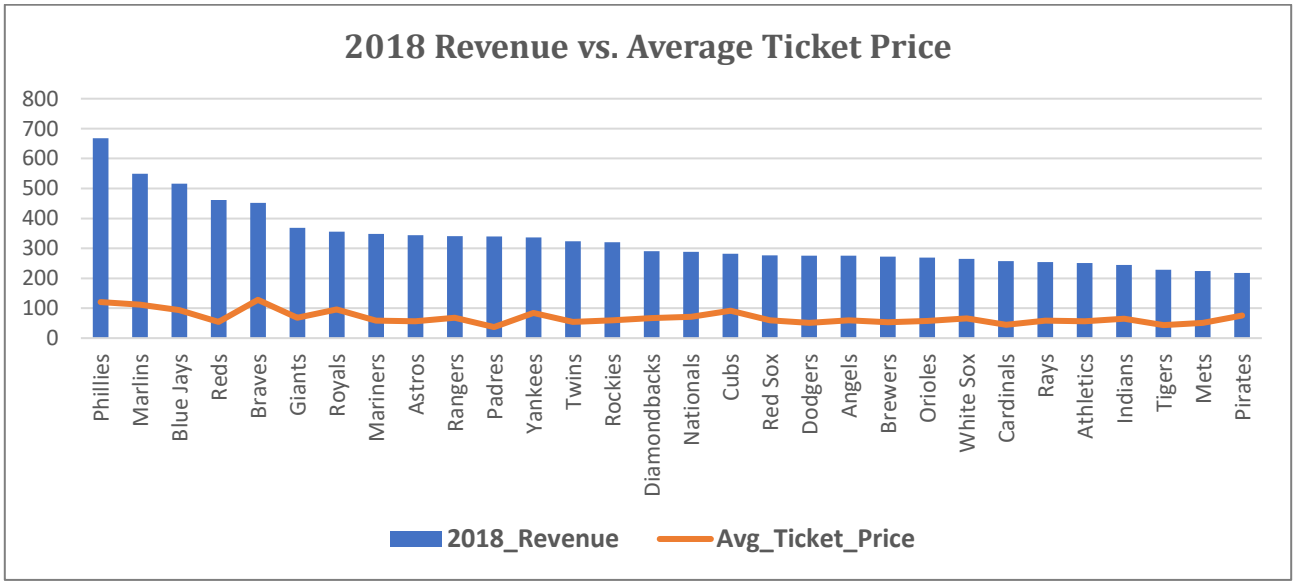


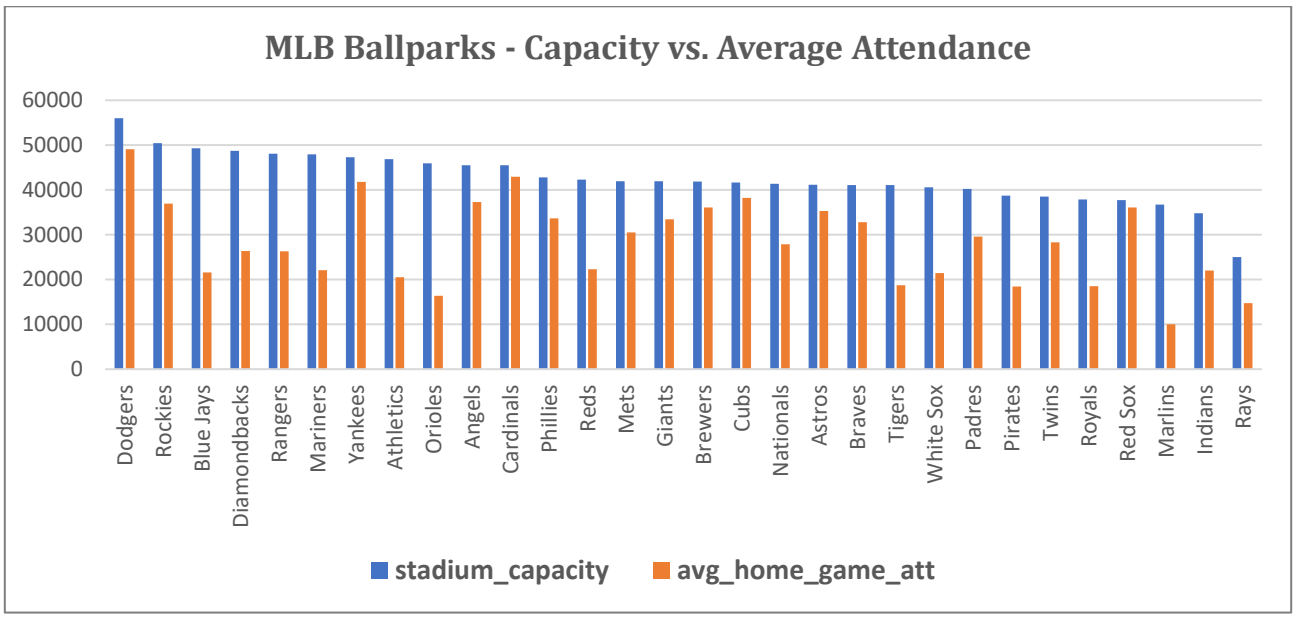
Total Payroll vs. Wins & Losses



The following bar charts contain data for all 30 MLB teams.






MLB Ballparks - Capacity vs. Average Attendance
60000
50000
40000
30000
20000
10000
0
Dodgers
Rockies
Blue Jays
Diamondbacks
Rangers
Mariners
Yankees
Athletics
Orioles
Angels
Cardinals
Phillies
Reds
Mets
Giants
Brewers
Cubs
Nationals
Astros
Braves
Tigers
White Sox
Padres
Pirates
Twins
Royals
Red Sox
Marlins
Indians
Rays
stadium_capacity
avg_home_game_att