**Final project – Business Analytics Fundamentals - DSCI 505 – Nick Macchio**

1) Define the problem: What is the business problem that you are attempting to solve. Why is Python a good way to solve this problem?

The Problem:

* In baseball, and in all sports, the main goal, since the beginning of time, has been to win. For MLB teams, in order to win consistently, you need to find out what makes your team successful. Baseball is unique in the sense that there are so many different factors that can lead to your team outscoring the opponents. By finding out which statistics have the biggest impact on winning, teams can attempt to put players on the field that give them the best chance to win.

Question:

* Which factors have the most impact on a Major League Baseball team's winning percentage?

Why python?

* Python works well for this project because of its ability to manipulate big data and build a number of different graphs that I need.

2) Describe the dataset:

* Before I get into describing the dataset I should mention that this dataset is from the Lahman Baseball Database. This database has 27 different csv files dating all the way back to 1871, in total this database has over 200,000 rows of data. This quote is straight from the website I found this data on.
* “This database was created by Sean Lahman, who pioneered the effort to make baseball statistics freely available to the general public. What started as a one man effort in 1994 has grown tremendously, and now a team of researchers have collected their efforts to make this the largest and most accurate source for baseball statistics available anywhere. This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2023.”
* The dataset I went with is the Teams.csv because it has all the information I need to answer my question above.

a) How many variables are in the data set?

* There are 48 variables in the original dataset
* Since this data dates back to 1871, there were a number of variables that were missing data. As well as some variables that just aren’t necessary, I took out 17 variables.
* I also created one variables (Winning Percentage) by taking wins divided by games
* This list is straight from the website, it describes all 48 of the variables in this dataset.
  1. yearID Year
  2. lgID League
  3. teamID Team
  4. franchID Franchise (links to TeamsFranchise table)
  5. divID Team's division
  6. Rank Position in final standings
  7. G Games played
  8. GHome Games played at home
  9. W Wins
  10. L Losses
  11. DivWin Division Winner (Y or N)
  12. WCWin Wild Card Winner (Y or N)
  13. LgWin League Champion(Y or N)
  14. WSWin World Series Winner (Y or N)
  15. R Runs scored
  16. AB At bats
  17. H Hits by batters
  18. 2B Doubles
  19. 3B Triples
  20. HR Homeruns by batters
  21. BB Walks by batters
  22. SO Strikeouts by batters
  23. SB Stolen bases
  24. CS Caught stealing
  25. HBP Batters hit by pitch
  26. SF Sacrifice flies
  27. RA Opponents runs scored
  28. ER Earned runs allowed
  29. ERA Earned run average
  30. CG Complete games
  31. SHO Shutouts
  32. SV Saves
  33. IPOuts Outs Pitched (innings pitched x 3)
  34. HA Hits allowed
  35. HRA Homeruns allowed
  36. BBA Walks allowed
  37. SOA Strikeouts by pitchers
  38. E Errors
  39. DP Double Plays
  40. FP Fielding percentage
  41. name Team's full name
  42. park Name of team's home ballpark
  43. attendance Home attendance total
  44. BPF Three-year park factor for batters
  45. PPF Three-year park factor for pitchers
  46. teamIDBR Team ID used by Baseball Reference website
  47. teamIDlahman45 Team ID used in Lahman database version 4.5
  48. teamIDretro Team ID used by Retrosheet
* The variables I decided to remove were:
  + 2 – lgID (Missing too much data)
  + 4 – franchID (Essentially the same as teamID)
  + 5 – divID (Missing too much data)
  + 7 – Ghome (Missing too much data)
  + 11 – DivWin (Missing too much data)
  + 12 – WCWIN (Missing too much data)
  + 13 – LgWin (Missing too much data)
  + 14 – WSWIN (Missing too much data)
  + 25 – HBP (Missing too much data)
  + 26 – SF (Missing too much data)
  + 42 – Park (Unnecessary to answering my question, also there have been so many different ballparks since 1871)
  + 43 – attendance (Missing too much data)
  + 44 – BPF (Unnecessary for this question as it relates to the park variable)
  + 45 – PPF (Unnecessary for this question as it relates to the park variable)
  + 46 – teamIDBR (very similar to teamID)
  + 47 – teamIDlahman45 (very similar to teamID)
  + 48 – teamIDretro (very similar to teamID)
* The variables I decided to create
  + WinPct – The variable is just the teams wins divided by the games played
    - The reason I decided to do this is because today, teams play 162 games each year but back in the day that number was not that high so I wanted to try to make something that could be the closest thing to equal for each team each year.
  + Runs\_per\_game – Exactly what it sounds like

b) What are the data types and levels of measurement for each of the variables?

* Data Types:
* Categorical Variables:
  + yearID
  + teamID
  + name
* Numerical Variables:
  + Team Performance Stats
    - G
    - W
    - L
    - Rank
  + Offensive Stats
    - R
    - AB
    - H
    - 2B
    - 3B
    - HR
    - BB
    - SO
    - SB
    - CS
  + Pitching Stats
    - ERA
    - CG
    - SHO
    - SV
    - IPouts
    - HA
    - HRA
    - BBA
    - SOA
    - ER
  + Fielding Stats
    - E
    - DP
    - FP
* Levels of measurement (for the variables I am using)
  + All of the categorical variables are Nominal.
  + All of the numerical variables are Ratio, except for Rank which is Ordinal.

c) How many observations are in the data set?

* There are 3045 observations in this data set.

d) Are there any missing values or outliers in the data set?

* There are a number of missing values just because of how old the dataset is. The variables with extensive missing data I removed because I believe it would ruin the model. I think the game of baseball obviously has changed quite a bit but if you look at all of the numbers from all time it can show how teams have either evolved or can show which factors have affected winning percentage the most over the last 152 years.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

* There are a lot of outliers because of just how long the data goes back
* The variables are on the left and the number of outliers for each variable is in the orange on the right.

e) How did you choose to handle the missing values and outliers?

* So for the missing values I took out a number of variables because of various reasons I stated above
* For the outliers I am going to leave them all because they are all statistically important for the data

f) What are some of the descriptive statistics for key variables?

* Because of the large amount of variables I have, and this question does say “for key variables” I am going to use my best judgement and only describe a handful of what I think are the important variables.
* Some of the “important” variables I am going with is
  + WinPct – Win percentage
  + R – Runs scored
  + ERA – Earned run average
  + FP – Fielding percentage
  + RA – Runs allowed
  + HR – Home runs
  + A screenshot of a computer

    Description automatically generatedA screenshot of a computer

    Description automatically generatedA screenshot of a computer

    Description automatically generatedA screenshot of a graph

    Description automatically generatedHRA – Home runs allowed

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a computer

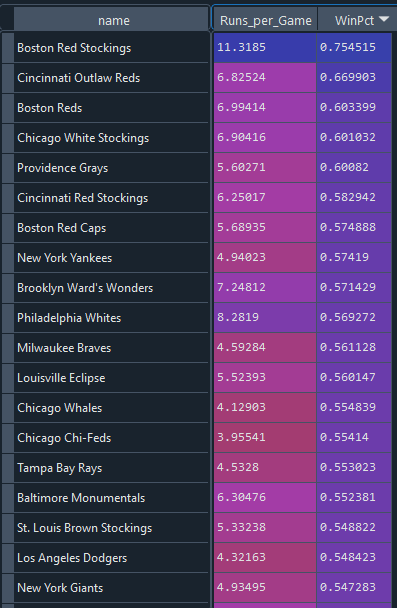
Description automatically generated

* WinPct (Winning Percentage)
* The mean being right around 50% makes sense because someone needs to win, and someone needs to lose. The reason it is slightly lower than 50% is because some games early on were cancelled and no one was awarded the win. The standard deviation being at 0.09 means that a lot of teams winning percentages stay close to the mean, this one seems pretty self-explanatory. The range of min and max is 0 to 0.87, all this means is that some years there are very good teams and some years there are teams that didn’t win any games, early on in baseball some teams played less than 10 games so that is very possible. Not nowadays, however.
* R (Runs Scored)
* Teams score an average of 682 runs in a season. The median is 692, which is very similar to the mean when stretched out over a whole season. The standard deviation is 138.7, which means there’s a big gap between teams. I think this is because they used to play less games and this number isn't a percentage so the teams that play more, score more, obviously. As I just mentioned with the variety of runs scored over a season compared to games played, the range is very big, 24 to 1220. The IQR is 615-766, which nowadays, is pretty accurate throughout a whole season.
* RA (Runs allowed)
* This stat basically mirrors the runs scored which obviously makes much sense.
* FP (Fielding Percentage)
* Fielding percentage has been very consistent throughout history with the mean at 0.97, and the median at 0.98. The standard deviation isn't much to note at 0.03, which means that if you're a major leaguer, for the most part you make a lot of the plays that should be made. The range goes from 0.76 to 0.99, with fielding percentage having the third most outliers, this could make sense if some teams had a relatively bad defense season. The IQR is 0.97 to 0.98, which means that generally, teams are good defensively.
* ERA (Earned Run Average)
* The average ERA is 3.85, meaning teams generally allow just under four earned runs per game. The median is 3.85 which is the same as the mean. The standard deviation of 0.76 shows there’s some differences in teams, even nowadays it makes sense because there are some teams whose strength is pitching and others whose weakness is pitching. The range is relatively large from 1.22 all the way up to 8. This is just the variance from a good pitching team and a bad one. The IQR is 3.38 to 4.38 which shows that most teams give up right around 4ish earned runs per game.
* HR (Home Runs)
* Teams hit an average of 107 home runs a season, with a median of 112. That’s very close when looking at it over a whole year. The standard deviation of 64.5 shows that some teams have hit a lot more than others, I think that today teams are hitting more home runs than ever, so it is easy to see this number being large compared to the earlier days of baseball. The range shows this very clearly, from 0 all the way to 307. The IQR is 47 to 157. I think that these numbers are being pulled lower because of the early days of baseball. If you look at just the last 20 years, I believe that it is a much higher number.
* HRA (Home runs allowed)
* This is the same as runs allowed in that it is mirroring the home runs hit.
* A screenshot of a computer

  Description automatically generatedA screenshot of a graph

  Description automatically generatedPivot tables
* This isn't the entire pivot table, but it is about 90% of the teams that have an all-time winning percentage over 50%. This table also has the ERA alongside them. There is a direct correlation that teams who have a lower ERA have a higher winning percentage on average. Now, even in this pivot table there may be some outliers, but I believe this is a good representation on how good pitching equals more wins. Of course there are always going to be some exceptions, but I think for the most part it paints a good picture.

A screenshot of a computer

Description automatically generated

* I did a couple of things here, first off it is the same teams based off their winning percentage. I did, however, make a new column of runs per game, I think this is pretty self-explanatory, how many runs per game they score, all time. Again, with ERA, you can see that in general, the more runs per game scored the higher the winning percentage.
* Correlation matrix’s
* A screenshot of a computer

  Description automatically generatedOffensive stats
* A screenshot of a computer screen

  Description automatically generatedDefensive stats
* I tried to put everything into one correlation matrix, but it was way too big, so I made two of them, one with offensive stats and the other with defensive stats. Both tables have the Rank, G, W, L, and WinPct.
* Offensive stats matrix
  + Runs Per Game: Medium positive correlation (0.42) – more runs equal more wins.
  + Hits: Small positive correlation (0.28) – consistent hitting supports winning, but obviously scoring helps more.
  + Walks: Small positive correlation (0.25) – getting on base helps them win.
* Defensive stats matrix
  + ERA: Moderate negative correlation (-0.46) – Better pitching equals more wins.
  + Shutouts: Moderate positive correlation (0.43) – When a shutout is pitched wins are more likely.
  + Fielding Percentage: Small positive correlation (0.21) – Good defense helps but isn’t as helpful as pitching.

g) Create data visualizations for variables in the data set.

* Because of how many variables I am working with I am only going to do 6 graphs to visualize my project.
* A blue dot diagram with white text

  Description automatically generated with medium confidence1. ERA vs Win Percentage scatterplot
* A blue graph with numbers

  Description automatically generated2. Runs per game histogram
* 3. Runs per game vs Win Percentage scatterplot

A blue dot diagram with white text

Description automatically generated

* 4. Fielding percentage vs Winning Percentage scatterplot
* 5. Runs allowed per season histogram

A blue dots on a white background

Description automatically generated

A blue graph with numbers

Description automatically generated

* A blue dotted graph with white background

  Description automatically generated6. Home runs vs Winning Percentage Scatterplot
* 1. ERA vs. WinPct Scatterplot
* This scatterplot shows how teams with better pitching and lower ERAs tend to win more games. It shows the importance of keeping the other team off the scoreboard.
* 2. Runs Per Game Histogram
* This histogram shows how many runs teams score in a game on average. Most teams are grouped around 4–5 runs per game, which feels pretty standard. Teams that score more than that or more than average have a better chance of winning.
* 3. Runs Per Game vs. WinPct Scatterplot
* This scatterplot connects scoring runs to winning games. It shows that teams scoring more runs per game usually have a better chance of winning percentages. Offense and scoring aren’t everything, but this graph shows that it is very important.
* 4. Fielding Percentage vs. WinPct Scatterplot
* This graph looks at defense and it shows that teams with a higher fielding percentage often times win more games, though it isn't as strong as pitching or scoring. Still, it does make sense that good defense helps win games .
* 5. Runs Allowed Per Season Histogram
* This histogram shows how many runs teams allow in a season. Most teams fall in the middle, but the fewer runs you allow, the better your chances are of winning. It goes back to the idea that defense and pitching go hand in hand for success.
* 6. Home Runs vs. WinPct Scatterplot
* This scatterplot checks if hitting home runs helps winning. While it’s not the biggest factor, you can see teams with more home runs generally do better. Hitting Home runs helps, but there needs to be more if you want to be successful throughout a whole season.

4) What do the descriptive statistics and visualizations that were created tell us about the data set? How can we use the results to “tell a story” about the data? What are the key findings and conclusions that you could present to the business to help answer their questions about the data from Part #1?

* In this project, I had a lot of variables, so I used my background in baseball and my best judgement to focus on some of the key variables that impact winning percentage. This project showed that ERA, and runs per game had the strongest relationships with winning percentage, so I mainly focused on these factors and closely related factors for my project.
* The story that I can tell from the data is that pitching is the most important factor relating to Winning Percentage. Teams that have lower ERAs win more games. Scoring also helps quite a bit, teams with a higher runs per game average also tend to win more. Lastly, defense is an important stat as it does help with winning but not nearly as much as offensive production and pitching.
* My suggestions for MLB teams would be to first focus on effective and consistent pitching, then focus on getting the best run production.

6) Results: What did you learn about the data? Describe how this exercise helped you become a better data analyst. What would you do differently? What were the biggest challenges when working with the data. Why is or isn’t Python a good tool to use for this type of analysis?

* After doing this project, I believe that it's helped me become a better data analyst because of the way I look at large datasets, as well as learning the best ways to clean big data. Throughout the whole semester working on python I was somewhat confident in how to create visualizations and how to clean data. What this project proved to me was that I can be flexible with my coding, and I can do a number of different things with it.
* If I had to do this again, I would have tried to find different things and different averages that could affect winning percentage. I also would have tried to maybe cut down the information from 1871 to maybe just post 1970 or something close to that. The reason I didn't do that is because I wanted to one, make it show all time what has been effective and two, I wasn't sure if there would have been enough data, 1000 rows, to do this project.
* The hardest part was narrowing down the variables to look at because there were so many options. Dealing with the missing data was somewhat tricky too because I needed to ask myself if that data was important or not. Also, the visualizations, I wish I knew a better way to show which data was more effective or somehow even if I could mix the data up to possibly tell a different story. I am not sure what kind of graph would even help but I wish there was something a little different I could have done there.
* Python was the perfect tool for this project because it made working with my big data pretty easy. Throughout the semester we did a lot of visualizations, and I love how, in python, there are multiple ways to do the same things so in python creating graphs was pretty easy. After using R for my data mining project, I realized I like Python way more. To me, it is just easier to work with and I think it is a bit more user friendly.

5) Code to solve the problem: Please copy/paste all Python code to solve the problem into a section of the document. Code should be commented to explain steps taken.

Code:

# Final project

#

# What factors have the most impact on a MLB team's winning percentage?

# Import pandas

import pandas as pd

# Calling in the dataset

teams = pd.read\_csv("C:/Users/nickg/Teams.csv")

# Creating a new dataset (teams1) without the variables that are missing data or

# irrelevant data in terms of helping me answer my question

# Variables removed:

# LgID, franchID, divID, Ghome, DivWin, WCWin, LgWin, WSWin, HBP, SF, park,

# attendance, BPF, PPF, teamIDBR, teamIDlahman45, teamIDretro

# Variables kept:

# yearID, teamID, Rank, G, W, L, R, AB, H, 2B, 3B, HR, BB, SO, SB, CS, RA, ER,

# ERA, CG, SHO, SV, IPouts, HA, HRA, BBA, SOA, E, DP, FP, name

teams1 = teams[["yearID", "teamID", "Rank", "G", "W", "L", "R", "AB",

"H", "2B", "3B", "HR", "BB", "SO", "SB", "CS", "RA", "ER", "ERA", "CG", "SHO",

"SV", "IPouts", "HA", "HRA", "BBA", "SOA", "E", "DP", "FP", "name"]]

# Adding one column to create my response variable

# (winning percentage = wins/games)

teams1["WinPct"] = teams1["W"]/teams1["G"]

# Adding another column for the sake of my pivot table

# how many runs per games

# (Runs per game = Runs/games)

teams1["Runs\_per\_Game"] = teams1["R"]/teams1["G"]

# quantiles to help find outliers (numerical) (Chapter 4)

# outliers = Q1 - (1.5\*IQR) = Lowerfence

# outliers = Q3 + (1.5\*IQR) = Upperfence

# Creating dataset with just numeric variables

num\_var = teams1[["Rank", "G", "W", "L", "R", "AB",

"H", "2B", "3B", "HR", "BB", "SO", "SB", "CS", "RA", "ER", "ERA", "CG", "SHO",

"SV", "IPouts", "HA", "HRA", "BBA", "SOA", "E", "DP", "FP", "WinPct"]]

# outliers\_dictionary to store outliers

outliers\_dictionary = {}

# Create for loop for Q1, Q3, IQR, lowerfence, upperfence, and number of outliers

for i in num\_var:

Q1 = teams1[i].quantile(0.25)

Q3 = teams1[i].quantile(0.75)

IQR = Q3 - Q1

lowerfence = Q1 - (1.5 \* IQR)

upperfence = Q3 + (1.5 \* IQR)

LF\_out = (teams1[i] < lowerfence).sum() # finding out the number of outliers

UF\_out = (teams1[i] > upperfence).sum() # same thing here

total\_outliers = LF\_out + UF\_out

outliers\_dictionary[i] = total\_outliers

# FYI this for loop took forever to make with many trial and seemily more errors

# It makes sense theres a lot of outliers because of the vast timeframe of this data

##################################################################

# Using .describe() to get the descriptive statistics

WinPct\_describe = teams1.WinPct.describe()

run\_describe = teams1.R.describe()

era\_describe = teams1.ERA.describe()

fp\_describe = teams1.FP.describe()

ra\_describe = teams1.RA.describe()

hr\_describe = teams1.HR.describe()

hra\_describe = teams1.HRA.describe()

# creating a pivot table to show which teams have had the highest winning %

# as well as tying it with ERA to show just how imporant pitching is to winning

piv\_table = teams1.pivot\_table(index = "name", values = ["WinPct", "ERA"],

aggfunc = "mean")

# making a second pivot table with runs per game instead of ERA

piv\_table2 = teams1.pivot\_table(index = "name", values = ["WinPct", "Runs\_per\_Game"],

aggfunc = "mean")

# Making 2 correlation matrix's with the numeric variables only

# first one is offensive stats only

off\_num\_var = teams1[["Rank", "G", "W", "L", "R", "AB",

"H", "2B", "3B", "HR", "BB", "SO", "SB", "CS", "Runs\_per\_Game", "WinPct"]]

corr\_matrix = off\_num\_var.corr()

# second on is defensive stats only

def\_num\_var = teams1[["Rank", "G", "W", "L", "RA", "ER", "ERA", "CG", "SHO",

"SV", "IPouts", "HA", "HRA", "BBA", "SOA", "E", "DP", "FP", "WinPct"]]

corr\_matrix2 = def\_num\_var.corr()

####################################################################

# Data vizualizations

# ERA vs WinPct scatterplot

era\_scat = teams1.plot.scatter("WinPct", "ERA")

# Runs per game histogram

rpg\_hist = teams1.Runs\_per\_Game.plot(kind = "hist", bins = 15)

# Runs per game vs WinPct scatterplot

rpg\_scat = teams1.plot.scatter("WinPct", "Runs\_per\_Game")

# Fielding percentage vs WinPct scatterplot

fp\_scat = teams1.plot.scatter("WinPct", "FP")

# runs allowed per year histogram

ra\_hist = teams1.RA.plot(kind = "hist", bins = 15)

# Home runs vs WinPCt scatterplot

hr\_scat = teams1.plot.scatter("WinPct", "HR")