Predicting ERA - Final Project

Jacob Koewler, Liam Bengston, Yehe Cui

2025-12-05

1 Overview

From the start, we knew we wanted to focus on sports analytics. We quickly settled on Baseball/MLB due to the vast amounts of public data that is easily accessible. We found all of our data on baseball savant. We also decided at this point our goal would be to find which baseball statistic is the best indicator of ERA. Once we had the data, We wrangled it in order to be able to create good data visualizations. From these, we will be able to determine which statistic is the best indicator of ERA.

Our data does meet the FAIR principles for Open Data. First off, our data is easily findable, as it is on a public website. Second, our data is accessible, as there is no authorization needed to view the data. Third, the data is Interoperable and Reusable because the data is well-organized and follows common baseball abbreviations for the Statistics. Our data also fulfills some of the CARE principles. First, our data is for the collective benefit, as this data can be used to learn more about these pitchers. Our data is also Ethical and Responsible, as there is no harm done in collecting this data. However, we had a hard time verifying whether our data satisfied Authority to Control.

2 Baseball Context

We are looking at 4 different indicators for Earned run average (ERA): WHIP, OBP, K%, and BB%. ERA is calculated by dividing the amount of runs that the pitcher allowed by the amount of innings pitched, all multiplied by 9. A lower ERA means that a pitcher gives up less runs. WHIP (Walks plus hits per inning pitched) is calculated how it sounds. The total amount of Walks plus the hits a pitcher gives up, all divided by the total innings pitched. The lower the WHIP, the less base runners a pitcher allows. OBP (On Base Percentage) is similar to WHIP, but with the addition of Hit by Pitch. It measures how often batters reaches base against a pitcher. For pitchers, A lower OBP is better. The next two are K% (Strikeout Percentage) and BB% (Walk Percentage). Each measures the percent of at-bats that end in Strikeouts or Walks. A lower BB% is better and a higher K% is better.

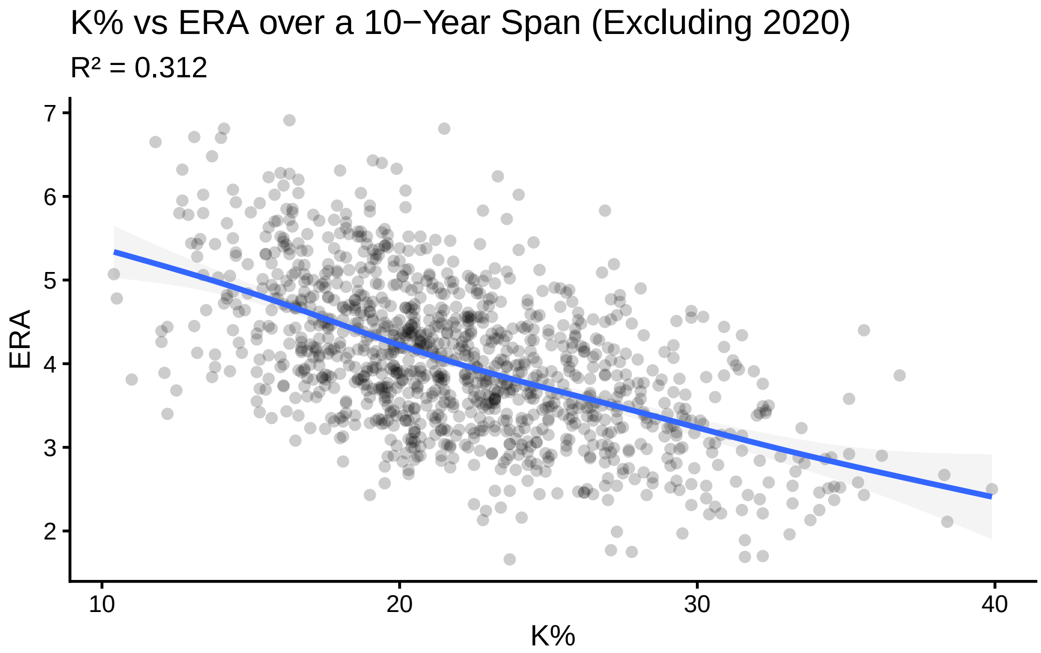
We are also excluding the 2020 season, due to the effects of the Pandemic. This led to a shortened 60 game season instead of the normal 162 games, which led to some statistics being not truly being representative of the real data.

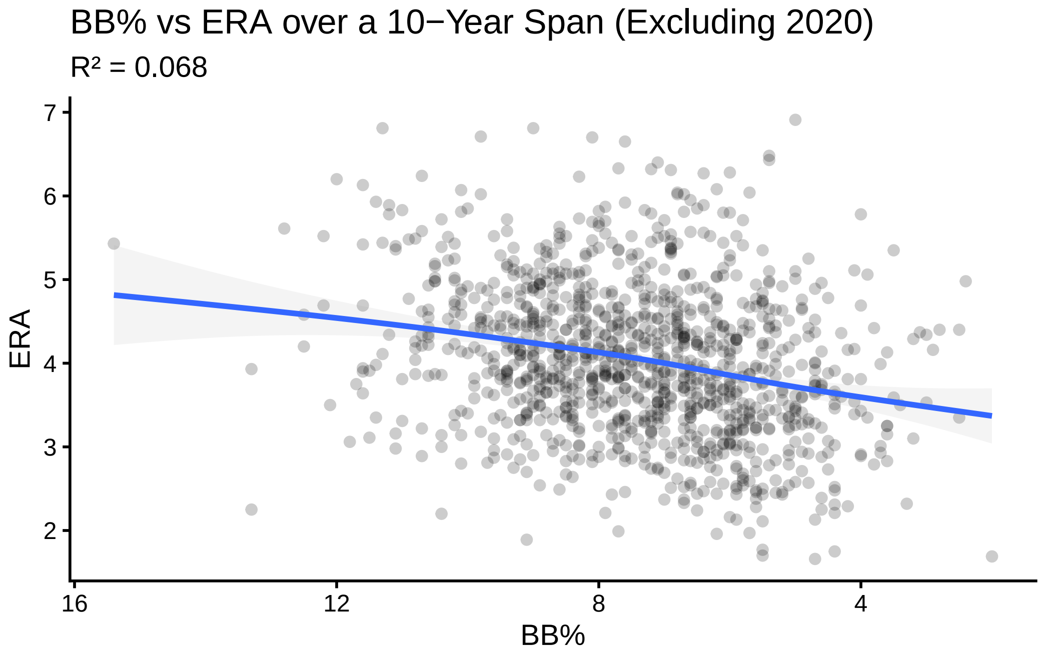
3 Data Wrangling

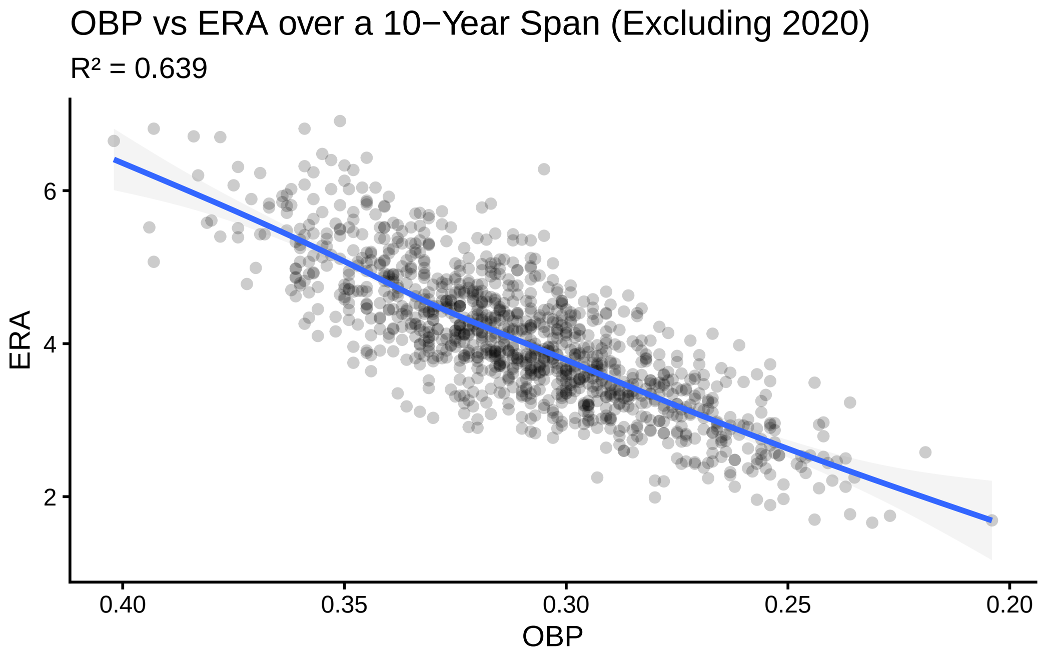
To wrangle the data, we first had to import the data and load packages. Second, we split the names column into two separate columns, one for first name and one for last name. Then, we merged them back into one column to transform the names from “Colon, Bartolo” into “Bartolo Colon”. Then, we renamed the columns, calculated WHIP, and then removed excess columns. Finally, we rounded our statistics and used the kable function to create a table. This is the sample data table from our cleaned data with 10 out of 1112 rows..

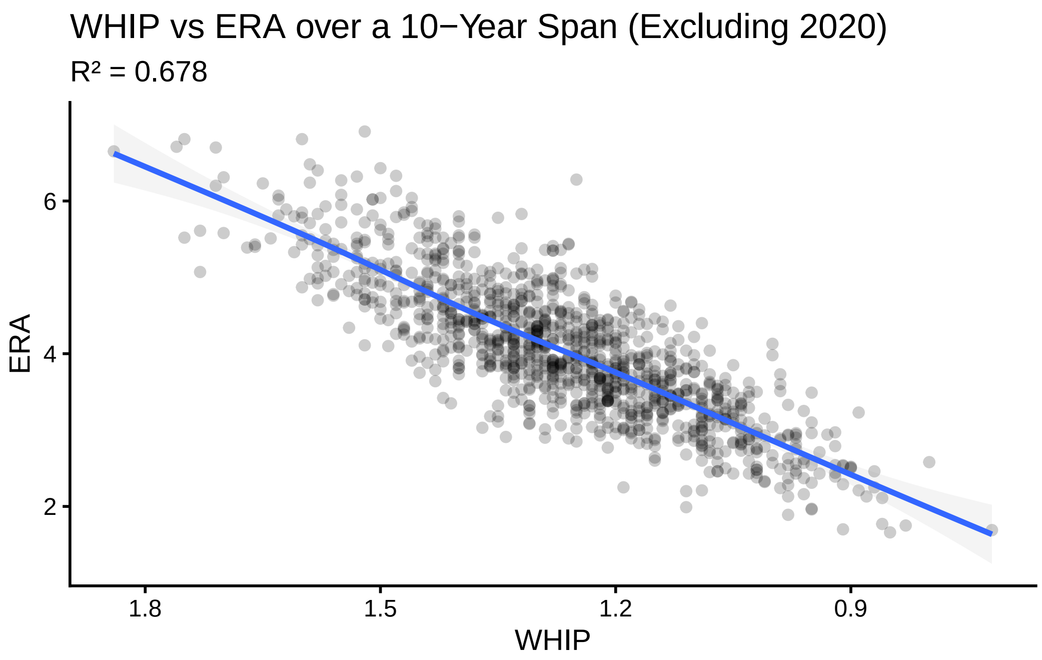
Name	Strikeout_Percentage	Walk_Percentage	On_Base_Percentage	WHIP	ERA
Bartolo Colon	16.7	2.9	0.304	1.24	4.16
A.J. Burnett	20.5	7.0	0.336	1.36	3.18
Tim Hudson	12.2	7.0	0.340	1.39	4.44
Mark Buehrle	11.0	4.0	0.311	1.25	3.81
CC Sabathia	18.9	6.9	0.338	1.42	4.73
Ryan Vogelsong	18.1	9.7	0.338	1.47	4.67
R.A. Dickey	14.3	6.9	0.303	1.20	3.91
Kyle Lohse	16.2	6.5	0.345	1.47	5.85
John Lackey	19.5	5.9	0.303	1.21	2.77
Jorge De La Rosa	21.1	10.3	0.325	1.36	4.17

4 Data Visualizations and Results




BB% vs ERA over a 10−Year Span (Excluding 2020)
R² = 0.068
ERA
7
6
5
4
3
2
16
12
8
4
BB%


OBP vs ERA over a 10−Year Span (Excluding 2020)
R² = 0.639
ERA
6
4
2
0.40
0.35
0.30
0.25
0.20
OBP



5 Code Appendix:

```
#| label: tbl-frequency-pitching-data
#| tbl-cap: "Frequency Table showing each pitcher and their stats for each season between 2015
#| tbl-alt: "Frequency Table showing each pitcher and their stats for each season between 2015

# Step 1: Load Packages
library(tidyverse)
library(rvest)
library(ggplot2)
library(dplyr)
library(readr)
library(janitor)
library(kableExtra)


# Step 2: Import Raw Data
Raw_Statistics <- read.csv("C:/College/GitHub Repos/Sec-1-FP-JacobKoewler-LiamBengston-YeheCui/

# Step 3: Split Name into First and Last Name
Raw_Statistics_Separated <- Raw_Statistics %>%
  separate_wider_delim( # Separate one name column into two (First and Last)
    cols = `last_name..first_name`, # original column
    delim = ",", # delimiter
    names = c("Last_Name", "First_Name"), # new column names
  )

# Step 4: Combine Names
Raw_Statistics_Names <- Raw_Statistics_Separated %>% # Turn names from "Colon, Bartolo" into "
  unite(
    col = "Name", # New Column name
    First_Name, # First name column
    Last_Name, # Last name column
    sep = " ") # Separate names with space

# Step 4: Rename Columns
Raw_Statistics_Renamed <- Raw_Statistics_Names %>% # rename columns for ease of use for those
  rename("Games_Pitched" = p_game,
         "Innings_Pitched" = p_formatted_ip,
         "Strikeout_Percentage" = k_percent,
         "Walk_Percentage" = bb_percent,
         "ERA" = p_era,
         "On_Base_Percentage" = on_base_percent)

# Step 5: Calculate WHIP
Raw_Statistics_WHIP <- Raw_Statistics_Renamed %>%
  mutate( # Calculate WHIP using mutate to create a new column
    "WHIP" = (walk + hit)/(Innings_Pitched) # use PEMDAS to ensure proper calculation
    )

# Step 6: Remove Extra Columns
Cleaned_ERA_Statistics <- Raw_Statistics_WHIP %>% # remove extra columns so only useful data r
  select(Name, Strikeout_Percentage, Walk_Percentage, On_Base_Percentage, WHIP, ERA)

# Step 7: Round WHIP
Cleaned_ERA_Statistics$WHIP <- round(Cleaned_ERA_Statistics$WHIP, digits = 2)

# Step 8: Slice the Top 10 Rows
Frequency_Table <- Cleaned_ERA_Statistics %>%
   slice_head(n = 10)

# Step 9: Create Frequency Tables
Frequency_Table %>%
  kable() %>%
  kableExtra::kable_classic()


#| label: code-data-wrangling

# Step 1: Load Packages
library(tidyverse)
library(rvest)
```

```r
library(ggplot2)
library(dplyr)
library(readr)

# Step 2: Import Raw Data
Raw_Statistics <- read.csv("C:/College/GitHub Repos/Sec-1-FP-JacobKoewler-LiamBengston-YeheCui/

# Step 3: Split Name into First and Last Name
Raw_Statistics_Separated <- Raw_Statistics %>%
  separate_wider_delim( # Separate one name column into two (First and Last)
    cols = `last_name..first_name`, # original column
    delim = ",", # delimiter
    names = c("Last_Name", "First_Name"), # new column names
  )

# Step 4: Combine Names
Raw_Statistics_Names <- Raw_Statistics_Separated %>% # Turn names from "Colon, Bartolo" into "
  unite(
    col = "Name", # New Column name
    First_Name, # First name column
    Last_Name, # Last name column
    sep = " ") # Separate names with space

# Step 4: Rename Columns
Raw_Statistics_Renamed <- Raw_Statistics_Names %>% # rename columns for ease of use for those
  rename("Games_Pitched" = p_game,
         "Innings_Pitched" = p_formatted_ip,
         "Strikeout_Percentage" = k_percent,
         "Walk_Percentage" = bb_percent,
         "ERA" = p_era,
         "On_Base_Percentage" = on_base_percent)


# Step 5: Calculate WHIP
Raw_Statistics_WHIP <- Raw_Statistics_Renamed %>%
  mutate( # Calculate WHIP using mutate to create a new column
    "WHIP" = (walk + hit)/(Innings_Pitched) # use PEMDAS to ensure proper calculation
    )

# Step 6: Remove Extra Columns
Cleaned_ERA_Statistics <- Raw_Statistics_WHIP %>% # remove extra columns so only useful data r
  select(Name, Strikeout_Percentage, Walk_Percentage, On_Base_Percentage, WHIP, ERA)

# Step 7: Round WHIP
Cleaned_ERA_Statistics$WHIP <- round(Cleaned_ERA_Statistics$WHIP, digits = 2)
```

```r
#| label: fig-nameLineGraph
#| fig-cap: "Attribute correlation with ERA"
#| fig-pos: "H"
#| fig-alt: "Correlation between K% and ERA"
#| echo: false

# Step 7:
r2 <- summary(
  lm(ERA ~ Strikeout_Percentage,
  data = Cleaned_ERA_Statistics))$r.squared

ggplot(
  data = Cleaned_ERA_Statistics,
  mapping = aes(
    x = Strikeout_Percentage,
    y = ERA,
  )
) +
  geom_point(alpha = 0.2) + # increased line size a little to improve readability
  labs(
    title = "K% vs ERA over a 10-Year Span (Excluding 2020)",
    subtitle = paste0("R² = ", round(r2, digits = 3)),
    x = "K%",
    y = "ERA"
  ) +
  geom_smooth(alpha = 0.1) + # a line of best fit to more clearly see the correlation
  theme_classic() # a visual polish adjustment

#| label: fig-nameLineGraph
#| fig-cap: "Attribute correlation with ERA"
#| fig-pos: "H"
#| fig-alt: "Correlation between BB% and ERA"
#| echo: false

# Step 8
r2 <- summary(
  lm(ERA ~ Walk_Percentage,
  data = Cleaned_ERA_Statistics))$r.squared

ggplot(
  data = Cleaned_ERA_Statistics,
  mapping = aes(
    x = Walk_Percentage,
    y = ERA,
  )
) +
  geom_point(alpha = 0.2) + # increased line size a little to improve readability
  labs(
    title = "BB% vs ERA over a 10-Year Span (Excluding 2020)",
    subtitle = paste0("R² = ", round(r2, digits = 3)),
    x = "BB%",
    y = "ERA"
  ) +
  geom_smooth(alpha = 0.1) + # a line of best fit to more clearly see the correlation
  theme_classic() + # a visual polish adjustment
  scale_x_reverse()

#| label: fig-nameLineGraph
#| fig-cap: "Attribute correlation with ERA"
#| fig-pos: "H"
#| fig-alt: "Correlation between OBP and ERA"
#| echo: false

# Step 9
r2 <- summary(
  lm(ERA ~ On_Base_Percentage,
  data = Cleaned_ERA_Statistics))$r.squared

ggplot(
  data = Cleaned_ERA_Statistics,
  mapping = aes(
    x = On_Base_Percentage,
    y = ERA,
  )
) +
  geom_point(alpha = 0.2) + # increased line size a little to improve readability
  labs(
    title = "OBP vs ERA over a 10-Year Span (Excluding 2020)",
    subtitle = paste0("R² = ", round(r2, digits = 3)),
    x = "OBP",
    y = "ERA"
  ) +
  geom_smooth(alpha = 0.1) + # a line of best fit to more clearly see the correlation
  theme_classic() + # a visual polish adjustment
  scale_x_reverse()

#| label: fig-nameLineGraph
#| fig-cap: "Attribute correlation with ERA"
#| fig-pos: "H"
#| fig-alt: "Correlation between WHIP and ERA"
#| echo: false

# Step 10
r2 <- summary(
  lm(ERA ~ WHIP,
  data = Cleaned_ERA_Statistics))$r.squared


ggplot(
  data = Cleaned_ERA_Statistics,
  mapping = aes(
    x = WHIP,
    y = ERA,
  )
) +
  geom_point(alpha = 0.2) + # increased line size a little to improve readability
  labs(
    title = "WHIP vs ERA over a 10-Year Span (Excluding 2020)",
    subtitle = paste0("R² = ", round(r2, digits = 3)),
    x = "WHIP",
    y = "ERA"
  ) +
  geom_smooth(alpha = 0.1) + # a line of best fit to more clearly see the correlation
  theme_classic() + # a visual polish adjustment
  scale_x_reverse()
```