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.

1.1 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.

2 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..

Table 1: Frequency Table showing each pitcher and their stats for each season between 2015 and 2025 (excluding 2020)

Name	K_rate	BB_rate	OBP	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

3 Data Visualizations and Results

3.1 Long descriptions for Figure 1

3.1.1 Long Description for K%

The image found in Figure 1a is a scatter plot illustrating the relationship between strikeout percentage (K%) and earned run average (ERA) over a ten-year span, excluding 2020. The horizontal axis represents K%, ranging from 10 to 40, while the vertical axis represents ERA, ranging from 2 to 7. Numerous small, gray circles are scattered across the chart, indicating individual data points. A blue line with a downward slope runs through the center of the plot, representing a trend line, with a gray shaded area around it indicating the confidence interval.

3.1.2 Long Description for BB%

The image found in Figure 1b is a scatter plot depicting the relationship between BB% (Base on Balls Percentage) and ERA (Earned Run Average) over a 10-year span, excluding the year 2020. The x-axis represents BB%, ranging from 16 to 4, while the y-axis represents ERA, ranging from 2 to 7. The plot contains numerous small, light gray dots scattered throughout, representing individual data points. A blue trend line traverses the plot diagonally from upper left to lower right, indicating a negative correlation between BB% and ERA. The trend line is surrounded by a gray shaded region that illustrates the confidence interval. The background of the plot is white, and the text is black.

3.1.3 Long Description for OBP

The image found in Figure 1c is a scatter plot showing the relationship between OBP (On-base Percentage) and ERA (Earned Run Average) over a ten-year span, excluding the year 2020. The x-axis represents OBP, ranging from 0.20 to 0.40, while the y-axis represents ERA, ranging from 2 to 6. Data points are represented by small gray dots scattered across the plot, showing a downward trend from left to right. A blue trend line runs diagonally from the top left to the bottom right, indicating a negative correlation between OBP and ERA. A shaded gray area around the line illustrates the confidence interval. The title of the plot is “OBP vs ERA over a 10-Year Span (Excluding 2020).”

3.1.4 Long Description for WHIP

The image found in Figure 1d is a scatter plot graph illustrating the relationship between WHIP (Walks plus Hits per Inning Pitched) and ERA (Earned Run Average) over a ten-year span, excluding 2020. The x-axis is labeled “WHIP” and ranges approximately from 0.9 to 1.8. The y-axis is labeled “ERA” and ranges from 2 to 6. A blue trend line, with a shaded confidence interval, slopes downward from left to right, indicating a negative correlation between WHIP and ERA. Numerous small, gray data points are scattered densely around the trend line.

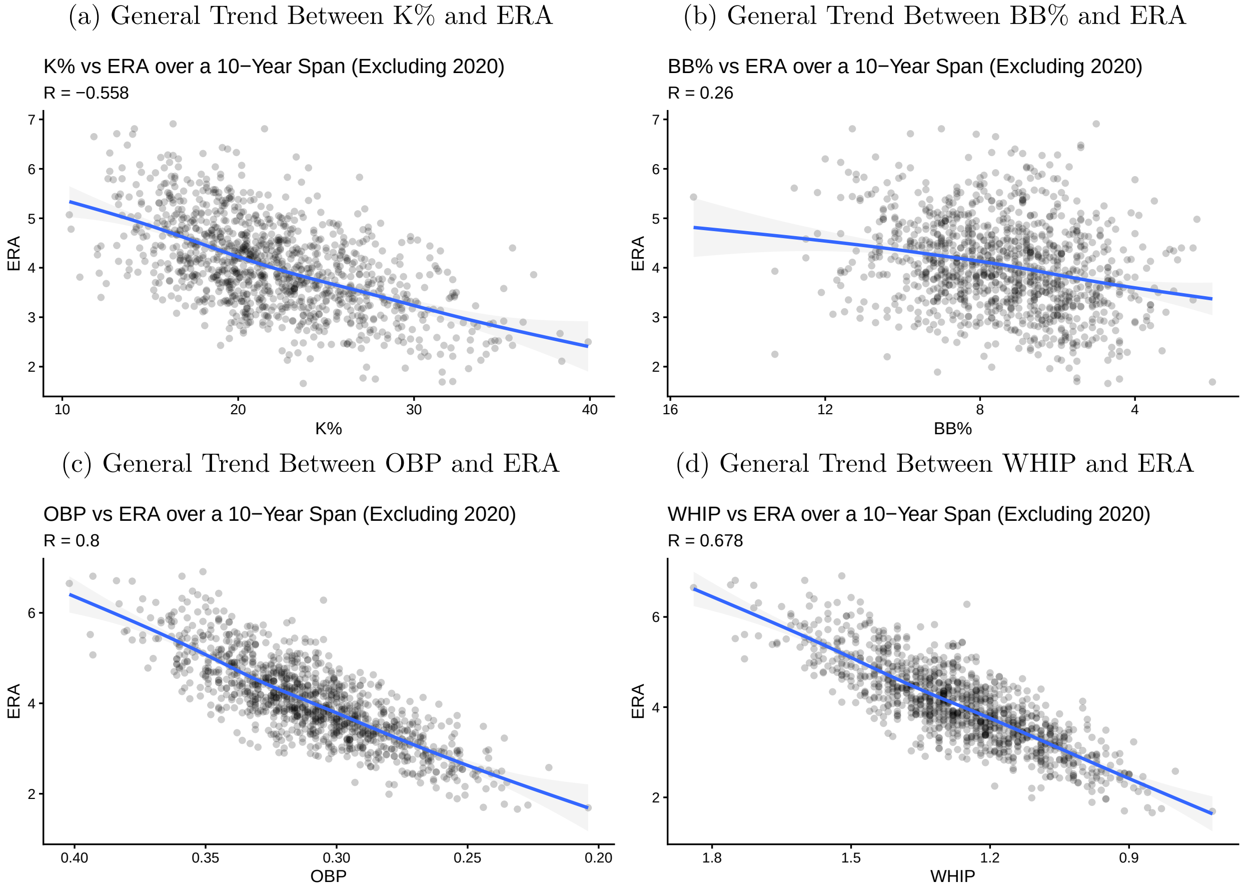
3.2 Long descriptions for Figure 2

The image is a correlation plot presented as a heatmap, depicting the relationships among five variables: K_rate, BB_rate, OBP, WHIP, and ERA. The plot is structured as a grid with rows and columns labeled with these variables. Each cell in the grid contains a numerical value representing the correlation coefficient between the corresponding pair of variables. The background color of each cell indicates the strength and direction of the correlation, using a gradient scale from blue to red. Blue shades represent positive correlations, while red shades denote negative correlations. The plot features a small color bar on the right, displaying a continuous range from -1 to 1, corresponding to the colors used in the plot. The diagonal cells show a perfect correlation of 1.00, displayed in dark blue. Other notable correlations include -0.64 between K_rate and OBP, and 0.99 between OBP and WHIP.

3.3 Results

Of the four statistics that we observed, WHIP and OBP were the most correlated with ERA. The correlation table however shows that WHIP and OBP are nearly identical, implying that they are describing the same thing. Both statistics are a type of measure about how many runners a pitcher lets on base. This means that ERA is heavily influenced by the amount of runners a pitcher lets on base. By whatever measure you chose, be it WHIP or OBP, both show a strong negative correlation with ERA.

Figure 1: Attribute Correlation with ERA



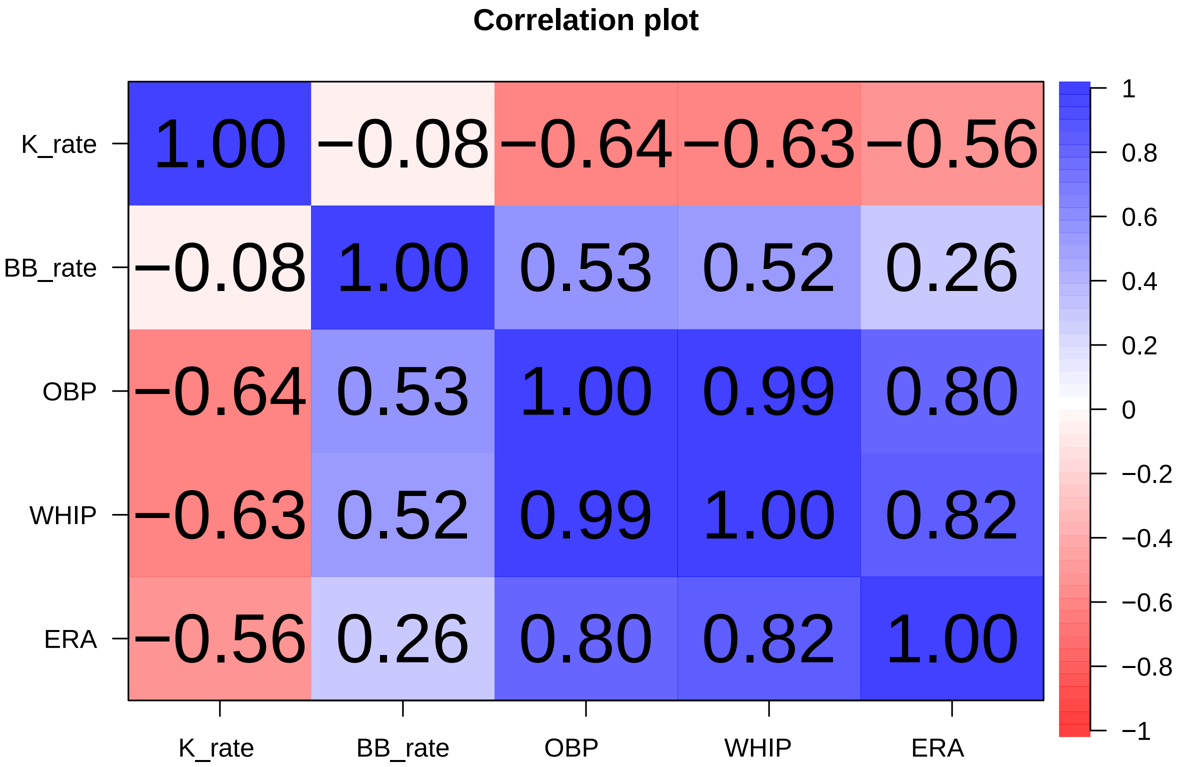
(a) General Trend Between K% and ERA

(b) General Trend Between BB% and ERA

(c) General Trend Between OBP and ERA

(d) General Trend Between WHIP and ERA

Figure 2: Correlation Table of the Attributes



	K_rate	BB_rate	OBP	WHIP	ERA
K_rate	1.00	−0.08	−0.64	−0.63	−0.56
BB_rate	−0.08	1.00	0.53	0.52	0.26
OBP	−0.64	0.53	1.00	0.99	0.80
WHIP	−0.63	0.52	0.99	1.00	0.82
ERA	−0.56	0.26	0.80	0.82	1.00

4 Code Appendix

```
# Step 1: Load Packages
library(tidyverse)
library(rvest)
library(googlesheets4)
library(janitor)
library(knitr)
library(kableExtra)
library(psych)


# Step 2: Import Raw Data

gs4_deauth()
Raw_Statistics <- read_sheet(
  ss = "https://docs.google.com/spreadsheets/d/1aQGzSOEpzus4ZRMdT7VMvGddumRQdRsXYek41cmyLLU/edi
  n_max = 1115 # read all lines
)

# 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 so it's 'first last'
Raw_Statistics_Names <- Raw_Statistics_Separated %>%
  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 for ease of use later on
Raw_Statistics_Renamed <- Raw_Statistics_Names %>%
  rename("GP" = p_game,
         "IP" = p_formatted_ip,
         "K_rate" = k_percent,
         "BB_rate" = bb_percent,
         "ERA" = p_era,
         "OBP" = on_base_percent)
```

```r
# Step 5: Calculate WHIP
Raw_Statistics_WHIP <- Raw_Statistics_Renamed %>%
  mutate( # Calculate WHIP using mutate to create a new column
    "WHIP" = (walk + hit)/(IP)
    )

# Step 6: Remove Extra Columns
Cleaned_ERA_Statistics <- Raw_Statistics_WHIP %>%
  select(Name, K_rate, BB_rate, OBP, 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()

# Step 10 K% Plot:
x <- Cleaned_ERA_Statistics$ERA
y <- Cleaned_ERA_Statistics$K_rate
r <- cor( x, y, method = "pearson")

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

# Step 11 BB% Plot
x <- Cleaned_ERA_Statistics$ERA
y <- Cleaned_ERA_Statistics$BB_rate
```

```r
r <- cor( x, y, method = "pearson")

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

# Step 12 OBP Plot

x <- Cleaned_ERA_Statistics$ERA
y <- Cleaned_ERA_Statistics$OBP
r <- cor( x, y, method = "pearson")

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

# Step 13 WHIP Plot

r <- 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 to improve readability
  labs(
    title = "WHIP vs ERA over a 10-Year Span (Excluding 2020)",
    subtitle = paste0("R = ", round(r, digits = 3)),
    x = "WHIP",
    y = "ERA"
  ) +
  geom_smooth(alpha = 0.1) + # a line of best fit
  theme_classic() + # a visual polish adjustment
  scale_x_reverse()

# Step 14 correlation table to see the relation between statistics

cors <- cor(Cleaned_ERA_Statistics[, -1])
corPlot(cors)
```