Work-In-Progress Presentation

Jacob, Liam, Yehe

Introduction

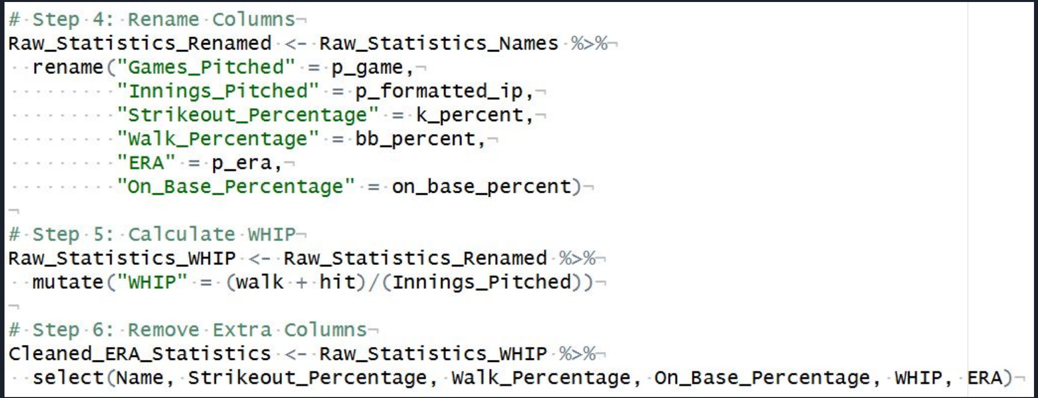
- The data we collected is from the website Baseball Savant.
- We are looking at 4 different variables to see which is the best indicator for Earned Run Average (ERA).
- Our Plan is create different data visualizations to see which is the best indicator of ERA.

Context

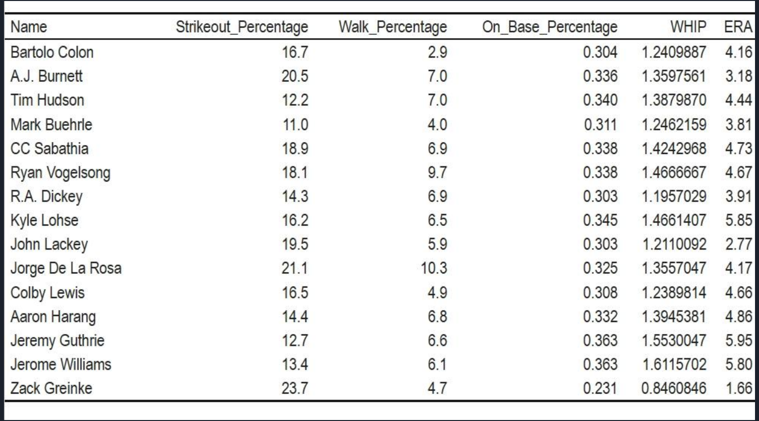
- ERA is calculated by dividing the amount of runs that the pitcher allowed by the amount of innings pitched, all multiplied by 9. A lower ERA means that a pitcher gives up less runs.
- WHIP (Walks plus hits per inning pitched) is calculated how it sounds. The total amount of Walks plus the hits a pitcher 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.

Data Wrangling and Table

- This is an example of our data wrangling, where our main focus was to transform the data into useable form, and we also need to calculate WHIP. This is the resulting table of the cleaned data.



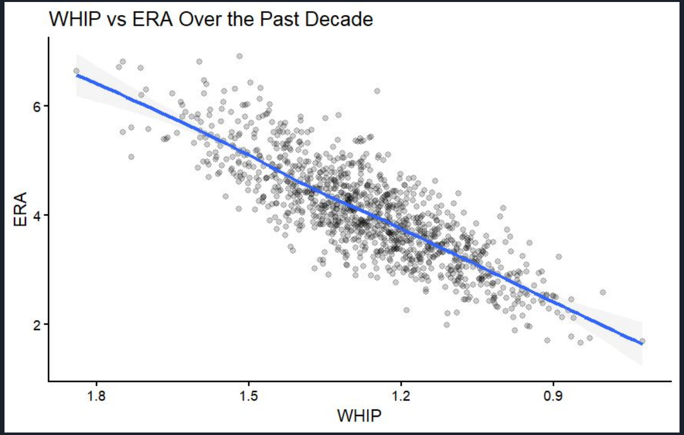
```
# Step 4: Rename Columns
Raw_Statistics_Renamed <- Raw_Statistics_Names %>%
  rename("Games_Pitched" = p_game,
         "Innings_Pitched" = p_formatted_ip,
         "Strikeout_Percentage" = k_percent,
         "Walk_Percentage" = bb_percent,
         "ERA" = p_era,
         "On_Base_Percentage" =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

# Step 5: Calculate WHIP
Raw_Statistics_WHIP <- Raw_Statistics_Renamed %>%
  mutate("WHIP" = (walk + hit)/(Innings_Pitched))

# Step 6: Remove Extra Columns
Cleaned_ERA_Statistics <- Raw_Statistics_WHIP %>%
  select(Name, Strikeout_Percentage, Walk_Percentage, On_Base_Percentage, WHIP, ERA)
```

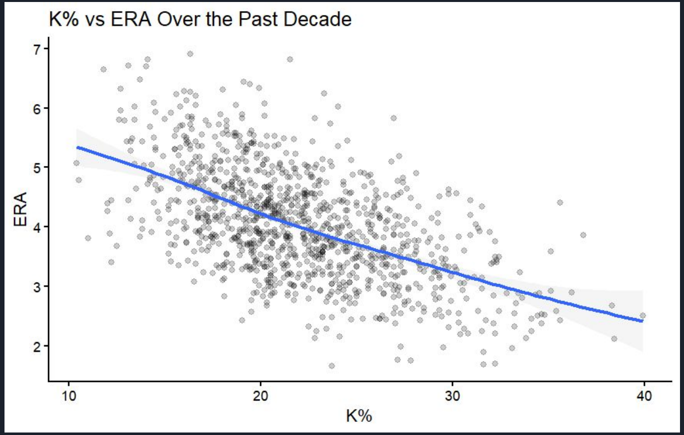


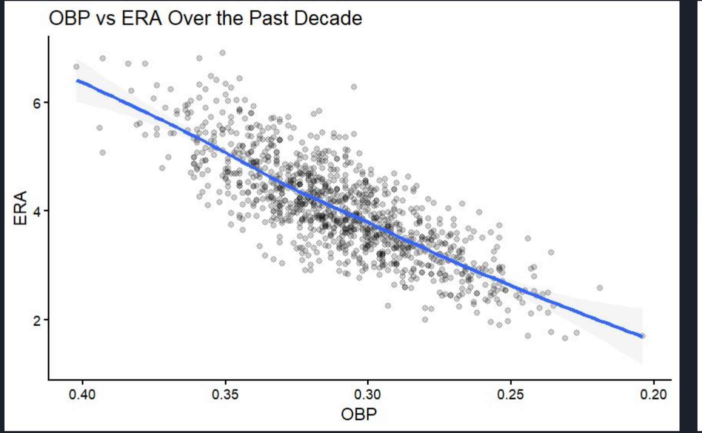
Name	Strikeout_Percentage	Walk_Percentage	On_Base_Percentage	WHIP	ERA
Bartolo Colon	16.7	2.9	0.304	1.2409887	4.16
A.J. Burnett	20.5	7.0	0.336	1.3597561	3.18
Tim Hudson	12.2	7.0	0.340	1.3879870	4.44
Mark Buehrle	11.0	4.0	0.311	1.2462159	3.81
CC Sabathia	18.9	6.9	0.338	1.4242968	4.73
Ryan Vogelsong	18.1	9.7	0.338	1.4666667	4.67
R.A. Dickey	14.3	6.9	0.303	1.1957029	3.91
Kyle Lohse	16.2	6.5	0.345	1.4661407	5.85
John Lackey	19.5	5.9	0.303	1.2110092	2.77
Jorge De La Rosa	21.1	10.3	0.325	1.3557047	4.17
Colby Lewis	16.5	4.9	0.308	1.2389814	4.66
Aaron Harang	14.4	6.8	0.332	1.3945381	4.86
Jeremy Guthrie	12.7	6.6	0.363	1.5530047	5.95
Jerome Williams	13.4	6.1	0.363	1.6115702	5.80
Zack Greinke	23.7	4.7	0.231	0.8460846	1.66

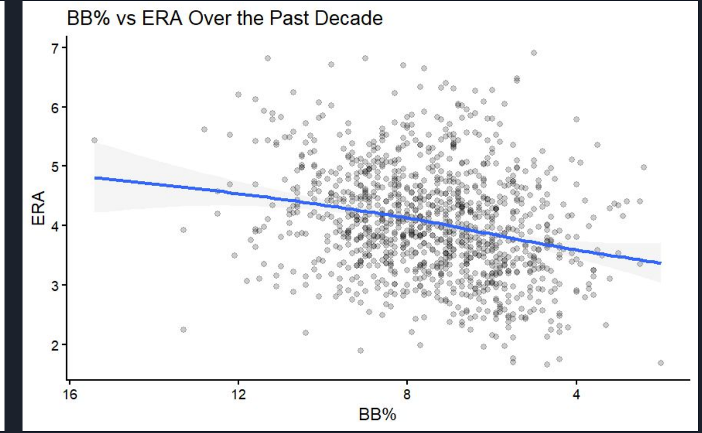
Data Visualizations

- K% is the only graph shown here where the value on the x-axis increases over the length of the graph


WHIP vs ERA Over the Past Decade


K% vs ERA Over the Past Decade


OBP vs ERA Over the Past Decade


BB% vs ERA Over the Past Decade

Challanges and Future Ideas

Challenges:

- We couldn’t find all of the statistics we wanted in Baseball Savant, so we had to calculate some of them manually in R.

Future Ideas

- Calculate the coefficient of determination for each data visualization to see which variable is the best indicator
- Simply the Table by rounding to two decimal places
- In the Quarto doc, combine the four data visualizations into one image (multiple frames)

Thanks!

Are there any Questions?