# Scaling Sports Analytics with R & Google Cloud

Alok Pattani
Data Science Developer Advocate, Google Cloud

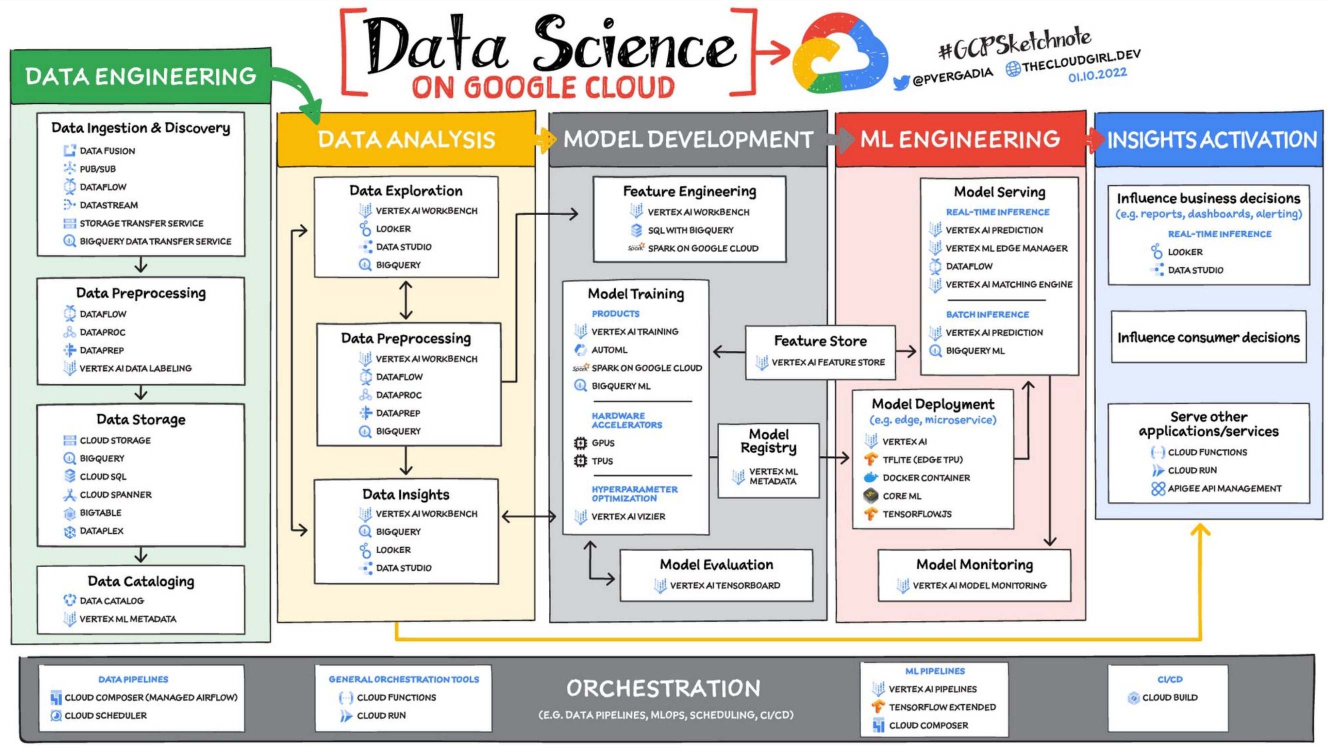
RStudio Sports Analytics Meetup - February 2022

# My Background

- BA/MA in Statistics at Boston University
- ESPN Statistics & Information Group (2006-2016)
  - Sports Analytics Team:  2010-2016
- Google (2016-Present)
  - Data Science Developer Advocate at Google Cloud Since 2019
- Other sports/data science consulting, etc.

# R & BigQuery for Data Science

# Google Cloud for Data Science




Turn data into insights - faster, easier, and at greater scale.

# BigQuery

**Serverless, highly scalable, and cost-effective data warehouse with customers ranging from TB to 100+ PB**

- Cloud-scale enterprise data warehouse
- Standard SQL(ANSI 2011) with DML Support
- Encrypted, durable, highly available

- Multi-cloud analytics
- Real-time insights
- Built-in machine learning
- Insights for everyone

# R and BigQuery

## Tasks I like to do in...

| R | BigQuery |
|---|---|
| • Getting data from packages, websites, APIs, etc.<br>• General data manipulation (cleaning, preprocessing)<br>• Exploratory data analysis<br>• Statistical analysis & modeling<br>• "Ad hoc" data visualization | • Data warehousing<br>• Create reusable "intermediate" data manipulation pieces (e.g. views, stored procedures)<br>• Modeling on very big data<br>• Storing analysis results for outputs (e.g. interactive dashboards) |

Use bigrquery package to interface between them

# Example Analysis with NCAA Basketball Data, R, & BigQuery

# NCAA Basketball Analysis Goal

**OBJECTIVE:** Create a rating system for NCAA basketball players.

**CRITERIA:**

- Use multiple player box score stats (e.g. points, assists, …)
- Represent player's contribution to winning
- Apply to men's and women's Division I
- Apply to current season and past few
- Adjust for schedule (level of competition)

# Why Is This Important or Useful?

“All-in-one” college player ratings could be used by…

- **Media/Fans**: “best player” debates, awards, general research
- **College Teams**: roster management, opponent scouting
- **Content Companies**: potential automated signals
- **Pro Teams**: evaluating draft prospects

# How Do We Do This?

**High-Level Overview**:

1. Pull public NCAA basketball team and player data from open-source R packages, upload to BigQuery.
2. Use established basketball analytics theory for initial player calculations, apply in BigQuery (SQL).
3. Read processed data back into R to implement modeling-based schedule adjustment and run final player calculations.
4. Push final results back into BigQuery.
5. Argue about player ratings!

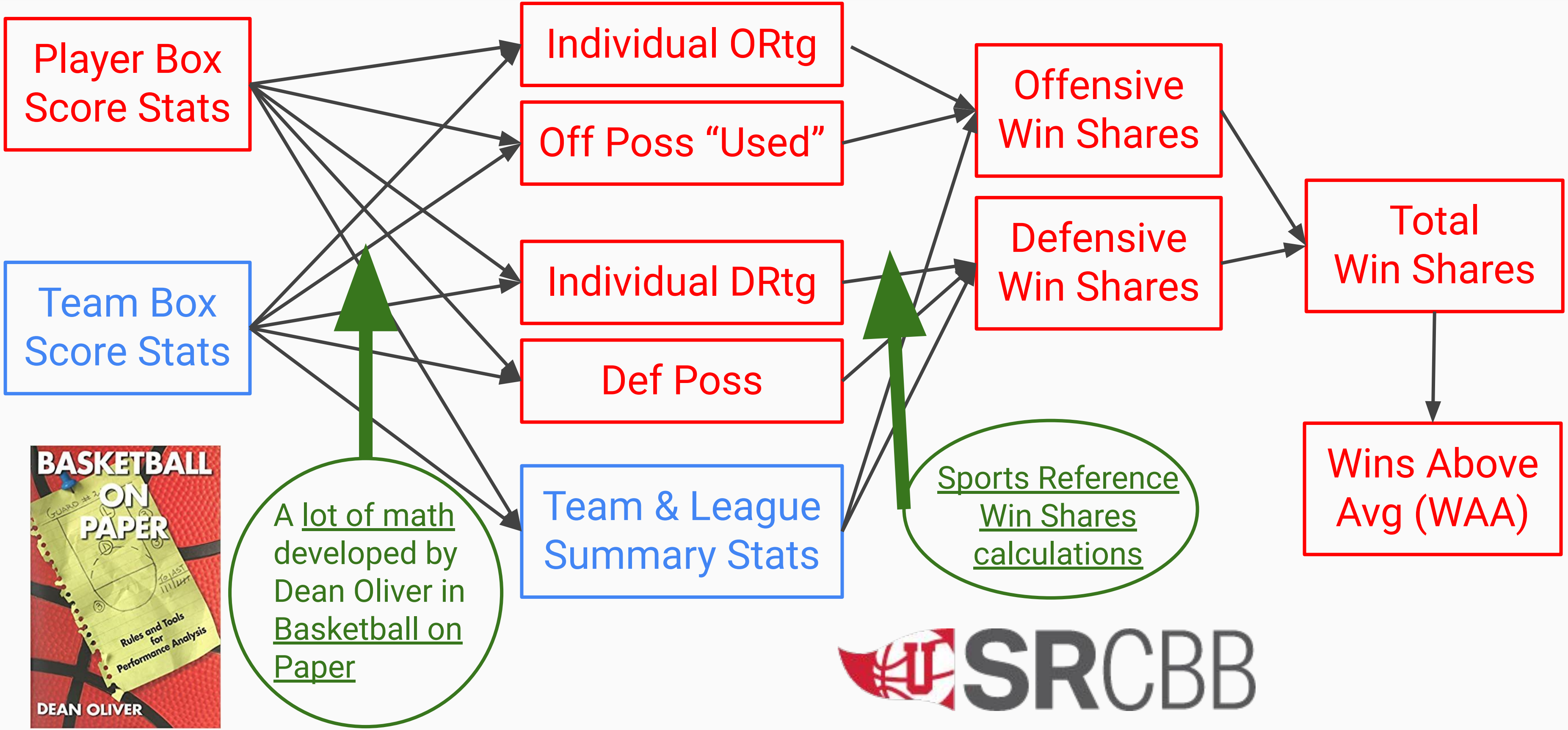
# Data Analysis Walkthrough (Part 1)

# Gathering NCAA Basketball Data

```r
library(tidyverse)
library(hoopR)
library(wehoop)

MBB_START_YEAR <- 2003
MBB_END_YEAR <- 2022

MBBTeams <- espn_mbb_teams()
MBBSchedule <- load_mbb_schedule(seasons = MBB_START_YEAR:MBB_END_YEAR)
MBBTeamBox <- load_mbb_team_box(seasons = MBB_START_YEAR:MBB_END_YEAR)
MBBPlayerBox <- load_mbb_player_box(seasons = MBB_START_YEAR:MBB_END_YEAR)

WBB_START_YEAR <- 2006
WBB_END_YEAR <- 2022

WBBTeams <- espn_wbb_teams()
WBBSchedule <- load_wbb_schedule(seasons = WBB_START_YEAR:WBB_END_YEAR)
WBBTeamBox <- load_wbb_team_box(seasons = WBB_START_YEAR:WBB_END_YEAR)
WBBPlayerBox <- load_wbb_player_box(seasons = WBB_START_YEAR:WBB_END_YEAR)
```

**MAJOR Thanks to Saiem Gilani!**

# Basketball Player Ratings Framework



# Let’s Look at RStudio and BigQuery…

```r
#### LOAD IN RELEVANT LIBRARIES AND SET SCRIPT-LEVEL OPTIONS ####
library(tidyverse)
library(lubridate)
library(janitor)

# Packages for obtaining men's and women's basketball data
library(hoopR)
library(wehoop)

# Interfacing with BigQuery
library(bigrquery)

options(tibble.width = Inf)


#### SET UP BIGQUERY PIECES FOR THIS SCRIPT ####
# Read in variables for Cloud access saved as system environment variables
CLOUD_AUTH_EMAIL <- Sys.getenv("DEFAULT_AUTH_EMAIL")
BIGQUERY_PROJECT <- Sys.getenv("DEFAULT_GOOGLE_CLOUD_PROJECT")

# Authorize using email
bq_auth(email = CLOUD_AUTH_EMAIL)

# Dataset within BigQuery where this data will go
BIGQUERY_DATASET <- 'ncaa_basketball'
```

```sql
PlayerTmGameStatsCalcs4 AS
(
  SELECT
    *,
    (scr_poss + fgx_poss + ftx_poss + tov) AS off_poss,

    SAFE_DIVIDE((scr_poss + fgx_poss + ftx_poss + tov), est_poss_on_floor)
      AS off_poss_pct,

    SAFE_DIVIDE(scr_poss, (scr_poss + fgx_poss + ftx_poss + tov))
      AS floor_pct,

    SAFE_DIVIDE(pprod, (scr_poss + fgx_poss + ftx_poss + tov)) * 100
      AS ortg,

    /* Start with team def efficiency, unless no defensive possessions */
    IF(def_poss = 0, NULL,
      tm_def_eff +
      /* Take tm_def_eff as is (no adj) if player-specific adjustment is null */
      IFNULL(0.2 * (100 * opp_pts_per_scr_poss * (1 - stop_pct) - tm_def_eff), 0)
      ) AS drtg

  FROM
    PlayerTmGameStatsCalcs3
)
```

# "Final" Results - Top 5 Players

## This Season

### Women

| season | athlete | team | waa |
|---|---|---|---|
| 2021-22 | Aliyah Boston | South Carolina | 5.65 |
| 2021-22 | Ayoka Lee | Kansas State | 5.43 |
| 2021-22 | Caitlin Clark | Iowa | 5.07 |
| 2021-22 | Shaylee Gonzales | BYU | 4.53 |
| 2021-22 | Katelyn Young | Murray State | 4.42 |

### Men

| season | athlete | team | waa |
|---|---|---|---|
| 2021-22 | Malachi Smith | Chattanooga | 4.15 |
| 2021-22 | Oscar Tshiebwe | Kentucky | 3.69 |
| 2021-22 | Keegan Murray | Iowa | 3.4 |
| 2021-22 | David Roddy | Colorado State | 3.14 |
| 2021-22 | Justin Bean | Utah State | 3.11 |

## Since 2014-15

### Women

| season | athlete | team | waa |
|---|---|---|---|
| 2016-17 | Kelsey Plum | Washington | 9.87 |
| 2015-16 | Breanna Stewart | UConn | 9.0 |
| 2019-20 | Sabrina Ionescu | Oregon | 8.71 |
| 2016-17 | Napheesa Collier | UConn | 8.45 |
| 2018-19 | Napheesa Collier | UConn | 8.29 |

### Men

| season | athlete | team | waa |
|---|---|---|---|
| 2014-15 | Frank Kaminsky | Wisconsin | 5.9 |
| 2018-19 | Matt Rafferty | Furman | 5.85 |
| 2015-16 | Thomas Walkup | Stephen F. Austin | 5.57 |
| 2014-15 | Delon Wright | Utah | 5.55 |
| 2017-18 | Jock Landale | Saint Mary's | 5.51 |

# Data Analysis Walkthrough (Part 2)

# Schedule Adjustment Theory

- Teams (and hence players) face varying levels of competition across 350+ Division I teams, 32 conferences
  - Also: home-court advantage
- Team/player stats could be product of who they play (& where)
  - Loose model representation:

```
game_stat ~ intercept + team_effect + opp_effect + home_adv + (error)
```

- glmnet library in R can fit ridge regression of this type
- Resulting regression coefficients provide team and home-court estimates for offensive & defensive efficiency (more details)

# Getting Schedule-Adjusted Player Stats

- Use adjusted team offensive and defensive efficiency as measure of opponent strength faced by players

- Adjust each player's game-level ORtg & DRtg based on home-court and opponent strength on *opposite side of the ball* (i.e. adjust ORtg for opponent def eff, vice versa):

```
player_adj_ortg = player_raw_ortg + home_adjustment + opp_def_adjustment
```

```
player_adj_drtg = player_raw_drtg + home_adjustment + opp_off_adjustment
```

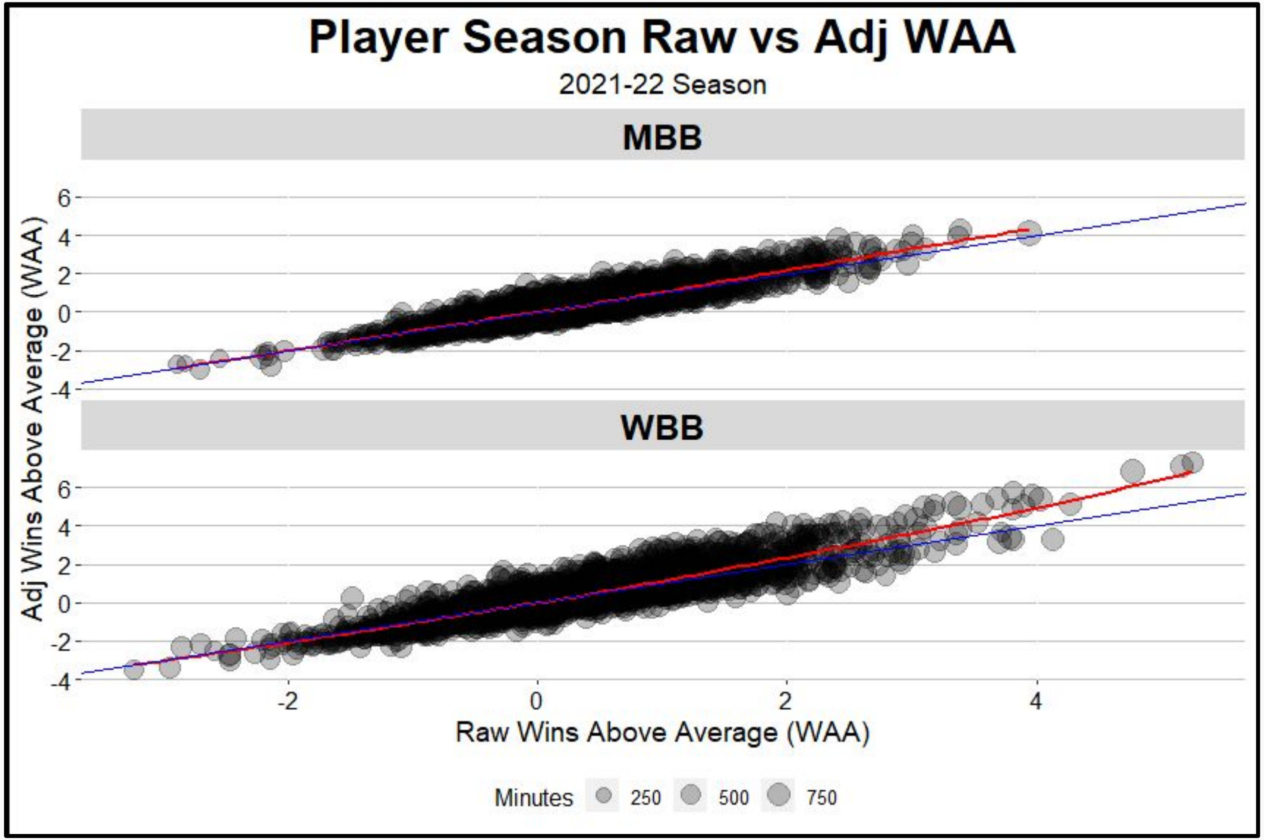
- Aggregate (adjusted) ratings and possessions to season level, follow prior procedure to get (adjusted) win shares, WAA, etc.

# Back into RStudio and BigQuery…

```r
GetTeamAdjRegressionResults <- function(tm_game_info_and_stat,
  regularization_regression_lambdas = 10 ^ seq(-5, 5, by = 0.1))
{
  # Get overall (weighted) average stat value, for use later
  ovr_avg_stat_value <- with(tm_game_info_and_stat,
    weighted.mean(stat_value, wt_value, na.rm = TRUE))

  # Get team-level total weights, for use later
  tm_total_wt_values <- tm_game_info_and_stat %>%
    group_by(sport_season_adj_tm_id) %>%
    summarize(.groups = "drop",
      tot_wt_value = sum(ifelse(!is.na(stat_value), wt_value, NA),
        na.rm = TRUE)
      )

  model_matrix <- with(tm_game_info_and_stat, cbind(tm_hca,
    model.matrix(~ sport_season_adj_tm_id + sport_season_adj_opp_id - 1)))

  model <- cv.glmnet(
    x = model_matrix,
    y = tm_game_info_and_stat$stat_value,
    family = "gaussian",
    weights = tm_game_info_and_stat$wt_value,
    alpha = 0, # This corresponds to ridge regression
    lambda = regularization_regression_lambdas,
    nfolds = 3, # Dropping down from default of 10 to speed up fitting
    intercept = TRUE
    )

  # cat(paste0("Best Lambda: ", model$lambda.min))

  # Get coef from model w/ lambda that made for best ridge regression fit
  model_coef <- predict(model, type = "coefficients", s = model$lambda.min)

  tidy_model_coef <- suppressWarnings(
    tidy(model_coef, return_zeros = TRUE)) %>%
    as_tibble() %>%
    mutate(
      coef_type = case_when(
        str_starts(row, fixed("(Intercept)")) ~ "Intercept",
        (row == "tm_hca") ~ "tm_hca",
        str_starts(row, "sport_season_adj_tm_id") ~ "tm",
        str_starts(row, "sport_season_adj_opp_id") ~ "opp",
        TRUE ~ NA_character_
        )
      ) %>%
    dplyr::select(coef_name = row, coef_type, coef_value = value)
```

```sql
/* GET (RAW) DATA FROM PLAYER SEASON ADVANCED STATS VIEW */
SELECT
  sport,
  `ncaa_basketball.get_season_name_from_year`(season, "END") AS season,
  athlete,
  team,
  ROUND(waa, 2) AS waa

FROM
  `ncaa_basketball.player_season_adv_stats`

WHERE
  sport = 'WBB' AND
  season >= 2014

ORDER BY
  waa DESC

LIMIT 10
;

/* GET ADJUSTED (& RAW) DATA FROM PLAYER SEASON SUMMARY TABLE */
SELECT
  sport,
  season_name AS season,
  athlete,
  team,
  ROUND(adj_waa, 2) AS adj_waa

FROM
  `ncaa_basketball.player_season_summary`

WHERE
  sport = 'WBB' AND
  season > 2000

ORDER BY
  adj_waa DESC

LIMIT 10
;
```

# Does Schedule-Adjusting Matter?



# “More Final” Results - Top 5 Players

## This Season

**Women**

| season | athlete | team | adj_waa |
|---|---|---|---|
| 2021-22 | Aliyah Boston | South Carolina | 7.36 |
| 2021-22 | Ayoka Lee | Kansas State | 7.08 |
| 2021-22 | Caitlin Clark | Iowa | 6.89 |
| 2021-22 | NaLyssa Smith | Baylor | 5.68 |
| 2021-22 | Elizabeth Kitley | Virginia Tech | 5.54 |

**Men**

| season | athlete | team | adj_waa |
|---|---|---|---|
| 2021-22 | Oscar Tshiebwe | Kentucky | 4.28 |
| 2021-22 | Malachi Smith | Chattanooga | 4.05 |
| 2021-22 | Tari Eason | LSU | 4.01 |
| 2021-22 | Keegan Murray | Iowa | 3.95 |
| 2021-22 | Collin Gillespie | Villanova | 3.76 |

## Since 2014-15

**Women**

| season | athlete | team | adj_waa |
|---|---|---|---|
| 2019-20 | Sabrina Ionescu | Oregon | 11.55 |
| 2016-17 | Kelsey Plum | Washington | 10.21 |
| 2015-16 | Breanna Stewart | UConn | 8.96 |
| 2019-20 | Ruthy Hebard | Oregon | 8.74 |
| 2016-17 | Napheesa Collier | UConn | 8.74 |

**Men**

| season | athlete | team | adj_waa |
|---|---|---|---|
| 2014-15 | Frank Kaminsky | Wisconsin | 7.49 |
| 2018-19 | Zion Williamson | Duke | 6.77 |
| 2018-19 | Cassius Winston | Michigan State | 6.69 |
| 2014-15 | Delon Wright | Utah | 6.49 |
| 2016-17 | Josh Hart | Villanova | 6.46 |

# “Loose Ends” and More Info

# Surfacing Large Data Outputs

- Google Sheets
  - Connected Sheets to directly access BigQuery data
- Data Studio: customizable dashboards and reports
- Looker: data experiences, business intelligence platform
- Shiny: interactive web applications in R
  - Publish to RStudio Connect, shinyapps.io

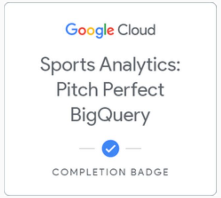
# More on R & Google Cloud

- Some ways to run R/RStudio ON Google Cloud:
  - RStudio Workbench for GCP (Professional)
  - RStudio Server (OSS) on Compute Engine (e.g. Linux installation)
  - Custom Docker container on GCP (e.g. to schedule R scripts)
  - R Jupyter notebook on Vertex AI Workbench (i.e. these instructions)

- Follow Mark Edmondson (@HoloMarkeD) for various R on Google Cloud-related packages, tutorials, etc.

# Other Google Cloud Sports Resources

- NCAA Basketball Analysis (2019-2020):
  - Medium blog with various analysis like this
  - 2020 insights dashboard (men's & women's)
- MLB Partnership Work
  - Automated game notes project
  - Recent talk on innovating MLB fan experience
- 2020 MIT SSAC talk on "Using Google Cloud to Take Sports Analytics to the Next Level"
- Analyzing Soccer Data with BigQuery Lab Series



# Summary

# Tools & Methods Takeaways

- R and RStudio work well with BigQuery and other Google Cloud tools.
- BigQuery is made for (very) large data storage & analytics.
- Regression is powerful, even when not explicitly building a prediction model.
- There are multiple ways to do things, with trade-offs to using different tools for various tasks.
- Data science is hard, but can be fun!

# Sports Analytics Takeaways

- There is a large (and increasing) number of open-source data resources in many sports (e.g. SportsDataverse).
- Having sports knowledge and using established work in the field are key “skills.”
- Adjusting for schedule can be important in evaluating teams and players, particularly in college sports.
- Aliyah Boston (South Carolina women) and Oscar Tshiebwe (Kentucky men) are good at basketball.

# Thank you!

Twitter: @AlokPattani
LinkedIn: Alok Pattani