Data Science – Final Project – 01

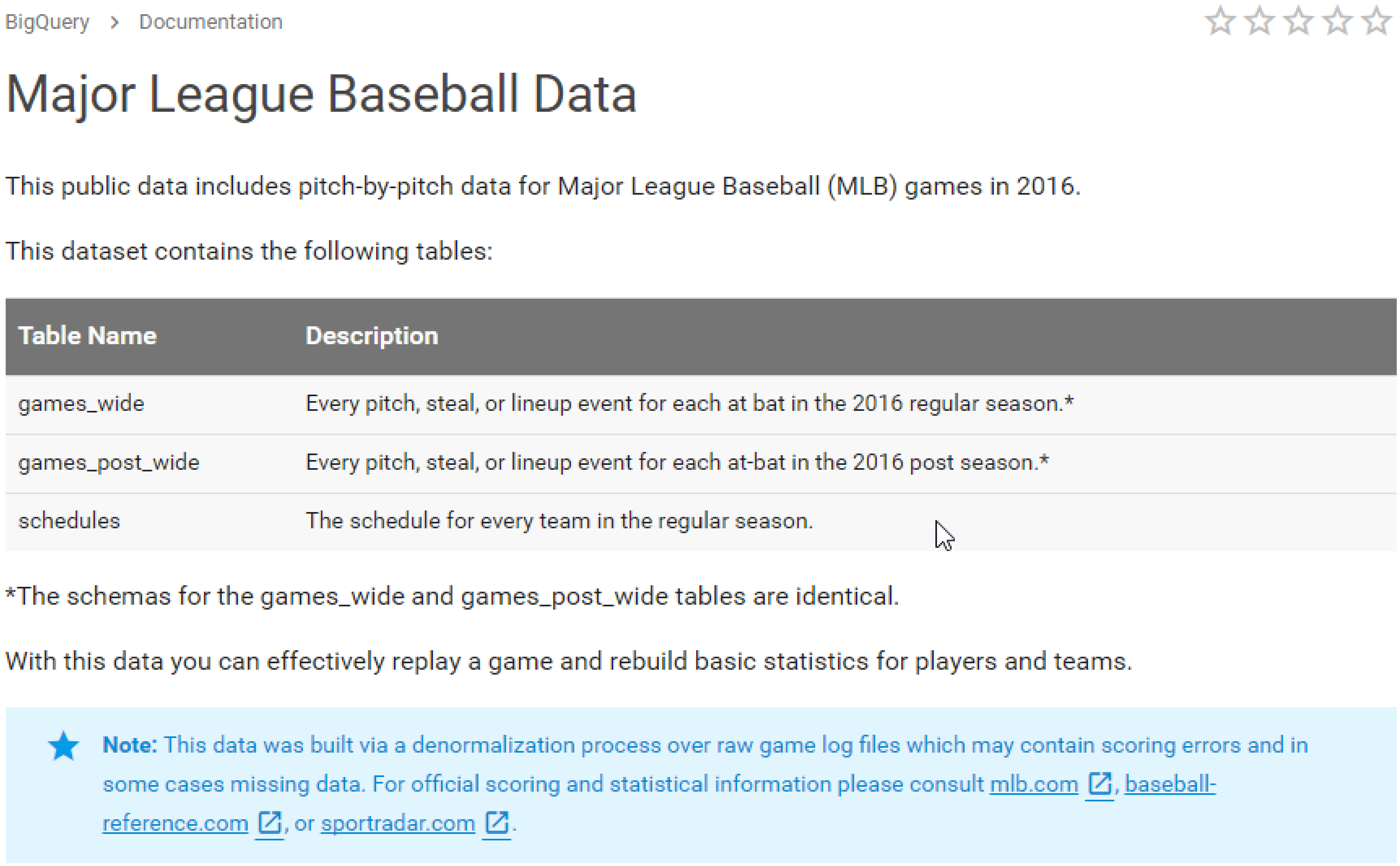
MITSOS TRIANTOPOULOS

05 January 2017

CONFIDENTIAL AND PROPRIETARY
Any use of this material without specific permission is strictly prohibited

WORKING, DRAFT
Last Modified 1/5/2017 11:53 PM GTB Standard Time
Printed 1/5/2017 11:55 PM GTB Standard Time

Idea #1 - Major League Baseball – 2016 – All regular season games



Major League Baseball Data

This public data includes pitch-by-pitch data for Major League Baseball (MLB) games in 2016.

This dataset contains the following tables:

Table Name	Description
games_wide	Every pitch, steal, or lineup event for each at bat in the 2016 regular season.*
games_post_wide	Every pitch, steal, or lineup event for each at-bat in the 2016 post season.*
schedules	The schedule for every team in the regular season.

*The schemas for the games_wide and games_post_wide tables are identical.

With this data you can effectively replay a game and rebuild basic statistics for players and teams.

Note: This data was built via a denormalization process over raw game log files which may contain scoring errors and in some cases missing data. For official scoring and statistical information please consult mlb.com, baseball-reference.com, or sportradar.com.

Last Modified 1/5/2017 11:53 PM GTB Standard Time

Printed 1/5/2017 11:55 PM GTB Standard Time

SOURCE:, Source

Dataset: an overabundance of variables/features

As expected, baseball has A LOT of data/stats
→ This database has 145 columns of data
I have highlighted the most "important" ones that I would use for my analysis

gameId	awayFinalRunsForInning	hitType	is_on_base	homeFielder1
seasonId	inningNumber	startingBalls	is_bunt	homeFielder2
seasonType	inningHalf	startingStrikes	is_bunt_shown	homeFielder3
year	inningEventType	startingOuts	is_double_play	homeFielder4
startTime	inningHalfEventSequenceNumber	**balls**	is_triple_play	homeFielder5
gameStatus	description	**strikes**	is_wild_pitch	homeFielder6
attendance	atBatEventType	**outs**	is_passed_ball	homeFielder7
dayNight	atBatEventSequenceNumber	rob0_start	homeCurrentTotalRuns	homeFielder8
duration	createdAt	rob0_end	awayCurrentTotalRuns	homeFielder9
durationMinutes	updatedAt	rob0_isOut	awayFielder1	homeFielder10
awayTeamId	status	rob0_outcomeId	awayFielder2	homeFielder11
awayTeamName	outcomeId	rob0_outcomeDescription	awayFielder3	homeFielder12
homeTeamId	outcomeDescription	rob1_start	awayFielder4	homeBatter1
homeTeamName	hitterId	rob1_end	awayFielder5	homeBatter2
venueId	hitterLastName	rob1_isOut	awayFielder6	homeBatter3
venueName	hitterFirstName	rob1_outcomeId	awayFielder7	homeBatter4
venueSurface	hitterWeight	rob1_outcomeDescription	awayFielder8	homeBatter5
venueCapacity	hitterHeight	rob2_start	awayFielder9	homeBatter6
venueCity	hitterBatHand	rob2_end	awayFielder10	homeBatter7
venueState	pitcherId	rob2_isOut	awayFielder11	homeBatter8
venueZip	pitcherFirstName	rob2_outcomeId	awayFielder12	homeBatter9
venueMarket	pitcherLastName	rob2_outcomeDescription	awayBatter1	lineupTeamId
venueOutfieldDistances	pitcherThrowHand	rob3_start	awayBatter2	lineupPlayerId
homeFinalRuns	pitchType	rob3_end	awayBatter3	lineupPosition
homeFinalHits	pitchTypeDescription	rob3_isOut	awayBatter4	lineupOrder
homeFinalErrors	pitchSpeed	rob3_outcomeId	awayBatter5	
awayFinalRuns	pitchZone	rob3_outcomeDescription	awayBatter6	
awayFinalHits	pitcherPitchCount	is_ab	awayBatter7	
awayFinalErrors	hitterPitchCount	is_ab_over	awayBatter8	
homeFinalRunsForInning	hitLocation	is_hit	awayBatter9	

Last Modified 1/5/2017 11:53 PM GTB Standard Time

Printed 1/5/2017 11:55 PM GTB Standard Time

Hypothesis - Plan

Use the 2016 data to build a model to:

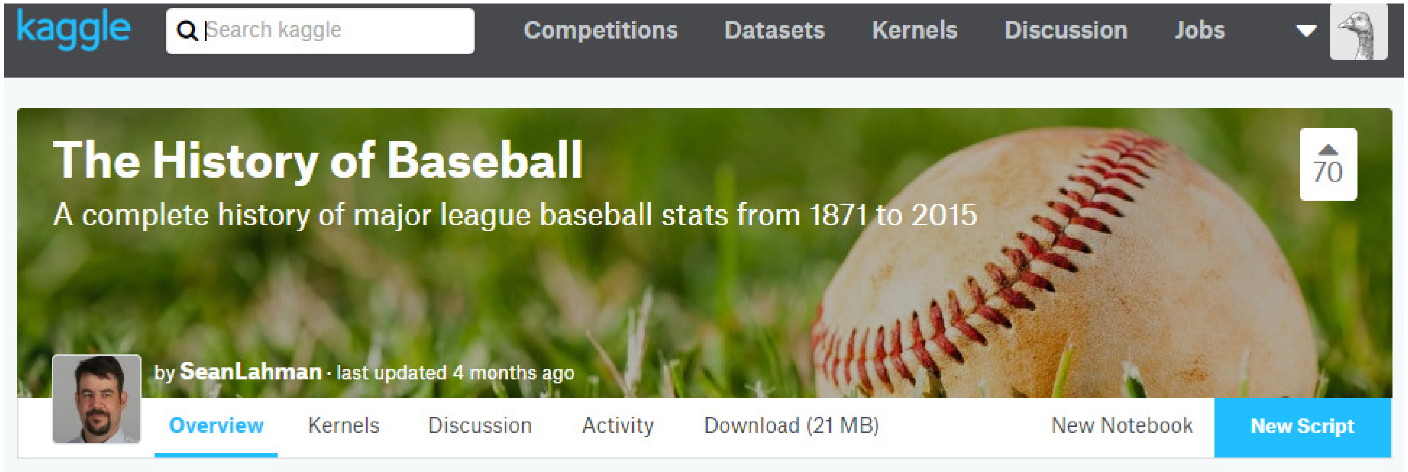
- Use variables such as homeFinalRuns, homeFinalHits, homeFinalErrors, awayFinalRuns, awayFinalHits, awayFinalErrors in the regular season

- Predict post-season performance

For example, a **hypothesis** can be:

- "The combination of runs, hits and errors in the home games can be used to predict relative performance of the post-season teams"

Then, use a different year/dataset from Kaggle to validate:



The History of Baseball

A complete history of major league baseball stats from 1871 to 2015

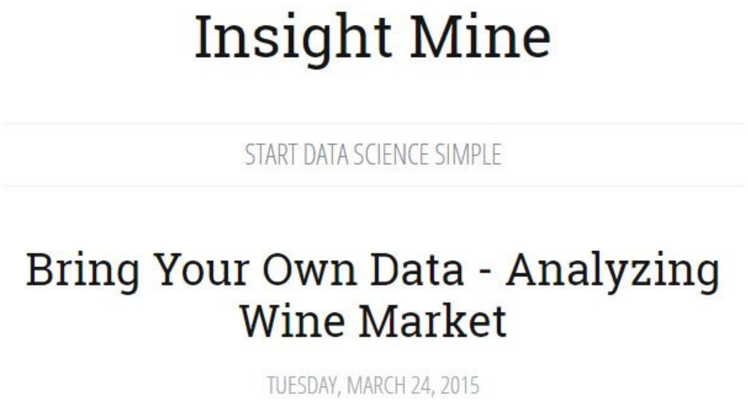
by SeanLahman · last updated 4 months ago

Last Modified 1/5/2017 11:53 PM GTB Standard Time

Printed 1/5/2017 11:55 PM GTB Standard Time

Idea #2 Find the perfect wine price/rating ratio

The idea for this project comes from an analysis I found online at http://insightmine.com/bring-your-own-data-analyzing-wine-market/



Insight Mine

START DATA SCIENCE SIMPLE

Bring Your Own Data - Analyzing Wine Market

TUESDAY, MARCH 24, 2015

Sample Variables:

- Wine Name
- Year (2001 - 2014)
- Grape Variety (Red, White, Sparkling, Rosé, Dessert)
- Region (wine producing region of the country)
- Country of origin
- # Reviews (customers gave to a specific bottle)
- Original Price
- Discounted Price
- Regular Price (if no discount was applied)
- Final Price (price at which wine was being sold)
- Rating Name (source of customer ratings)
- Rating Score (overall customer rating score)

The data scientist took the data from wine.com in order to answer "my analysis question is: *What drives wine rating score and, hence, increases chances of a bottle to be sold?"*

I would like to do a similar analysis using the data from:

- snooth.com (Open API), or
- cellartracker.com (need to ask for access)

A question I would like to answer is: **what is the best price/rating ratio to maximize a buyer's "bang for your buck"**

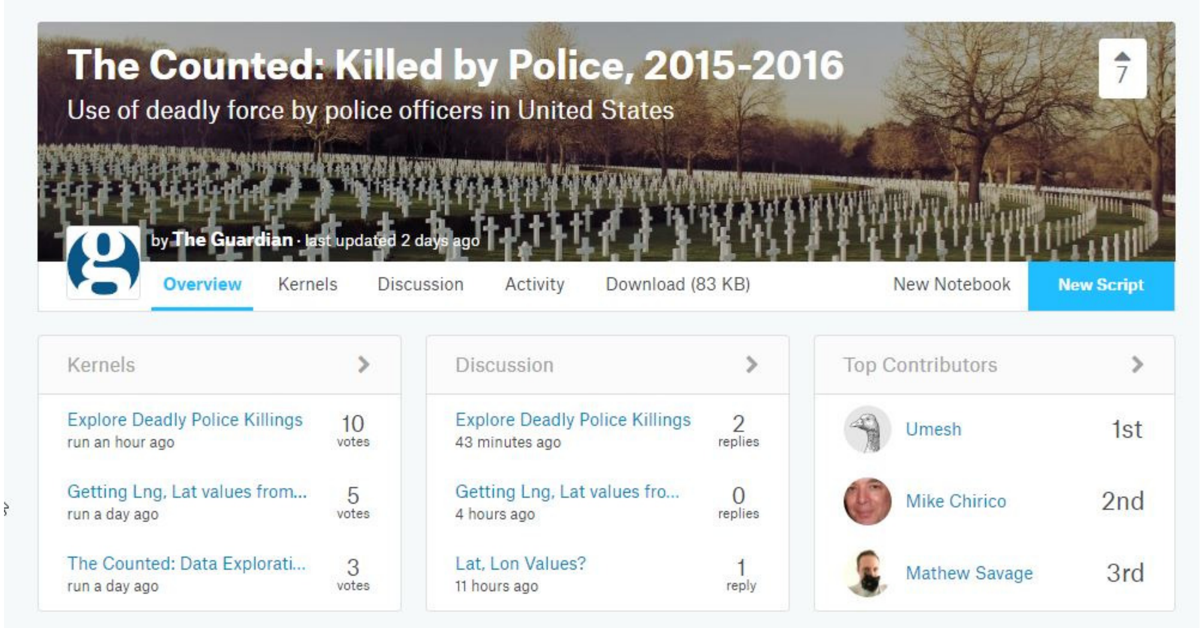
Last Modified 1/5/2017 11:53 PM GTB Standard Time

Printed 1/5/2017 11:55 PM GTB Standard Time

Source

Idea #3 Explore the newly added police deaths data

The newly (Jan 5th, 2017) released data from The Guardian regarding the deaths from police shootings can be a very interesting area of exploration.



The Counted: Killed by Police, 2015-2016

Use of deadly force by police officers in United States

by The Guardian · last updated 2 days ago

Kernels

Explore Deadly Police Killings run an hour ago	10 votes
Getting Lng, Lat values from... run a day ago	5 votes
The Counted: Data Explorati... run a day ago	3 votes

Discussion

Explore Deadly Police Killings 43 minutes ago	2 replies
Getting Lng, Lat values fro... 4 hours ago	0 replies
Lat, Lon Values? 11 hours ago	1 reply

Top Contributors

Umesh	1st
Mike Chirico	2nd
Mathew Savage	3rd

uid
name
age
gender
raceethnicity
armed
month
day
year
streetaddress
city
state
classification
lawenforcementagency

Last Modified 1/5/2017 11:53 PM GTB Standard Time

Printed 1/5/2017 11:55 PM GTB Standard Time

Some of the questions we can ask are:

- Where and when are police-related deaths occurring
- Given one’s gender, race, and age where should they be more “careful”

Source