

MONEYBALL

Guided By: Stefan Willi Hart

*Presented By: Nikhil Raikar
Chery Jacob
Vamshi Bhushanaboina
Pooja Ajit Kumar*

TABLE OF CONTENTS

- Business Case
- Project Requirements
- Project Plan
- Concept
- Show Case
- Potential Sales Volume of the Project
- Lessons learned
- Discussion

BUSINESS CASE

- Lot of Sport companies these days want to predict the future outcome of their club successes.
- Their goal is to bring in good players who can make big contributions to the club.
- They want to make affordable decisions so as to bring in a quality athletes to the club and inreturn get the best results.
- In this project we have designed strategies to help those companies make better decisions in terms of their players and in terms of the club success.

BUSINESS CASE

- In this project we have focused on a baseball project named after a baseball movie called MoneyBall.
- The main goal of this project was to make "MoneyBall" predictions about a players by taking their past performance data.
- With the help of data mining methods and techniques, this project will help make better decisions in selecting an affordable player by analysing their past data in terms of the ability, skill and commitment level of a player.

PROJECT REQUIREMENTS

- Data ingestion of various baseball statistics of the last 25 years for which the data has been gathered using the below dataset.
- Dataset : <http://seanlahman.com/baseball-archive/statistics>
- Implementation of 5 evaluation algorithms.
- Tools used: SAP HANA Database, SAP Predictive Analysis, Python(For Statistical computing), Javascript, HTML and SAPUI5(UI Framework).

PROJECT PLAN

Phase 1 - Requirement

- Understanding of Project Objectives.
- Understanding the requirements which needs to be implemented.
- Collection of the data set and analyse on how the data has been structured.
- Perform data transformation process.
- Loading data into the HANA

PROJECT PLAN

Phase 2 - Implementation

- Data in the HANA is further minimized by association process.
- Research on how our dataset can be correlated and associated.
- Calculation view of HANA is used for the association.
- Associated data is further sent to predictive analysis.
- Design of the user interface to showcase the results.
- Implementation of the user Interface in SAP UI5

PROJECT PLAN

Phase 3 - Evaluation

- Research on which algorithms suitable for the data.
- Research on the implementation details of the algorithms.
- Implement the algorithms on the dataset.
- Evaluation of the results.
- Export the results back to the SAPUI5 User Interface and populate the results

PROJECT PLAN

The Discussed Project Plan in the previous slides has been carried out successfully on the following dates,

Gitlab: https://gitlab.com/nikhilraikar88/Project_MoneyBall

Task	Estimated Complete Date	Done
Cleaning of data	17/12/2017	✓
Methods to load data	22/01/2017	✓
Functional Design Document	22/01/2017	✓
Design of User Interface	8/02/2017	✓
Research on algorithms	11/02/2017	✓
Implementation of algorithms	17/02/2017	✓
Evaluation Of results	24/02/2017	✓

CONCEPT

- Data is cleaned and various tables are aggregated to form four main column views,
 - Batting
 - Fielding
 - Pitching
 - Salary
- Calculate new columns to get the necessary player information.
- Create predicted columns to find the best players.
- Create SAP UI5 app to display results.

PROCESSING LOGIC

Calculation Logic

- The different column views such as Batting, Fielding, Pitching and Salary data are added into the SAP HANA catalog.
- Sort and derive important data using SQL.
- CUBE type calculation view with star join to merge all the different column views on the database.
- Data is then loaded into the SAP predictive analysis tool.
- After the calculation, the data is populated back to the catalog and processed further.

EXAMPLES

Adding the column views

.hdbtable

```
import = [  
  {  
    table = "gbi-student-015.MoneyBall::Fielding";  
    schema = "MONEYBALL_PROJECT";  
    file = "gbi-student-015.MoneyBall:Fielding.csv";  
    header = false;  
  }  
];
```

.hdbti

```
table.schemaName = "MONEYBALL_PROJECT";  
table.tableType = COLUMNSTORE;  
table.columns = [  
  { name="playerID"; sqlType=NVARCHAR;length=30;},  
  {name= "G_SUM" ; sqlType=INTEGER;},  
  {name="GS_SUM" ; sqlType=DOUBLE;},  
  {name=" InnOuts_SUM" ; sqlType=DOUBLE;},  
  {name="PO_SUM" ; sqlType=INTEGER;},  
  {name="A_SUM" ; sqlType=INTEGER;},  
  {name="E_SUM" ; sqlType=INTEGER;},  
  {name="DP_SUM" ; sqlType=INTEGER;}  
];
```






EXAMPLES

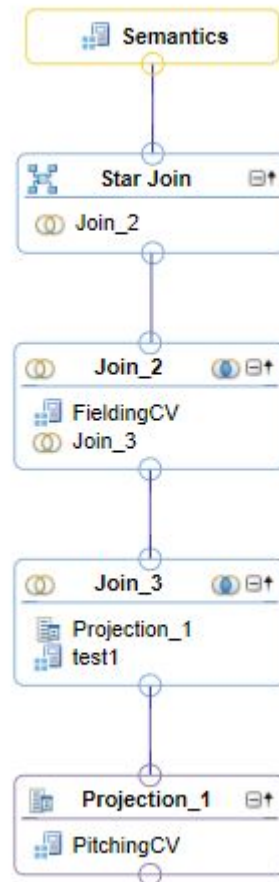
Queries executed to retrieve top 1000 players

```
SELECT TOP 1000
    "WeightedAvg",
    "Salary",
    "playerID",
    "PredictedValues"
FROM "MONEYBALL_PROJECT"."ABCSAP"
where "PredictedValues" = 'A'
Order by "WeightedAvg" DESC;
```

EXAMPLES

CUBE calculation view

		Name	Definition: SP
		SP	GENERAL
		OBP	
		BABIP	
		BattingAvg	
		EQA	
			EXPRESSION
			<p>Name * <input type="text" value="SP"/></p> <p>Data Type * <input type="text" value="DOUBLE"/></p> <p>Length <input type="text"/> Scale <input type="text"/></p> <p>Value <input type="text" value="Column Engine"/> <input type="button" value="Expression Editor"/></p> <p>$(\text{double}(\text{"H_SUM"}) + (2 * (\text{double}(\text{"2B_SUM"}))) + (3 * (\text{double}(\text{"3B_SUM"}))) + (4 * (\text{double}(\text{"HR_SUM"})))) / \text{float}(\text{"AB_SUM"})$</p>



BATTING METRICS

- Total bases refer to the number of bases gained by a batter through his hits.

$$TB = 1B + 2 * 2B + 3 * 3B + 4 * HR$$

- Slugging percentage tells us the power of the hitter. Slugging percentage is calculated as Total Bases divided by At Bats.

$$SP: "TB"/"AB_SUM"$$

- On Base Percentage: It helps us identify how many times the batter can hit the ball perfectly from his base of the bat.

$$OBP: (Hits + Base\ on\ Ball + Hit\ By\ Pitch) \div (At\ Bats + Base\ on\ Balls + Hit\ by\ Pitch + Sacrifice\ Flies)$$

- Batting average on balls in play (abbreviated **BABIP**) measures how many of a batter's balls in play go for hits

$$BABIP = (Hits + home_runs) / (at_bats + strikeouts + sacrifice_flies)$$

PITCHING METRICS

- We have firstly used this to get how effective the pitcher is.

Pitcher: (Walks (BB) + Hits) Divided by Innings Pitched

- We then calculate the walks for 9 innings (Traditional Length of a game) by using this formula. It is the average number of bases on balls, (or **walks**) given up by a pitcher per nine **innings** pitched.

$$BB/9 = 9 \times (\text{Walks} / (\text{Innings Pitched} + (\text{Outs (Partial Innings) Pitched} / 3)))$$

- Earn Run Average: It gives us the mean of earn runs given by the pitcher per 9 innings. In Baseball, Component ERA is calculated as,

$$\text{Component ERA} = (((\text{Hits} + \text{Walks} + \text{Hit by Pitch}) \times \text{PTB}) / (\text{Batters Faced by Pitcher} \times (\text{Innings Pitched} + (\text{Outs (Partial Innings) Pitched} / 3)))) \times 9)$$

FIELDING METRICS

- A players fielding worthiness is calculated using the below formulas. This will give us a players fielding capability on the ground.

$$\text{Fielding Percentage (FPCT)} = (\text{putouts} + \text{assists}) / (\text{putouts} + \text{assists} + \text{errors})$$

- We have also calculated the fielders total chances,so know how accurate he is on the ground.

$$\text{TC} = \text{putouts} + \text{assists} + \text{errors}.$$

SAP EXPERT ANALYTICS

- The dataset with the new calculated columns is loaded into SAP predictive analysis tool.
- Perform ABC clustering. The best players are found in cluster A.
- Perform K-Means clustering. We found that the clusters 1,4 contained the best players.
- The results are sent back to SAP HANA catalog.

APPLICATION LOGIC

- OData service is defined to expose the data defined by the entity sets.
- We expose the calculation views “AGG”, “ABC” and “K MEANS” created in the previous layer using the OData services.

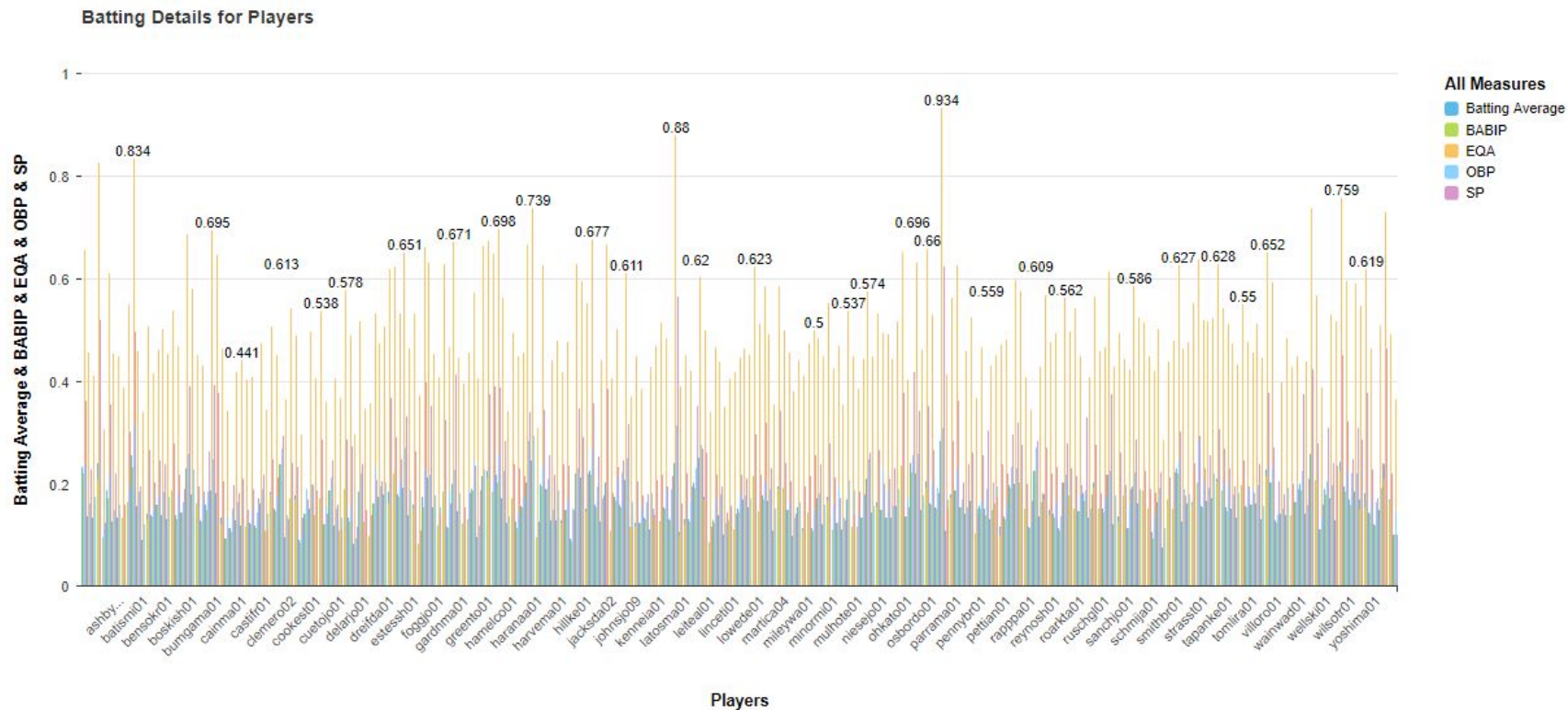
.xsodata

```
service {  
  "gbi-student-015::AGG" as "MoneyBall"  
  key("playerID")  
  aggregates always;  
  
  "gbi-student-015::FeatureAllExp" as "MoneyBallAll"  
  key("playerID")  
  aggregates always;  
  
  "gbi-student-015::ABCSAP" as "ABC"  
  key("playerID")  
  aggregates always;  
  
  "gbi-student-015::KMEANSCV" as "KMEANS"  
  key("playerID")  
  aggregates always;  
}
```

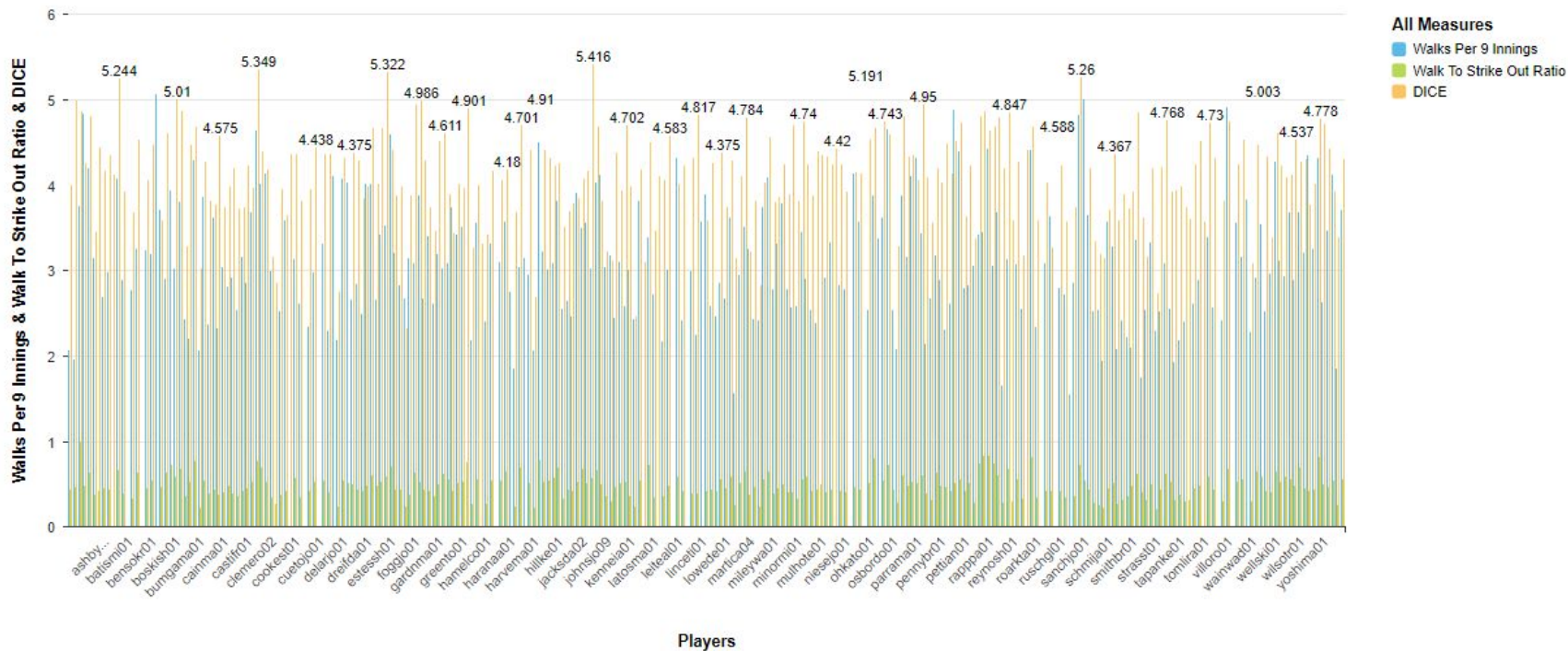
PRESENTATION LOGIC

- SAPUI5 web application uses HTML for defining the structure and layout of a Web document .
- The content is designed and defined by using XML and JavaScript respectively.
- Different tabs for Predictive and Descriptive analysis.
- The best players are presented in the form of graphs and tables.

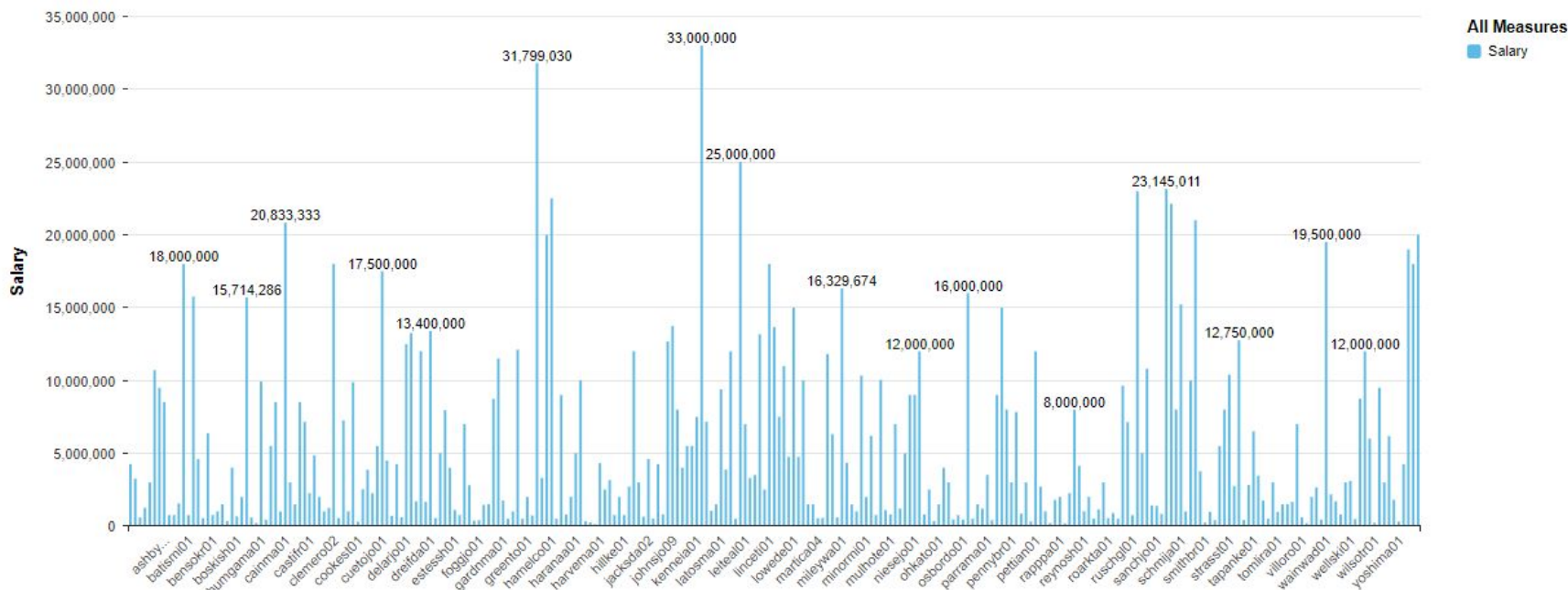
DESCRIPTIVE ANALYSIS



Pitching Details for Players



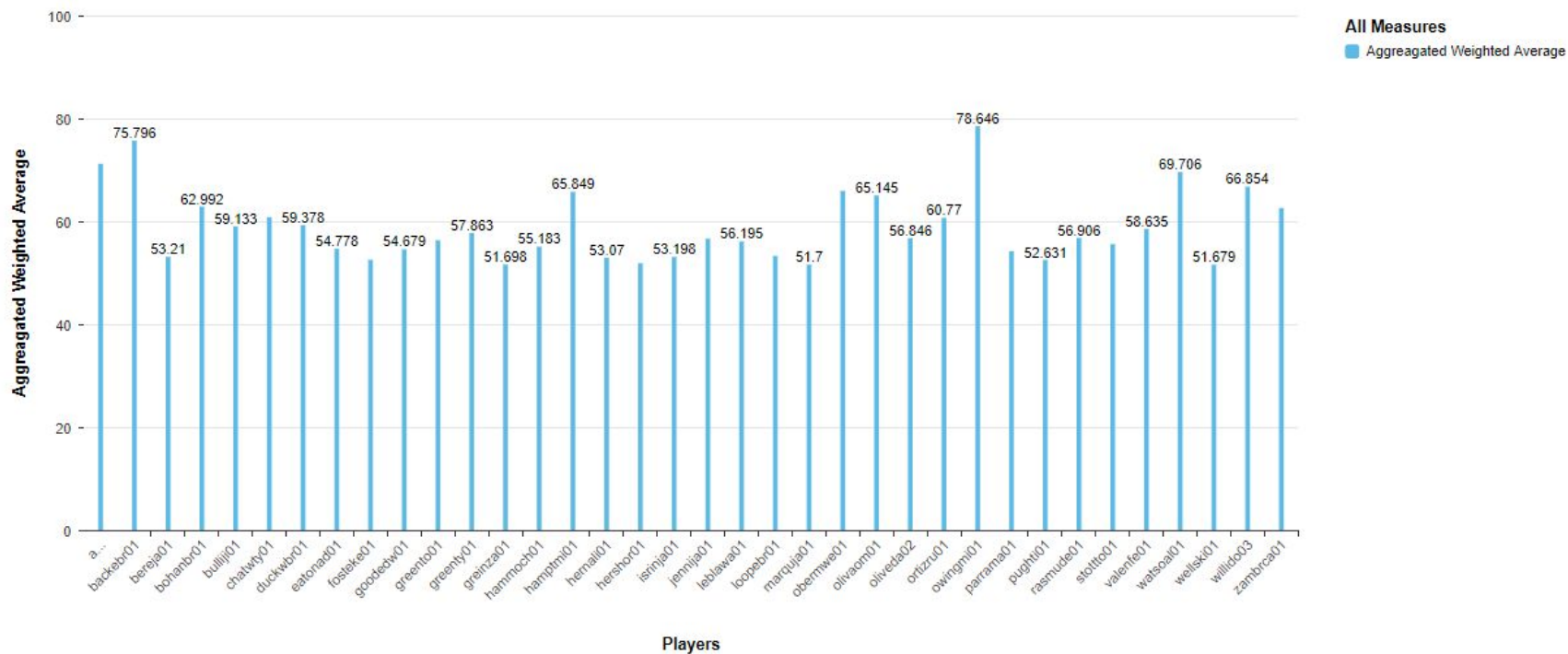
Salary Details for Players



PREDICTIVE ANALYSIS

OVERVIEW **ABC ANALYSIS** K-MEANS ANALYSIS ABC DATA SET KMEANS DATA SET BATTING ANALYSIS PITCHING ANALYSIS SALARY ANALYSIS

Batting Details for Players

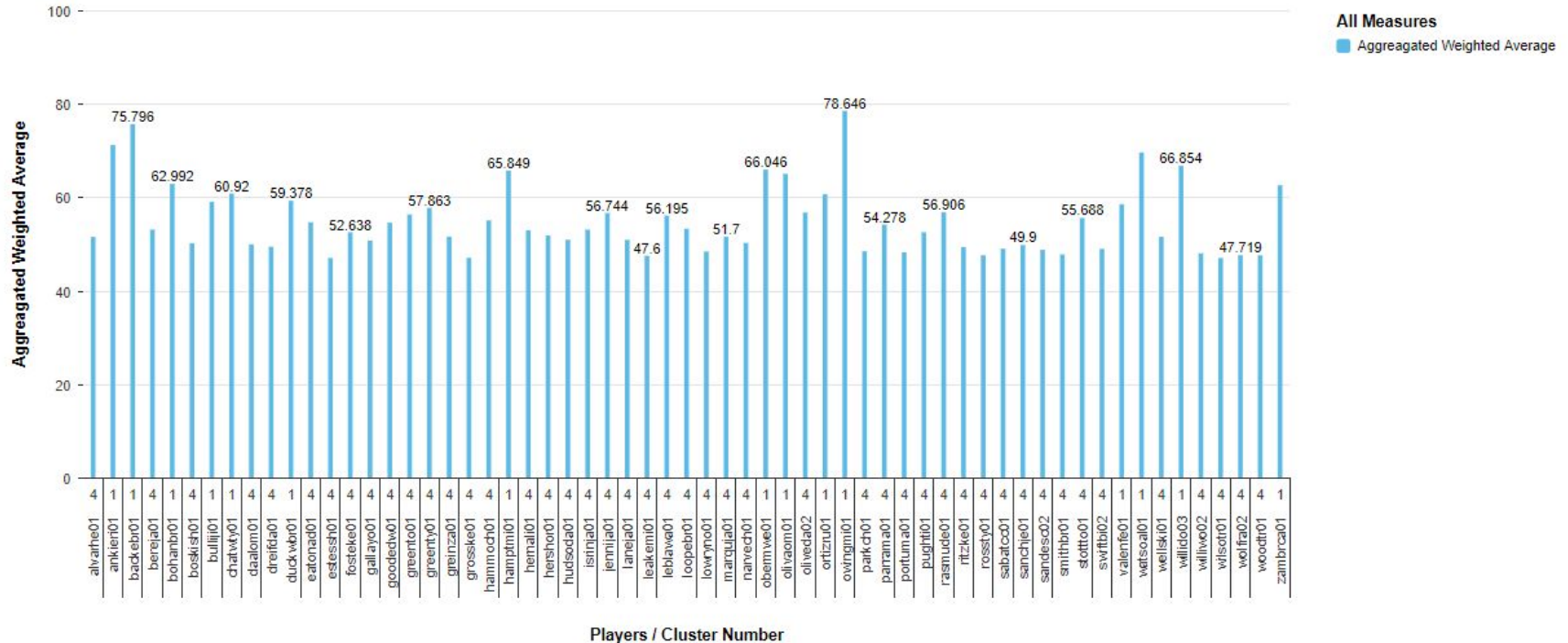


OVERVIEW ABC ANALYSIS K-MEANS ANALYSIS **ABC DATA SET** KMEANS DATA SET BATTING ANALYSIS PITCHING ANALYSIS SALARY ANALYSIS

Player_ID	Weighted Average
ankieri01	71.31770325
backebr01	75.79646301
bereja01	53.20979309
bohanbr01	62.99171829
bulliji01	59.13348007
chatwty01	60.92000961
duckwbr01	59.37818527
eatonad01	54.7778511
fosteke01	52.63790512
goodedw01	54.67861938
greento01	56.44312668
greenty01	57.86274338
greinza01	51.69783783
hammoch01	55.18291855
hamptmi01	65.8493042
hernali01	53.07039642
hersh01	51.96118546
isrinja01	53.19799423
jennija01	56.74443436
leblawa01	56.19529724

OVERVIEW ABC ANALYSIS **K-MEANS ANALYSIS** ABC DATA SET KMEANS DATA SET BATTING ANALYSIS PITCHING ANALYSIS SALARY ANALYSIS

Batting Details for Players



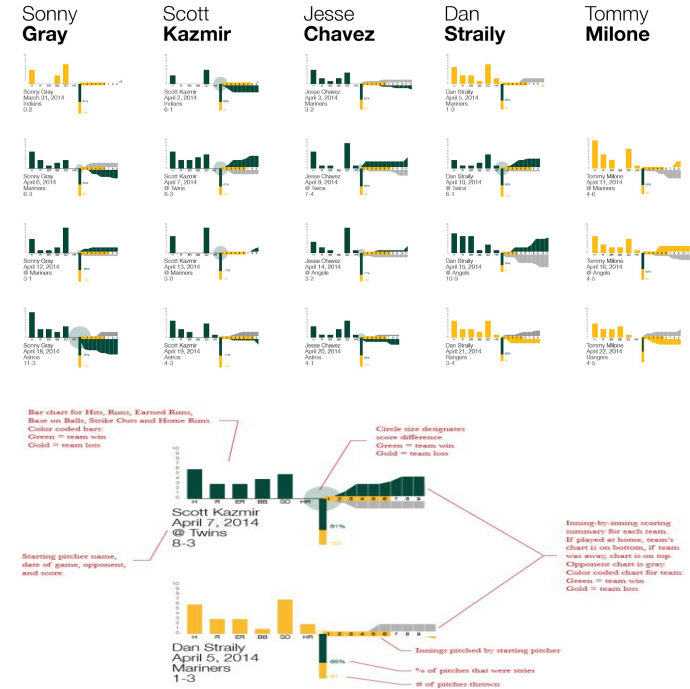
OVERVIEW ABC ANALYSIS K-MEANS ANALYSIS ABC DATA SET **KMEANS DATA SET** BATTING ANALYSIS PITCHING ANALYSIS SALARY ANALYSIS

Player_ID	Weighted Average
alvarhe01	51.64188385
ankieri01	71.31770325
backebr01	75.79646301
bereja01	53.20979309
bohanbr01	62.99171829
boskish01	50.26540375
bulliji01	59.13348007
chatwty01	60.92000961
daalom01	50.04416656
dreifda01	49.53017044
duckwbr01	59.37818527
eatonad01	54.7778511
estessh01	47.1322403
fosteke01	52.63790512
gallayo01	50.8664093
goodedw01	54.67861938
greento01	56.44312668
greenty01	57.86274338
greinza01	51.69783783
grosske01	47.17708969

POTENTIAL SALES VOLUME

How can these analysis help raise sales volume in business?

- The main goal behind this is value and consistency separate top players from the pack.
- Using the formulas we find “VALUE” in players that nobody else can see.
- We make affordable smart decisions.
- Teams such as Pittsburgh Pirates, Houston Astros, Milwaukee Brewers, Boston Red Sox, Oakland Athletics have used these analysis to help raise their business and make smart decisions.



LESSONS LEARNED

- Subjectivity is the key.
- Product knowledge increasingly trump's relationship selling.
- Data Cleaning is important.
- Descriptive Analysis will help us get a picture of where the business presently stands.
- Predictive analysis is one of the best techniques to help improve the performance of any businesses.
- The world is dealing with Big Data and SAP is one of the best tools to help you streamline your processes, giving you the ability to use live data to predict customer trends.



REFERENCES

- <https://archive.sap.com/discussions/thread/3745888>
- <https://archive.sap.com/discussions/thread/3954130>
- <https://help.sap.com/viewer/b3d0daf2a98e49ada00bf31b7ca7a42e/2.0.02/en-US/ae6d2dd6612f4f4fb29cdc01a4690377.html>
- <https://www.fangraphs.com/library/pitching/babip/>
- https://help.sap.com/doc/86fb8d26952748debc8d08db756e6c1f/2.0.00/en-us/sap_hana_predictive_analytics_library_pal_en.pdf
- https://en.wikipedia.org/wiki/Equivalent_average
- <https://captaincalculator.com/sports/baseball/>
- <https://www.miniwebtool.com/batting-average-calculator/>
- <http://vizthinker.com/baseball-data-visualization-experiment/>

THANK YOU! :)

Questions?