Hogeschool van Amsterdam

Capstone Big-Data in urban technology

Batting

Klaas Berbee

Thierry Ligeon

supervised by

Dr. J.R. Helmus

November 6, 2017

Contents

This page is intentionally left blank.

1 Capstone Batting

Klaas Berbee & Thierry Ligeon

Voor de capstone van de minor Big-Data in urban technology aan de Hogeschool van Amsterdam moeten wij de Batting dataset gaan bestuderen. Als eerste beginnen we met het globaal inzicht krijgen van de dataset.

1.1 Context van de dataset

De Batting dataset bevat data over o.a.: teams, spelers en managers. Ook bevat de dataset informatie over hun salarissen en skills. Verder bevat deze dataset informatie over hoe de individuele wedstrijden verlopen zijn, welke activiteiten hierin zijn voorgekomen en door wie deze zijn uitgevoerd.

Het doel is om deze dateset om te zetten naar nuttige informatie en dit vervolgens conclusies uit kunnen trekken. Teams kunnen bijvoorbeeld een inzicht krijgen in welke factoren een grote rol spelen in de prestatie van het team.

1.1.1 Belanghebbende voor deze dataset

Kansspel spelers
: Er kan een voordeel voor kansspelspelers ontstaan zodra zij bezitten over voorspelmodellen voor komende wedstrijden.

Teamcoaches
: De coach kan uit de dataset de sterke en zwakke punten van de spelers inzien. Hierop kan hij zijn trainingen specificeren op bepaalde aspecten voor zijn spelers. Daarnaast kunnen de sterke en zwakke punten van de competitie bekeken worden om zo beter voor te kunnen bereiden oo aankomende wedstrijden.

Junior spelers
: Beginnende spelers kunnen hun idolen en vaardigheden vergelijken en bekijken bij welke club zij getraind hebben. Dit zou een aanleiding kunnen zijn om een club te kiezen

Stadions
: Met deze dataset kan de hoeveelheid bezoekers voorspeld worden. Hierop kunnen prijzen, hoeveelheid beveiliging en promotie op worden aangepast.

1.1.2 Hoe de dataset tot stand gekomen is

Deze dataset is initieel opgezet door Sean Lahman in 1994. De dataset bevat data van 1871 tot 2012[**?**].

1.2 Data importeren via mySQL

Verbinding maken met de mySQL database.

Vervolgens maken we verbinding met de locale mySQL server. Vervolgens drukken we alle tabellen van de database af.

```
my_db <- dbConnect(MySQL(),
                   user='klaasberbee',
                   password='119OKVU',
                   dbname='Batting',
                   host='localhost')
```

```
print(dbListTables(my_db))
```

```
##  [1] "AllstarFull"         "Appearances"         "AwardsManagers"
##  [4] "AwardsPlayers"       "AwardsShareManagers" "AwardsSharePlayers"
##  [7] "Batting"             "BattingPost"         "CollegePlaying"
## [10] "Fielding"            "FieldingOF"          "FieldingOFsplit"
## [13] "FieldingPost"        "HallOfFame"          "HomeGames"
## [16] "Managers"            "ManagersHalf"        "Master"
## [19] "Parks"               "Pitching"            "PitchingPost"
## [22] "Salaries"            "Schools"             "SeriesPost"
## [25] "Teams"               "TeamsFranchises"     "TeamsHalf"
```

Hier kunnen we zien dat de Batting dataset uit 27 verschillende tabellen bestaat. Uit de ERD analyse was gebleken dat ..... de belangrijkste tabellen zijn.

```
dbTables <- dbListTables(my_db)

allTables <-
  lapply(dbTables, function(table) {
    dbGetQuery(my_db, paste("select * from", table))
  })
names(allTables) <- paste("db", dbTables, sep = "_")
list2env(allTables, envir = .GlobalEnv)
```

```
## <environment: R_GlobalEnv>
```

Doormiddel van de onderstaande code kan er uitgezocht worden wat de Primary Keys zijn.

```
## [1] playerID    yearID      gameNum     gameID      teamID      lgID
## [7] GP          startingPos
## <0 rows> (or 0-length row.names)
```

```
##  [1] yearID    teamID    lgID      playerID  G_all     GS        G_batting
##  [8] G_defense G_p       G_c       G_1b      G_2b      G_3b      G_ss
## [15] G_lf      G_cf      G_rf      G_of      G_dh      G_ph      G_pr
## <0 rows> (or 0-length row.names)
```

```
## [1] playerID awardID  yearID   lgID     tie      notes
## <0 rows> (or 0-length row.names)
```

```
## [1] playerID awardID  yearID   lgID     tie      notes
## <0 rows> (or 0-length row.names)
```

```
## [1] awardID    yearID     lgID       playerID   pointsWon  pointsMax
## [7] votesFirst
## <0 rows> (or 0-length row.names)
```

```
## [1] awardID    yearID     lgID       playerID   pointsWon  pointsMax
## [7] votesFirst
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   stint    teamID   lgID     G        AB
##  [8] R        H        2B       3B       HR       RBI      SB
## [15] CS       BB       SO       IBB      HBP      SH       SF
## [22] GIDP
## <0 rows> (or 0-length row.names)
```

```
##  [1] yearID   round    playerID teamID   lgID     G        AB
##  [8] R        H        2B       3B       HR       RBI      SB
```

```
## [15] CS       BB       SO       IBB      HBP      SH       SF
## [22] GIDP
## <0 rows> (or 0-length row.names)
```

```
## [1] playerID schoolID yearID
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   stint    teamID   lgID     POS      G
##  [8] GS       InnOuts  PO       A        E        DP       PB
## [15] WP       SB       CS       ZR
## <0 rows> (or 0-length row.names)
```

```
## [1] playerID yearID   stint    Glf      Gcf      Grf
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   stint    teamID   lgID     POS      G
##  [8] GS       InnOuts  PO       A        E        DP       PB
## [15] WP       SB       CS       ZR
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   teamID   lgID     round    POS      G
##  [8] GS       InnOuts  PO       A        E        DP       TP
## [15] PB       SB       CS
## <0 rows> (or 0-length row.names)
```

```
## [1] playerID    yearid      votedBy     ballots     needed      votes
## [7] inducted    category    needed_note
## <0 rows> (or 0-length row.names)
```

```
## [1] year.key   league.key team.key   park.key   span.first span.last
## [7] games      openings   attendance
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   teamID   lgID     inseason G        W
##  [8] L        rank     plyrMgr
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   teamID   lgID     inseason half     G
##  [8] W        L        rank
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID     birthYear    birthMonth   birthDay     birthCountry
##  [6] birthState   birthCity    deathYear    deathMonth   deathDay
## [11] deathCountry deathState   deathCity    nameFirst    nameLast
## [16] nameGiven    weight       height       bats         throws
## [21] debut        finalGame    retroID      bbrefID
## <0 rows> (or 0-length row.names)
```

```
## [1] park.key   park.name  park.alias city       state      country
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   stint    teamID   lgID     W        L
##  [8] G        GS       CG       SHO      SV       IPouts   H
## [15] ER       HR       BB       SO       BAOpp    ERA      IBB
## [22] WP       HBP      BK       BFP      GF       R        SH
## [29] SF       GIDP
## <0 rows> (or 0-length row.names)
```

```
##  [1] playerID yearID   round    teamID   lgID     W        L
##  [8] G        GS       CG       SHO      SV       IPouts   H
```

```
## [15] ER       HR       BB       SO       BAOpp    ERA      IBB
## [22] WP       HBP      BK       BFP      GF       R        SH
## [29] SF       GIDP
## <0 rows> (or 0-length row.names)
```

```
## [1] yearID   teamID   lgID     playerID salary
## <0 rows> (or 0-length row.names)
```

```
## [1] schoolID  name_full city      state     country
## <0 rows> (or 0-length row.names)
```

```
## [1] yearID       round        teamIDwinner lgIDwinner   teamIDloser
## [6] lgIDloser    wins         losses       ties
## <0 rows> (or 0-length row.names)
```

```
##  [1] yearID         lgID           teamID         franchID
##  [5] divID          Rank           G              Ghome
##  [9] W              L              DivWin         WCWin
## [13] LgWin          WSWin          R              AB
## [17] H              2B             3B             HR
## [21] BB             SO             SB             CS
## [25] HBP            SF             RA             ER
## [29] ERA            CG             SHO            SV
## [33] IPouts         HA             HRA            BBA
## [37] SOA            E              DP             FP
## [41] name           park           attendance     BPF
## [45] PPF            teamIDBR       teamIDlahman45 teamIDretro
## <0 rows> (or 0-length row.names)
```

```
## [1] franchID   franchName active     NAassoc
## <0 rows> (or 0-length row.names)
```

```
##  [1] yearID lgID   teamID Half   divID  DivWin Rank   G      W      L
## <0 rows> (or 0-length row.names)
```

```
## Observations: 5,148
## Variables: 8
## $ playerID    <chr> "gomezle01", "ferreri01", "gehrilo01", "gehrich01"...
## $ yearID      <int> 1933, 1933, 1933, 1933, 1933, 1933, 1933, 1933, 19...
## $ gameNum     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ gameID      <chr> "ALS193307060", "ALS193307060", "ALS193307060", "A...
## $ teamID      <chr> "NYA", "BOS", "NYA", "DET", "CHA", "WS1", "NYA", "...
## $ lgID        <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "A...
## $ GP          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0,...
## $ startingPos <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "", "...
```

```
##     playerID yearID teamID lgID inseason   G   W  L rank plyrMgr
## 1  vanhage01   1892    BLN   NL        1  11   1 10   12       Y
## 2  waltzjo99   1892    BLN   NL        2   8   2  6   12       N
## 3  hanlone01   1892    BLN   NL        3 133  43 85   10       Y
## 4   wardjo01   1892    BRO   NL        1 158  95 59    2       Y
## 5  seleefr99   1892    BSN   NL        1 152 102 48    1       N
## 6  ansonca01   1892    CHN   NL        1 147  70 76    7       Y
## 7  comisch01   1892    CIN   NL        1 155  82 68    4       Y
## 8  tebeapa01   1892    CL4   NL        1 153  93 56    1       Y
## 9  chapmja01   1892    LS3   NL        1  54  21 33   11       N
## 10 pfefffr01   1892    LS3   NL        2 100  42 56    9       Y
## 11 powerpa99   1892    NY1   NL        1 153  71 80    6       N
```

```
## 12 wrighha01   1892    PHI   NL        1 155  87 66    3       N
## 13 buckeal99   1892    PIT   NL        1  29  15 14    6       N
## 14 burnsto01   1892    PIT   NL        2  60  27 32    4       Y
## 15 buckeal99   1892    PIT   NL        3  66  38 27    4       N
## 16 glassja01   1892    SLN   NL        1   4   1  3    9       Y
## 17 striccu01   1892    SLN   NL        2  23   6 17    9       Y
## 18 crookja01   1892    SLN   NL        3  62  27 33    9       Y
## 19  goregе01   1892    SLN   NL        4  16   6  9   11       Y
## 20 carutbo01   1892    SLN   NL        5  50  16 32   11       Y
## 21 barnibi01   1892    WAS   NL        1   2   0  2    7       N
## 22 irwinar01   1892    WAS   NL        2 108  46 60    7       N
## 23 richada01   1892    WAS   NL        3  43  12 31   12       Y
```

```
##     playerID yearID teamID lgID inseason half  G  W  L rank
## 1  hanlone01   1892    BLN   NL        3    1 56 17 39   12
## 2  hanlone01   1892    BLN   NL        3    2 77 26 46   10
## 3  vanhage01   1892    BLN   NL        1    1 11  1 10   12
## 4  waltzjo99   1892    BLN   NL        2    1  8  2  6   12
## 5   wardjo01   1892    BRO   NL        1    1 78 51 26    2
## 6   wardjo01   1892    BRO   NL        1    2 80 44 33    3
## 7  seleefr99   1892    BSN   NL        1    1 75 52 22    1
## 8  seleefr99   1892    BSN   NL        1    2 77 50 26    2
## 9  ansonca01   1892    CHN   NL        1    1 71 31 39    8
## 10 ansonca01   1892    CHN   NL        1    2 76 39 37    7
## 11 comisch01   1892    CIN   NL        1    1 77 44 31    4
## 12 comisch01   1892    CIN   NL        1    2 78 38 37    8
## 13 tebeapa01   1892    CL4   NL        1    1 74 40 33    5
## 14 tebeapa01   1892    CL4   NL        1    2 79 53 23    1
## 15 chapmja01   1892    LS3   NL        1    1 54 21 33   11
## 16 pfefffr01   1892    LS3   NL        2    1 23  9 14   11
## 17 pfefffr01   1892    LS3   NL        2    2 77 33 42    9
## 18 powerpa99   1892    NY1   NL        1    1 74 31 43   10
## 19 powerpa99   1892    NY1   NL        1    2 79 40 37    6
## 20 wrighha01   1892    PHI   NL        1    1 77 46 30    3
## 21 wrighha01   1892    PHI   NL        1    2 78 41 36    5
## 22 buckeal99   1892    PIT   NL        1    1 29 15 14    6
## 23 buckeal99   1892    PIT   NL        3    2 66 38 27    4
## 24 burnsto01   1892    PIT   NL        2    1 47 22 25    6
## 25 burnsto01   1892    PIT   NL        2    2 13  5  7    4
## 26 carutbo01   1892    SLN   NL        5    2 50 16 32   11
## 27 crookja01   1892    SLN   NL        3    1 47 24 22    9
## 28 crookja01   1892    SLN   NL        3    2 15  3 11   11
## 29 glassja01   1892    SLN   NL        1    1  4  1  3    9
## 30  goregе01   1892    SLN   NL        4    2 16  6  9   11
## 31 striccu01   1892    SLN   NL        2    1 23  6 17    9
## 32 barnibi01   1892    WAS   NL        1    1  2  0  2    7
## 33 irwinar01   1892    WAS   NL        2    1 74 35 39    7
## 34 irwinar01   1892    WAS   NL        2    2 34 11 21   12
## 35 richada01   1892    WAS   NL        3    2 43 12 31   12
```

```
## [1] 179
```

References

[1] Sean Lahman, Batting database explanation 2012,
http://seanlahman.com/files/database/readme2012.txt