STAATLICH
ANERKANNTE
HOCHSCHULE

Data Curation and Modeling (DM2)

Athlete Dataset

Presented by:
Rohit Bewoor (11011831)

On 14.11.2019

HOCHSCHULE
SRH HEIDELBERG
Intelligence in Learning

DAS CORE PRINZIP

Content

STAATLICH
ANERKANNTE
HOCHSCHULE

- Introduction to Data
- Business Questions
- Cleaning Process
- Schema Creation
- Demo
- Github Link: https://github.com/rbewoor/DataManagement2

Introduction to Data

STAATLICH ANERKANNTE HOCHSCHULE

- CSV file
- 15 columns
- 271145 data rows
- Information in each row about particular Athlete participating during a particular Olympic Games, in one or more sports and medal won (if any)

Column Name	Description (Guessed from looking at the data)	Type
ID	Unique ID assigned to an Athlete	Number
Name	Athletes name	String
Sex	Athletes sex	String
Age	Athletes age (years)	Number
Height	Height (most probably in centimeters)	Number
Weight	Weight (most probably in kilograms)	Number
Team	Team that athlete played for (usually a country name)	String
NOC	National Olympic Committee name that team belongs to	String
Games	Combination of the Year and the Season	String
Year	Year of the Olympic Games	Number
Season	Season of the games: Summer/ Winter	String
City	City hosting the Event	String
Sport	Name of the sport in which athlete participated	String
Event	The actual event name that corresponds to the sport	String
Medal	Type of medal, if any, won by athlete in the Event	String

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
1	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NA
2	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NA
3	Gunnar Nielsen Aaby	Male	24	NA	NA	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NA
4	Edgar Lindenau Aabye	Male	34	NA	NA	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold

Introduction to Data

- Data Dirtiness at start:
  - Duplicates (entire row)
  - Completeness: Missing values (row and subsets of meaningful information)
  - Accuracy (e.g. Games, Year, Season not matching)
  - Inconsistency (same athlete different Names, Sex coded in different ways)
  - Junk values (e.g. Height 1982.5, Weight 7.466.666.667)

Column	Dimension	Justification from the data observed
Dataset Level	Consistency (full record redundancy)	13625 duplicates found – addressing this first reduces the overall dirtiness of columns substantially and values given below are based on this step being carried out first.
		5 duplicates found during Second and final round of de-duplication using same logic as above. Done right at the end after all the columns have been cleaned.
Dataset Level	Accuracy (full record level)	Many rows found with only ID and Name fields populated and ALL other crucial column data missing. Decided to remove these rows as they do not contribute to analysis in any way.
ID	None	Repetition allowed, each athlete has unique ID
Name	Consistency (value)	Name for one athlete (ID#5) is slightly different
Sex	Completeness	Missing values (5 values)
	Consistency (value)	E.g. "Female", "female", "F" represent the feminine sex. Similar values found for men.
Age	Completeness	Missing values (8820 values)
	Accuracy (reality)	Non-numeric values (e.g. "AUT", "fin", "male", "USA")
	Validity	Assumed business rule (must be Integer value) – not clean as numerical values already integers but junk values present
Height	Completeness	Missing values as data is not available (54117 values)
	Accuracy (reality)	Value of 1982.5 is treated as impossible height.
	Validity (precision)	Decimal values present (business rule allows only Integer values)
Weight	Completeness	Missing values as data is not available (56422 values)
	Accuracy (reality)	Junk values (7.466.666.667 and 7.733.333.333) – possible typo
	Validity (precision)	Decimal values present (business rule allows only Integer values)
Team	Completeness	Missing values (3 values)
NOC	Completeness	Missing values (133 values)
	Conformity	Possibly dirty as Team names did not make sense to their corresponding NOC and unfamiliarity with coding methodology. Attempted to check against NOC list from Olympics site; unable to get full list and check all values
Games	Completeness	Missing values (128 values)
Year	Completeness	Missing values (130 values)
Season	Completeness	Missing values (129 values)
City	Completeness	Missing values (128 values)
Sport	Completeness	Missing values (131 values)

Business Questions

STAATLICH
ANERKANNTE
HOCHSCHULE

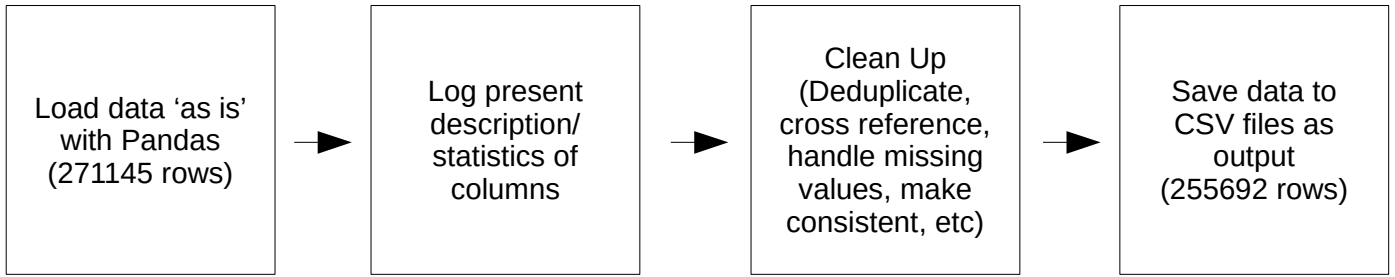
- Predict whether an athlete will win any medal based on certain features.
  - Winning = FunctionOf (Age, Height, Weight, EventType, whether Won Medal before)

- For a particular Sport and Sex of the athlete, over time, track the average Height, Weight, Age.
  This will be attempted for all participants and for only those winning medals.

Cleaning Process

STAATLICH ANERKANNTE HOCHSCHULE

- Input: original CSV data
- Outputs:
  - CSV file with cleaned data with missing values marked as MISSING
  - CSV file with coding of -1 for missing values for columns: Age, Height, Weight



Cleaning Process

STAATLICH
ANERKANNTE
HOCHSCHULE

- Using only Pandas
- Implemented Logging
- Steps:
  - Load Data in dataframe with keep_default_na=False for full control of finding missing values
  - Inserted MySN (as unique serial number for tracking later)
  - Trimmed and lowercase
  - Tagged with MISSING any cells with missing data
  - Deleted duplicate rows (13980 found)
  - Deleted rows with all meaningful data missing (NOC, Games, Year, Season, City, Sport, Event, Medal)

Cleaning Process

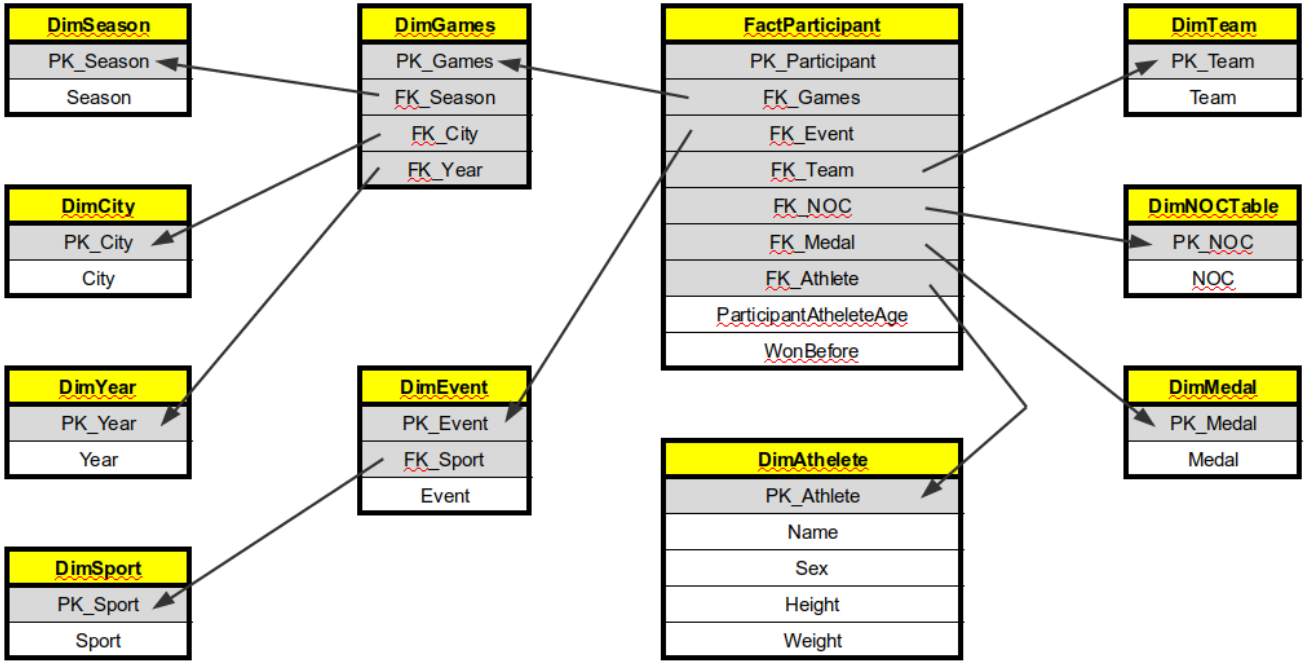
STAATLICH ANERKANNTE HOCHSCHULE

- Steps - continued:
  - For all columns, used cross referencing to find correct value for MISSING data.
    For each ID, found frequency for each unique value and used the maximum frequency value.
  - Sex: inconsistent values made uniform
  - Height, Weight: Corrected values for incorrect or absurd values. Floored all values to integer.
  - Year, Season: Using Games column as ground truth corrected inconsistent values
  - Sport: Event name corresponds with start of any Sport name. Used regex to find check if any rows inconsistent after the cross referencing.
  - Medal: Insert word NO where no medal (gold/ silver/ bronze) explicitly mentioned
  - Final removal of any row duplicates (90 found)
- Total rows remaining at end of cleaning process: 255692 (from 271145 earlier)

Schema Creation

STAATLICH ANERKANNTE HOCHSCHULE

- Snowflake Schema
  - 10 Dimension Tables
  - 1 Fact Table
- Derived Fields:
  - WonBefore: Value 1 if particular athlete has won any medal before. Else value is 0.
- PK: Primary Key
- FK: Foreign Key



Schema Implementation

STAATLICH
ANERKANNTE
HOCHSCHULE

- Input: Cleaned Data CSV file with missing values replacement for Age, Height, Weight.

- Output:
  - 11 CSV files
  - One CSV file per Dimension and Fact table.

Demo

Demo Run

STAATLICH
ANERKANNTE
HOCHSCHULE

1) Cleaning Process:
   - Script: curationVer11.ipynb
   - Input File: Athlete_Events_ORIGINAL.csv
   - Output Files: Athlete_Events_CLEANDED.csv and Athlete_Events_CLEANDED_MissingReplaced.csv
2) Schema Creation:
   - Script: scemaCreationVer3.ipynb
   - Input File: Athlete_Events_CLEANDED_MissingReplaced.csv
   - Output Files: 11 CSV files
3) Github Link: https://github.com/rbewoor/DataManagement2

STAATLICH
ANERKANNTE
HOCHSCHULE

Q & A