Luleå University of Technology (LTU)

Department of Computer Science, Electrical and Space Engineering

Digital Services and Systems

COURSE: D0025E

# Data Mining Assignment Number 1: CORRELATIONS

GROUP NO.: 2, MEMBERS: Ebbe BÄCKLINDER, Sophia DONATO,
Simon JOHANSSON, Brandon LIU, Patrizio MANDELLI

LP1-2025

# Abstract

This work has the objective of exploring and analysing data of two professional organizations (FIFA[1] and NBA[2]) in order to identify correlations between different features. After selecting the two organizations, two datasets focusing on male players were created from available information, using relevant attributes such as the age, height, weight, ... The dataset were then imported in Altair RapidMiner and Knime, for preprocessing, correlation matrix building and results evaluation. This work shows the importance of data preprocessing, and feature understanding/selection to allow for a critical interpretation of correlations to better understand the dynamic of the organizations.


# Highlights

This section explains shortly the key points of the assignment.

- Theory :
  The CRISP-DM Process was thoughtfully followed for this assignment, highlighting the importance of a standardized framework for data mining. The different steps of this process, with their interactions, are shown in Figure 1 and explain in more details in the following sections. In this assignment, highlights are particularly made on data understanding (to


Figure 1: CRISP-DM Process

  interpret better the results and select the correct attributes), on data preparation (cleaning, selecting relevant attributes, removing/replacing missing attributes) and on the correct understanding of the correlation concept as a tool to identify relationships between features.

- Practice :
  The key practical takeaways for this assignment are the selection of relevant professional organization (along with the collection of relevant datasets) and the building of working processes in Altair RapidMiner and Knime.

- Others :
  Some other key points are the importance of using enough records to have relevant statistics and the reflection on the discovered correlations.

[1] https://drive.google.com/file/d/17BqMT-W2QPERWnVWbn2Yt89skl6i2KmX/view?usp=sharing
[2] https://drive.google.com/file/d/1bYqAaHEnVF5cQnP4OfWOREkL-8FyW1up/view?usp=sharing

# Contents

**1 Phase I: Business Understanding** **4**

**2 Phase II: Data Understanding** **4**

**3 Phase III: Data Preprocessing** **5**

**4 Phase IV: Modeling** **5**

**5 Phase V: Evaluation** **5**

**6 Phase VI: Deployment** **6**

**7 Tools Insights** **6**
7.1 RapidMiner Insights . . . . . 6
7.1.1 Ease of Use : . . . . . 6
7.1.2 Efficiency : . . . . . 7
7.1.3 Costs : . . . . . 7
7.2 Knime Insights . . . . . 7
7.2.1 Ease of Use : . . . . . 7
7.2.2 Efficiency : . . . . . 7
7.2.3 Costs : . . . . . 7

**8 Answers to Assignment Questions** **8**
8.1 What correlations exist? . . . . . 8
8.2 How strong are they ? . . . . . 8
8.3 Are they surprising to you. If so, why? . . . . . 8
8.4 What other attributes would you like to add? . . . . . 8
8.5 Are there any you would like to eliminate now? Why? . . . . . 8

**9 Conclusion** **9**

# 1 Phase I: Business Understanding

The purpose of this assignment is to understand and look at the correlation within an organization. The goal is to understand how different attributes relate to each other and help sport organizations understand their players' value. Sports organizations get a large data of players, who they are going to recruit or make their team better. However, it is not easy to know what factors help a player perform better or what helps the team be better. We made a comparison on the player statistics in the NBA and FIFA dataset. In the FIFA data set we took physical and performance attributes such as pace,age, weight, wage. This could help the organization recruit the best players, remove the worst players, and help the players negotiate salaries. In the NBA dataset we used attributes such as height, weight, age, assist and points to identify the possible relationships. By comparing these attributes from both datasets,the organizations can gain insight into which factors benefit the performance of a player, use this information to develop, recruit, and make strategy decisions.

# 2 Phase II: Data Understanding

The two chosen datasets are relative to male players in the NBA league and in the 2023 version of FIFA. These were available on Kaggle, downloaded and uploaded on GoogleDrive (see Abtract for links) for your convenience. The following section presents some interesting facts about the different attributes in both datasets.

- **NBA dataset**
  The following features were available : age, height, weight, points scored, season, college, country, ... Here are some insights about some attributes :
  - *Age:* follows a fairly even distribution with most players in their mid-20s
  - *Height and weight:* smooth distribution consistent with typical basketball player physiques
  - *Points scored:* varies widely across players, with a small number of high scorers skewing the distribution

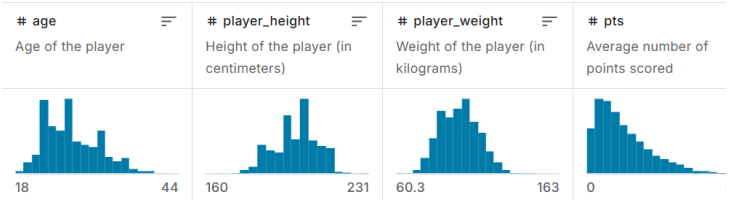


Figure 2: NBA data graphics

- **FIFA dataset**
  The following features were available : name, position, wage, value, age, height, weight, league name, club name, ... Here are some insights about some attributes :
  - *Overall score attribute:* normal curve with most players having an overall of 67 and a potential of 70.
  - *Wage:* most players have a weekly wage between 500€ and 11990€, with a few rare/good players having a bigger wage.
  - *Age:* the age follows a distribution that leans to the young ages, most players being around 21.
  - *Shooting and defending* : in these two distributions, you can see 2 peaks of most present values : one peak in the really low values and one peaks in the really high values. This corresponds to the position of the player. Indeed, if a player is a forward, he will be good at shooting but bad at defending (eg Messi has a score of 89 in shooting

and 34 in defending). It is the opposite if the player is a defender. If the player is a midfielder, he will tend to be average in both.

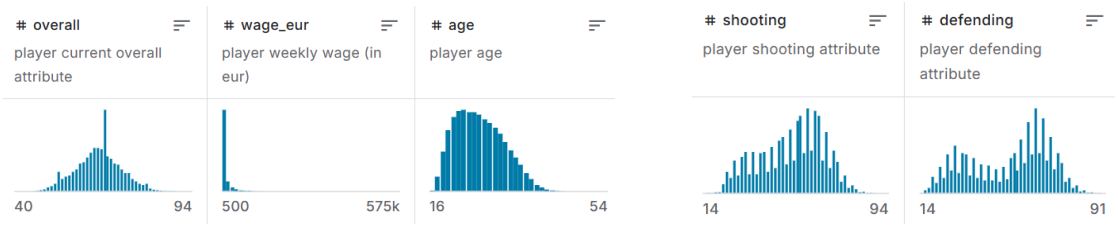


Figure 3: FIFA data graphics

# 3 Phase III: Data Preprocessing

The data processing was quite straightforward for this assignment. Indeed, available datasets were clean and clear, and only a few steps of preprocessing were required :

- **FIFA dataset:**
  - Selected attributes : overall score, wage, age, height, weight and pace
  - Normalisation : already consistent (wages in euro, weights in kg, heights in cm)
  - Missing values : pace and wage columns had some missing values, however, since the number of records was a lot bigger than the expected one, a dropping method was chosen for the objects with missing attributes.
  - Record count : 16407 (after dropping and selecting the 2023 FIFA version)
  - Only numeric attributes were kept for the correlation analysis.
- **NBA dataset:**
  - Selected attributes: age, height, weight, points
  - Normalisation: already consistent (heights in cm, weights in kg, points as totals).
  - Missing values: none of the selected columns has missing values.
  - Record count: 12566 rows
  - Only numeric attributes were kept for the correlation analysis

# 4 Phase IV: Modeling

The adopted model is a *linear correlation matrix*, which makes it possible to analyze the degree of association between pairs of variables.

The matrix is square and symmetric, with elements $R_{ij}$ ranging between $-1$ and 1:

- values close to $+1$ indicate a strong positive correlation (the variables increase together);
- values close to $-1$ indicate a strong negative correlation (one increases while the other decreases);
- values near 0 indicate no linear relationship.

This approach allows us to identify dependencies, redundancies, and possible groups of variables within the observed data.

# 5 Phase V: Evaluation

The evaluation phase is relatively straightforward in this case because it just consist in checking that the correlations matrices are correctly build and that they seem logical with respect to the data used and, in case of incoherence, change the business objective or the data preparation.
The correlation matrices for the NBA and FIFA players extracted from Altair AI and Knime can be seen, respectively in Figures 4 and 5 . Explanations are given in the Chapter 8.

| RowID | age | player_height | player_weight | pts |
|---|---|---|---|---|
| age | 1 | -0.008 | 0.064 | 0.011 |
| player_height | -0.008 | 1 | 0.822 | -0.055 |
| player_weight | 0.064 | 0.822 | 1 | -0.025 |
| pts | 0.011 | -0.055 | -0.025 | 1 |

(a) NBA correlation matrix results in Knime

| Attributes | age | player_height | player_weight | pts |
|---|---|---|---|---|
| age | 1 | -0.008 | 0.064 | 0.011 |
| player_height | -0.008 | 1 | 0.822 | -0.055 |
| player_weight | 0.064 | 0.822 | 1 | -0.025 |
| pts | 0.011 | -0.055 | -0.025 | 1 |

(b) NBA correlation matrix results in RapidMiner

Figure 4: NBA results

| RowID | overall | wage_eur | age | height_cm | weight_kg | pace |
|---|---|---|---|---|---|---|
| overall | 1 | 0.605 | 0.445 | 0.061 | 0.157 | 0.178 |
| wage_eur | 0.605 | 1 | 0.152 | 0.043 | 0.067 | 0.13 |
| age | 0.445 | 0.152 | 1 | 0.053 | 0.207 | -0.201 |
| height_cm | 0.061 | 0.043 | 0.053 | 1 | 0.74 | -0.396 |
| weight_kg | 0.157 | 0.067 | 0.207 | 0.74 | 1 | -0.352 |
| pace | 0.178 | 0.13 | -0.201 | -0.396 | -0.352 | 1 |

(a) FIFA correlation matrix results in Knime

| Attributes | overall | wage_eur | age | height_cm | weight_kg | pace |
|---|---|---|---|---|---|---|
| overall | 1 | 0.605 | 0.445 | 0.061 | 0.157 | 0.178 |
| wage_eur | 0.605 | 1 | 0.152 | 0.043 | 0.067 | 0.130 |
| age | 0.445 | 0.152 | 1 | 0.053 | 0.207 | -0.201 |
| height_cm | 0.061 | 0.043 | 0.053 | 1 | 0.740 | -0.396 |
| weight_kg | 0.157 | 0.067 | 0.207 | 0.740 | 1 | -0.352 |
| pace | 0.178 | 0.130 | -0.201 | -0.396 | -0.352 | 1 |

(b) FIFA correlation matrix results in RapidMiner

Figure 5: FIFA results

# 6 Phase VI: Deployment

The results that we got can be used to understand players' roles such as shorter players contribute to more assists, taller players contribute to score more. This can support future scouting or training for other teams.

# 7 Tools Insights

The development in the following section is done for the FIFA dataset but the same reasoning can be applied for the NBA dataset and all workflows from both datasets are included in the submission along with this report.

## 7.1 RapidMiner Insights

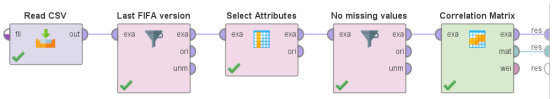


Figure 6: RapidMiner workflow for the FIFA dataset

- Read CSV: imports the Dataset.
- Last FIFA version: filters the dataset to include only the examples with the latest version of FIFA.
- Select Attributes: keeps the relevant columns (overall score, age, height, weight, wage, pace).
- No missing values: removes the examples containing attributes with missing values.
- Correlation matrix: computes the Correlation Matrix

### 7.1.1 Ease of Use :

One of the key strengths of RapidMiner is its user-friendly, drag-and-drop interface, which simplifies the process of building and deploying machine learning models. Working with the FIFA dataset was efficient because of the intuitive nature of RapidMiner's visual workflows.

### 7.1.2 Efficiency :

Although RapidMiner is excellent for prototyping and experimenting with different algorithms, its efficiency depends largely on the size of the dataset and the resources allocated during processing.

### 7.1.3 Costs :

The cost of using RapidMiner depends on whether you are using the free version or a paid version. For enterprise-level use or large-scale deployments, there are licensing and cloud-based service fees to consider. For us students, when we connect our student e-mail to RapidMiner, we get the full license included.

## 7.2 Knime Insights

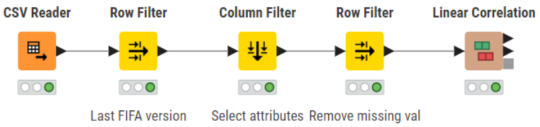


Figure 7: Knime workflow for the FIFA dataset

- CSV Reader: imports the dataset.
- First Row filter: filters the dataset to include only the examples with the latest version of FIFA.
- Column filter : keeps the relevant attributes (overall score, age, height, weight, wage, pace).
- Second Row filter : removes the examples containing attributes with missing values.
- Linear correlation : computes the Correlation Matrix

### 7.2.1 Ease of Use :

Using Knime was quite easy because it has a simple drag-and-drop interface, meaning we didn't need to write a lot of code. We could easily build the workflow by dragging and connecting different tools to clean the data and train the model. Most of the tools were straightforward to use, and Knime offers many pre-built nodes for common tasks, like handling missing data or applying machine learning models. Although there was a bit of a learning curve at first, once we got the hang of it, the process became quite simple and intuitive.

### 7.2.2 Efficiency :

Knime performed well in terms of processing speed. The workflow execution time for training models and performing evaluations was efficient, even with a dataset of over 18,000 players and numerous features. The parallel processing capabilities in Knime helped speed up the data transformation and modeling steps, particularly when working with larger datasets.

### 7.2.3 Costs :

One of the advantages of Knime is that it is open-source and completely free to use for individual users. We used the free version, which provided all the necessary functionality for building and testing the FIFA player performance prediction model. There were no licensing costs involved in our case. However, for larger teams or more enterprise-scale deployments, there would be a need to use Knime Server, which comes with additional costs depending on the number of users and the computational resources required.

# 8 Answers to Assignment Questions

Each question is answered for both datasets. The matrices can be visualised in Figures 4a, 4b, 5a and 5b.

## 8.1 What correlations exist?

- **NBA:** All correlations exists, we have all numerical values.
- **FIFA:** All correlation exists, even if some of them are really low, so almost non existent. Obviously, correlation between the attributes themselves is always of 1.

## 8.2 How strong are they ?

- **NBA:** Height and weight show a strong positive correlation, while the correlations between points, age, height, and weight are all weak or negligible, either slightly positive or slightly negative.
- **FIFA:** The strongest correlations are between wage and overall score, between age and overall score, weight and height, pace and height, pace and weight (darker blue cells in the RapidMiner matrix in Fig 5b).

## 8.3 Are they surprising to you. If so, why?

- **NBA:** The strong height–weight correlation is expected since taller players usually weigh more. The weak correlation between points and body size is not too surprising because scoring depends more on role, position, and usage than just physical traits. Age also shows little relationship, which makes sense.
- **FIFA:** The correlations seem pretty logical, indeed we can see strong positive correlation between overall score and wage (better player gain more), between overall score and age (the more you play, the best you get) but this is also a bit surprising because older players are supposed to be less good, between height and weight (taller often means bigger), and between pace and height/weight (the smaller, lighter you are, the faster you are).

## 8.4 What other attributes would you like to add?

- **NBA:** I would like to add the number of games played and the net rating of every match. Games played could help show whether more opportunities on the court translate to higher point production, while net rating could reveal if players who contribute to overall team success also tend to score more or less. Both variables provide more context for understanding performance beyond just physical attributes.
- **FIFA:** Attributes about shooting, defending, dribbling could be interesting to add to have more physical values and see how it is linked to the age, overall score, .. The number of goals scored could also be an interesting data.

## 8.5 Are there any you would like to eliminate now? Why?

- **NBA:** Yes, height and weight could be eliminated since they are strongly correlated and can be summarized in a single fraction, such as a weight-to-height ratio, which would capture the same information more efficiently.
- **FIFA:** Height and weight don't seemed really important since they are strongly only correlated to each other and not the overall score nor the wage.

# 9 Conclusion

In this assignment, we have analysed and compared the attributes of players from both the NBA and FIFA datasets, with the goal of understanding correlations that could help sports organizations make better decisions. By preprocessing the data and focusing on key numerical features, we can build correlation matrices that give us meaningful relationships. The strongest correlation in the NBA dataset was observed between height and weight, while the correlation between age,point, height, and weight is all weak or negligible. That could show the importance of role,skill and game context beyond body size. In the FIFA dataset, the strongest correlations are between wage and overall score, age and overall score, weight and height, pace and height,pace and weight. These findings show both logical patterns that taller players weigh more, better players are earning more, and the more you play, the better you get. The evaluation results show that the correlation was expected, and in general, we learned that correlation analysis is a valuable tool for identifying dependencies. We recommend sports organizations to use them to optimize scouting, player development, and salary negotiations.