



# IFN701 Project- Project proposal plan

## EUROPEAN SOCCER DATABASE: A DATA ANALYSIS PROJECT

Harmandeep Kaur Bhullar(N9784098)

**Supervisor**  
Dr. Guido Zuccon

## Table of content

Serial number	Title	Page no.
1.	Introduction	3
1.1.	Background and Context of the Problem	3
1.2.	Motivation and research gap	3
1.3.	Significance of problem	4
1.4.	Aim and objective of project	4
1.5.	stakeholders	4
1.6.	Outcome of project	5
2.	Project methodology	5
2.1.	Tasks breakdown	6
3.	Project management approach	7
3.1.	MoSCoWs prioritization	7
3.2.	Time frame	8-10
3.3.	Gantt chart	9
4.	Communication plan	10-11
5.	Risk management	11-12
6.	Ethics	12
	References	13
	Appendix A	14
	Appendix B	15

# 1. Introduction

This report is about the project planning for the data analysis and predictive modelling to be performed on a database containing data about European soccer teams and matches. The main objective of this project is to analyze the provided data to identify interesting facts about game results and players, e.g. deriving an understanding of the players characteristics. This analysis will be done by acquiring the dataset from Kaggle and using the R language.

## 1.1 Background and Context of the Problem

Soccer is the most popular and prevalent sport worldwide. According to a recent A.T. The global sport events market (which includes ticketing, marketing and media revenues) for all sports was worth €45 billion in 2009. Soccer remains king as always with global worth of €20 billion every year. Particularly in Europe, it is €16 billion which is almost half of the market. Moreover, this game is the most widely covered and viewed event in the world with the total of 28.8 billion viewers. So, soccer is greatest source of money making in entertainment industry. (Reference: H. Collignon & N. Sultan. The Sports Market: Major trends and challenges in an industry full of passion. Retrieved from: <https://www.atkearney.com/documents/10192/6f46b880-f8d1-4909-9960-cc605bb1ff34> )

On the flip side, according to Sportradar director Darren Small, the worldwide betting industry for sports is worth about \$700 billion to \$1 trillion every year for both legal and illegal betting markets. The interesting factor is that almost 70 percent is contributed from soccer betting. Therefore, many data- analysis companies are making lots of money by predicting the results of matches and performance of players by analysing the data which is collected from matches. It can be expected that this market will continue to rise to trillions in next years which will make the betting industry major part of data companies. (reference: Mr Neutral (January 21, 2014). Soccer Betting: A Worldwide Gambling Industry Worth Billions. Retrieved from: <http://www.caughtoffside.com/2014/01/21/soccer-betting-a-worldwide-gambling-industry-worth-billions/>)

## 1.2 Motivation and research gap

A lot of data acquired regarding the matches. Available data can be used to perform different type of analysis tasks. However, to make the sense of using the data is to find out the facts to answer the question **why** (i.e. understanding the behavior and characteristics of players by analyzing the factors behind actions) instead of **what** (who will win or loss). Because by understanding the factors of performance of players in manner of technical, physical and tactical strengths and weakness can help to reach the correct prediction about the results of matches and performance of teams and players. Moreover, by tracking the past performance over the time period of players can be used to analyze the trends of performance of players. This information can be stored and sell on demand according to time period, player, match, speed etc. Also, this analysis are displayed graphically to track the performance of team or player over the certain time period.

- **How this project helps to solve this problem:** As above discussed s always very hard to predict the soccer match results and it is not obvious like best team always win or team with top/best players is a best team. So, this project will use the dataset from

kaggle platform which is about about more than 25,000 matches, 10,000+ players from 11 European Countries with their lead championship from seasons 2008 to 2016.

I will use this dataset to answer the following research questions:

- Understand the characteristics of players to find the strengths and weaknesses of players.
- Track the performance of players over time and understand the trends pf performance.
- Analyze and find the factors affects match results.
- Predict the results of matches.

### 1.3 Significance of problem

- By understanding the characteristics of players will help the football clubs and scouting to identify strengths and weaknesses of their teams and opposite teams so that they can use these insights to improve the training level and increase the awareness of team against opposition to prepare them for next challenges.
- By identifying the weaknesses of players, Coaches can implement new fitness training programme to improve the performance of players.
- Deep comparisons of different players characteristics and abilities can be useful for clubs to make decision of choosing players and development or selection of teams and players.
- The trend of performance by tracking the performance of particular players can give idea about the future performance of that player.
- Moreover, many professional betting agencies are making a lot of money by predicting match results. So, it means betting odds are calculated to increase the profits and minimize their risks (wrong predictions). So, by analysing the factors which affects the results of matches, teams and players performance can be used to predict the results of matches and performance of team and players in which betting industries are highly interested in.

### 1.4 Aim and objective of project

Task	Priority
Literature review to understand the problem and identify the gap	Must have
Data analysis to answer the research questions	Must have
Report writing	Must have
Initial prediction model to predict the results of matches and performance of team and players	Could have

### 1.5 Stakeholders:

Stakeholder	Role
Student	Driving force behind project responsible for development of project.
Supervisor	Give right directions and Suggestions for successful development of project
Project Coordinator	Determines project progress

## 1.6 Outcome of project

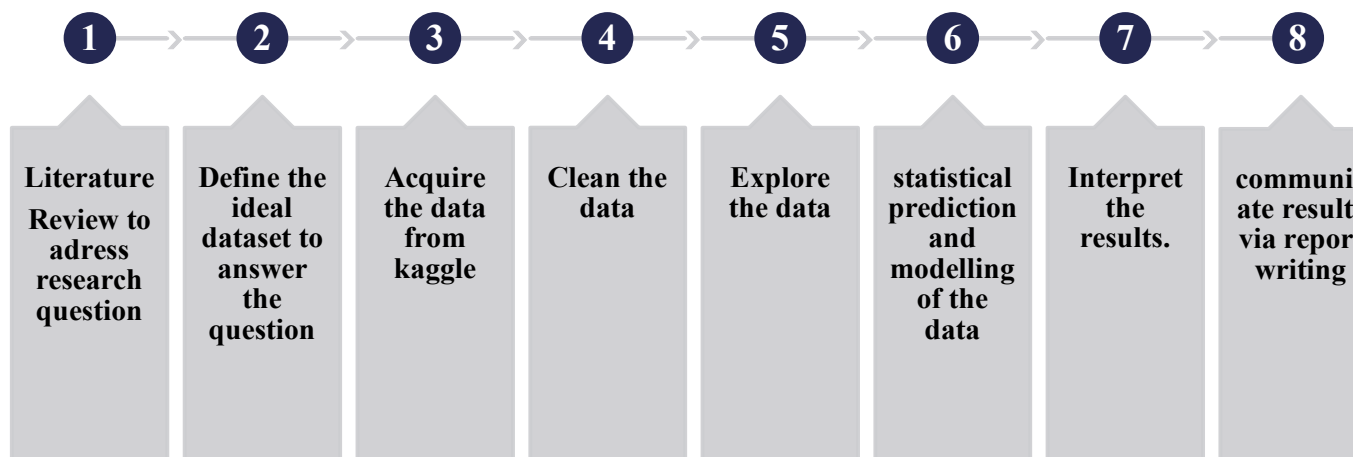
The tangible results of this data analysis project will be as listed below:

- **Data analysis report:** This data analysis report gives the detailed theory about the outcomes derived from the projects which includes analysis and visualisations done in R studio.
- **R markdown file:** Analysis and visualisations are completed using R. An R markdown file will be produced; this file will include all the source code for the visualisations and analysis.

The outcome of this project can be used by football clubs, scouting, coaches and betting agencies to determine the players performance, to improve training techniques, to find trends of performance of players over the certain time period and to predict the results of matches.

## 2. Project methodology

The project development approach includes following 8 steps to execute this project; these are the core steps in any Data Science pipeline:



**Step 1:** First of all, there is need to do literature review to address the research gap which helps to create the research questions of interest. The research questions are selected by considering the available resources and time. For example, the prediction models can be developed by team of two or more and may require more than 13 weeks time period to develop and test successfully. So, research questions are finalised by discussing the problem with project supervisor.

**Step 2:** The next stage is to select the appropriate dataset which is used for data analysis and visualisations in order to answer the questions and to gain the insights from it. The

dataset should be relevant and accurate enough to answer the research questions accurately and effectively.

**Step 3:** The dataset for this project is acquired from the public platform “Kaggle” which is popular for data analysis competitions where data is uploaded by some researchers or business people. This dataset is in form of CSV files. There are total 8 tables about the European soccer matches.

**Step 4:** Data cleaning is a very crucial step for data analysis project as it is very important to derive accurate and reliable results. Here, data is cleaned by removing unwanted data. By identification missing or incomplete data points which can be handled by removal or filling the missing data points. The noisy data or error prone data can be handled by using techniques like binning, normalisation etc.

**Step 5:** The next step is data exploration where various methods are explored for developing hypothesis, plotting patterns and clustering etc. in this step interesting patterns can be derived by analysis and visualisation of dataset.

**Step 6:** In this step, Statistical prediction and modelling of the data is done in R studio to produce the outcome to understand the characteristics of players and uncover the factors and trends of match results. Here, initial prediction model can be developed to predict the results of matches by using various methods like regression, clustering and pattern matching.

**Step 7 and 8:** Finally, the outputs and observations from Analysis with R markdown file is interpreted as written report to explain and communicate the results and answer the targeted questions.

## 2.1 Tasks breakdown

Phase	Tasks	Outcome	Duration
Phase 1 initialization (step1, 2 & 3)	<ul style="list-style-type: none"> <li>• Project management team</li> <li>• Literature review</li> <li>• Finalize research question</li> <li>• Define project scope and objective</li> <li>• Join slack group for continuous communication</li> <li>• Make a repository in GitHub</li> <li>• Project Proposal</li> <li>• Download R</li> <li>• Acquire dataset</li> </ul>	<ul style="list-style-type: none"> <li>• Project agreement</li> <li>• Repository in GitHub</li> <li>• Project proposal presentation</li> <li>• Project proposal plan report</li> </ul>	5 weeks (week 1 – week 5)
Phase 2: analysis and visualization of data (step 3, 4, 5 & 6)	<ul style="list-style-type: none"> <li>• Clean data</li> <li>• Explore the data</li> <li>• Analysis and visualization</li> <li>• Statistical modelling and prediction model</li> </ul>	R markdown file Analysis report	6 weeks (week 6- week 11)

Phase 3: consultation	Review and feedback from supervisor	Feedback and suggestions	1 week (week 12)
Phase 4: final	Final Data analysis report	Final report	Throughout from week 6-week 13) and submit on week 13

### 3. Project management approach

This project will be managed by DSDM approach (Dynamic Systems Development Method). This approach supports the iterative and incremental development by keeping the time, quality and cost fixed but features are variable for optimal and flexible solution which helps to control the risk. Moreover, an active communication and involvement of stakeholders through out the development period to give feedback and review the deliverable. I have selected this approach because of following reasons:

- Delivery on time: This project is needed to complete on time to meet the academic requirements. So, timeboxing will allow me to deliver on time.
- Never compromise quality: By applying this approach ensures the quality which can be achieved by MOSCOWS rule and regular meetings with supervisors and project coordinates to get regular feedbacks through reviewing the deliverables which leads to deliver expected results and quality.
- Build incrementally from firm foundations: This approach will support the concept of developing the firm foundation for this project before committing to any significant development. It will advocates understanding the scope problem first which is required to be solved with proposal od solution with not deep details so that the project becomes complex by overdetailed plan.

#### 3.1 MoSCoW Prioritization for Scope:

Prioritization	deliverable
Must have (60%)	<ul style="list-style-type: none"> <li>• data analysis and visualization to understand the characteristics of players</li> <li>• track the performance of players over time period to figure out trends</li> <li>• report writing</li> </ul>
Should have (20%)	<ul style="list-style-type: none"> <li>• Uncover the factors which affects the results of matches and performance of teams and players</li> </ul>
Could have (20%)	<ul style="list-style-type: none"> <li>• Initial prediction of model to predict the results of matches and performance of teams and players</li> </ul>

### 3.2 Time frame

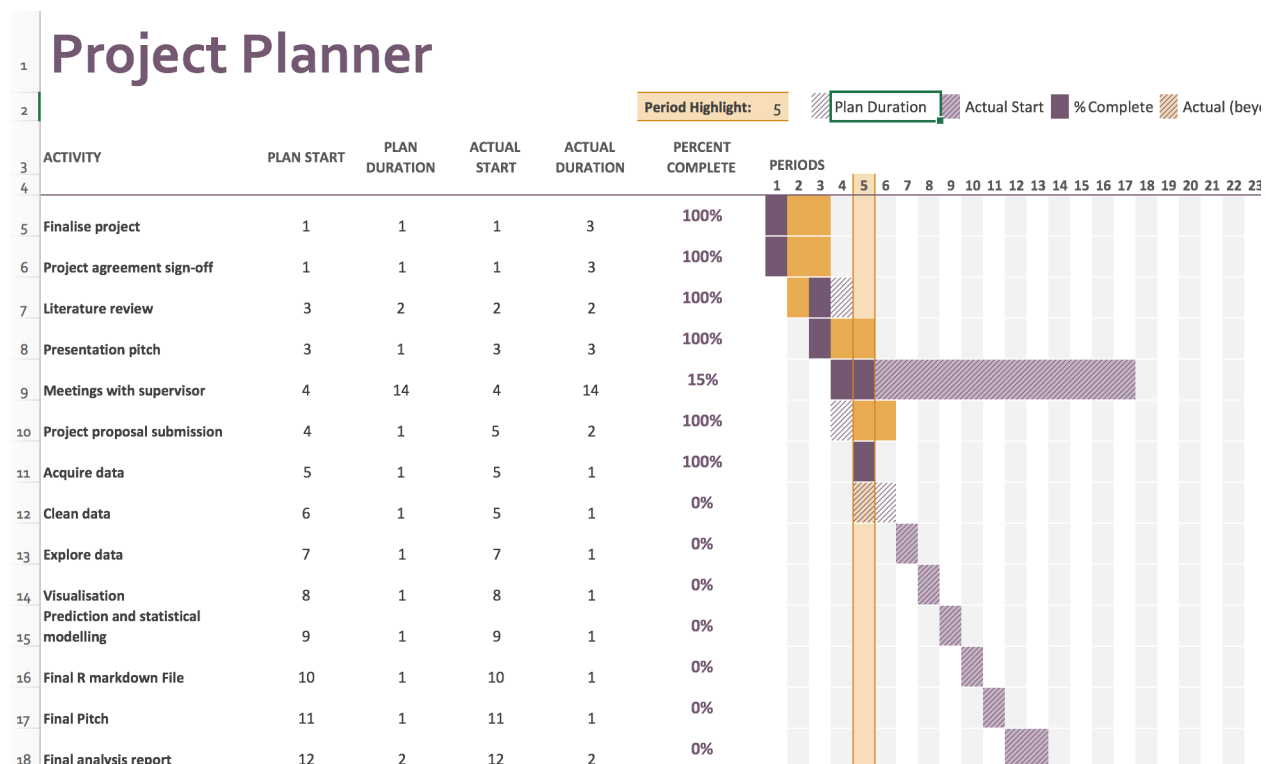
Total timeframe for this project	13 weeks
Number of timeboxes	6
Time period of each timebox	2 weeks
Number of increments	3
Time period of each increment	1 month (4 weeks)



Duration	Period	Planned tasks	Outcome
<b>Increment 1</b>			
<b>Timebox 1</b>			
Week 1		<ul style="list-style-type: none"> <li>meeting with supervisor</li> <li>create repository on gitHub</li> </ul>	<ul style="list-style-type: none"> <li>Install Rstudio</li> <li>Repository on GitHub</li> </ul>
Week 2		<ul style="list-style-type: none"> <li>Project selection</li> </ul>	<ul style="list-style-type: none"> <li>Project allocated</li> </ul>
<b>Timebox 2</b>			
Week 3		<ul style="list-style-type: none"> <li>Sign off the project agreement</li> <li>Prepare project presentation</li> </ul>	<ul style="list-style-type: none"> <li>Agreement Submission</li> <li>Acquire the Dataset</li> <li>Join Slack group</li> <li>Literature review</li> <li>Project proposal pitch</li> </ul>
Week 4		Start working on project proposal	Note: got extension for project proposal as special reason late project agreement was done
<b>Increment 2</b>			
<b>Timebox 3</b>			
Week 5		<ul style="list-style-type: none"> <li>Project proposal report</li> <li>Acquire dataset</li> <li>Clean the data</li> </ul>	<ul style="list-style-type: none"> <li>Submission of project proposal report</li> <li>Download dataset from “Kaggle”</li> </ul>
Week 6		<ul style="list-style-type: none"> <li>Data Analysis</li> </ul>	<ul style="list-style-type: none"> <li>Data cleaning by coding in R</li> </ul>
<b>Timebox 4</b>			
Week 7		<ul style="list-style-type: none"> <li>Analyse and Explore the data</li> </ul>	<ul style="list-style-type: none"> <li>Data exploration by coding in R</li> </ul>
Week 8		<ul style="list-style-type: none"> <li>Visualize the data</li> </ul>	<ul style="list-style-type: none"> <li>Data visualization via R coding</li> </ul>
<b>Increment 3</b>			
<b>Timebox 5</b>			
Week 9		<ul style="list-style-type: none"> <li>Prediction modelling</li> <li>Registration for week 12 presentation</li> <li>Work on presentation</li> </ul>	<ul style="list-style-type: none"> <li>Data modelling and statistical prediction by coding in R</li> <li>upload slides on GitHub to seek feedback</li> </ul>
Week 10		<ul style="list-style-type: none"> <li>Project presentation</li> <li>R markdown and update on GitHub to seek feedback from supervisor</li> </ul>	<ul style="list-style-type: none"> <li>Improvements in R markdown coding according to received feedback</li> </ul>

Timebox 6			
Week 11		<ul style="list-style-type: none"> <li>Documentation of data analysis in report</li> <li>Submit the draft on GitHub for review and feedback</li> </ul>	<ul style="list-style-type: none"> <li>Data Analysis report</li> <li>Project plan presentation</li> </ul>
Week 12	week 13	<ul style="list-style-type: none"> <li>Final report</li> </ul>	<ul style="list-style-type: none"> <li>Final report submission</li> </ul>

### 3.3 Gantt Chart:



## 4. Communication plan

Communication type	Purpose of communication	Way	frequency	Who is involved
Weekly meeting	Feedback and review of current timebox and planning for the next timebox	Face to face	forth night or as needed	<ul style="list-style-type: none"> <li>Student</li> <li>Supervisor</li> </ul>

Collaborative working	To work closely with the Supervisor in the assigned Lab at University and take feedback.	Face to face / slack/Git Hub	Weekly	<ul style="list-style-type: none"> <li>• Student</li> <li>• Supervisor</li> </ul>
Notifications status	To Do tasks and keep track of all planned tasks	Slack	As needed	<ul style="list-style-type: none"> <li>• <b>Supervisor</b></li> <li>• Student</li> </ul>
Quick chat	Handle unexpected issues if faced or to deal with any kind of emergency situation	Email/ slack	Frequently as required	<ul style="list-style-type: none"> <li>• Supervisor</li> <li>• Student</li> </ul>
Review and feedback	To get feedback for each increment for improvements	Face to face	After deliver every increment	<ul style="list-style-type: none"> <li>• Supervisor</li> <li>• Student</li> </ul>

## 5. Risk management

Risk type	Risk description	Mitigation step
Low	Lack of availability of supervisor due to emergency condition or busy schedule	<ul style="list-style-type: none"> <li>• Set a schedule for weekly meeting.</li> <li>• Direct communication over slack channel</li> </ul>
High	Solution does not meet requirements/unexpected solution	Continuous incremental delivery of solutions and weekly feedback from supervisor/gitHub

Medium	Missing data points of interest or noisy data or error prone data	Techniques like binning, normalizing, filling missing values etc.
High	Project does not meet deadlines	Follow project plan and keep few days in fortnight reserved for risk handling.

## 6. Ethics

There was no need of ethical clearance for this project.

## Reference

- H. Collignon & N. Sultan. The Sports Market: Major trends and challenges in an industry full of passion. Retrieved from: <https://www.atkearney.com/documents/10192/6f46b880-f8d1-4909-9960-cc605bb1ff34>
- Mr Neutral (January 21, 2014). Soccer Betting: A Worldwide Gambling Industry Worth Billions. Retrieved from: <http://www.caughtoffside.com/2014/01/21/soccer-betting-a-worldwide-gambling-industry-worth-billions/>
- C.-H. Kang, J.-R. Hwang, and K.-J. Li. Trajectory analysis for soccer players. In Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, pages 377–381. IEEE, 2006.
- Tobias Schreck, Member, IEEE, Daniel A. Keim, Member, IEEE, and Oliver Deussen. Feature-Driven Visual Analytics of Soccer Data.
- Jamal Jokar Arsanjani & Marco Helbich, & Amin Tayyebi & Amit Birenboim (1 January 2017). How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects. Retrieved from: [www.mdpi.com/journal/data](http://www.mdpi.com/journal/data)
- The DSDM Agile Project Framework (2014 Onwards)(n.d.). Agile Business Consortium. Retrieved from <https://www.agilebusiness.org/content/principles>
- V Dhar, Data Science and Prediction, Communications of the ACM, vol 56, No 12, December 2013.

## Appendix A:

### Feedback from presentation:

Comments	changes
Scope and research questions needed to be more clear and defined	Scope and research questions are discussed with supervisor for more clarification
Significance of problem is needed to be defined accurately	Literature review helped to Strengthen the significance by adding context and explanation.

### Feedback from project supervisor:

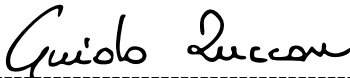
Comments	Changes
Avoid much use of adjectives such as tremendous, enormous.	Modification done.
References should be in APA format	Changed to APA style
Add MoSCoWs rule	MoSCoWS implemented
Proof reading and correct grammar mistakes	Proof read and changes made

## Appendix B: Recommended Template to obtain supervisor sign-off

---

I, Guido Zuccon <name of supervisor>, confirm that I have gone through the project plan made by Harmandeep Kaur Bhullar <student name> holding student ID number: N9784098 the project titled: “European Soccer Database: A Data Analysis Project” for IFN701 <unit code>

I confirm that I have been consulted in deriving this project proposal and that I approve of the suggested scope and tasks described in this project plan and that I am satisfied with the identified risk mitigation and communication plans articulated here

  
-----

Supervisor signature

27/08/2017  
-----

Date