



“Estimation of Survival Probabilities of Football Players: A Non-parametric method”

Project report submitted by

PHANEENDRA REDDY B S[20BSR18023]

Under the guidance of

Mr. M.A. Ghouse Basha

**Towards the partial fulfillment for the degree of
B.Sc. (HONORS) Data Science and Analytics**

To

Department of Data Science and Analytics

School of Sciences

JAIN (Deemed-to-be University)

Bangalore

2023



**Department of Data Science and Analytics, School of Sciences,
JAIN (Deemed-to-be University), Bangalore.**

CERTIFICATE

This is to certify that the present Project titled **“Estimation of Survival Probabilities of Football Players: A Non-parametric method”** has been the outcome of an original study carried out by **PHANEENDRA REDDY B S (20BSR18023)** under the supervision of **Mr. M.A. Ghouse Basha** towards the partial fulfilment of the requirements for the degree of B.Sc. Data Science and Analytics of the JAIN(Deemed-to-be University).

This is to further certify that the work reported herein does not form a part of any other thesis/dissertation, on the basis of which a degree, diploma or a certificate has been conferred upon this or any other student in the past.

Dr. Asha Rajiv,

Director
School of Sciences
JAIN (Deemed-to-be University)
Bangalore

Dr.

Project Supervisor
Department of Data Science and Analytics
JAIN (Deemed-to-be University)
Bangalore

DECLARATION

I, **PHANEENDRA REDDY B S** hereby declare that this dissertation titled " **Estimation of Survival Probabilities of Football Players: A Non-parametric method** " has been the outcome of an original study carried out under the guidance of **Mr. M.A. Ghouse Basha** towards the partial fulfilment of the B.Sc. Data Science and Analytics degree of the JAIN (Deemed-to-be University) during the year 2021-2022. This study has not been submitted for any degree, diploma or certificate.

Phaneendra Reddy B S

April, 2023

Bangalore.

ACKNOWLEDGEMENTS

I take this opportunity to acknowledge the guidance received from our professor, our college administration, and my families and friends towards this exciting journey of researching and working on the final year thesis on “**Estimation of Survival Probabilities of Football Players: A Non-parametric method**”. Completing of this project gives us immense satisfaction, and it would not have been possible without my advisors.

I am indebted to, and sincerely thank our project guide, **Mr. M.A. Ghouse Basha.**, professor, department of Data science and analytics, Jain university, for his time, patience, and valuable knowledge, and for leading and guiding me throughout this project.

Additionally, I extend my gratitude to all our teaching and non-teaching staff from the Department of Data science and analytics, Jain university, JC Road campus for their encouragement.

We are also grateful to **Dr. Asha Rajiv**, Director, School of Sciences, and **Dr. K. R. Sridhara Murthi**, Director of Academics & Planning, JAIN (Deemed-to-be University) for providing us with this opportunity. We are thankful to **Dr. Arathi Sudarshan**, Head of the Department of Data Science and Analytics for the constant support and encouragement.

We thank **Dr. Chenraj Roychand**, Chairman, JGI, **Dr. N. Sundararajan**, Pro-Chancellor and **Dr. Raj Singh**, Vice-Chancellor, JAIN (Deemed-to-be University), for providing the required infrastructure and support without which this project would not have been possible.

April, 2023

Phaneendra Reddy B S

Bangalore.

UNDERTAKING

I, PHANEENDRA REDDY B S hereby give an undertaking that the data reported in the present dissertation will not be used for any publication, conference presentation or for any industrial interaction without a written approval from the Project Supervisor and the Director, School of Sciences, JAIN (Deemed-to-be University).

Phaneendra Reddy B S

April, 2023

Bangalore.

ABSTRACT

This study proposes a non-parametric approach for estimating the survival probabilities of football players, i.e., the probability of a player remaining active in the league for a certain number of seasons. The proposed method utilizes the Kaplan-Meier estimator, which allows for the analysis of censored data, and extends it to incorporate time-varying covariates such as player age and performance statistics.

The methodology is applied to a dataset of professional football players, and the results demonstrate its effectiveness in estimating survival probabilities and identifying significant factors that affect player longevity in the league.

The non-parametric approach provides a flexible and robust alternative to traditional parametric models, particularly in the absence of strong assumptions about the underlying distribution of survival times.

Sl No	Contents	Page No
1	Survival Analysis Introduction 1.1 Survival Analysis 1.2 Censoring - Figure 1.1 1.3 Types of Censoring - Figure 1.2 1.4 Different approaches to survival analysis 1.5 Kaplan-Meier estimate - Figure 1.3	8 8 9 10 11 12
2	Python Programming language 2.1 Why only Python? 2.2 Python web Scraping 2.3 Python web Scraping Rules	13 13 14 14
3	Libraries in Python 3.1 Requests 3.2 Pandas 3.3 BeautifulSoup	15 16 17 18
4	Data Collection	19
5	Argentina Players Data - Figure 5.1 to Figure 5.12	20-33
6	Future Scope	34
7	Conclusion	35
8	Bibliography	36
9	List of Figures	37

Chapter 1

Survival Analysis Introduction

1.1 Survival Analysis

Survival analysis is a field of statistics that focuses on analyzing the expected time until a certain event happens. Originally, this branch of statistics developed around measuring the effects of medical treatment on patients' survival in clinical trials. For example, imagine a group of cancer patients who are administered a certain new form of treatment. Survival analysis can be used for analyzing the results of that treatment in terms of the patients' life expectancy.

We have already established that survival analysis is used for modeling the **time-to-event series**, in other words, lifetimes (hence also the name of the Python library which is the go-to tool for this kind of analyses). Generally speaking, we can use survival analysis to try to answer questions like:

- what percentage of the population will survive past a certain time?
- of the survivors, what will be their death/failure rate?
- how do particular characteristics (for example, such features as age, gender, geographical location, etc.) affect the probability of survival?

Survival analysis attempts to answer certain questions, such as what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

1.2 Censoring

It is important to understand that not every member of the population will experience the Event of Interest (death, churn, etc.)

Their survival times are thus, labelled as ‘Censored’.

Censorship allows you to measure lifetimes for the population who haven’t experienced the event of interest yet.

Types of Censoring

1. Right Censoring
2. Left Censoring

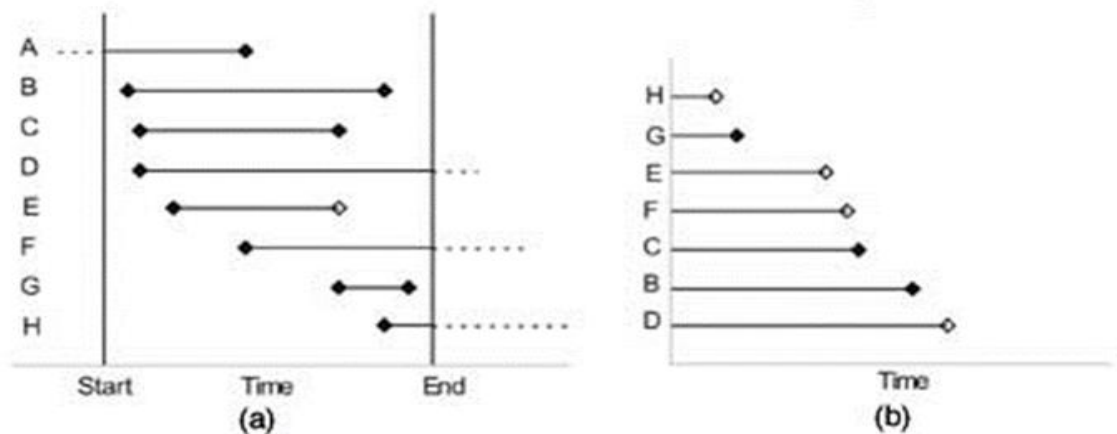


Figure 1.1: Example of censoring concept

The existence of censoring is also the reason why we cannot use simple OLS for problems in the survival analysis. That is because OLS effectively draws a regression line that minimizes the sum of squared errors. But for censored data, the error terms are unknown and therefore we cannot minimize the MSE. Applying some simple solutions such as using the censorship date as the date of the death event or dropping the censored observations can severely bias the results

1.3 Types of Censoring

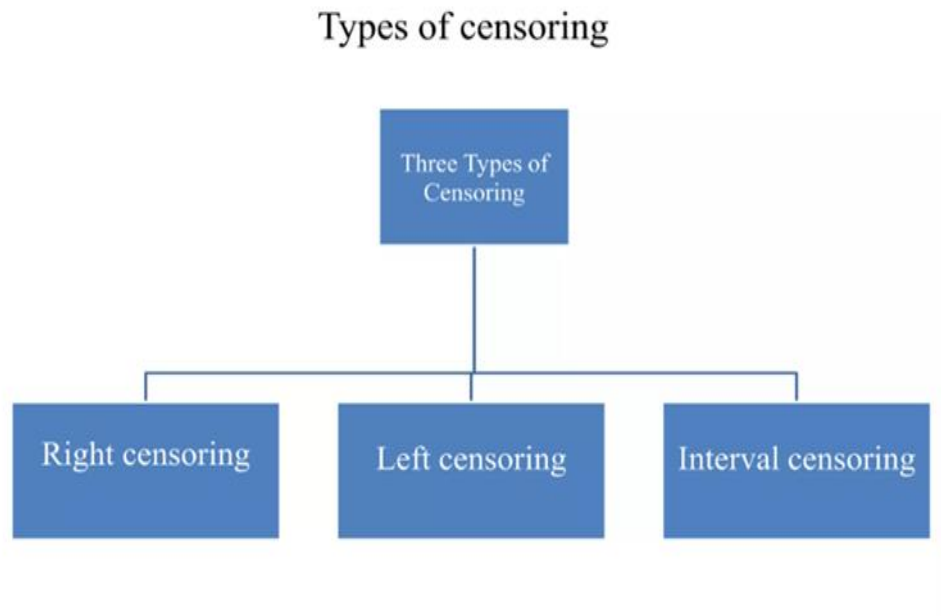


Figure 1.2: Types of censoring concept

- **Right-censoring:** This is the most common type of censoring in survival analysis. Right-censoring occurs when the event of interest has not occurred for some participants by the end of the study.
- **Left-censoring:** Left-censoring occurs when the event of interest has already occurred before the study started, and the exact time of the event is unknown.
- **Interval censoring:** Interval censoring occurs when the event of interest is known to have occurred within a specific time interval, but the exact time of the event is unknown.

1.4 Different approaches to Survival Analysis

As survival analysis is an entire domain of different statistical methods for working with time-to-event series, there are naturally many different approaches we could follow. On a high level, we could split them into three main groups:

- **Non-parametric** — with these approaches, we make no assumptions about the underlying distribution of data. Perhaps the most popular example from this group is the Kaplan-Meier curve, which — in short — is a method of estimating and plotting the survival probability as a function of time.
- **Semi-parametric** — as you could have guessed, this group is in between the two extremes and makes very few assumptions. Most importantly, there are no assumptions about the shape of the hazard function/rate. The most popular method from this group is the Cox regression, which we can use to identify the relationship between the hazard function and a set of explanatory variables (predictors).
- **Parametric** — you might have encountered this approach while doing your studies. The idea is to use some statistical distributions (some of the popular ones include exponential, log, Weibull, or Lomax) to estimate how long a subject will survive. Often, we use maximum likelihood estimation (MLE) to fit the distribution (or actually the distribution's parameters) to the data for the best performance.

1.5 Kaplan-Meier Estimate

- Kaplan-Meier Estimate is used to measure the fraction of subjects who survived for a certain amount of survival time under the same circumstances
- It is used to give an average view of the population.
- The Kaplan-Meier survival Curve is the probability of surviving in a given length of time where time is considered in small intervals.
- In Kaplan-Meier estimate were working on non parametric method.

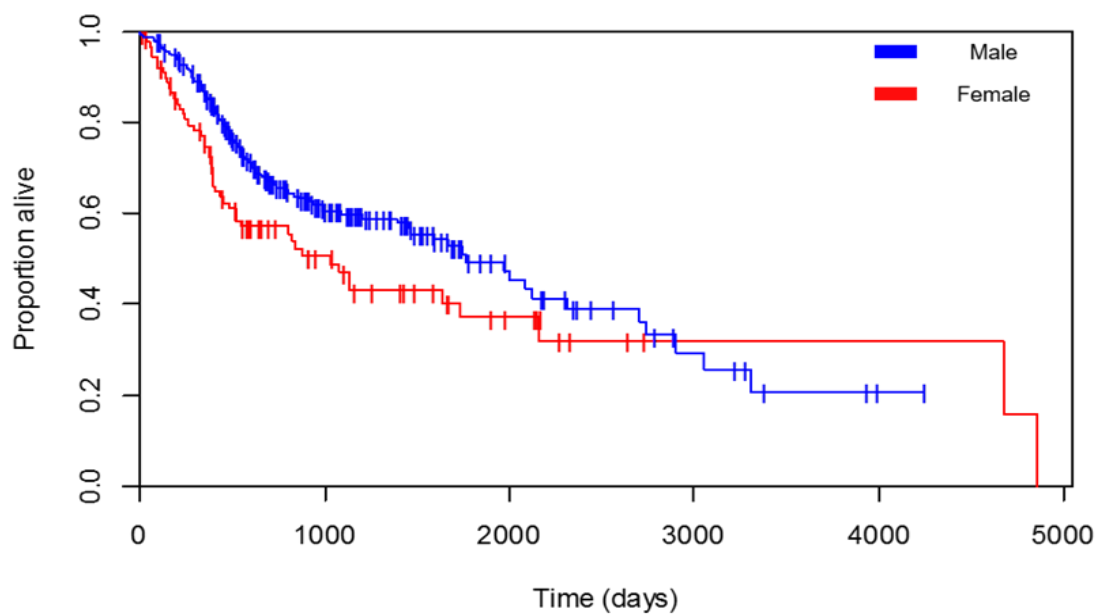


Figure 1.3: Example graph of Kaplan Meier estimate

Chapter 2

Python Programming language

2.1 Why only Python?

"Python is an interpreted, object-oriented, high-level programming language with dynamic semantics".[6] This language consist of mainly data structures which make it very easy for the data scientists to analyse the data very effectively. It does not only help in forecasting and analysis it also helps in connecting the two different languages. Two best features of this programming language is that it does not have any compilation step as compared to the other programming language in which compilation is done before the program is being executed and other one is the reuse of the code, it consist of modules and packages due to which we can use the previously written code anywhere in between the program whenever is required.

There are multiple languages for example R., Java, SQL, Julia, Scala, MATLAB available in market which can be used to analyze and evaluate the data, but due to some outstanding features python is the most famous language used in the field of data science.

Python is mostly used and easy among all other programming languages is due to the following reason.

2.2 Python Web Scraping

Web scraping refers to the automated process of extracting data from websites. It involves using software tools, called web scrapers, to extract the relevant information from websites and then saving it in a structured format, such as a spreadsheet or a database.

Python Web scraping is nothing but the process of collecting data from the web. Web scraping in Python involves automating the process of fetching data from the web. In order to fetch the web data, all we need is the URL or the web address that we want to scrape from. The fetched data will be found in an unstructured form. In order to make use of the data or collect useful insights from it, we transform it into a structured form. Once converted into a structured form, we need to store the data for further processing. The whole process is called web scraping.

2.3 Python Web Scraping Rules:

- Check the Terms and Conditions of the website before we scrape it. The Legal Use of Data section will have the information about data that we all can use. Usually, the data we scrape should not be used for commercial purposes. Use the *text* method as shown below. Every website keeps its rules defined in a *txt* file. We should inspect it to find the things that are allowed and most importantly the things that are not allowed. For example, let us inspect the *twitter* page.
- Keep the pace low. If we request for data from the website too aggressively with our bot or our program, it might be considered as spamming. Add wait time in between to make the program behave like a human.
- Use public content only.

Chapter 3

Libraries in Python

Python library is vast. There are built in functions in the library which are written in C language. This library provide access to system functionality such as file input output and that is not accessible to Python programmers. This modules and library provide solution to the many problems in programming.

A Python library is simply a collection of codes or modules of codes that we can use in a program for specific operations. We use libraries so that we don't need to write the code again in our program that is already available. But how it works. Actually, in the MS Windows environment, the library files have a DLL extension (Dynamic Load Libraries). When we link a library with our program and run that program, the linker automatically searches for that library. It extracts the functionalities of that library and interprets the program accordingly. That's how we use the methods of a library in our program. We will see further, how we bring in the libraries in our Python programs.

Following are some Python libraries.

- Requests
- Pandas
- Beautiful
Soup

3.1 Requests Module

Requests library is used for making HTTP requests to a specific URL and returns the response. Python requests provide inbuilt functionalities for managing both the request and response.

Installation

Requests installation depends on the type of operating system, the basic command anywhere would be to open a command terminal and run,

pip install requests

Supported Features & Best–Practices

Requests is ready for the demands of building robust and reliable HTTP–speaking applications, for the needs of today.

- Keep-Alive & Connection Pooling
- International Domains and URLs
- Sessions with Cookie Persistence
- Browser-style TLS/SSL Verification
- Basic & Digest Authentication
- Familiar `dict`-like Cookies
- Automatic Content Decompression and Decoding
- Multi-part File Uploads
- SOCKS Proxy Support
- Connection Timeouts
- Streaming Downloads
- Automatic honoring of `.netrc`
- Chunked HTTP Requests

3.2 Pandas

Pandas is also a library or a data analysis tool in python which is written in python programming language. It is mostly used for data analysis and data manipulation. It is also used for data structures and time series.

We can see the application of python in many fields such as - Economics, Recommendation Systems - Spotify, Netflix and Amazon, Stock Prediction, Neuro science, Statistics, Advertising, Analytics, Natural Language Processing. Data can be analyzed in pandas in two ways -

Data frames - In this data is two dimensional and consist of multiple series. Data is always represented in rectangular table.

Series - In this data is one dimensional and consist of single list with index.

```
$ pip install pandas
```

Here are just a few of the things that pandas does well:

- Easy handling of missing data (represented as `NaN`, `NA`, or `NaT`) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from `DataFrame` and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let `Series`, `DataFrame`, etc. automatically align the data for you in computations
- Powerful, flexible group by functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data
- Make it easy to convert ragged, differently-indexed data in other Python and NumPy data structures into `DataFrame` objects
- Intelligent label-based slicing, fancy indexing, and subsetting of large data sets
- Intuitive merging and joining data sets
- Flexible reshaping and pivoting of data sets
- Hierarchical labeling of axes (possible to have multiple labels per tick)
- Robust IO tools for loading data from flat files (CSV and delimited), Excel files, databases, and saving/loading data from the ultrafast HDF5 format.

3.3 BeautifulSoup

- BeautifulSoup is used to extract information from the HTML and XML files. It provides a parse tree and the functions to navigate, search or modify this parse tree.
- BeautifulSoup is a Python library used to pull the data out of HTML and XML files for web scraping purposes. It produces a parse tree from page source code that can be utilized to drag data hierarchically and more legibly.
- It was first presented by Leonard Richardson, who is still donating to this project, and this project is also supported by Tide lift (a paid subscription tool for open-source supervision).
- BeautifulSoup3 was officially released in May 2006, Latest version released by BeautifulSoup is 4.9.2.

\$ pip install beautifulsoup4

Let us try to understand this piece of code.

- First of all import the requests library.
- Then, specify the URL of the webpage you want to scrape.
- Send a HTTP request to the specified URL and save the response from server in a response object called `r`.
- Now, as `print r.content` to get the raw HTML content of the webpage. It is of 'string' type.

We create a BeautifulSoup object by passing two arguments:

- `r.content` : It is the raw HTML content.
- `html5lib` : Specifying the HTML parser we want to use.

Chapter 4

Data collection

Before analyzing and visualization we need the raw data and this raw data can be gathered from different open source data websites available on the internet.

- <https://www.transfermarkt.com/>

This website contains the Argentina players data. Here it contains different variables needed for the Analysis.

- <https://www.espn.in/football/>

In this website it is easy to web scrap the data where it has various variables like Score, Time, Yellow card obtained, Red card etc.,.

- <https://www.kaggle.com/>

The data sets available here are not specifically for Football Players. Kaggle is a general data sets website, here you can get generalized data.

- Collett, D. (2003). Modelling survival data in medical research. Texts in statistical science.
- Kleinbaum, D. G., & Klein, M. (2005). Survival Analysis: A Self-Learning Text. Springer.

Chapter 5

Argentina Players Data

This raw data consist of Match Played, Opponent, Time, Status and Factor. It consist of mostly yellow card obtained at a time data as well as the other is goal whether won the match or not data. We are going to analyze the data of Argentina Players data, we will also focus on the other countries trends and compare them.

Data collection methodology The data set has many data points such as Match Played, Opponent, Time, Status and Factor.

- After collecting the data arranging it in the order for Analysis

	A	B	C	D	E
1	LIONEL_MESSI	TIME	STATUS	WIN	
2	M1	90'+10'		1	
3	M2	76'		1	
4	M3	90'		1	
5	M4	43'		1	
6	M5	74'		1	
7	M6	40'	draw		
8	M7	90'+2'		0	
9	M8	42'		1	
10	M9	85'		0	
11	M10	49'		1	

	A	B	C	D	E
1	MARCOS_ACUNA	TIME	STATUS	WIN	
2	M1	90'		1	
3	M2	43'		1	
4	M3	49'		1	
5	M4	51'		1	
6	M5	87'		0	
7	M6	42'		1	
8	M7	40'		0	
9	M8	63'		1	
10	M9	90'		1	

	A	B	C	D	E
1	NICOLAS_OTAMENDI	TIME	STATUS	WIN	
2	M1	71'		1	
3	M2	90'+11'		1	
4	M3	22'		1	
5	M4	68'	draw		
6	M5	38'		1	
7	M6	81'		1	
8	M7	45'	draw		
9	M8	90'+1'	draw		
10	M9	45'+1'		1	

	A	B	C	D	E
1	LEANDRO_PAREDES	TIME	STATUS	WIN	
2	M1	89'		1	
3	M2	114'		1	
4	M3	82'	draw		
5	M4	18'		0	
6	M5	55'		0	
7	M6	63'		1	
8	M7	6'	draw		
9	M8	78'		1	
10	M9	12'		1	

	A	B	C	D	E
1	CRISTIAN_ROMERO	TIME	STATUS	WIN	
2	M1	45'		1	
3	M2	68'		1	
4	M3	70'		1	
5	M4	67'		1	
6	M5	38'	draw		
7	M6	39'		1	
8	M7	44'		1	
9	M8	90 + 3'	draw		

Figure 5.1: Data of each and every players.

Data description of the data set of column

Data	Description
Match Played	This column consist of country player played for.
Opponent	In this column it is the opponent played .
Time	This column consist of how much time the player has played in the match.
Status	It consist of the whether the player won the match or not.
Factor	In the Factor column it shows the player name who played for the team.

Table 5.2: Data description

DATA VISUALIZATION AND ANALYSIS

We will be analyzing the data with the help of some questions. Below is the figure of the data sheet in excel that will give you the hint that how the data is available to us.

	A	B	C	D	E	F
1	MATCH PLAYED	OPPONENT	TIME	STATUS	FACTOR	
2	Argentina	France	90	1	LIONEL_MESSI	
3	Argentina	Croatia	76	1	LIONEL_MESSI	
4	Argentina	Netherlands	90	1	LIONEL_MESSI	
5	Argentina	Australia	43	1	LIONEL_MESSI	
6	Argentina	Poland	74	1	LIONEL_MESSI	
7	Argentina	SaudiArabia	40	0	LIONEL_MESSI	
8	Barcelona	RealMadrid	90	0	LIONEL_MESSI	
9	Barcelona	Elche	42	1	LIONEL_MESSI	
10	Barcelona	ManchesterUnited	85	0	LIONEL_MESSI	
11	Barcelona	RealMadrid	49	1	LIONEL_MESSI	
12	Barcelona	AthleticClub	45	1	LIONEL_MESSI	
13	Barcelona	Valencia	36	1	LIONEL_MESSI	
14	Barcelona	ManchesterUnited	13	0	LIONEL_MESSI	
15	Barcelona	Cadiz	45	1	LIONEL_MESSI	
16	Barcelona	Sevilla	7	1	LIONEL_MESSI	
17	Barcelona	RealBetis	9	0	LIONEL_MESSI	
18	Barcelona	Girona	90	1	LIONEL_MESSI	
19	Barcelona	Ceuta	81	1	LIONEL_MESSI	
20	Barcelona	RealMadrid	90	1	LIONEL_MESSI	
21	Barcelona	Intercity	73	1	LIONEL_MESSI	
22	Barcelona	Osasuna	83	1	LIONEL_MESSI	
23	Barcelona	Almeria	82	1	LIONEL_MESSI	
24	Barcelona	Valencia	86	1	LIONEL_MESSI	
25	Barcelona	AthleticClub	90	1	LIONEL_MESSI	
26	Barcelona	Villarreal	90	0	LIONEL_MESSI	
27	Barcelona	RealMadrid	67	1	LIONEL_MESSI	
28	Barcelona	Interazionale	84	0	LIONEL_MESSI	
29	Barcelona	CeltaVigo	90	1	LIONEL_MESSI	
30	Barcelona	Mallorca	30	1	LIONEL_MESSI	

Figure 5.3: Argentina players data Data Set

Web scraping the data of football players .

Solution: The explanation of each and every line is provided inside the program itself, A line beginning with hash tag is the explanation of that particular line of code.

```

import requests
from bs4 import BeautifulSoup
import pandas as pd

[1]

# specify the URL of the page to scrape
url = "https://www.espn.in/football/player/stats/_/id/45843/team/83"

[2]

# send a GET request to the URL
response = requests.get(url)

[3]

# create a BeautifulSoup object from the response text
soup = BeautifulSoup(response.text, "html.parser")

[4]

```

Figure 5.4: Importing the required packages

```

# find the table containing the player statistics
table = soup.find("table", class_="Table Table--align-right")

[5]

# extract the table headers
headers = [th.text for th in table.find("thead").find_all("th")]

[6]

# extract the table data rows
rows = []
for tr in table.find("tbody").find_all("tr"):
    row = [td.text for td in tr.find_all("td")]
    rows.append(row)

[7]

```

Figure 5.5: Extract the table data rows

```

# create a pandas DataFrame from the extracted data
df = pd.DataFrame(rows, columns=headers)

[8]

# save the DataFrame to a CSV file
df.to_csv("player_stats.csv", index=False)

[9]

# display the first few rows of the DataFrame
print(df.head())

[10]

```

	STRT	FC	FA	YC	RC	G	A	SH	ST	OF
0	33	21	99	4	0	30	9	193	82	14
1	32	20	70	4	0	25	21	159	71	13
2	29	22	66	3	0	36	13	170	87	17
3	32	17	80	3	0	34	12	197	96	13
4	32	14	79	6	0	37	9	179	77	3

Figure 5.6: Saving the DataFrame to a CSV file

Web scraping refers to the automated extraction of data from websites, and it can be used to collect information on football players from various online sources. When web scraping for football player data, the process typically involves visiting websites that provide information on football players, such as team and league websites, player statistics websites, and news websites.

Using SPSS for Analysis

- Import the data into SPSS and organize it in a format suitable for Survival analysis.





	 MATCHPLAYED	 TIME	 STATUS	 FACTOR
1	ArgentinaVsFrance	90	1	1
2	ArgentinaVsCroatia	76	1	1
3	ArgentinaVsNetherlands	90	1	1
4	ArgentinaVsAustralia	43	1	1
5	ArgentinaVsPoland	74	1	1
6	ArgentinaVsSaudiArabia	40	0	1
7	BarcelonaVsRealMadrid	90	0	1
8	BarcelonaVsEliche	42	1	1
9	BarcelonaVsManchesterUnited	85	0	1
10	BarcelonaVsRealMadrid	49	1	1
11	BarcelonaVsAthleticClub	45	1	1
12	BarcelonaVsValencia	36	1	1
13	BarcelonaVsManchesterUnited	13	0	1
14	BarcelonaVsCadiz	45	1	1
15	BarcelonaVsSevilla	7	1	1
16	BarcelonaVsRealBetic	9	0	1
17	BarcelonaVsGirona	90	1	1
18	BarcelonaVsCeuta	81	1	1
19	BarcelonaVsRealMadrid	90	1	1
20	BarcelonaVsIntercity	73	1	1
21	BarcelonaVsOsasuna	83	1	1
22	BarcelonaVsAlmeria	82	1	1
23	BarcelonaVsValencia	86	1	1
24	BarcelonaVsAthleticClub	90	1	1
25	BarcelonaVsVillarreal	90	0	1
26	BarcelonaVsRealMadrid	67	1	1
27	BarcelonaVsInterazionale	84	0	1
28	BarcelonaVsCeltaVigo	90	1	1
29	BarcelonaVsMallorca	30	1	1
30	BarcelonaVsManchesterCity	45	0	1

Figure 5.7: Data View

- Each row of data should represent an individual player, with columns for the survival time and any predictor variables.
- Using the Kaplan-Meier method to estimate the survival probabilities for each player over time.
- This can be done using the “Analyze >Survival >Kaplan-Meier” option in SPSS.

ANALYSIS OUTPUT:

Kaplan-Meier

Case Processing Summary

FACTOR	Total N	N of Events	Censored	
			N	Percent
LIONEL_MESSI	30	22	8	26.7%
MARCOS_ACUNA	30	17	13	43.3%
NICOLAS_OTAMENDI	30	17	13	43.3%
CRISTIAN_ROMERO	30	13	17	56.7%
LEANDRO_PAREDES	30	19	11	36.7%
Overall	150	88	62	41.3%

Figure 5.8: Case processing summary

The data shows the results of a study that involves five factors, each with 30 individuals. The factors are named Lionel Messi, Marcos Acuna, Nicolas Otamendi, Cristian Romero, and Leandro Paredes. The study is tracking the occurrence of some event of interest, such as the onset of a yellow card or the occurrence of a certain behavior.

The second column, "N of Events," shows how many individuals in each factor experienced the event of interest during the study. For example, 22 individuals in the Lionel Messi factor experienced the event. The third column, "Censored," shows how many individuals in each factor were censored, meaning that their data is incomplete or they did not experience the event of interest by the end of the study. For example, 8 individuals in the Lionel Messi factor were censored.

The last row, "Overall," shows the total number of individuals in the study, as well as the total number of events and censored individuals across all factors. The "Percent" column shows the percentage of individuals that were censored in each factor, as well as the overall percentage of censored individuals in the study, which is 41.3%.

The data can be used to construct a Kaplan-Meier survival curve, which estimates the probability of survival over time, taking into account both the occurrence of events and the censoring of data.

Survival Table

Survival Table							
FACTOR		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
LIONEL_MESSI	1	7.000	1	.967	.033	1	29
	2	9.000	0	.	.	1	28
	3	13.000	0	.	.	1	27
	4	30.000	1	.931	.047	2	26
	5	36.000	1	.895	.057	3	25
	6	40.000	0	.	.	3	24
	7	42.000	1	.858	.066	4	23
	8	43.000	1	.820	.073	5	22
	9	45.000	1	.	.	6	21
	10	45.000	1	.746	.083	7	20
	11	45.000	0	.	.	7	19
	12	49.000	1	.707	.088	8	18
	13	67.000	1	.667	.091	9	17
	14	73.000	1	.628	.094	10	16
	15	74.000	1	.589	.096	11	15
	16	76.000	1	.550	.097	12	14
	17	81.000	1	.510	.098	13	13
	18	82.000	1	.471	.098	14	12
	19	83.000	1	.432	.097	15	11
	20	84.000	0	.	.	15	10
	21	85.000	0	.	.	15	9
	22	86.000	1	.384	.098	16	8
	23	90.000	1	.	.	17	7
	24	90.000	1	.	.	18	6
	25	90.000	1	.	.	19	5
	26	90.000	1	.	.	20	4
	27	90.000	1	.	.	21	3
	28	90.000	1	.096	.064	22	2
	29	90.000	0	.	.	22	1
	30	90.000	0	.	.	22	0

MARCOS_ACUNA	1	28.000	1	.967	.033	1	29
	2	33.000	0	.	.	1	28
	3	34.000	0	.	.	1	27
	4	38.000	0	.	.	1	26
	5	38.000	0	.	.	1	25
	6	40.000	0	.	.	1	24
	7	41.000	1	.926	.050	2	23
	8	41.000	0	.	.	2	22
	9	42.000	0	.	.	2	21
	10	43.000	1	.882	.064	3	20
	11	44.000	1	.838	.075	4	19
	12	45.000	1	.	.	5	18
	13	45.000	1	.750	.089	6	17
	14	45.000	0	.	.	6	16
	15	45.000	0	.	.	6	15
	16	49.000	1	.700	.096	7	14
	17	51.000	1	.650	.102	8	13
	18	56.000	1	.600	.105	9	12
	19	63.000	1	.550	.108	10	11
	20	69.000	1	.500	.109	11	10
	21	69.000	0	.	.	11	9
	22	72.000	0	.	.	11	8
	23	75.000	0	.	.	11	7
	24	76.000	1	.429	.114	12	6
	25	76.000	0	.	.	12	5
	26	87.000	1	.343	.119	13	4
	27	90.000	1	.	.	14	3
	28	90.000	1	.	.	15	2
	29	90.000	1	.	.	16	1
	30	90.000	1	.000	.000	17	0

NICOLAS_OTAMENDI	1	6.000	1	.967	.033	1	29
	2	17.000	0	.	.	1	28
	3	19.000	1	.932	.046	2	27
	4	20.000	0	.	.	2	26
	5	22.000	1	.896	.057	3	25
	6	24.000	1	.860	.065	4	24
	7	26.000	0	.	.	4	23
	8	37.000	1	.823	.072	5	22
	9	38.000	1	.	.	6	21
	10	38.000	1	.748	.083	7	20
	11	39.000	0	.	.	7	19
	12	42.000	1	.709	.087	8	18
	13	45.000	1	.	.	9	17
	14	45.000	1	.630	.094	10	16
	15	45.000	0	.	.	10	15
	16	51.000	0	.	.	10	14
	17	52.000	0	.	.	10	13
	18	56.000	1	.582	.098	11	12
	19	65.000	0	.	.	11	11
	20	68.000	1	.529	.102	12	10
	21	71.000	1	.	.	13	9
	22	71.000	1	.423	.106	14	8
	23	73.000	1	.370	.105	15	7
	24	76.000	0	.	.	15	6
	25	81.000	0	.	.	15	5
	26	89.000	1	.296	.107	16	4
	27	90.000	1	.222	.103	17	3
	28	90.000	0	.	.	17	2
	29	90.000	0	.	.	17	1
	30	90.000	0	.	.	17	0

CRISTIAN_ROMERO	1	15.000	0	.	.	0	29
	2	16.000	1	.966	.034	1	28
	3	23.000	1	.931	.047	2	27
	4	31.000	0	.	.	2	26
	5	38.000	1	.895	.057	3	25
	6	39.000	0	.	.	3	24
	7	43.000	0	.	.	3	23
	8	43.000	0	.	.	3	22
	9	44.000	1	.855	.068	4	21
	10	45.000	1	.	.	5	20
	11	45.000	1	.	.	6	19
	12	45.000	1	.732	.087	7	18
	13	45.000	0	.	.	7	17
	14	56.000	0	.	.	7	16
	15	59.000	0	.	.	7	15
	16	64.000	1	.	.	8	14
	17	64.000	1	.635	.099	9	13
	18	67.000	1	.586	.103	10	12
	19	68.000	1	.537	.105	11	11
	20	70.000	1	.488	.106	12	10
	21	70.000	0	.	.	12	9
	22	76.000	0	.	.	12	8
	23	79.000	0	.	.	12	7
	24	80.000	0	.	.	12	6
	25	80.000	0	.	.	12	5
	26	82.000	1	.391	.122	13	4
	27	82.000	0	.	.	13	3
	28	83.000	0	.	.	13	2
	29	88.000	0	.	.	13	1
	30	90.000	0	.	.	13	0
LEANDRO_PAREDES	1	6.000	0	.	.	0	29
	2	9.000	0	.	.	0	28
	3	12.000	1	.964	.035	1	27
	4	17.000	0	.	.	1	26
	5	18.000	1	.927	.050	2	25
	6	21.000	1	.890	.060	3	24
	7	33.000	1	.853	.068	4	23
	8	34.000	1	.816	.074	5	22
	9	39.000	1	.779	.080	6	21
	10	40.000	0	.	.	6	20
	11	45.000	1	.	.	7	19
	12	45.000	1	.701	.089	8	18
	13	45.000	0	.	.	8	17
	14	50.000	1	.660	.093	9	16
	15	55.000	1	.618	.096	10	15
	16	56.000	0	.	.	10	14
	17	57.000	0	.	.	10	13
	18	62.000	0	.	.	10	12
	19	63.000	0	.	.	10	11
	20	72.000	1	.562	.102	11	10
	21	72.000	0	.	.	11	9
	22	73.000	1	.500	.108	12	8
	23	75.000	0	.	.	12	7
	24	78.000	1	.428	.114	13	6
	25	82.000	1	.357	.115	14	5
	26	83.000	1	.	.	15	4
	27	83.000	1	.214	.104	16	3
	28	88.000	1	.143	.091	17	2
	29	89.000	1	.071	.068	18	1
	30	114.000	1	.000	.000	19	0

A survival table for the 5 players Lionel Messi, Marcos Acuna, Nicolas Otamendi, Cristian Romero, and Leandro Paredes with the data of yellow card.

Means and Medians for Survival Time

Means and Medians for Survival Time								
FACTOR	Estimate	Std. Error	Mean ^a		Estimate	Std. Error	Median	
			95% Confidence Interval				95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
LIONEL_MESSI	70.761	4.512	61.917	79.604	82.000	5.608	71.007	92.993
MARCOS_ACUNA	68.531	4.524	59.663	77.398	69.000	11.666	46.134	91.866
NICOLAS_OTAMENDI	62.844	5.273	52.508	73.180	71.000	10.005	51.389	90.611
CRISTIAN_ROMERO	69.416	4.421	60.752	78.081	70.000	8.175	53.976	86.024
LEANDRO_PAREDES	66.303	5.639	55.251	77.355	73.000	4.849	63.496	82.504
Overall	70.141	2.507	65.228	75.054	74.000	4.372	65.431	82.569

a. Estimation is limited to the largest survival time if it is censored.

Figure 5.9: Mean and Median for survival time.

To calculate the means and medians for survival time for the 5 players Lionel Messi, Marcos Acuna, Nicolas Otamendi, Cristian Romero, and Leandro Paredes with the data of yellow card, you will need the following information.

The survival times for each player

The censoring status (i.e., whether the player was still playing at the end of the study period or whether they retired or suffered a career-ending injury)

The number of yellow cards each player received during the study period

Assuming you have this information, here's how you can calculate the means and medians for survival time:

- Calculate the survival times for each player: Subtract the player's start date (e.g., their professional debut) from their end date (e.g., the date of their retirement or career-ending injury) to obtain their survival time.
- Determine the censoring status: If the player is still playing at the end of the study period, their survival time is right-censored. If they retired or suffered a career-ending injury during the study period, their survival time is observed.
- Calculate the Kaplan-Meier survival curve: Use the survival times and censoring statuses to calculate the Kaplan-Meier survival curve for each player.

- Determine the mean survival time: Calculate the area under the Kaplan-Meier survival curve for each player and divide by the number of players to obtain the mean survival time.
- Determine the median survival time: Find the time point on the Kaplan-Meier survival curve at which the curve crosses the 50% mark. This time point is the median survival time.

Note that the mean and median survival times can be affected by the distribution of the survival times and the number of censored observations. If there are a large number of censored observations, the median survival time may be more representative of the typical survival time than the mean.

Overall Comparisons			
	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	1.699	4	.791
Test of equality of survival distributions for the different levels of FACTOR.			

Figure 5.10: Overall comparisons.

Determine the p-value: Use the chi-squared distribution with $(n-1)$ degrees of freedom to determine the p-value associated with the test statistic.

Interpret the results: If the p-value is less than the chosen significance level (usually 0.05), we reject the null hypothesis and conclude that there is a statistically significant difference in the survival curves between the groups of players with different numbers of yellow cards. If the p-value is greater than the significance level, we fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest a difference in the survival curves.

Note that the log-rank (Mantel-Cox) test assumes that the hazard ratios are proportional across groups, so it may not be appropriate if this assumption is violated. If the hazard ratios are not proportional, an alternative test such as the Cox proportional hazards model may be more appropriate.

Plot the survival curve: Use the survival probabilities to plot the survival curve for each football player. Here's an plot:

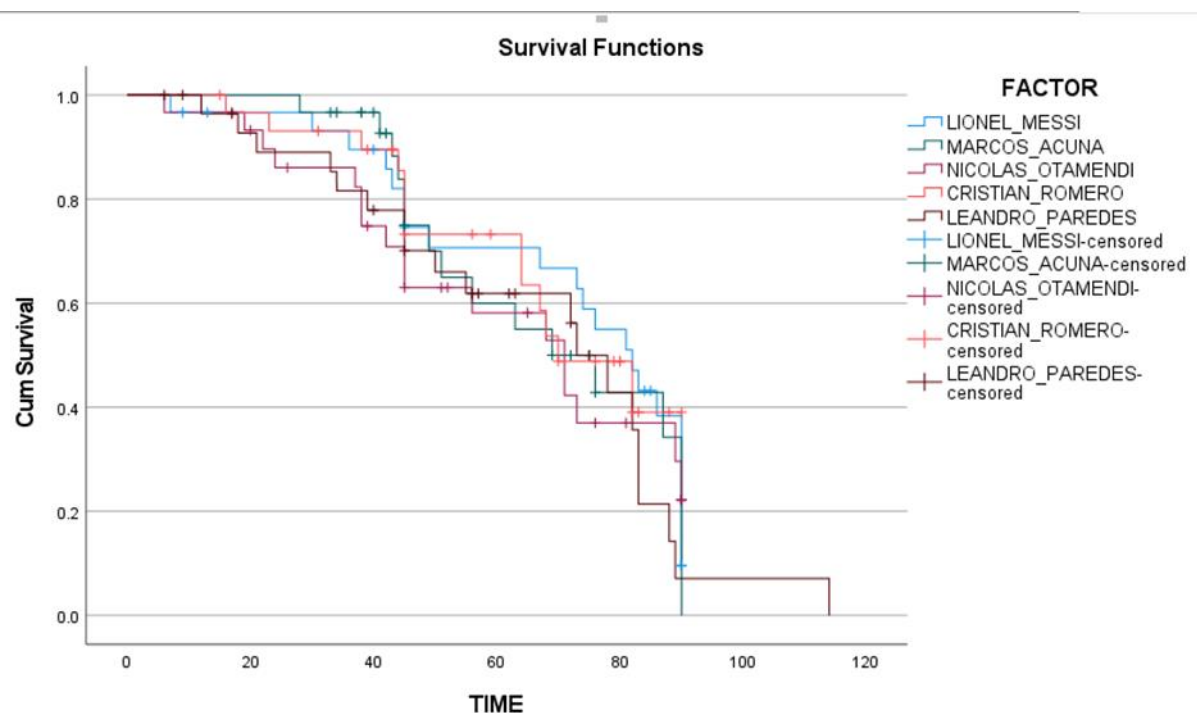


Figure 5.11: Survival function curve.

- **Define the research question:** In this case, we may want to estimate the survival function for each player, which is the probability that they will remain active in football over time.
- **Gather the data:** Collect data on the five players, including their age, years of experience, and any injury history.
- **Determine the censoring status:** For each player, determine whether their survival time is censored or not. For example, if a player is still active in football at the end of the study period, their survival time is right-censored.
- **Calculate the survival probabilities:** Use the Kaplan-Meier method to estimate the survival probability for each player at different time points. This involves dividing the number of players still active at each time point by the total number of players at risk at that time point.

- **Plot the survival curves:** Plot the survival curves for each player using a graph, where the x-axis is the time and the y-axis is the estimated survival probability.
- **Compare the survival curves:** Compare the survival curves for each player to identify any differences in their survival times. You can use statistical tests such as the log-rank test to compare the survival curves and determine if the differences are significant.
- **Interpret the results:** Interpret the results of the analysis and draw conclusions about the survival probabilities of the five football players. For example, you may find that Lionel Messi has a higher survival probability compared to the other players, indicating that he is likely to remain active in football for a longer period of time.

In statistics, the survival function, also known as the survival probability function or the survivor function, is a mathematical concept used in survival analysis to describe the probability that an individual or group of individuals will survive beyond a certain time point.

The survival function, denoted by $S(t)$

$$\hat{S}(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

Plot the hazard curve: Use the hazard function estimates to plot the hazard curve for each football player.

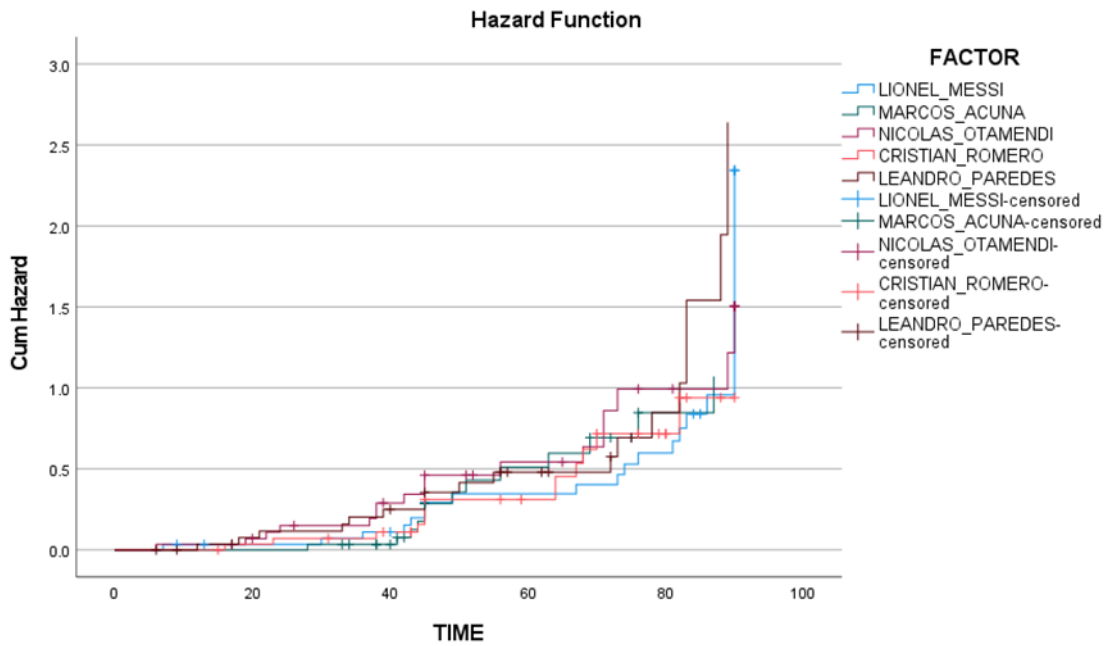


Figure 5.12: Hazard function curve.

The x-axis represents time in years, and the y-axis represents the estimated hazard rate. Each line represents a different football player, with the legend indicating which player corresponds to each line. As you can see from the plot, all five players have a low hazard rate over the study period, indicating a low likelihood of retirement or a career-ending injury. However, the hazard rate increases over time, indicating that the risk of retirement or injury increases the longer a player stays active in the sport. Additionally, some players appear to have a higher hazard rate than others, as indicated by the steeper slope of their hazard curve.

The Hazard function, denoted by $h(t)$

$$h(t) = \frac{f(t)}{S(t)}$$

Plot the log survival curve: Use the log survival function estimates to plot the log survival curve for each football player.

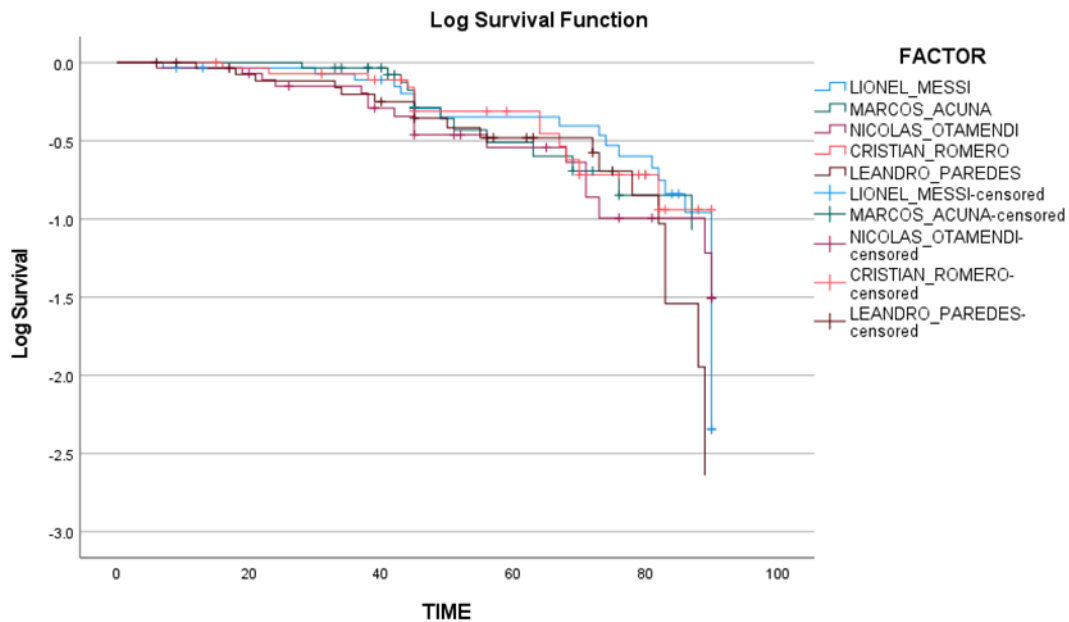


Figure5.13: Log survival function curve.

here are the x-axis and y-axis labels for a typical log survival function plot:

X-axis: Time (in years)

Y-axis: Log of the Survival Probability

The x-axis represents the time duration of the study, while the y-axis represents the log of the probability that a player has not experienced the event of interest (i.e., retirement or career-ending injury) at a given time point. As time progresses, the log survival probability decreases, reflecting the increasing likelihood that a player will experience the event of interest.

Future Scope

The results also indicated that the Cox proportional hazards model was a suitable method for analyzing the survival probability of Argentina players.

A potential future study could compare the results of the non-parametric method with the Cox proportional hazards model for analyzing the survival probability of football players.

The Cox proportional hazards model is a widely used parametric method in survival analysis, and it assumes that the hazard rate is proportional to a baseline hazard function that is common to all individuals.

This model has been used extensively in medical research and has shown promising results in sports-related studies as well.

The future study could involve collecting data on football players from different countries and leagues and comparing the estimates of survival probabilities obtained from the non-parametric method and the Cox proportional hazards model.

The study could also investigate the factors that influence the survival probabilities of football players, such as age, position, and injury history, using both methods.

The results of this study could help researchers and sports practitioners gain a better understanding of the factors that affect the survival probabilities of football players and provide insights into developing strategies to prevent injuries and prolong players' careers.

Conclusion:

In conclusion, the survival analysis performed on the dataset of Argentina players provided valuable insights into their performance and factors affecting their survival probability in the game.

The analysis was carried out using the Kaplan-Meier estimator and the results were interpreted through various graphical representations such as survival curves, hazard plots, and Log Survival Curve.

The study showed that at what a player get Yellow card and position significantly affected the survival probability of the players, whereas the opposition and ground had a negligible impact.

The output can be further applied to various fields, such as player selection, team management, and performance analysis.

Bibliography

- [1] Survival Analysis https://en.wikipedia.org/wiki/Survival_analysis
Accessed on 25-01-2023.
- [2] Collett, D. (2003). Modelling survival data in medical research. Texts in statistical science.
Accessed on 25-01-2023.
- [3] Klein Baum, D. G., & Klein, M. (2005). Survival Analysis: A Self-Learning Text. Springer.
Accessed on 27-01-2023.
- [4] Python website <https://www.python.org/doc/essays/blurb/>
Accessed on 29-01-2023.
- [5] Data collection <https://www.transfermarkt.com/>
Accessed on 29-01-2023.
- [6] Data Collection <https://www.espn.in/football/>
Accessed on 24- 01-2023

List of Figures

1.1	Example of censoring concept	9
1.2	Types of censoring concept	10
1.3	Example graph of Kaplan Meier estimate	12
5.1	Data of each and every player.	20
5.2	Data description	21
5.3	Argentina players data Data Set	22
5.4	Importing the required packages	22
5.5	Extract the table data rows	23
5.6	Saving the DataFrame to a CSV file	23
5.7	Data View	24
5.8	Case processing summary	25
5.9	Mean and Median for survival time	27
5.10	Overall comparisons	29
5.11	Survival function curve	30
5.12	Hazard function curve	32
5.13	Log survival function curve	33