**DATA SCIENCE PROJECT REPORT**

(Project Semester August-December 2020)

***LA-LIGA 2018 ANALYSIS***

Submitted by

SHANKAR KUMAR

Registration No-11811192

Programme and Section- Data Science (KM077)

Course Code INT217

Under the Guidance of

**Komal Arora : 17783**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

****

**CERTIFICATE**

This is to certify that Shankar Kumar bearing Registration no. 11811192 has completed Data Science project of INT217 titled, **“LA-LIGA 2018 ANALYSIS”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Shankar Kumar, student of Data Science under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 27/11/2020 Signature

Registration No. 11811192 Shankar Kumar

**ACKNOWLEDGEMENT**

A project work is a combination of views, ideas, suggestions and contribution of many people. Thus, one of the pleasant parts of writing the report is to thank those who have contributed towards its fulfilment.

I consider it as great privilege to have esteemed **Lecturer Ms. Komal Arora** as my project guide. I take this opportunity to express my sincere gratitude to him through constant advice and constructive criticism nourished my interest in the subject and provided a free and pleasant atmosphere to work against all odd situations. I avail this opportunity to extend my heart full thanks and deep respect to faculty member for their able guidance during this project.

My gratitude to all those, who **responded to my questionnaire** in a well-defined manner and helped me acquiring knowledge.

I would like to communicate a deep sense of gratitude to all these people without whom my project would not have been such a great learning experience.

*Shankar Kumar*

KM077

Reg no: 11811192

Lovely Professional University

**Table of Content**

|  |
| --- |
| Sr. No. Contents |
| 1. [Introduction](#Intro) |
| 1. [Objective/ Scope of The Analysis](#Obj) |
| 1. [Source of dataset](#Source) |
| 1. [ETL Process](#ETL) |
| 1. [Analysis on Dataset](#ANALYSIS) (for each analysis) |
| i) Introduction |
| ii) General Description |
| iii) Specific Requirement, functions and formulas |
| iv) Analysis Results |
| v) Visualization |
| 6. [List of Analysis with Results](#LIst) |
| 7. [References](#REFRENCES) |
| 8. [Bibliography](#REFRENCES) |

**Introduction**

La Liga is the name of the Spanish football league and has existed since 1929. The first division is called Primera División (which is commonly referred to La Liga) and the second is called Segunda División. Since 1997 there are 20 clubs participating in the top league of La Liga.

La Liga is one of the most popular professional sports leagues in the world, with an average attendance of 26,933 for league matches in the 2018–19 season. This is the sixth-highest of any domestic professional sports league in the world and the third-highest of any professional association football league in the world, behind the Bundesliga and the Premier League, and over the another two so-called "Big Five" European leagues, Serie A and Ligue 1.Also La Liga is the sixth wealthiest professional sport league in the world, after the NFL, the MLB, the NBA, the Premier League, and the NHL by revenue.

It has also produced the continent's top-rated club more times (22) than any other league in that period, more than double that of second-placed Serie A (Italy), including the top club in 10 of the 11 seasons between 2009 and 2019; each of these pinnacles was achieved by either Barcelona or Real Madrid.

**This dataset has the stats of all the 570 players in the season 2018-2019 in La Liga**.

The 2018–19 La Liga season, also known as La Liga Santander for sponsorship reasons, was the 88th since its establishment. The season began on 17 August 2018 and finished on 19 May 2019.

Fixtures for the 2018–19 season were announced on 24 July 2018. This was the first La Liga season using VAR.

Barcelona were the defending champions after winning La Liga the previous season for the 25th time. Huesca, Rayo Vallecano and Valladolid join as the promoted clubs from the 2017–18 Segunda División. They replaced Málaga, Las Palmas and Deportivo La Coruña who were relegated to the 2018–19 Segunda División.

Barcelona were crowned as league winners for the second season in a row on 28 April, after defeating Levante 1–0 with 3 games to spare.

Year after year, the analysis of actions and patterns occurring in a football match is becoming more complex.

Technology is mainly responsible for the avalanche of new kinds of datasets that analysts and data scientists working in football clubs have to deal with. In this way, every action occurring on the pitch is recorded and categorized, from passes to goals, but also tackles, shots, fouls,corners and dribbles. At the same time, the position of all players (including the referees) and the ball is recorded at rates up to 25 frames per second, which allows obtaining not only the position of players in real-time but also their speeds, accelerations and total distances covered.

The availability of these datasets has resulted in a diversity of new kinds of methodologies and metrics to understand what is happening on the pitch.

New points of view have arisen, such as evaluating the control of the pitch , measuring the area covered by the convex hull and tracking the evolution of the passing networks between players. Furthermore, new metrics have been defined to quantify the performance of specific actions such as the expected goal (xG) parameter, which quantifies the quality of a shot, or the post-shot expected goals (PSxG), defined for evaluating goalkeepers.

However, in this project, we investigated the overall analysis of all the team as well as their players. We analyzed the 380 matches of the 2018/2019 season of the Spanish national football league “LaLiga".

Our analysis focused on five issues , goals , defence , match participation , passing and fouls.

**Objective/ Scope of The Analysis**

This dataset has the stats of all the 570 players in the season 2018-2019 in La Liga.

Our goal is to analyse these stats and provide the overall analysis of the La-Liga teams as well as their Players based on this dataset.

List of Objectives for this project are :-

* GOALS ANALYSIS
* DEFENSIVE ANALYSIS
* PASSING ANALYSIS
* MATCH PARTICIPATION ANALYSIS
* FOULS ANALYSIS

These Analysis will gives an idea about :

* Performance of forwads of teams
* Defensive ability of different teams
* Performance of midfielders of different teams
* Players who less prone to injuries
* Teams who committed more mistakes

**Source of dataset**

* The dataset is taken from the Kaggle with the name ‘LaLiga Player Stats’.

[**https://www.kaggle.com/thegreatcoder/laliga-player-stats**](https://www.kaggle.com/thegreatcoder/laliga-player-stats)

* This includes the compiled stats of all the 570 players with goals, assists, appearances , etc. in the season 2018-2019 in La Liga**.**
* **Usage Information**
* License : CC0: Public Domaininfo
* Visibility : Public
* **Provenance**
* Sources : Sports bomb
* Collection methodology : Data extraction using Table Capture
* **Maintainers**
* Dataset owner: Prabhat
* **Updates**
* Expected update frequency : Annually
* Last updated : 2020-10-05
* Date created : 2020-10-04
* Current version : Version 2

**ETL Process**

ETL is the process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area and then finally, loads it into the Data Warehouse system.

Diagram

Description automatically generated

**ETL is a 3-step process :-**

**Step 1) Extraction**

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

Data warehouse needs to integrate systems that have different

DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, Point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others.

Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

Three Data Extraction methods :-

* Full Extraction
* Partial Extraction- without update notification.
* Partial Extraction- with update notification

Irrespective of the method used, extraction should not affect performance and response time of the source systems.

**Step 2) Transformation**

Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as direct move or pass through data.

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.

**A picture containing polygon

Description automatically generated**

Following are Data Integrity Problems:

* Different spelling of the same person like Jon, John, etc.
* There are multiple ways to denote company name like Google, Google Inc.
* Use of different names like Cleaveland, Cleveland.
* There may be a case that different account numbers are generated by various applications for the same customer.
* In some data required files remains blank
* Invalid product collected at POS as manual entry can lead to mistakes.

**Validations are done during this stage :**

* Filtering – Select only certain columns to load
* Using rules and lookup tables for Data standardization
* Character Set Conversion and encoding handling
* Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
* Data threshold validation check. For example, age cannot be more than two digits.
* Data flow validation from the staging area to the intermediate tables.
* Required fields should not be left blank.
* Cleaning ( for example, mapping NULL to 0 or Gender Male to "M" and Female to "F" etc.)
* Split a column into multiples and merging multiple columns into a single column.
* Transposing rows and columns,Use lookups to merge data
* Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically reject the row from processing).

**Step 3) Loading**

Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

**ETL process can also use the pipelining concept** i.e. as soon as some data is extracted, it can transformed and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed. The block diagram of the pipelining of ETL process is shown below:

Diagram

Description automatically generated

**ETL Tools: Most commonly used ETL tools are Sybase, Oracle Warehouse builder, CloverETL and MarkLogic.**

I have used Tableau prep as ETL tool.

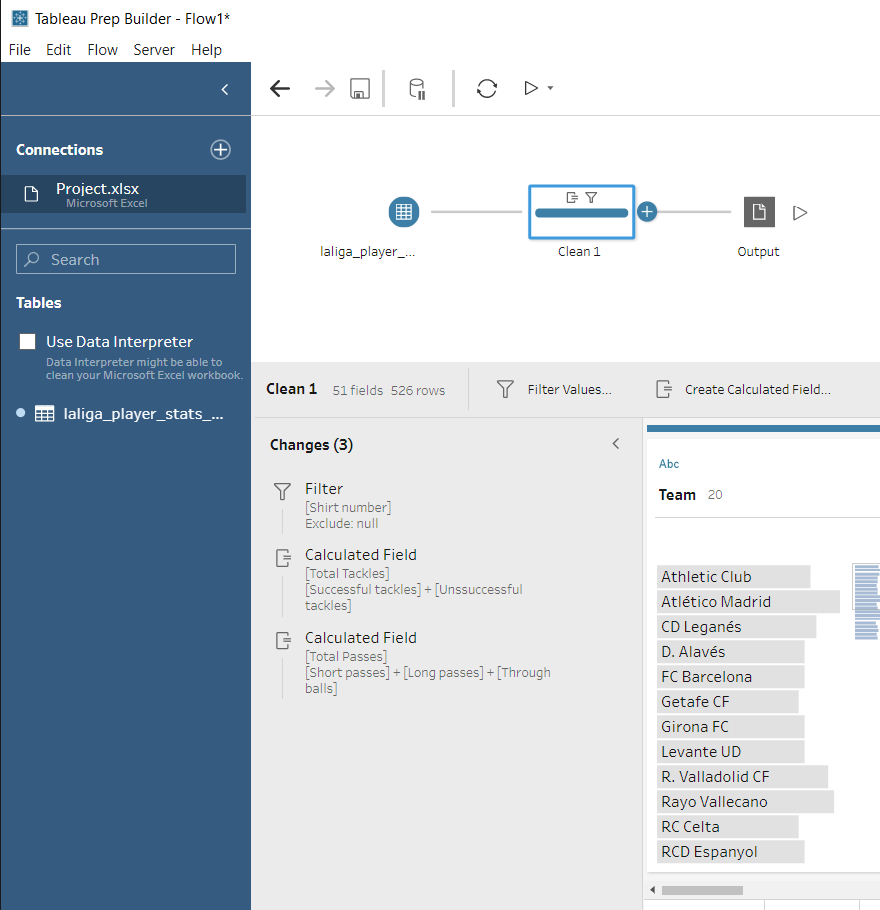
The Raw Dataset were having 62 columns and 570 rows.It contained a lot of null values and certain unwanted columns.

So, We have to do the preprocessing the data first before we could analyse the data.

* I have removed the null values in the dataset as they were the players which didn’t participated in the LaLiga 2018-19 season.
* I have also created two calculated fields in tableau prep which combined the total passess and tackles made by each players.
* I have also removed certain rows which didn’t have any significance in our analysis in excel.
* I have used pivot table and pivot chart for the analysis of objectives in Ms-Excel.

Graphical user interface

Description automatically generated



**Analysis on Dataset**

1. **GOALS ANALYSIS :-**
2. **Introduction : -**

* This Analysis gives an overview of the goals scored by different players as well as total goals scored by each team.
* In this analysis I have used columns minutes played, shots on target , goals scored , penalty scored and assists.

1. **General Description :-**

Minutes played :- This columns provides info. about the no. of minutes played by each players.

Shots on target :- This column tells about the no. of shots player has fired on goal-post.

Goals scored :-This column tells the no. of goals scored by each player.

Penalties scored :- This column tells the no. of penalties scored by each player.

Assists :- This column tells the no. of assists provided by the player.

* This objective has analysed all the minutes played, shots on target , goals scored , penalty scored and assists provided by the players.
* It also analyse the no of shots attempted on target and goals scored from that by each teams.
* This analysis also provodes slicers to filter any value (if we want) based on the teams ,position of player and percentage of games played by the players.

1. **Specific Requirement, functions and formulas :-**

* Pivot table with player name as row and minutes played , shots on target , goals scored , penalty scored and assists as values field.
* Pivot table with Teams as row and shots on target and goals scored as value field.
* 2 D column Pivotchart is used to analyse teams goals scored performance.

1. **Analysis Results :-**

* Out of all the teams , only Barcelona and real madrid players were able to fire 200+ shots on target.
* Barcelona were the best attacking team with highest shots on target as well as the highest no of goals scored in Laliga 2018-19 season.
* Messi of team Barcelona has highest no. of shots on target as well as highest no. goals in the LaLiga 2018-19 season.

**Table

Description automatically generated**

* **Stats of Barcelona and real madrid attacking players are :-**

**Graphical user interface, table

Description automatically generated**

1. **Visualization :-**

**Chart, bar chart

Description automatically generated**

**Chart, bar chart, treemap chart

Description automatically generated**

1. **DEFENSIVE ANALYSIS :-**
2. **Introduction : -**

* This Analysis gives an overview of defensive ability of different players as well as defensive strength of teams.
* In this analysis I have used the columns total takles , recoveries , interceptions , clearance and goals conceded while on pitch by each player.

1. **General Description :-**

Total Tackles :-This column tells about the total tackles made by each player including successful and unsuccessful tackles.

Interceptions :- This columns tells about the interceptions made by each player .

Recoveries :- This column tells about the ball recovered by each player.

Clearances :-This column tells abut the no of clearance made by each player.

Goals conceded while player on pitch :-This columns tells about the no of goals scored by opponents while he was on the pitch.

* This objective has analysed all the takles , recoveries , interceptions , clearance and goals conceded while on pitch by each player.
* It also analyse the total tackles and clearance by each teams.
* This analysis also provides slicers to filter any value(if we want) based on team and position of player.

1. **Specific Requirement, functions and formulas :-**

* Made a column total tackle by using calculated field in tableau prep.
* Pivot table with player name as row and takles ,recoveries , interceptions ,clearance and goals conceded while on pitch as values field.
* Pivot table with Teams as row and Total tackles and clearance as value field.
* 2 D column Pivotchart is used to analyse teams defensive strength of teams.

1. **Analysis Results :-**

* Atletico madrid had made the most no of tackles.
* D. Alaves had made the most no. of clearance.
* Atletico madrid were the most defensive teams in the La-Liga 2018-19 as they had the most no. of tackles as well as almost same ratio of clearance .

**Table

Description automatically generated**

Defensive Stats of Atletio madrid players :-

Graphical user interface, application

Description automatically generated

1. **Visualization :-**

**Chart, bar chart

Description automatically generated**

1. **PASSING ANALYSIS :-**
2. **Introduction : -**

* This Analysis gives an overview of passing ability of different players as well as of teams.
* In this analysis I have used the columns short , long , through ball passes and assists by each player.

1. **General Description :-**

Short passes :- This column represents the no. of short passess provided by each player.

Long passes :- This column represents the no. of long passess provided by the player.

Through balls :- This column represents the no. of through balls provided by the player.

Total Passes :- This column tells the total no. of passess provided by each player (which is sum of the short passes ,long passes and through balls) in the whole La liga season.

* This objective has analysed all the short , long , through ball passes and assists by each player.
* It also analyse the total passes by each teams.
* This analysis also provides slicers to filter any value(if we want) based on team and position of player.

1. **Specific Requirement, functions and formulas :-**

* Made a column total passes by using calculated field in tableau prep.
* Pivot table with player name as row and short , long , through ball passes and assists as values field.
* Pivot table with Teams as row and Total passes as value field.
* 2 D column Pivotchart is used to analyse teams Passing ability of teams.

1. **Analysis Results :-**

* Barcelona and real madrid teams had the best passing ability in the LaLiga 2018-19 as they were the only teams with 22000+ passes.
* **Table

  Description automatically generated**Out of all the teams in the LaLiga 2018-19, Barcelona players have the best passing ability.

**Graphical user interface

Description automatically generated**Graphical user interface, table

Description automatically generated

1. **Visualization :-**

**Chart, pie chart

Description automatically generated**

**Chart, pie chart

Description automatically generated**

1. **MATCH PARTICIPATION ANALYSIS :=**
2. **Introduction : -**

* This Analysis gives an overview of match participation of different players in the LaLiga 2018-19.
* In this analysis I have used the columns minutes played by each player,percentage of games where player are substituted ,percentage of games played ,percentage of game where player has started the game.

1. **General Description :-**

Minutes played :- This columns provides info. about the no. of minutes played by each players.

Percentage of games played :- This column tells the percent of games played by each players.

Percentage of games started :- This column tells the percent of games started by the player.

Percentage of games where substituted :- This column tells the percentage of games where player had been substituted.

* This objective has analysed minutes played by each player,percentage of games where player are substituted ,percentage of games played ,percentage of game where player has started the game.
* It also analyse the total players played by each teams.
* This analysis also provides slicers to filter any value(if we want) based on team and position of player.

1. **Specific Requirement, functions and formulas :-**

* Pivot table with player name as row and minutes played by each player,percentage of games where player are substituted ,percentage of games played ,percentage of game where player has started the game.as values field.
* Pivot table with Teams as row and Total players used in LaLiga 2018-19 as value field.
* 2 D column Pivotchart is used to analyse the total no. of players used by each team.

1. **Analysis Results :-**
2. **Table

   Description automatically generated**Atletico madrid and real Sociedad had used most no. of players in th LaLiga 2018-19.

**Graphical user interface, table, Excel

Description automatically generated**

**Graphical user interface, application, table, Excel

Description automatically generated**

1. **Visualization :-**

**Chart, bar chart

Description automatically generated**

1. **FOULS ANALYSIS :-**
2. **Introduction : -**

* This Analysis gives an overview of fouls committed and suffered by different players in the LaLiga 2018-19.
* In this analysis I have used the columns the fouls committed , fouls suffered , yellow cards and red cards given to each player.

1. **General Description :-**

Fouls suffered :- This column represents the no of fouls suffered by each player.

Fouls committed :- This column represents the no. of fouls committed by each player.

Yellow Cards :- This column represents the no. of yellow cards earned by the player.

Red Cards :- This column reprensents the no. of red cards earned by the player.

* This objective has analysed all the fouls committed , fouls suffered , yellow cards and red cards given to each player.
* It also analyse the total fouls committed and fouls suffered by each teams.
* This analysis also provides slicers to filter any value(if we want) based on team and position of player.

1. **Specific Requirement, functions and formulas :-**

* Pivot table with player name as row. and fouls committed , fouls suffered , yellow cards and red cards as values field.
* Pivot table with Teams as row and Total fouls suffered and total fouls committed as value field.
* 2 D column Pivotchart is used to analyse the fouls committed and suffered by each team.

1. **Analysis Results :-**

* Getafe had committed the most no. of fouls and were the only team who had committed 600+ fouls in LaLiga2018-19 season.
* Barcelona has suffered most no. of fouls followed by their rival real madrid.

**Table

Description automatically generated**

**Graphical user interface, table

Description automatically generatedGraphical user interface

Description automatically generated**

1. **Chart, bar chart

   Description automatically generatedVisualization :-**

**Chart, bar chart

Description automatically generated**

**List of Analysis with Results**

1. **Goals Analysis** :-

* Barcelona were the best attacking team with highest shots on target as well as the highest no of goals scored in Laliga 2018-19 season.

**Chart, bar chart

Description automatically generated**

1. **Defensive Analysis :-**

* Atletico madrid were the most defensive teams in the La-Liga 2018-19 as they had the most no. of tackles as well as almost same ratio of clearance.

**Chart, bar chart

Description automatically generated**

1. **Passing Analysis :-**

* Out of all the teams in the LaLiga 2018-19, Barcelona players have the best passing ability.

**Chart, pie chart

Description automatically generated**

1. **Match Participation Analysis :-**

* Atletico madrid and real Sociedad had used most no. of players in th LaLiga 2018-19.

**Chart, bar chart

Description automatically generated**

1. **Fouls Analysis :-**

* Getafe had committed the most no. of fouls and were the only team who had committed 600+ fouls in LaLiga2018-19 season.
* **Chart, bar chart

  Description automatically generated**Barcelona has suffered most no. of fouls followed by their rival real madrid.

**References and Bibliography**

* [**https://www.kaggle.com/thegreatcoder/laliga-player-stats?select=laliga\_player\_stats\_english.csv**](https://www.kaggle.com/thegreatcoder/laliga-player-stats?select=laliga_player_stats_english.csv)
* [**https://www.guru99.com/etl-extract-load-process.html**](https://www.guru99.com/etl-extract-load-process.html)
* [**https://www.stitchdata.com/etldatabase/etl-process/**](https://www.stitchdata.com/etldatabase/etl-process/)
* [**https://en.wikipedia.org/wiki/2018%E2%80%9319\_La\_Liga**](https://en.wikipedia.org/wiki/2018%E2%80%9319_La_Liga)
* [**https://en.wikipedia.org/wiki/La\_Liga**](https://en.wikipedia.org/wiki/La_Liga)
* **Microsoft Excel 2016 Bible: The Comprehensive Tutorial Resource by John Walkenbach, Wiley**