

Project Proposal

Requirements:

- Code to read in data
 - df =
- Description of:
 - Questions
 - **Domain:** Attackers.
 - Predict player market value (in euros) based on individual performance using linear regression. Assess our prediction using an SVM to determine if players are over/under valued.
 - Variables
 - Goals, assists, minutes played, club, age, time in league, nationality, position
 - Methods used
 - Regression analysis (linear, logistic), knn nearest neighbors, k means clustering, SVM for determining undervalue or overvalued by taking the linear regression model finding what their value should be based off the slope and seeing if predicted value is on par or some deviation and then categorize over/under by deviation.

Data:

- <https://github.com/dcaribou/transfermarkt-datasets>
 - Clean, structured and automatically updated football (soccer) dataset built from Transfermarkt data -- 79,000+ games, 37,000+ players, 1,800,000+ appearances and more, refreshed weekly.

Project Proposal

Group 11:

Aiden Rasmussen, Angela Lee, Eloy Park, Grace Im, Keshav Aditya, Solomon Zimmermann

Reading in Data

```
df = pd.read_csv('/Users/solomon/Downloads/Spring2026/STAT 451/Group Project/soccer_dataSet/players.csv.gz')
df.head(10)
```

	player_id	first_name	last_name	name	last_season	current_club_id	player_code	country_of_birth	city_of_birth	country_of_citizenship	...	age
0	10	Miroslav	Klose	Miroslav Klose	2015	398	miroslav-klose	Poland	Opole	Germany	...	35
1	26	Roman	Weidenfeller	Roman Weidenfeller	2017	16	roman-weidenfeller	Germany	Diez	Germany	...	34
2	65	Dimitar	Berbatov	Dimitar Berbatov	2015	1091	dimitar-berbatov	Bulgaria	Blagoevgrad	Bulgaria	...	33
3	77	NaN	Lúcio	Lúcio	2012	506	lucio	Brazil	Brasília	Brazil	...	32
4	80	Tom	Starke	Tom Starke	2017	27	tom-starke	East Germany (GDR)	Freital	Germany	...	31
5	109	NaN	Dedê	Dedê	2013	825	dede	Brazil	Belo Horizonte	Brazil	...	30
6	123	Christoph	Metzelder	Christoph Metzelder	2012	33	christoph-metzelder	Germany	Haltern	Germany	...	29
7	132	Tomas	Rosicky	Tomas Rosicky	2015	11	tomas-rosicky	CSSR	Praha	Czech Republic	...	28
8	162	Marc	Ziegler	Marc Ziegler	2012	79	marc-ziegler	Germany	Blieskastel	Germany	...	27
9	215	Roque	Santa Cruz	Roque Santa Cruz	2015	1084	roque-santa-cruz	Paraguay	Asunción	Paraguay	...	26

10 rows x 26 columns

Description:

Our group has decided to predict the market value of professional soccer players, specifically those in an attacking position. Using various machine learning techniques, we hope to accurately predict players' market values.

By using attacking player performance data, we aim to accurately predict a player's market value in euros and determine whether they are over- or undervalued relative to our prediction. To achieve this, we have selected these specific variables to help our models: *Goals, Assists, Minutes played, Age, Club, Time in League, Nationality, and Position*.

We will apply a variety of techniques; however, our Linear Regression model will be the primary tool for predicting market value based on performance metrics. A Support Vector Machine will then help us evaluate whether a player's actual value deviates from that prediction and categorize them as over- or undervalued based on the amount of deviation. From there, we will also use KNN and K-Means clustering to improve our predictions by comparing instances and grouping players with similar performance levels.

We initially sourced our information from the open-source [Transfermarkt Dataset](#) repository on GitHub. This is a comprehensive database, updated weekly, that gathers soccer data scraped from [Transfermarkt.com](#), one of the most reputable sites for player valuations and professional soccer stats. This dataset covers players, clubs, competitions, and much more, giving us a wide range of data to analyze.