**CCT College Dublin**

**Assessment Cover Page**

| **Module Title:** | Strategic Thinking |
| --- | --- |
| **Assessment Title:** | CA 3 - Capstone Project Final Submission |
| **Lecturer Name:** | James Garza |
| **Student Full Name:** | Kavi Patak |
| **Student Number:** | sba22391 |
| **Assessment Due Date:** | 12 May 2024 |
| **Date of Submission:** | 12 May 2024 |

****

**Declaration**

| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |
| --- |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Leveraging Machine Learning and Data Science

# for Competitive Advantage:

Estimating Football Player Market Values

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

by Kavi Patak

sba22391

Strategic Thinking (M1)

Lecturer: James Garza

CCT College, Dublin

May 12, 2024

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Table of Contents

[**Leveraging Machine Learning and Data Science**](#_3c6aa5ca8gi2) [**for Competitive Advantage: Estimating Football Players Market Value**](#_3iddv1ojtzjk)

[**Table of Contents 3**](#_ytyrg5mbj69k)

[**Introduction 4**](#_bdll0xhgx8or)

[**Problem Domain and Objectives 4**](#_iypgj6hbkao9)

[**Scope and Methodology 3**](#_k6trt3hy8gen)

[Consolidating and Characterising the Data 3](#_gltajlzf0k87)

[Exploratory Data Analysis (EDA) 5](#_cibvyqre24q5)

[Player Market Values Over Time 6](#_631arkk1y85s)

[Current Player Market Distribution 7](#_qyrf6ypqturk)

[Player Market Values by Domestic Competition 8](#_3h1iattw34b)

[Player Market Values by Position 9](#_9pi8w4z2ahcj)

[Player Market Values by Age 10](#_y3434t518ebd)

[Data Preprocessing 12](#_j1b1br31tmax)

[Feature Engineering 13](#_ej0u64o7esau)

[Model Selection 14](#_c8mmoqinq687)

[Training and Validation 15](#_c7br2kfle4b1)

[Evaluation Metrics 16](#_j095i2mly2f4)

[Tune Hyperparameters 16](#_inwe6h5orjjp)

[Producing Visualisations 16](#_itju3rh2sn73)

[Boundaries and Limitations 16](#_oi6ihiip8kwz)

[Results 17](#_yc9qwrfs2as)

[Model Evaluation 18](#_hw6xb2aabci7)

[Predictive Analytics and Interpretability 20](#_yl49ho5wnmn)

[Deployment 22](#_ytgddd7ozdrv)

[**Conclusion 22**](#_gwcfdkaxoadm)

[**Appendix A 23**](#_g1n2gsdqrtbz)

[**Appendix B 34**](#_imhrhjf7knau)

[**Appendix C 38**](#_niomfvb2ai5p)

[**Appendix D 40**](#_xt6xy04mv456)

[**References 42**](#_kya26suve4eo)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Introduction

Organisations have historically implemented enterprise systems such as enterprise resource planning in attempts to gain a competitive advantage (Goundar, 2021). The advent of the internet and proceeding developments in technologies like the Internet of Things(IoT), cloud computing, block chain, big data, Machine Learning (ML) and Artificial Intelligence (AI), have heavily influenced enterprises systems of today, and opened new avenues for conducting business.

Organisations have realised the potential value that resides in data and thus search for ways of utilising this valuable asset, it is here that Business Intelligence (BI) has become an important concept (Agarwal & Dhar 2014, cited in Persson and Sjöö, 2017). BI is an organisation's ability to effectively use the information it collects from daily enterprise (Vidal-García et al., cited in Niu et al., 2021). By identifying emerging opportunities, highlighting potential risks, providing useful insights and supporting decision making, ensure BI plays a significant role in optimising organisational effectiveness (Zhao et al., cited in Niu et al., 2021).

The role of analytics in football has evolved over the past decade, and will continue to do so. Technological developments continue to improve the volume and quality of data available to the world’s leading clubs, as well as the ability to derive insight from them. The opportunity exists for clubs of all sizes to use analytics to build a sustainable competitive advantage, something which will be most evident in the area where they invest most: the transfer market.

Having previously assessed the potential for a club to gain a competitive advantage through AI by means of Porter's Five Force Framework (Patak, 2023), this research will focus on leveraging machine learning and data analytics in player scouting and recruitment. More precisely, this report will outline the proposed steps in developing a football player value assessment model using machine learning techniques, and in doing so, aid a football club in making more objective and data driven transfer decisions.

This research project will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach in iterating through the stages of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

# Problem Domain and Objectives

The first phase of the CRISP-DM lifecycle calls for an understanding of the domain and business objectives, and extracting from this the requirements and goals for this project (Wirth and Hipp, 2000).

Football is a highly lucrative sport that depends heavily on the services of football players, the main suppliers to a football club (Patak, 2023). It is therefore no surprise that the majority of a club's expenditure is on player transfers and wages, with smaller clubs unable to compete financially with their larger counterparts (Metelski, 2021). Furthermore, the consequences of poor scouting and recruitment can be devastating to a club, regardless of size, and can have adverse affects not only from a business perspective (Depken and Globan, 2020), but to squad harmony, on pitch result, and even to a clubs reputation.

This project aims to level the playing field, at least financially in the transfer market, by developing a machine learning model to estimate the market value of football players based on various features such as age, goals, assists and other contributing factors.

Hypothesis: A fair and accurate player market value can be estimated through ML from the available features within the selected dataset.

In addition to this core objective, this project aims to provide insights into promising players who may be undervalued in the market. In-depth analyses shall be conducted on player performance data in specific positions, based on certain physical attributes, as well as exploring the relationship between a players age, value and statistics.

Finally, this project will compare the market value within different domestic leagues and investigate any changes in player market values over time.

# Scope and Methodology

Having developed a business understanding of the task at hand, and following the CRISP-DM methodology, the remaining steps in the development lifecycle call for Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

The CRISP\_DM methodology is “Agile” in nature, and unlike the traditional linear Waterfall lifecycle, the sequence of phases is not strict (Wirth and Hipp, 2000). In identifying a viable use case for this capstone project, both the Business Understanding and Data Understanding stages have been initiated as one is intrinsically linked to the other. A brief exploration of the datasets is conducted to confirm that it will satisfy its need.

Continuing on from that and adhering to the CRISP\_DM lifecycle, the following processes will be carried out and discussed in this report:

* Consolidating and Characterising the Data,
* Exploratory Data Analysis (EDA),
* Data Preprocessing,
* Feature Engineering,
* Model Selection,
* Training and Validation,
* Defining Evaluation Metrics,
* Tuning Hyperparameters,
* Producing Visualisations,
* Boundaries and Limitations,
* Results
* Model Evaluation,
* Predictive Analytics and Interpretability,
* Deployment,

### Consolidating and Characterising the Data

The selected dataset is called “Football Data from Transfermarkt” provided by Kaggle and available at: <https://www.kaggle.com/datasets/davidcariboo/player-scores/versions/284>

This dataset consists of nine CSV files with information on football competitions, games, clubs, players and player appearances. Each dataset contains a vast amount of observations (entries or rows) and numerous features (columns, attributes or variables). So too for missing and duplicate values. A more detailed overview of each file's feature names, data types, unique value counts and statistical makeup of numerical features, and missing values can be found in appendix A. A schema of the database design with attributes of the entity and the ID’s that are used to join them can also be found in the appendix. EDA and initial preprocessing is performed on individual datasets before merging the required data for further processing and modelling. The appearances dataset contains valuable performance statistics which needed to be collated along with game and player attributes. Following failed attempts at merging the data, the researcher is fortunate to find a coding solution which is modified for use. The original code is provided by Luis Gasper Cordeiro and found on David Coxon’s kaggle webpage (Cordeiro, n.d., in Coxen, n.d.). The consolidated datasets general information can be seen in Figure 1 below.

|  |
| --- |

Figure 1. Merged data general information

### Exploratory Data Analysis (EDA)

EDA is a process of examining the available data to discover patterns of correlation, identify trends, and gain insights of interest. It aids in spotting anomalies, testing hypotheses and clarifying any assumptions (Suresh Kumar Mukhiya and Ahmed, 2020). Initial EDA is performed by utilising Pandas Profiling which generates a comprehensive profile report on the data. The report consists of an overview of the datasets statistics, missing values, duplicates, attributes, as well as the correlation between variables. It is a quick and easy way to assess the data, especially with large datasets. This process aids in identifying outliers within the players dataset, with two player height values considerably lower than the mean for that attribute. These values were thus replaced with the correct values sourced from the transfermarkt website. On this second iteration, and on the consolidated data, we analyse the frequency distribution of ‘market\_value\_in\_eur’, the target variable more closely. Plotting the distribution using a Seaborn histogram (Figure 2) we find the data to be right skewed. Three transformations in Box Cox, square root and log are applied in attempts to achieve a Gaussian distribution, with the latter being the most successful (Figure 2). Univariate analysis continues by investigating the distribution of all numerical variables. Apart from a player's height, they too do not follow a normal distribution. Boxplots are created which highlight the existence of outliers, while also displaying the variance and spread of quantitative variables (Figure 2). The findings of this additional analysis proves invaluable in selecting the most suitable preprocessing methods for optimal ML modelling results. With the ultimate objective of evaluating a players market value, thorough EDA using numerous graphs, including scatterplots, barcharts, histograms, pie charts and boxplots are leveraged in plotting different categorical variables against player market value. The transfer market itself is first analysed for changes in market values over time and to establish the players market distribution by position, footedness and age group.

|  |
| --- |

Figure 1. Frequency distribution of Market Value (target variable)

|  |  |
| --- | --- |

Figure 2. Log Transformed distribution of Market Value (left) abd Box plot of variance and outliers (right)

#### Player Market Values Over Time

|  |
| --- |
|  |

Figure 3. Individual player market values over time

For the individual player transfer market values above (Figure 3), we see a strong positive relationship between the market value and year from 2004 to 2019 with a steady increase until falling off in the last three years.

|  |
| --- |

Figure 4. Averaged player market value changes over time

The averaged player market values over the same time is nonlinear with peaks in 2005, 2018 and 2023 and lows in 2003 and 2015 (Figure 4).

#### Current Player Market Distribution

|  |  |
| --- | --- |

Figure 5. Current market density by position (left) and sub position (right)

The pie chart above (Figure5 left) illustrates that the market of currently playing players is evenly distributed between Attack, Midfield and Defence. There are fewer players in the goalkeeper position which is to be expected as only one goalkeeper is ever selected in a starting eleven.

|  |  |
| --- | --- |

Figure 6. Current market density distribution by footedness (left) and age group (right)

We gather from the pie chart above left that the large majority of players are right footed. Only 25.67% are left footed which, along with the 3.96% for both footed players, may contribute to a higher market value due to scarcity.

As one would expect, approximately 67% of the player market is of players aged between 22 and 32 which is generally considered a players prime years. 21.91% consists of younger, up and coming players, with roughly 10% of players in the twilight of their professional careers.

#### Player Market Values by Domestic Competition

|  |  |
| --- | --- |

Figure 7. Boxplot of individual market values by domestic competition on the left and scatterplot of market value by age group on the right

The highest valued players currently play in LaLiga (Spain - ES1), Ligue 1 (France - FR1), Serie A (Italy - IT1), Premier League (England - GB1) and Bundesliga (Germany - L1). All five leagues also contain players whose market values far exceed the average within those leagues. This box plot also highlights the variance and outliers with regards to market value within each league (Figure 7 left). From the scatterplot on the top right we see that the highest valued players are generally under 27 years of age across all domestic competitions (Figure 7 right).

|  |  |
| --- | --- |

Figure 1.8.Individual player market values for each domestic competition

Attackers and Midfielders are valued the highest with Goalkeepers generally valued the lowest. Left footed players are highly valued relative to their market density. Three of the top seven highest valued players are left footed. Additionally they all apply their trade in the English Premier League (GB1) (Figure 8 right).

#### Player Market Values by Position

|  |  |
| --- | --- |

Figure 9. Individual player market values by position

|  |  |
| --- | --- |

Figure 10. Averaged player market values, by position and foot (left) and position and age group (right).

Left footed attackers seem to be valued the highest on average (Figure 10 left). This is little surprise following earlier analysis of current players and their footing which showed that only 25.67% of current players are left footed. There is clearly a shortage and demand for such players. The highest valued age category appears to be between the ages of 22 and 26. However, in the goalkeeping department, age and experience is slightly more valued (Figure 1 right).

#### Player Market Values by Age

|  |
| --- |

Figure 11. Individual player market values by age

We garner that the highest valued players at around 100 million euro or more are aged between 20 and 27 years (Figure 11).

|  |  |
| --- | --- |

Figure 12. Individual player market values by age, footedness (left) and position (right)

|  |  |
| --- | --- |

Figure 13. Averaged player market values by age, footedness (left) and position (right)

Across all domestic leagues, players of all footedness average a value between the ages of 22 and 32. The average market value of players is around 5 million or under across all leagues (Figure 13 left). The highest average market value across all leagues is for attackers aged between 22 and 26 with an average value of around 6 million euro. Midfielders are valued higher on average between the ages of 27 and 32 years old (Figure 13 right).

### Data Preprocessing

Data preprocessing is a crucial step in any Data Science projects lifecycle. It involves a number of operations and transformations being applied to raw data to make it suitable for analysis and modelling. The aim is to enhance the quality of the data by addressing missing or inconsistent values, handling various data types, including type casting, number formatting and label encoding categorical variables. Transforming the data using scaling, standardisation or normalisation techniques, and conducting feature engineering. Preprocessing is applied in an iterative manner throughout this project. As previously discussed, inconsistencies within the player height attribute are identified and replaced. The player's dataset contains 10919 missing market values. The player valuations dataset does not contain any. Original thoughts were to replace the missing values from the players dataset with those from the player valuations. On further investigation, this proved futile as the players' valuation market values are inconsistent, having been evaluated at different time periods. It was thus decided to impute missing values with the mean. A more refined and robust mean market value was aggregated by grouping players by position and age group, using a custom defined python function of conditional statements to replace the missing values. Improved EDA through univariate and bivariate analysis of variables highlighted the flaws in this process, due to the variance in the data and existence of multiple outliers. Two new imputation methods are applied on this second iteration. The first using the median values which are not influenced by extreme outliers, and the second using a KNNImputer which is a more robust imputation method based on using the mean values of k-nearest neighbours, making it less sensitive to outliers.

|  |
| --- |

Figure 14. Barchart of missing values for the players dataset

The same logic and steps are followed in imputing missing values for highest market value, player heights and remaining contract days attributes (Appendix B). As the foot feature is categorical, the mode is used to impute missing values based on a player's position and sub-position. EDA highlights the inclusion of retired players within the data, who’s attributes, mainly market values, are inaccurate based on their current age, performance and the current market. Retaining them in the dataset would only introduce noise. They are therefore removed along with any features that would not aid in the Machine Learning (ML) process. Unique identifiers such as player names, player id’s, date of births, demographics and unrelated features like images and urls are all removed. Duplicate entries too, are identified and removed. At this stage, the dataset still contains valuable categorical features including, a player position and sub-position, preferred foot, age-group, current club name and competition id. As the order or rank of these categories has no significance, one-hot label encoding is applied to create a new binary feature for each category, and assign a value of 0 or 1 depending on whether the player belongs to that category.

On the first run through, and for comparative purposes, multiple datasets are created based on the features, and the preprocessing applied to them. Following one-hot label encoding, a smaller dataset of a select few categorical features (to avoid the curse of dimensionality) is created, a dataset of all categorical features, and a dataset with features based on Pearson's Correlation. As the data contains outliers, a Robust Scaler is the first choice for scaling as it removes the median, and scales the data according to the Interquartile Range (IQR) (Nalcin, 2022). MinMax Scaler is also applied for comparative purposes, with new datasets created accordingly. It is worth noting that scaling is only applied to nonbinary numerical features so as to retain the value and impact of categorical features on the ML models.

However, not all machine learning models require this processing, with many producing exceptional results on categorical type data. With new insight and knowledge gained on models such as CatBoost Regressor and LightGBM, additional datasets are created on the second iteration, with categorical variables type cast from object data types to categorical. It is also important to mention that preprocessing is only applied following the split of data for training and testing. A step that was originally overlooked. By isolating the training from test data, we reduce the risk of data leakage, thereby allowing for more reliable and reproducible ML model results.

### Feature Engineering

Feature engineering is an invaluable process in developing and enriching ML models. It includes feature generation, feature extraction and feature selection. Feature generation entails creating new features from the existing features within the dataset. These new features are engineered to provide additional information that may be relevant and useful to the problem at hand.

A feature for a player's age is created by casting the date of birth feature to datetime and subtracting it from an instantiated object of the current datetime, before dividing by 365.25 for years. A feature for remaining contact days is created in a similar manner, by calculating the difference between the current time, and the contract expiration date feature. A feature to categorise players into age groups is created using a python function of conditional statements.

Further feature engineering includes creating features for goals, assists, yellow and red cards, minutes played, goals for, goals against and clean sheets for each of the last five years, as well as an accumulated total feature for each of these attributes. This was performed whilst merging the appearance, games and players datasets. Finally, features are engineered from the above features for minutes per a goal and minutes per an assist with the resulting NaN (not a number) and inf (infinity) values replaced with 1e400 to represent infinity. Unfortunately, these features are ultimately removed as they are too large for processing.

Feature extraction is the process of reducing the size or dimensionality of a large dataset while feature selection involves choosing a subset of the most relevant features, by removing redundant or irrelevant rows that may introduce noise or lead to overfitting (Rahul Kumar, 2019). To aid in feature selection, statistical testing is applied. Pearson's Correlation coefficients are generated to measure the correlations between independent variables to the target variable, ‘market\_value\_in\_eur’. Correlation coefficient values, and heatmap visualisations, can be found in appendix C. It was surprising to find that many of the categorical features for player position and age group do not rank high for correlation to the target variable. Playing in the Premier League (GB1), or for one of the bigger clubs in Europe, does however correlate high to market value. There is no universal correlation coefficient threshold value for retaining features, this is rather determined through domain knowledge and per use case. A refined dataset of the highest correlated feature was created on the first iteration. We found no improvement in model results, and chose to proceed with all numerical features on this latest iteration. Feature selection to combat overfitting can always be performed at a later stage, if required , additionally, many models accommodate for this, such as Ridge and Lasso regression through regularisation (Anwar, 2021). For feature selection of the categorical attributes, Analysis of Variance (ANOVA), or two

way factorial Analysis of Variance to be precise, is tested (qualtrics, n.d.). ANOVA is an extension of the t-test for independent samples, for more than two groups. It measures the difference in mean between the different categories of the independent variable with respect to the dependent variable (qualtrics, n.d.). The null hypothesis (H0) of ANOVA is that there is no statistical difference among the group means. If a difference exists, it has occurred “by chance”. On the other hand, the alternative hypothesis (Ha) is that at least one group differs significantly from the overall mean of the target variable (Kalyvas, 2024). ANOVA assumes a normal distribution, as such, is tested against the log transformation of the target variable. With all resulting p-values below a 0.05 (5%) significance level, we conclude that they all influence the target variable, and are thus retained for modelling.

### Model Selection

Model selection is choosing the appropriate machine learning algorithm based on the objective at hand. The objective is to predict the market value of a player, which is continuous data, making this a regression problem. As we are measuring the relationships between multiple independent variables to the target variable, it is a multiple regression problem (Indeed Career Guide, n.d.). Furthermore, we are using labelled data making this a supervised ML application. Many supervised ML algorithms offer classification and regression variants, with some excelling in one over the other. It is important to note that when developing a ML model, it is extremely unlikely to produce an optimal model on a first attempt. It is about trial and error, and continuous development and improvement. A linear regression model is implemented as a baseline model, followed by Kernel Ridge (KR), with different kernels applied for multiple linear regression on non linear data. Random Forest Regressor (RFR) and Gradient Boosting Regressor (GBR) are also selected, based on positive results found through literature reviews of similar applications (www.kaggle.com, n.d.).

For modelling on the latest datasets, we proceed in a more intuitive and efficient manner. PyCaret is an open-source, low-code, machine learning library in Python, that can automate machine learning workflows (Gitbook.io, 2023). We leverage this tool on all of the new and improved datasets to establish which are the most promising models for further optimisation. As many as 25 baseline models are tested. We find that the CatBoost Regression (CBR), indicated by CBR in Figure 15, returns the most promising and consistent scores overall, for all but two datasets which contain the categorical data type attributes. Although this model, along with LightGBM, excel at handling such data types, they need to be specified when initiating the model. In truth, and in comparison to results from the first iteration, the baseline results below are surprisingly good, across all datasets. LightGBM Regresor requires manual setup, and will be optimised along with CatBoost Regressor.

CatBoost or Categorical Boost in full, is a gradient boosting model that is designed to handle categorical features and even missing values, thereby removing the need for extensive preprocessing. This is done using an innovative method called ordered boosting. CatBoost is an ensemble model of decision trees, with each subsequent tree seeking to eliminate the errors of the previous one (GeeksforGeeks, 2023). LightGBM or Light Gradient Boosting Machine in full, is as the name suggests, another gradient boosting algorithm. It too incorporates tree based learning techniques and introduces the concept of leaf-wise tree growth, which chooses the leaf with the maximum gain to grow the tree. This reduces the memory requirements while enhancing model efficiency. LighGBM leverages two innovative techniques, Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), to overcome the limitations of traditional models (Technology, 2023).

|  |
| --- |

Figure 15. Bar chart of the best baseline models based on R² scores

### Training and Validation

This step requires splitting the dataset into a training and testing set. The train set which allows the model to learn, and the test set on which we test how well the model has learned. It is not uncommon for ML practitioners to split the dataset into three subsets, one each for training, validation/development and testing. The development set can be used to assess different models performance, tune parameters and minimise overfitting ( Rahul Kumar, 2019). Multiple instances of different training to testing set ratios are measured on the first application, including 70:30, 80:20 and 90:10, while studying the training error, validation error and evaluation metrics. We find the best results occurring on an 80:20, train:test ratio, and as such, apply this split before preprocessing the updated datasets.

In selecting and training a model, it is important to be aware of overfitting, underfitting, bias and variance. Overfitting occurs when you fit a model too precisely to the particularities of the training set, thereby obtaining a model that works well on the training set but is not able to generalise to new data. Essentially, the model is memorising the training data instead of learning the relationships within it (docs.aws.amazon.com, n.d.). This can occur when a model is too complex and learns the noise in the data instead of the underlying pattern. Having a complex model and too many features can also lead to high variance and low bias. Underfitting on the other hand occurs when a model is too simple and does not capture the underlying pattern in the data (Simplilearn.com, n.d.). An underfitting model with too few features will have low variance and high bias. Bias refers to the difference between the expected value of a model's prediction and the actual value. When a model makes test predictions, bias leads it to make inaccurate estimates (Hali, 2019). Variance refers to how much a model is dependent on the training data. It refers to the amount by which a model's prediction may vary for different training sets (Hali, 2019). Finding the right balance between bias and variance is essential in developing a model that generalises well. Generalisation refers to the ability of a model to perform well on new data. This means that the model is not underfitting nor overfitting and can be achieved by balancing the complexity of a model with the amount of training data available. Cross validation is used to estimate the ability of a model to generalise to new data. K-fold cross-validation is the process of splitting the data into k-many folds. A series of training and testing is then applied, holding one split of the data as the test set and the remainder as the training set. The test and training split are alternated k times, each time using a different fold for validation. An average score is calculated based on each fold to determine the overall performance of a given model. Cross validation is more stable and reliable than using a split in training and test set (Müller and Guido, 2017).

### Evaluation Metrics

Here we will define metrics to evaluate the performance of each model. For regression problems, Sklean provides Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared (R²) for more advanced metrics, with R² being more intuitive, and the preferred choice in evaluating regression models (Müller and Guido, 2017). Without delving too deep into the mathematics behind each metric, MAE measures the average absolute differences between predicted and actual values. MSE calculates the average of the squared differences between predicted and actual values. RMSE is the square root of MSE while R² measures the proportion of variance in the target variable that is predictable from the independent variables. R² ranges from 0 to 1 with a score of 1 indicating a perfect fit. For, MAE. MSE and RMSE, lower values indicate better performing models. As models are trained on different dataset, some with scaling or transformations applied, comparing their MAE, MSE, and RMSE becomes difficult. Cross Validated Mean R² (CV Mean R²) will be the determining metric for this project.

### Tune Hyperparameters

Hyper parameter tuning is the process of selecting the optimal set of hyperparameters for a machine learning model. This is an important step in the development of a model, as the choice of parameters can have a significant impact on performance. Manual, Random and Grid Search are all utilised in optimising models. For GBR applying the suggested values from Random Search did not contribute to a better model. Hyperparameters such as n\_estimators, learning\_rate, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf, n\_iter\_no\_change, subsample and validation\_fraction are all adjusted for GBR in attempts to counter overfitting while optimising CV Mean R². For KR, alpha, gamma, degree, coef0 and the kernel are all adjusted in optimising the model with positive results. For CBR, hyperparameter tuning through GridSearchCV was unsuccessful overall. Multiple values for depth, iterations, 12\_leaf\_reg, learning\_rate and early\_stopping\_rounds are specified in attempts to optimise the model and counter overfitting. Although improvements are made, the model performs comparably well with default settings. Parameters tuned for LightGBM include n\_estimators, learning\_rate, num\_leaves and max\_depth.

### Producing Visualisations

Visualisations provide a clear and intuitive way to present data, making complex patterns, trends, outliers and relationships easier to understand, without the need for attentive processing (McQuaid, n.d.).

Graphics and plots are leveraged throughout the development process, in EDA, feature engineering, modelling and finally to compare and communicate the research findings effectively. Open source libraries like Matplotlib, Seaborn and Yellowbricks are utilised using Python code in an interactive Jupyter Notebook environment, in developing and documenting the project.

### Boundaries and Limitations

It is important to note that each club is unique and so to their structure, ambitions and vision. What is right for one club may not be for another. Some aspire to win the league while others to avoid relegation. Other business models may depend solely on selling players for profit (Sloane, cited in Van den Berg, 2011). Hence aligning and framing scouting requirements with the vision of a club is crucial, as is the recognition that talent id is one small part of squad evolution and building. Another challenge may reside in harnessing the data and bringing together different data sources. There are multiple factors that contribute to evaluating a player's value, some of which may not be available within the selected dataset. Aspects such as exchange rates and social or economical factors like war, recession or a worldwide pandemic may prove challenging to account for (Metelski, 2021).

### Results

|  |
| --- |

Figure 16. Table of Model Results

|  |
| --- |

Figure 17. Barchart of the Best Model Results for R² scores

The best performing models on the new datasets, based on R² scores, can be seen above (Figure 17), and all metrics can be easily compared to the first iteration results, and LR baseline model. The first LR model's results are poor, which is expected, given that it is a linear model being applied for multiple linear regression on nonlinear data. The model is underfitting with poor training and testing results. RFR produced an overfitting model with a high training score of 0.9597334032950553 and a low test score of 0.7572457452441076, relative to training. Kernel Ridge on a polynomial kernel exhibited surprisingly good results on the first iteration, with a training R² of 0.944062, and test R² score of 0.943771. On further investigation, the results were indeed too good to be true, having mistakenly trained the model on the complete dataset. Its CV Mean R² score has however marginally increased from 0.704091 to 0.7221912831991801 on better processed data. These results are still on the mean imputed datasets.

GBR provided a slightly better CV Mean R² of 0.7213522752382111. However the model is also clearly overfitting to the training data with a considerable difference in training to testing results.

From the barchart in Figure 17, we appreciate the improvements made on Cross Validated R² scores from the earlier models. Both CBR and LightGBM algorithms have provided consistently higher results, predominantly on the median imputed datasets, with categorical data types attributes which have not been normalised or label encoded.

LightGBM has produced the best results. The model achieves a CV Mean R² score of 0.823 which increases to 0.852 when optimised and trained on the log transformation of the target variable. In comparison to the baseline model, we have achieved an 84.5% improvement.

### Model Evaluation

|  |  |
| --- | --- |

Figure 18. Residuals Plot (left) and Prediction Error graph (right) for the Best LightGBM Regressor Model

|  |
| --- |

Figure 19. Learning curve for the Best LightGBM Regressor Model

|  |
| --- |

Figure 20. Feature Importance for the Best LightGBM Regressor Model

In evaluating regression models, a residual is a measure of how far away a point is vertically from the regression line (Gohar, 2020). They are the difference between our predicted player market values and their actual market values (i.e. predicted — actual values), and a residuals plot displays the relationship of errors, with predictions, as well as the distribution of these errors.

From the residuals plot in Figure 18, we find little noticeable difference or pattern in error for lower market values on the left. The model appears to predict well for lower market values, densely located around the origin. However, as predicted values increase, we see a steady increase in prediction errors for the test data displayed in purple. As our data is transformed, we are unable to utilise Yellowbricks library for creating the plot, which would also provide its frequency distribution. We can nevertheless establish a normal distribution by eye, based on the even or symmetrical spread of purple dots about the origin line (Gohar, 2020).

Through the prediction error plot in Figure 18, we visualise the variance in the model. The prediction error plot displays the actual targets from the data against the predicted values generated by the model. The closer the broken lines align, the better the model is performing. We find the model performs best when predicting around 14.

Although we could have determined an overfitting model from the training and testing R² scores in Figure 16, the learning curves in Figure 19 do offer hope for future improvement, as they indicate an increase in scores with an increase in training instances (Brownlee, 2019).

To conclude the evaluation of the model, we leverage an inbuilt LightGBM function for feature importance. The horizontal barchart in Figure 20 provides direct insight into the contributing features in descending order of importance. We find that the model considers a players club name, age, remaining contract days, height and sub position the most, when making a prediction on their market value. This is valuable information for presentation to stakeholders, for model explainability and informed decision making.

### Predictive Analytics and Interpretability

Once satisfied with the training and performance of a model, it can be used to make predictions on new data. In interpreting the results both globally and locally, we utilise the SHAP library, designed for explainable AI. SHAP Produces shapley values which help us understand how a model's features impact its predictions. It uses a game theoretic approach that measures each player's contribution to the final outcome.

The SHAP summary plots in Figure 21,provide a visual representation of the absolute SHAP values for each feature, showing their importance relative to the model's output. The x-axis of the summary plot typically displays the average absolute SHAP value for each feature across all instances in the dataset. Therefore, the values on the x-axis represent the magnitude of the impact of each feature on the model's predictions. The beauty and advantage of this SHAP plot is that values on the x-axis are usually displayed in the same metric as the target variable. However, as this model is developed on the log transformation of the target variable, 'market\_value\_in\_eur', it may not be as insightful. Nonetheless, in the plot on the left, we understand that the model is most influenced by a player's current club domestic competition, current club name, their total goals scored, goals scored in the year 2022, age and total games played. Both graphics display features in order of importance. The visual on the right provided additional insight. Each point represents an entry or player from the dataset, while the colour of each point represents the value of the corresponding feature, with red indicating higher value and blue indicating lower value to the model's prediction. Taking age and remaining contract days as examples, through the visual we garner that the lower a player's age, the higher the predicted market value, while the higher a player's remaining contract days, the higher the predicted market value.

In Figure 22 and 23, SHAP values add up to the difference between the expected model output and the actual output for a given input. By doing so, SHAP values provide an accurate and local interpretation of the model's prediction for a given input (Readthedocs.io, 2024). For a random player, in Figure 22 we find a measurable impact of features to the model's prediction. As with the previous plots, ideally this value would be in euros, the same metric as the target variable. We see that the age feature value at 36 is bringing down the predicted market value while the current club name being Real Madrid is contributing to a higher predicted market value.

|  |  |
| --- | --- |

Figure 21. SHAP Feature Importance for the Best LightGBM Regressor Model

| Figure 22. SHAP force for local interpretation |
| --- |
|  |

Figure 22. SHAP force for local interpretation on the Best LightGBM Regressor Model

### Deployment

|  |
| --- |

Figure 23. LightGBM Regressor Model for PLayer Market Value Prediction Pipeline

The Deployment phase within the CRISP-DM methodology involves implementing the solutions derived from modelling into practical use. It marks the culmination of the data mining process by applying these solutions and models into operational settings. This includes integrating them with existing systems, providing documentation, and offering support to end-users to derive business value. Although perhaps premature, a pipeline of the model has been generated.

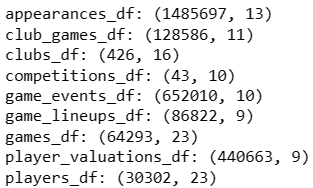
# Conclusion

Player scouting and market value analysis are key areas where ML can provide valuable knowledge and insights to the football industry. This capstone project aims to contribute to this by empowering clubs and stakeholders with data driven decision making capabilities for a competitive advantage. By leveraging machine learning on the “Football Data from Transfermarkt'' dataset, the project aligns with the evolving landscape of sports analytics, providing practical solutions to real world challenges in the football industry. In adhering to the CRISP\_DM life cycle, the application of this project has proven challenging, due to data complexities, time restraints, and computational limitations. The results however, are quite promising, with as much as an 84.5% improvement on select baseline metrics. The project still provides plenty of scope for improvement, at each step of the development process. Further statistical analysis can be conducted for additional insight into the characteristics of the datasets. Different techniques can be applied in preprocessing, including alternative imputation and scaling methods. Additional feature engineering, including dimensionality reduction through PCA, and feature generation, by resolving the infinity problem can be achieved. If anything, whilst modelling, this iteration has highlighted the importance of individually garnering, tuning and optimising a model. The most successful model is produced with the LightGBM algorithm. A model that does not even rank in the top ten, when comparing baseline models for further optimisation using Pycarot. This highlights the hidden potential for improvement within each model, and emphasises the need for continual research and improvement.

Four Jupyter notebooks documenting the process and python code are created, one each for EDA, Preprocessing and Modelling Iterations 1 and 2. Unfortunately the Preprocessing notebook is too large to be pushed to the Github repository. A pdf of the notebook has been uploaded instead.

Github: <https://github.com/KaviCCT/Strategic_Thinking_CA2>

# Appendix A



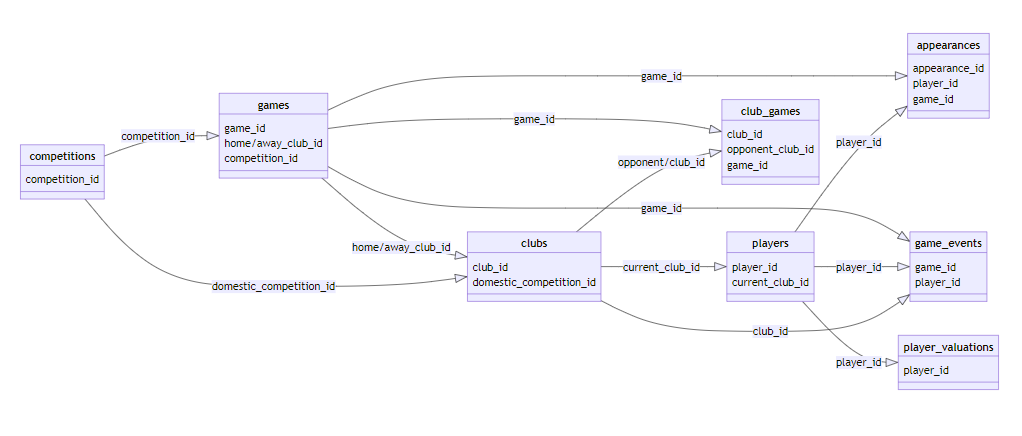
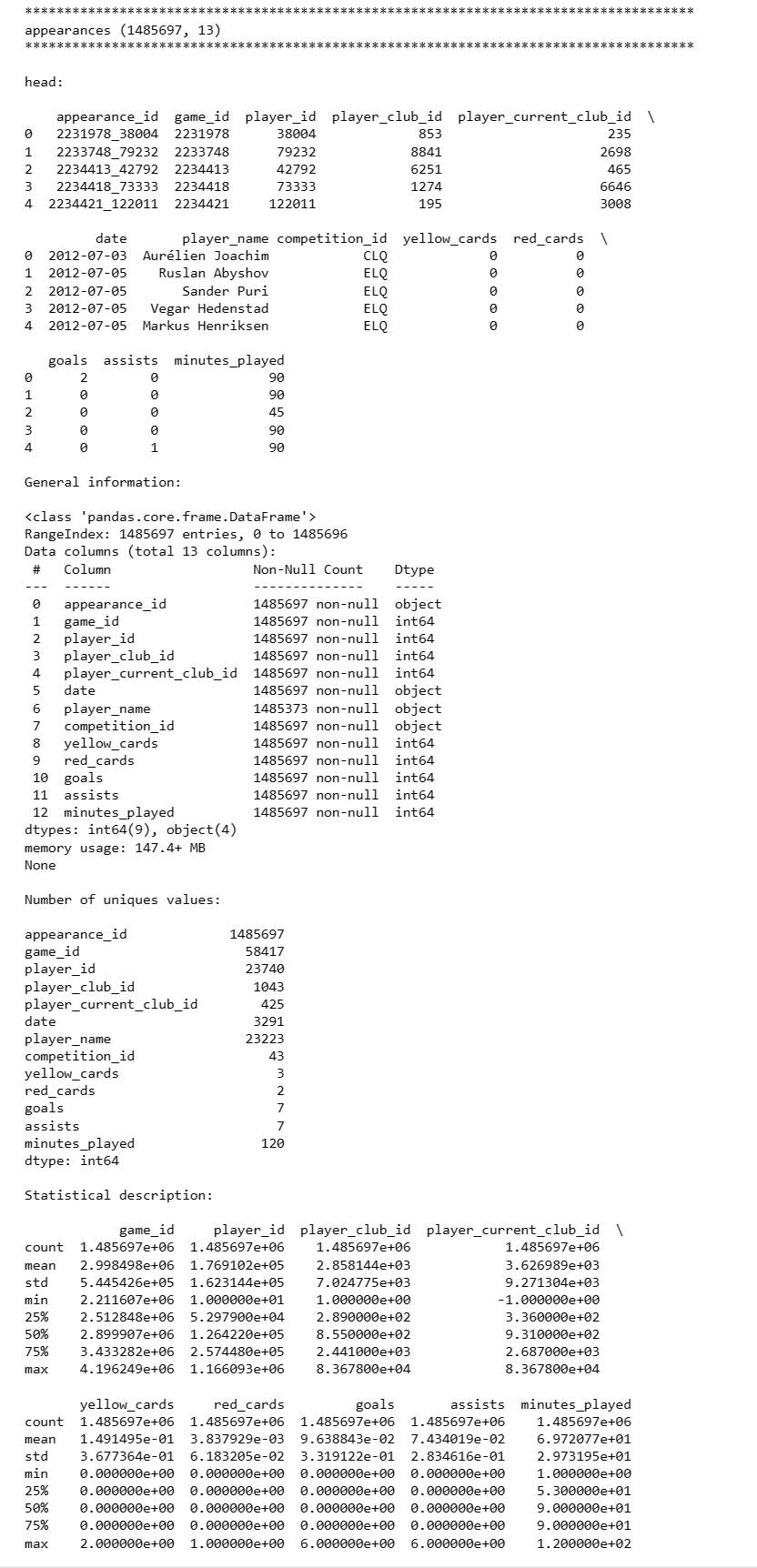
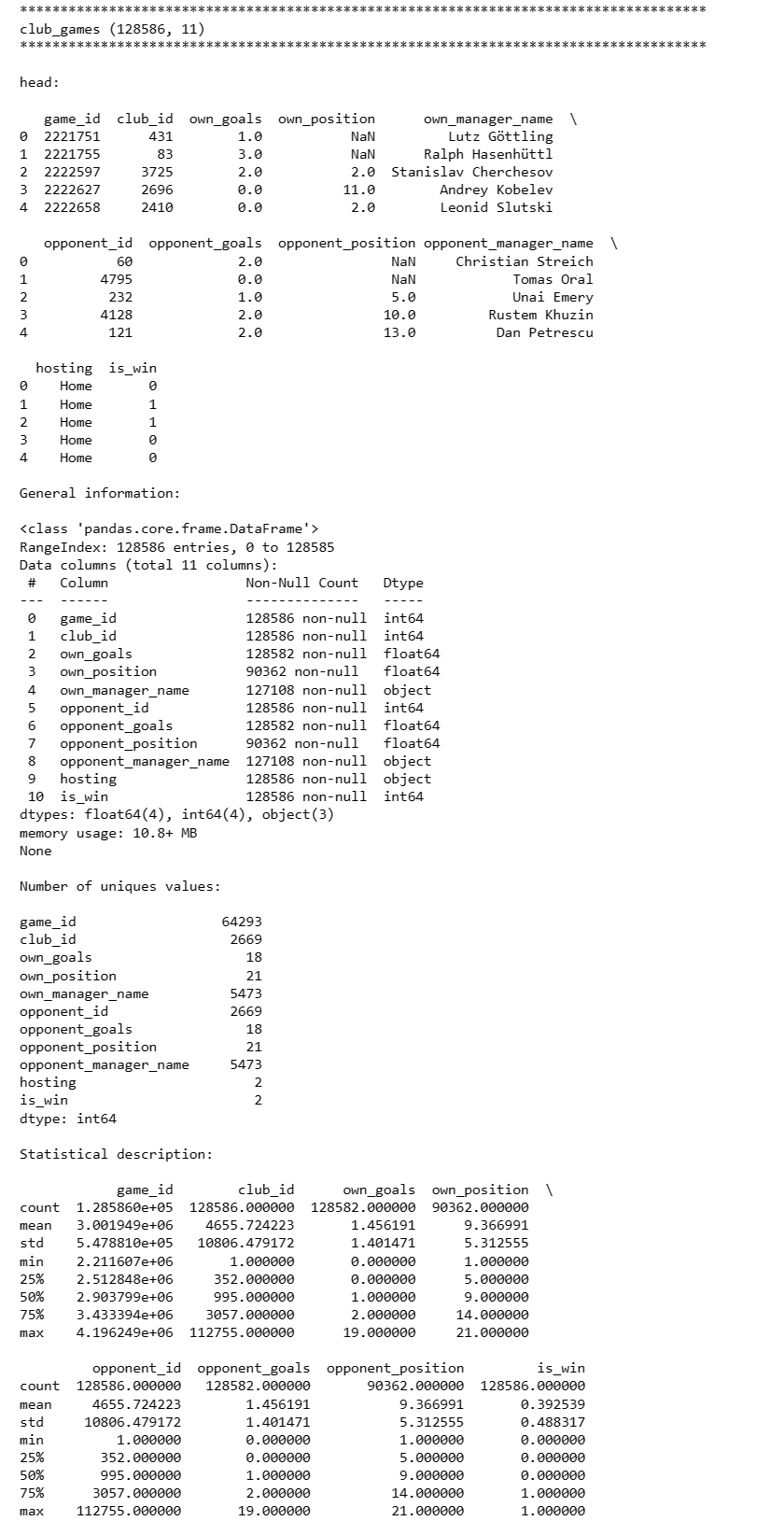
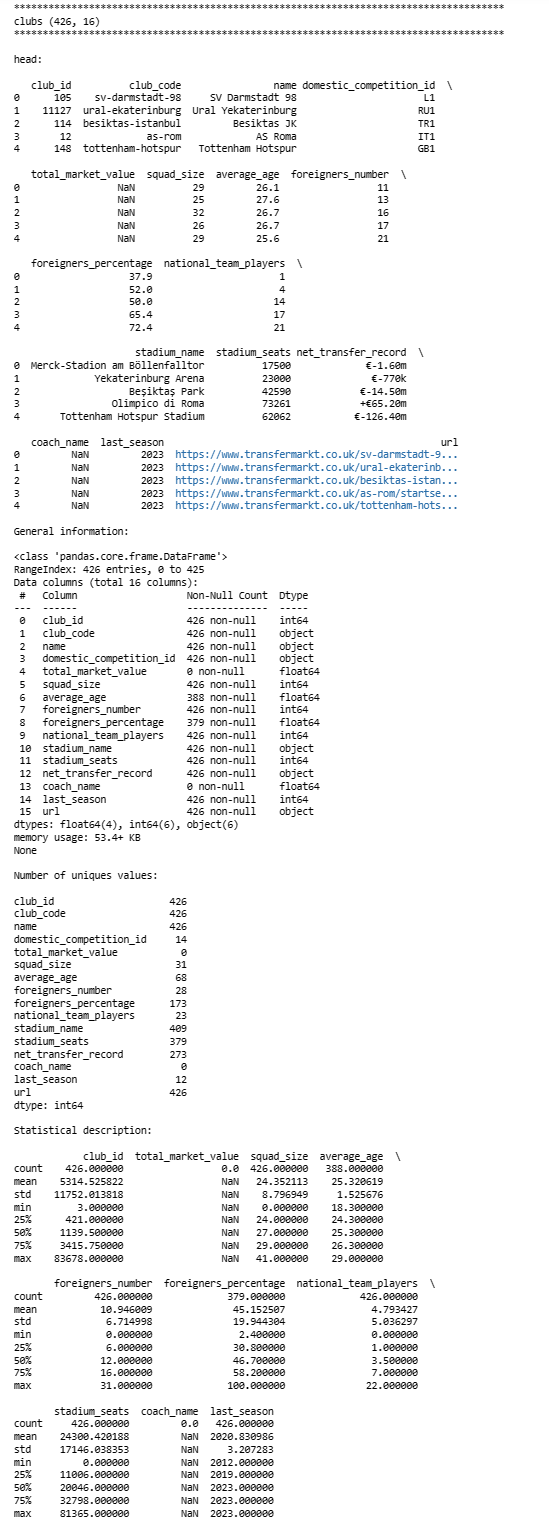


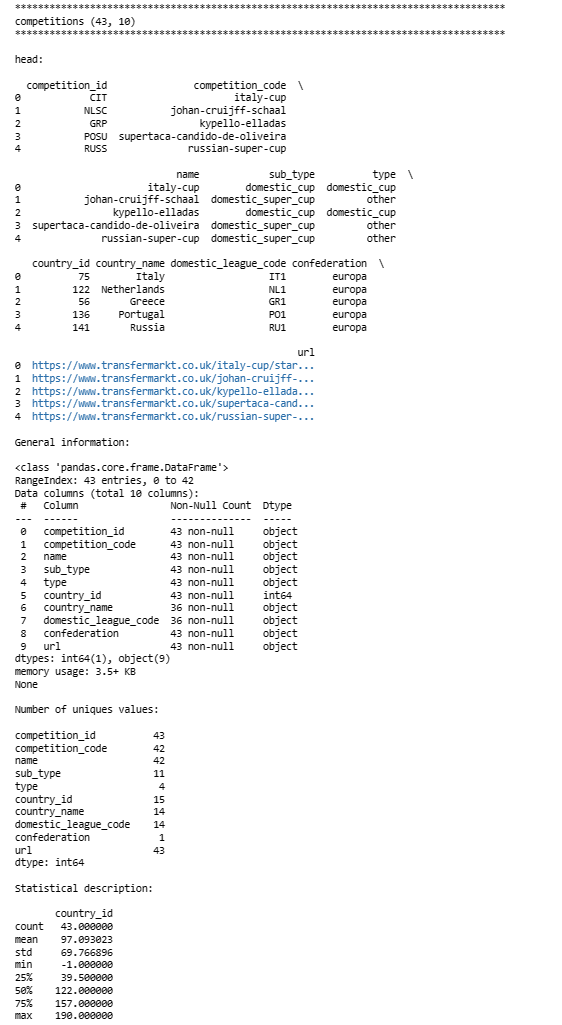


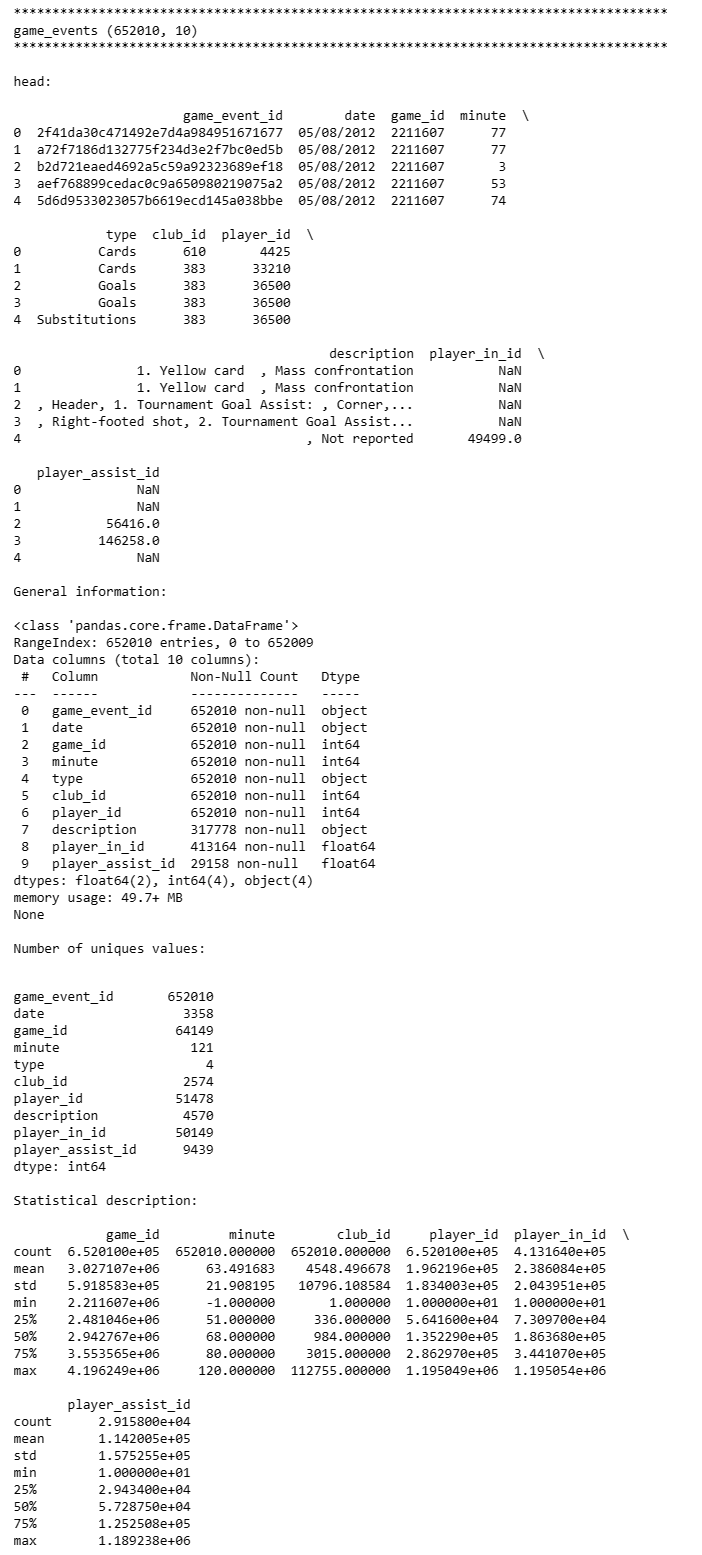
Figure 1.1. Dataset names, shapes and missing values



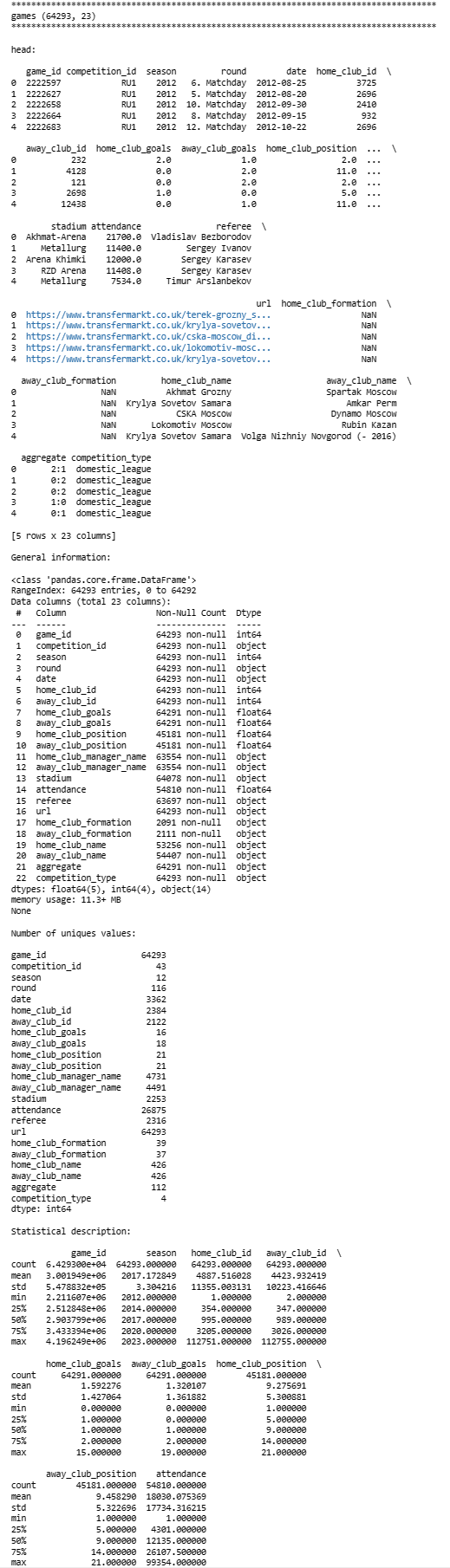




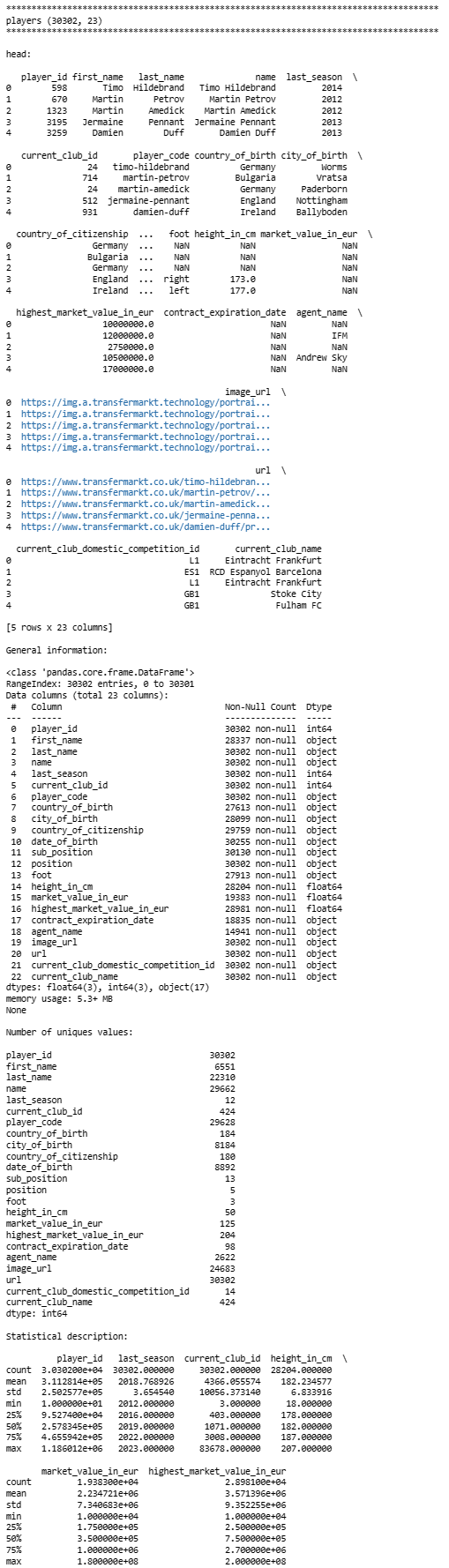












# Appendix B

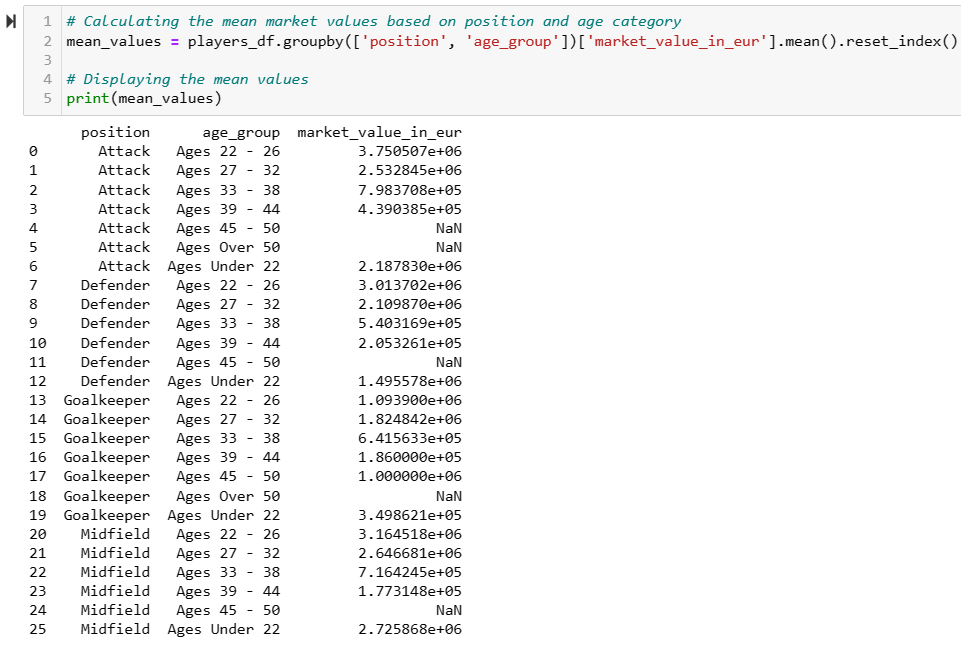
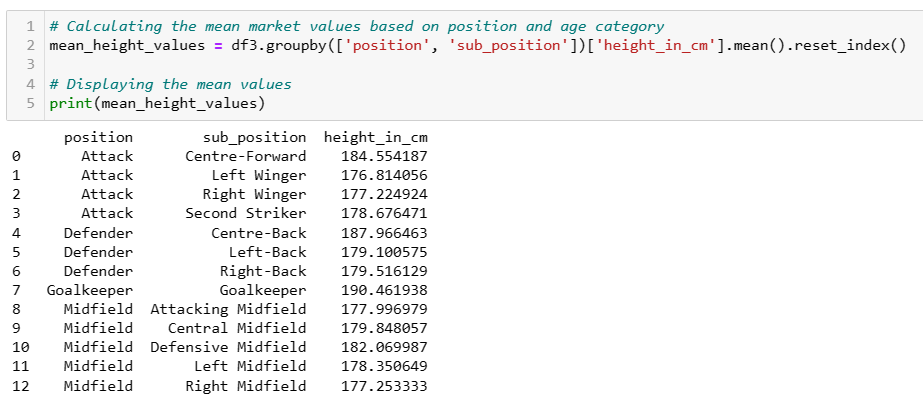
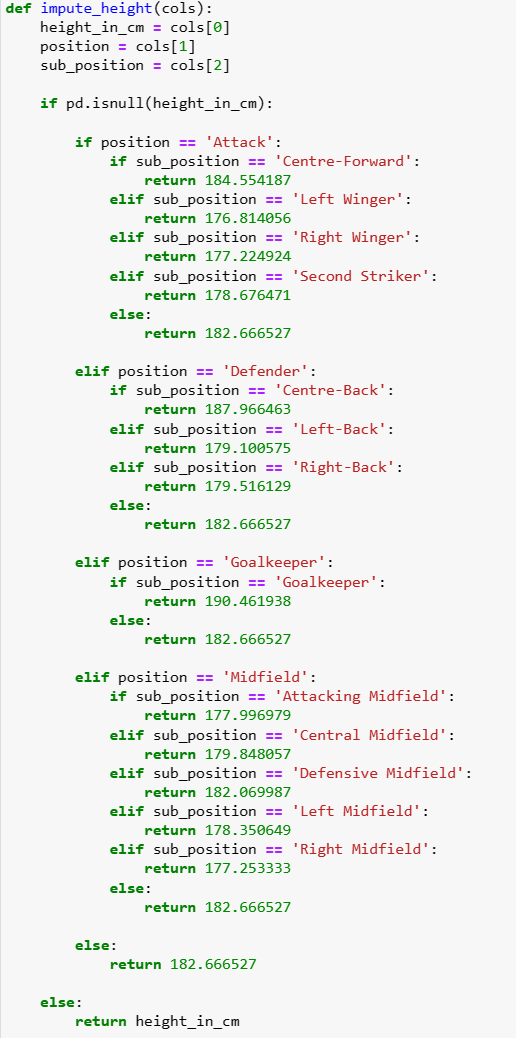
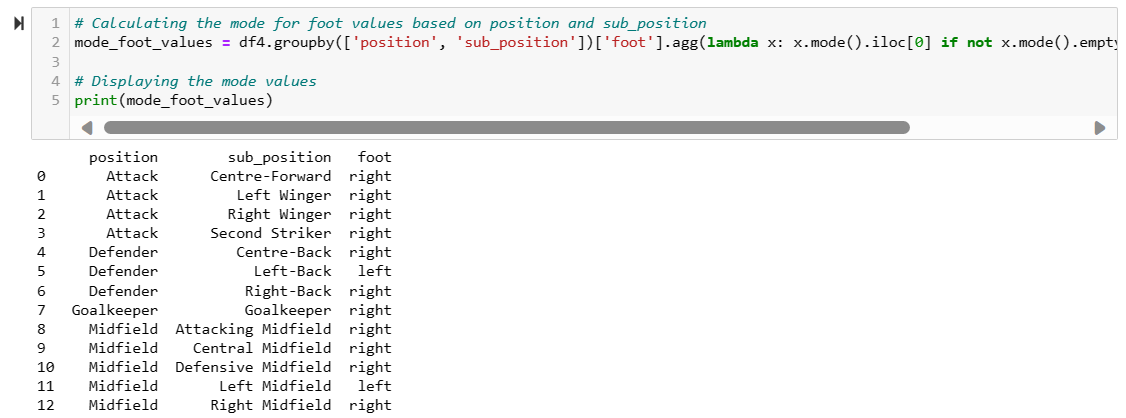


Figure 3.1. Mean values grouped by position and age group (left), Function for imputing mean values (right)











# Appendix C

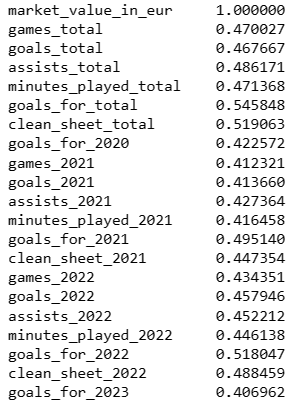
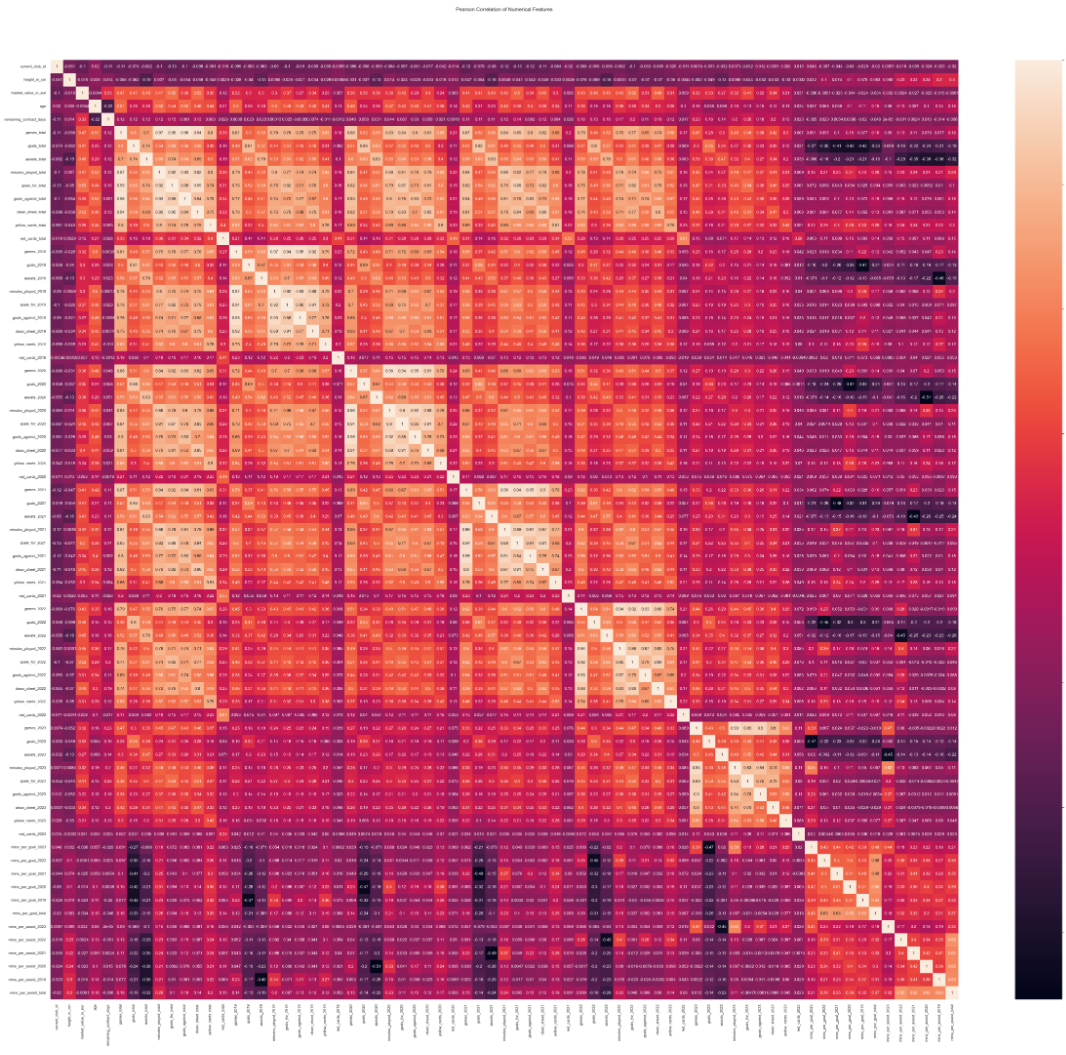
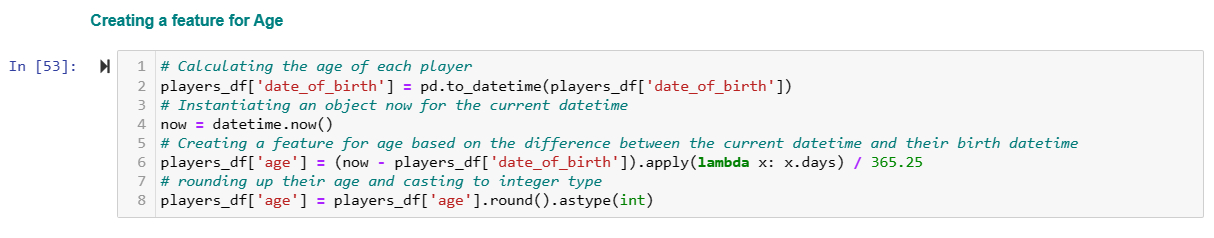
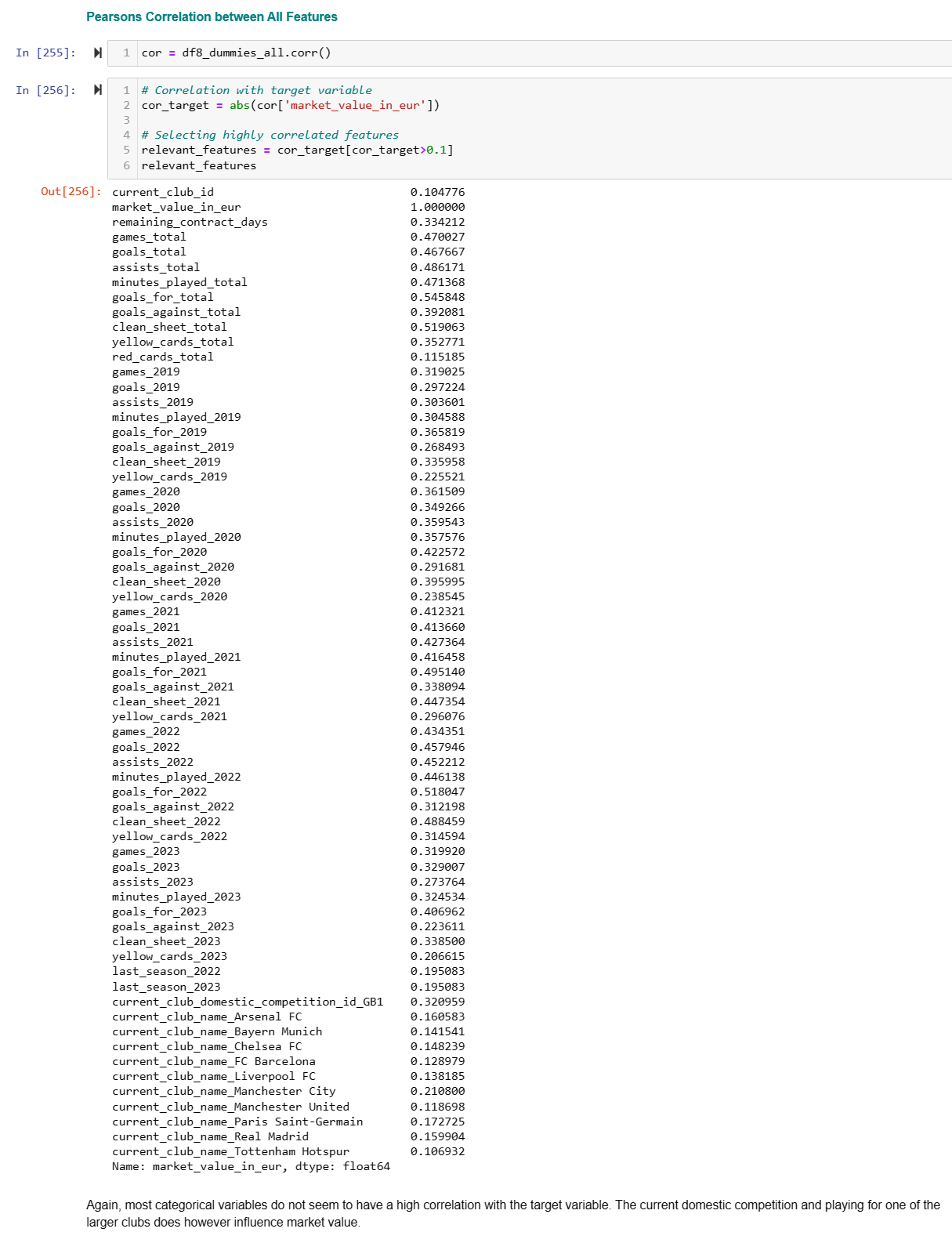


Figure 3.2. Results for Pearson Correlation on Numerical Features

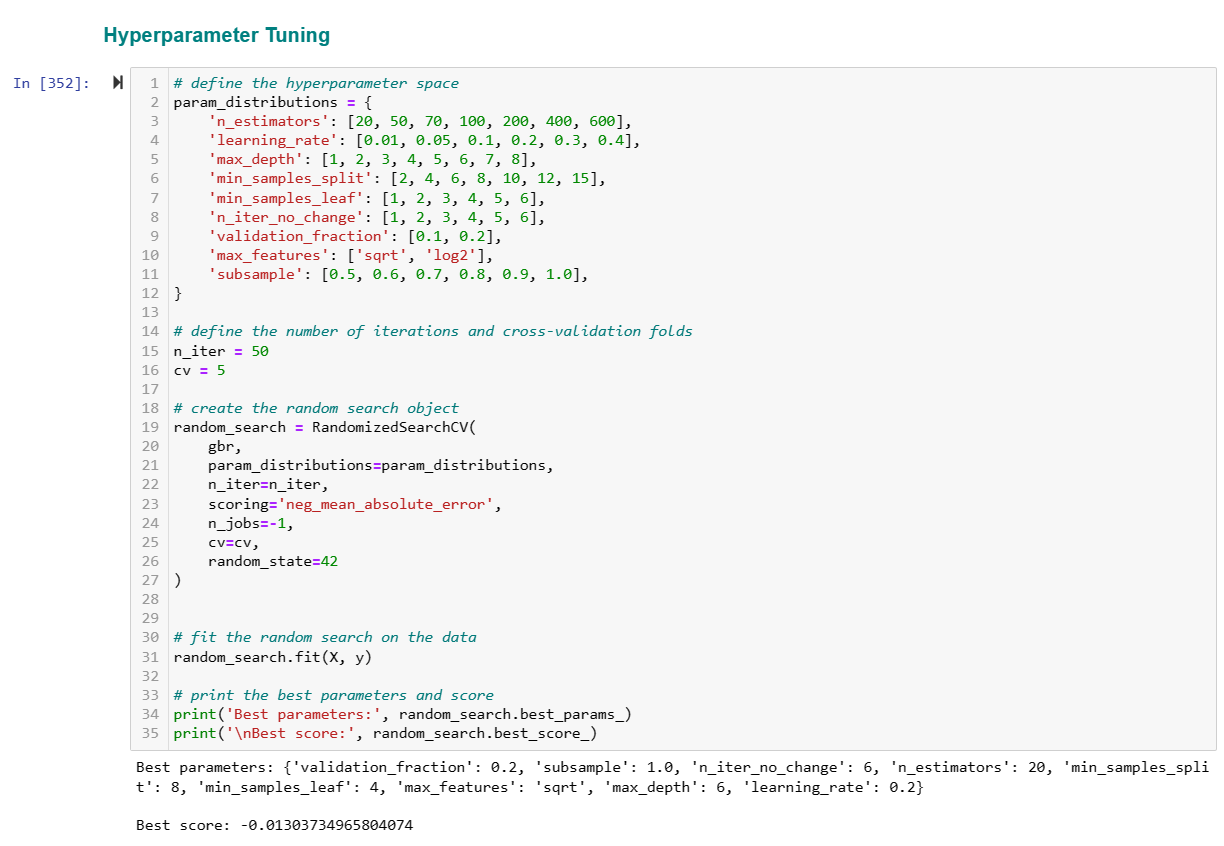






# Appendix D





|  |  |
| --- | --- |

Residuals and Prediction Error plots for LightGBM(prior to log transformation)

|  |
| --- |

Learning curve for LightGBM Model (prior to log transformation)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# References

Agnellutti, C. (2014) Big Data: an Exploration of Opportunities, Values, and Privacy Issues. New York: Nova Science Publishers, Inc (Internet Theory, Technology and Applications).

Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e020mww&AN=811106&site=eds-live>

Ahmed, T. (n.d.). CCT College Dublin: Log in to the site. [online] moodle.cct.ie.

Available at: <https://moodle.cct.ie/course/view.php?id=2637> [Accessed 20 Dec. 2023].

Analytics Vidhya. (2020). KNNImputer | Way To Impute Missing Values. [online]

Available at: <https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/>.

Anwar, A. (2021). A Beginner’s Guide to Regression Analysis in Machine Learning. [online] Medium.

Available at: <https://towardsdatascience.com/a-beginners-guide-to-regression-analysis-in-machine-learning-8a828b491bbf>.

ASQ (2019). What are Histograms? Analysis & Frequency Distribution | ASQ. [online] Asq.org.

Available at: <https://asq.org/quality-resources/histogram>.

Brownlee, J. (2019a). A Gentle Introduction to Learning Curves for Diagnosing Machine Learning Model Performance. [online] Machine Learning Mastery.

Available at: <https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/>.

Brownlee, J. (2019b). Overfitting and Underfitting With Machine Learning Algorithms. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/>.

Campbell, F. (2019). Data Scraping – Considering the Privacy Issues. [online] Fieldfisher.

Available at: <https://www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues>.

Codecademy. (n.d.). Seaborn Styling, Part 2: Color. [online]

Available at: <https://www.codecademy.com/article/seaborn-design-ii>.

Cordeiro, L. (n.d.). Market Value EDA 📈⚽💸. [online] kaggle.com.

Available at: <https://www.kaggle.com/code/luisgasparcordeiro/market-value-eda/notebook> [Accessed 22 Dec. 2023].

Coxen, D. (n.d.). Football Transfer Market EDA & Basic Modelling. [online] kaggle.com.

Available at: <https://www.kaggle.com/code/davidcoxon/football-transfer-market-eda-basic-modelling> [Accessed 19 Dec. 2023].

Depken, C.A. and Globan, T. (2020). Football Transfer-Fee Premiums and Europe’s Big Five: Online Appendix. SSRN Electronic Journal. doi:<https://doi.org/10.2139/ssrn.3617260>.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

docs.aws.amazon.com. (n.d.). Model Fit: Underfitting vs. Overfitting - Amazon Machine Learning. [online]

Available at: <https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>.

Garza, J. (2024). Class Lecture: Cars Project. [online] accounts.google.com. Available at: <https://drive.google.com/drive/folders/1JHFnYE0FJ-OAAJZIJbmH2uezZBjCERgC>.

GeeksforGeeks. (2023). Regression using CatBoost. [online]

Available at: <https://www.geeksforgeeks.org/regression-using-catboost/>.

Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, [online] 63(1), pp.3–42. doi:<https://doi.org/10.1007/s10994-006-6226-1>.

Gitbook.io. (2023). PyCaret 3.0 | Docs. [online]

Available at: <https://pycaret.gitbook.io/docs>.

Gohar, U. (2020). How to use Residual Plots for regression model validation? [online] Medium.

Available at: <https://towardsdatascience.com/how-to-use-residual-plots-for-regression-model-validation-c3c70e8ab378#:~:text=Residuals>.

Goundar, S. (2021). Enterprise systems and technological convergence : research and practice. [online] Charlotte, NC:

<https://eds.p.ebscohost.com/eds/ebookviewer/ebook/ZTAyMG13d19fMjczMDYwNF9fQU41?sid=997bf805-38e3-47f1-bfb3-813e10171b3f@redis&vid=3&format=EB>

Hali, B. (2019). Understanding Bias-Variance Trade-Off in 3 Minutes. [online] Medium.

Available at: <https://towardsdatascience.com/understanding-bias-variance-trade-off-in-3-minutes-c516cb013513>.

Indeed Career Guide. (n.d.). 13 Types of Regression Analysis (Plus When To Use Them). [online]

Available at: <https://www.indeed.com/career-advice/career-development/regression-types>.

Kalyvas, V. (2024). Predicting Car Prices with Machine Learning (Step 1 — Analysis). [online] Medium.

Available at: <https://python.plainenglish.io/predicting-car-prices-with-machine-learning-step-1-the-analysis-76024196f92b> [Accessed 3 May 2024].

Marsja, E. (2020). How to use Square Root, log, & Box-Cox Transformation in Python. [online] Erik Marsja.

Available at: <https://www.marsja.se/transform-skewed-data-using-square-root-log-box-cox-methods-in-python/>.

Matplotlib (2012). Matplotlib: Python plotting — Matplotlib 3.1.1 documentation. [online] Matplotlib.org.

Available at: <https://matplotlib.org/>.

McQuaid, D. (n.d.). Week 1 Overview. [online] moodle.cct.ie.

Available at: <https://moodle.cct.ie/course/view.php?id=80>.

Metelski, A. (2021). Factors affecting the value of football players in the transfer market. Journal of Physical Education and

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Sport, 21(2). doi:<https://doi.org/10.7752/jpes.2021.s2145>.

Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Beijing: O’reilly.

Nalcin, S. (2022). StandardScaler vs. MinMaxScaler vs. RobustScaler: Which one to use for your next ML project? [online]

OpenAi (2022). ChatGPT . [online] chatgpt.com. Available at: <https://chatgpt.com>.

Medium. Available at: <https://medium.com/@onersarpnalcin/standardscaler-vs-minmaxscaler-vs-robustscaler-which-one-to-use-for-your-next-ml-project-ae5b44f571b9>.

Niu, Y., Ying, L., Yang, J., Bao, M. and Sivaparthipan, C.B. (2021). Organisational business intelligence and decisionmaking using big data analytics. Information Processing & Management, [online] 58(6), p.102725. doi:<https://doi.org/10.1016/j.ipm.2021.102725>.

Pannucci, C.J. and Wilkins, E.G. (2011). Identifying and Avoiding Bias in Research. Plastic and Reconstructive Surgery, [online] 126(2), pp.619–625. doi:<https://doi.org/10.1097/prs.0b013e3181de24bc>.

Patak, K. (2023). Strategic Analysis of Emerging Technology for Competitive Advantage: Artificial Intelligence in the Football Industry. [online] Available at: <https://drive.google.com/file/d/1g7y_U2cSYDq771WJZoka4AvgIkdOvIW7/view?usp=sharing> [Accessed 27 Oct. 2023].

Persson, J. and Sjöö, E. (2017). Business Intelligence - its impact on the decision making process at higher education institutions

qualtrics (n.d.). What is ANOVA (Analysis Of Variance). [online] Qualtrics. Available at: <https://www.qualtrics.com/uk/experience-management/research/anova/#:~:text=ANOVA%20stands%20for%20Analysis%20of>.

Rahul Kumar (2019). Machine learning quick reference : quick and essential machine learning hacks for training smart data models. Packt Uuuu-Uuuu.

Readthedocs.io. (2024). Catboost tutorial — SHAP latest documentation. [online] Available at: <https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Catboost%20tutorial.html>

Saxena, S. (2020). Everything you Should Know about p-value from Scratch for Data Science. [online] Analytics Vidhya. Available at: <https://medium.com/analytics-vidhya/everything-you-should-know-about-p-value-from-scratch-for-data-science-f3c0bfa3c4cc>.

Scikit-learn.org. (2018). Compare the effect of different scalers on data with outliers — scikit-learn 0.20.3 documentation. [online] Available at: <https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html> .

Scikit-learn.org. (2019). 1. Supervised learning — scikit-learn 0.21.3 documentation. [online]

Available at: <https://scikit-learn.org/stable/supervised_learning.html#supervised-learning>.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Simplilearn.com. (n.d.). The Complete Guide on Overfitting and Underfitting in Machine Learning. [online]

Available at: <https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting>.

Solon Barocas and Selbst, A.D. (2015). Big data’s disparate impact. Ssrn Elibrary.

Statistics Solutions. (n.d.). ANOVA. [online]

Available at: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/anova/#:~:text=ANOVA%20assumes%20that%20the%20data>.

Suresh Kumar Mukhiya and Ahmed, U. (2020). Hands-on exploratory data analysis with Python : perform EDA techniques to understand, summarize, and investigate your data smartly. Birmingham: Packt Publishing.

Technology, T. (2023). Light GBM Light and Powerful Gradient Boost Algorithm. [online] Medium.

Available at: <https://medium.com/@turkishtechnology/light-gbm-light-and-powerful-gradient-boost-algorithm-eaa1e804eca8>.

Van den Berg, E. (2011). The Valuation of Human Capital in the Football Player Transfer Market. [online]

Available at: [Link](https://d1wqtxts1xzle7.cloudfront.net/6177574/berg__e._van_den__296332_-libre.pdf?1390844876=&response-content-disposition=inline%3B+filename%3DThe_Valuation_of_Human_Capital_in_the_Fo.pdf&Expires=1698586960&Signature=Y6SAWE8dDwH7y7ufin5qpNIwpIKoAHkam~bVdYd-Hua7-sofcQOPoHTyyeB7eOFRp67FiQm34FDIHNAkmGzy5jy2dCtstGfCy-8hzjPIuRFj3jVXdf7cLMHrI7TOJWzVugUFtYKNJe3zngLuydJ0TyT3JaHXJdCtVmgL4KsQjDnOtkbBy6Ajb-78Sj1~EtsJJWnHRcmeTRqq-M6mW~Wl7bTnP-mQHlM4ZopEqv3Z-nTEoCpqpIb4gZvI-mlF4LDSFA8DnZNO6l9TsUJz1CKpnMvbbsKm~9y0L4c4Ha3dEDZ4eyvGLfxzaKQF7IC72DFXJCKoPXNfzuzRfpb5uYTASg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)

wiki.creativecommons.org. (n.d.). CC0 FAQ - Creative Commons. [online]

Available at: <https://wiki.creativecommons.org/wiki/CC0_FAQ>.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining . Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, (Vol. 1, pp. 29-39).

World Population Review (2022). Most Popular Sport by Country 2020. [online] worldpopulationreview.com.

Available at: <https://worldpopulationreview.com/country-rankings/most-popular-sport-by-country>

www.footballbenchmark.com. (n.d.). Football Benchmark - Methodology and limitations of published information. [online] Available at: <https://www.footballbenchmark.com/methodology/player_valuation>.

www.kaggle.com. (n.d.). Football Data from Transfermarkt. [online]

Available at: <https://www.kaggle.com/datasets/davidcariboo/player-scores/versions/284/data> [Accessed 29 Oct. 2023].

www.scikit-yb.org. (n.d.). Residuals Plot — Yellowbrick v1.5 documentation. [online]

Available at: <https://www.scikit-yb.org/en/latest/api/regressor/residuals.html#:~:text=The%20residuals%20plot%20shows%20the>.

www.transfermarkt.co.uk. (n.d.). Aïssa Boudechicha - Player profile 23/24. [online] Available at: <https://www.transfermarkt.co.uk/aissa-boudechicha/profil/spieler/592398> [Accessed 16 Dec. 2023].

www.transfermarkt.co.uk. (n.d.). Genar Fornés - Player profile 23/24. [online] Available at: <https://www.transfermarkt.co.uk/genar-fornes/profil/spieler/628490> [Accessed 5 Dec. 2023].

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**