## 3.4. Data analysis

### **Choosing the Dataset**

Choosing the most appropriate data set which needs to toil over is important and it should be the first task. The primary purpose of this is to examine the characteristics of online players and find out the category or rank they belong to. This analysis is to find the known defined category of the player. Also, we decided to get feature columns as defined columns. Considering all the above factors it is best to choose a labeled data set. Since unsupervised learning requires additional stages and procedures to develop a prediction model, the ability to perform supervised machine learning is an advantage of selecting a labeled dataset.

### **Choosing the Tools and Language**

When it comes to translating static knowledge to programming, the two most popular approaches exist. The most popular choice is Python, yet R has emerged as the language of choice among data scientists in the field. The R programming language is an effective statistical programming script for analyzing big data sets, visualizing data, and developing new statistical models. We chose R language's native tool, R studio, even if the web has Jupiter Notebook and other well-known applications, to continue the analysis.

### **Data Pre-Processing**

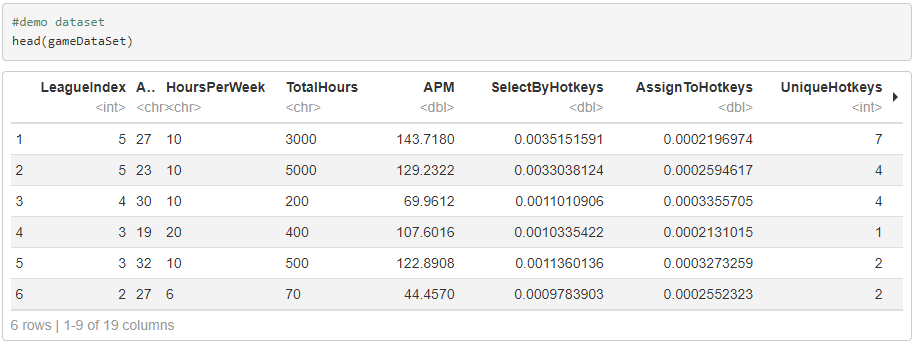
Graphical user interface, application, Word

Description automatically generated with medium confidenceThe first thing to do with the notebook is to install and load the fundamental libraries which need to perform common tasks. Additional libraries will be installed and loaded as proceed.

Graphical user interface, application

Description automatically generated with medium confidenceAfter spending time browsing through the open APIs and the dataset, we spotted a dataset for the game “SkillCraft” that contained the exact feature columns we were looking for. It is a CSV file that contains 3395 rows of data that works well to train and test a model. All column names are formatted the same manner to make data processing and applying R functions to the loaded data frame smoother in the future.

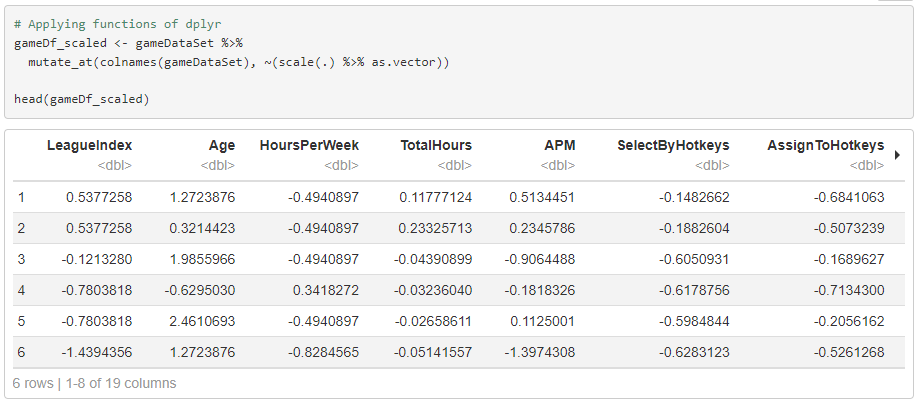
After loading the dataset, the first few rows of the dataset were shown to give an overview of the dataset.



To train an effective and accruable prediction model, some human judgments must be made based on the dataset. The “LeagueIndex” column has no meaning as a feature column as it does not fall under player characteristics. All column types are converted to “numeric” type for ease of manipulation across functions.

Chart

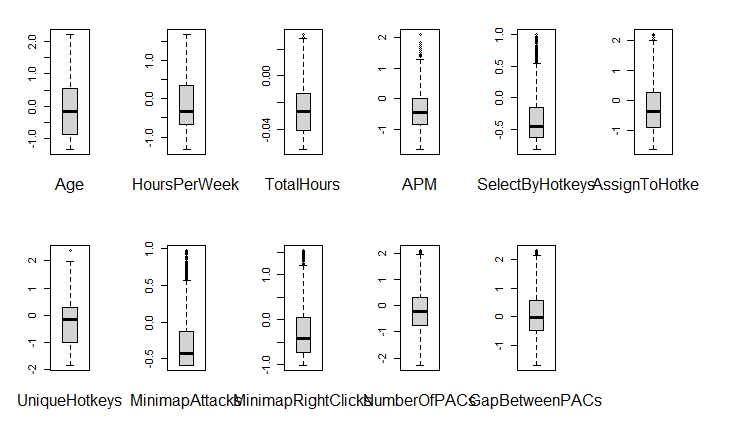
Description automatically generatedOutliers are the next crucial component. One of those statistical problems that everyone is aware. Outliers have a significant impact on bulk of parametric statistics, including means, standard deviations, correlations, and all statistics built on these. Some outliers in a dataset represent natural population variance and ought to be left alone. They are known as true outliers. Other outliers are harmful and ought to be eliminated because they indicate measurement mistakes, faults in data entry or processing, or inadequate sample. Outliers can be found using a few custom R functions written there, and they can be eliminated using a different function. Before that need to give an explanation on “boxplot” which helps to visualize the outliers in the dataset.

Removing outliers from the data set is hard hitting as the data is in different ranges. A possible solution is to normalize and scale the dataset. Scaling a data set means that you transform your data to fit a specific scale. Scaling merely modifies the data's range. A more drastic change is normalization. Normalization is the process of transforming your observations into something that can be compared to a normal distribution. It is easier to proceed with the normalization now that we have made some changes, such as keeping all column types to a single common type. Following graph is the demonstrate of scaled dataset.

Then the final deal with the outlier is to keep the true outliers as they are and remove the others from the dataset using custom functions already written. After removing the outlier data set can be plotted again using "boxplot".

Chart, box and whisker chart

Description automatically generated

The remaining outliers can be presented in a graph with each individual class.

### **Data Processing**

Graphical user interface, application

Description automatically generated with medium confidenceSince the dataset is clear to be proceed, the next step is break down the feature columns, target column. “LeagueIndex” is easily recognized as the target column. It would be clearer for the process in the future by separate the target column and assign it a meaningful name for the class. We may select the ideal supervised learning machine learning model for this since we have specified, precisely defined classes as the target column.

Chart, treemap chart

Description automatically generatedThe Random Forest Classification approach is ideal for this use case since we have established classes. There should be an equal number of value rows in each feature column in the dataset. If not, there will be a complication when applying PCA and training the model. To prove this, use "missmap.R" to render the missing map.

The dataset is divided into train and test subsets as the last step before being trained to model. 70% of the whole dataset will be used for training, while the remaining 30% will be used for testing.

### **Training the Classification Model**

Graphical user interface, text, application, email

Description automatically generatedThe chosen model has been trained using the training dataset, and it took “14.56417 seconds” to complete the process. Below is a summary of the trained model.

Chart

Description automatically generatedUsing R’s “varImpPlot” function, a dot-chart of variable importance as calculated by a Random Forest can be presented. As shown below, the two grids will form.

The first graph illustrates how much the MSE will rise if a variable is given values by random permutation. However, the second graph, the Gini Index, which compares RSS before and after a variable split, is used to determine the node purity.

is necessary to assess whether the model is over fitted after training the dataset as normal. We must use the train dataset to forecast the outcomes in order to investigate it. We must acknowledge that our dataset is over fitted if the model provides 100% or unusually accurate predictions for the training dataset. If not, the trained model is more robust and can analyze the test dataset.

### **Prediction using Model**

Graphical user interface, application, Word

Description automatically generatedPredicting the model accuracy using the testing dataset is the last step in the analysis. After making that prediction, we may visualize the outcomes using different charts and a confusion matrix.

Chart, bar chart

Description automatically generatedTable

Description automatically generated