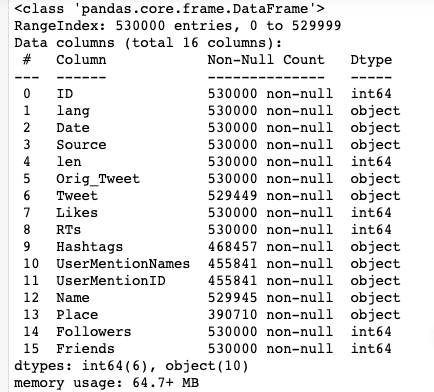
**Data acquisition**

The **FIRST** data acquired in this chapter is the DecTreeAssign1.dat data provided by Dr. Santago. The data provided in the DecTreeAssign1.dat include one class variable together with 30 features named by Fxx where xx is the feature id (00, 01, ...., 29). The class variable is a binary categorical variable with data values “B” and “M”. The remaining features are all continuous variables for the purpose of decision tree building.

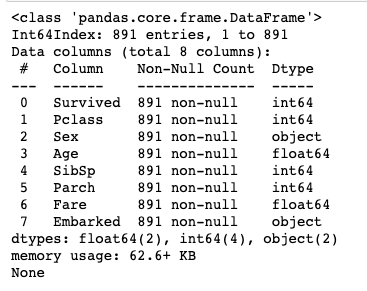
The **FOURTH** dataset is called FIFA.csv dataset that I found on Kaggle. This dataset contains a random collection of 530k tweets starting from the Round of 16 till the World Cup Final that took place on 15 July, 2018 & was won by France. Multiple features are available in this dataset as shown in the figure below:



Despite the fact that we have multiple features at hand, the specific feature of my interest is the “Tweet” feature which is essentially the content of each tweet. This feature is categorical and I will further process this feature as will be demonstrated later.

<https://www.kaggle.com/rgupta09/world-cup-2018-tweets>

The **SECOND** data I used is the Titanic data I downloaded from Kaggle and preprocessed (all missing data imputed). The dataset has 891 rows and 8 columns. Among them, the dataset has 7 features to predict whether or not a person survived on titanic. The categorical features are “Sex” and “Embarked”; while the continuous features are “Pclass”, “Age”, “SibSp”, “Parch”, and “Fare”. Here shown below is a summary of the dataset.



<https://www.kaggle.com/c/titanic>

The third dataset I used is the credit card data I downloaded from Kaggle. This dataset has 284807 rows and 31 columns. Among the features, there are 28 explanatory variables labeled V1 to V28, which is how I will base my anomaly detection on. In addition, there is the variable ‘Time’ and ‘Amount’ which I won’t consider in my analysis. There is also the ‘Class’ variable which indicates whether or not a data point is an anomaly (so this study is supervised).