efrei
PARIS
École d'ingénieurs du numérique

# LEAGUE OF LEGENDS

## Big Data Processing and Visualization: Tools and Platforms


Eliot Babet

Tancrède Martin

Alexis Ribat

**Eliot Babet**
**Tancrède Martin**
**Alexis Ribat**

**Eliot Babet**
**Tancrède Martin**
**Alexis Ribat**

# SUMMARY

**Eliot Babet**
**Tancrède Martin**
**Alexis Ribat**

# I. Where does our data come from?

League of Legends is the most popular video game of the decade developed by Riot Game and played by a huge number of players. It is played **5 players vs 5 players** in competitive game where players are ranked depending on the result of their own games. The dataset we will use has been collected from the **Riot Games API** itself. After playing a game of League of Legends, any player can check a report of what was done this game (number of kills, total damage, who won, etc). Our dataset is the resume of **1000 games** at Challenger tier (which is the very top level) from the Brazilian server. We found this dataset using google dataset search.

*Figure 1: Map of the summoner's rift*

The dataset is composed of 1001 lines, including the titles, with each line being a game. We had at the beginning 552 columns, each column being a statistic from the game. Some statistics are not interesting because:

- Some statistics need another knowledge to be efficiently used (What is an “unreal kill”? We don’t know because this is a specifical word chose by developers, and we can’t ask them)
- Some statistics are outdated (references on a game mode that no longer exists)
- Some statistics are simply not relevant (queueId is an entire column of the number 420)

After this part of pre-processing, we only have 515 Columns. We deeply understand the meaning of each of these columns.

515 columns are still a lot, so we went deep further into features selections using a correlation matrix between features. Each player has 49 columns each, which is a lot. Using correlation, we reduced our columns to only 5. By example, we realized that the number of deaths is highly related to the longest time spent living. The damage dealt is divided into true, magical and physical damage, but we do not need these details, so we only keep the global damage dealt. We finally chose these 5 features: Longest time spent living, Total damage dealt to champions, Damage dealt to objectives (towers and neutral monsters giving buffs to whoever kill execute them), Vision scores (players can place cameras to see a place from far away, it’s highly strategical) and Total minion killed (weak enemies that give some resources when you kill them). Since it is still a lot of columns, we deleted the columns that were not related with players themselves. It remains 10 times the 5 columns, one for each player. We decided to ignore the second team and focus and the first one. We also decided to use the mean of each players statistics, because it makes sense in a team game. We called our last columns:

['gameId','AverageLongestTimeSpentLiving','AverageTotalDamageDealtToChampions','AverageDamageDealtToObjectives','AverageVisionScore','AverageTotalMinionsKilled','win']

We chose to work on this type of datasets because all three of us play (or used to play) League of Legends daily, so all these data have a meaning for us, and we can easily imagine any application and knowledge we could extract from the data.

The goal of our data analysis project is to analyse the different victory conditions that any players can put in practice to try to win more frequently. By example, the more you kill, the more you win, but with more unintuitive facts.

# II. Reduction dimensionality

For this part, we decided to focus only on the most important features. It is explained in part 1. Therefore, we chose 6 components which where:

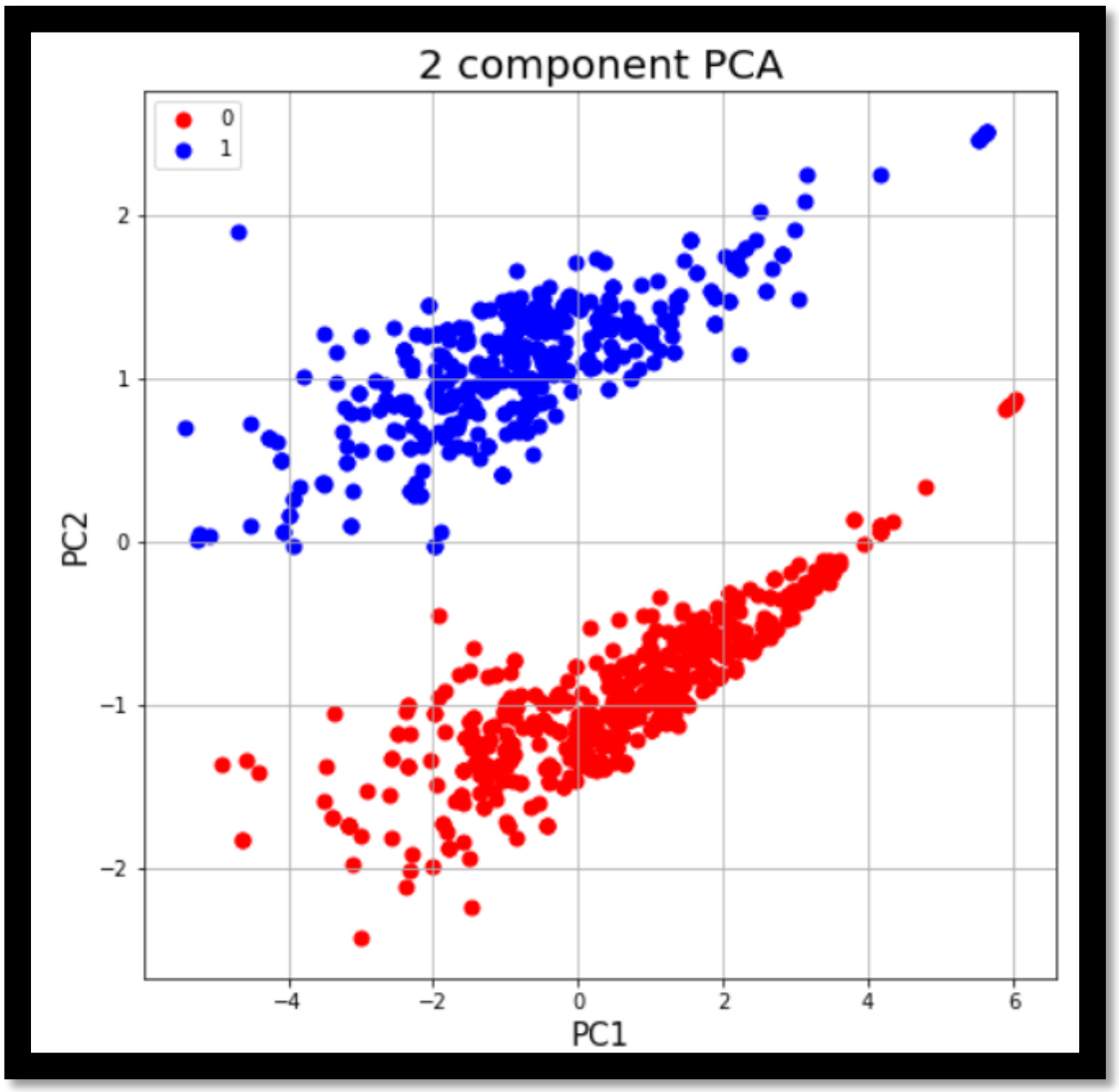
1. 'gameId'
2. 'AverageLongestTimeSpentLiving'
3. 'AverageTotalDamageDealtToChampions'
4. 'AverageDamageDealtToObjectives'
5. 'AverageVisionScore'
6. 'AverageTotalMinionsKilled'

Then, we decided to try 3 different models (PCA, Random Forest and Isomap) in order to reduce the dimensionality of the data and to plot the results to extract knowledge from it.

- **PCA**

We chose PCA because we have numerical variables and a lot of variables. PCA is a good way to obtain correlation between the old and new variables, and to visualize this correlation.

After plotting the data, we had:




In our case, after visualizing the data, we wanted to know the explained variance in order to know how much information could be attributed to each of the principal components. We used *pca.explained_variance_ratio_* function and we obtained:

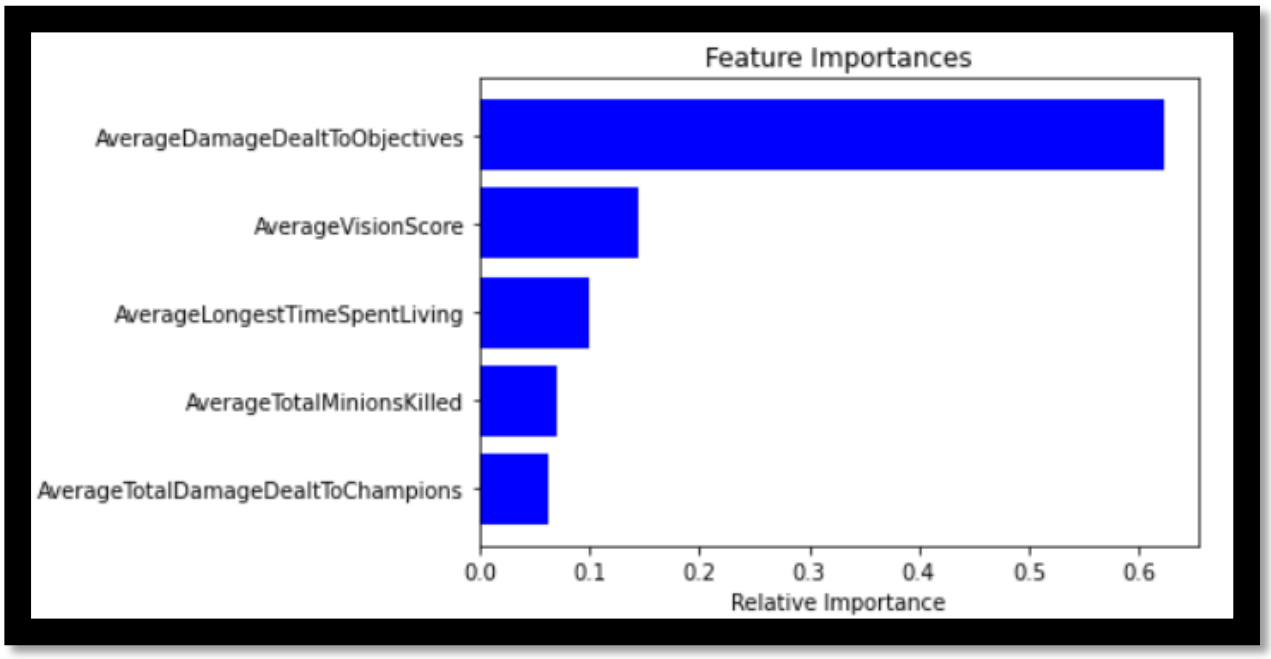
```
Entrée [9]:  1 pca.explained_variance_ratio_

Out[9]: array([0.62184183, 0.20458246, 0.09969966, 0.03030483, 0.02696653,
               0.01660469])
```

Which mean that two components contain 80% of the information.

- **Random Forest**

We also used the Random Forest algorithm to reduce dimensionality. We dropped the column 'win' since it is only two cases variables of 1 and 2. After fitting the model, the graph looked like that:




Here we have an interesting graph. We knew that DamageDealtToObjectives data was important, but not that important. In fact, it is one of the most useful data that is the most predictive attributes.

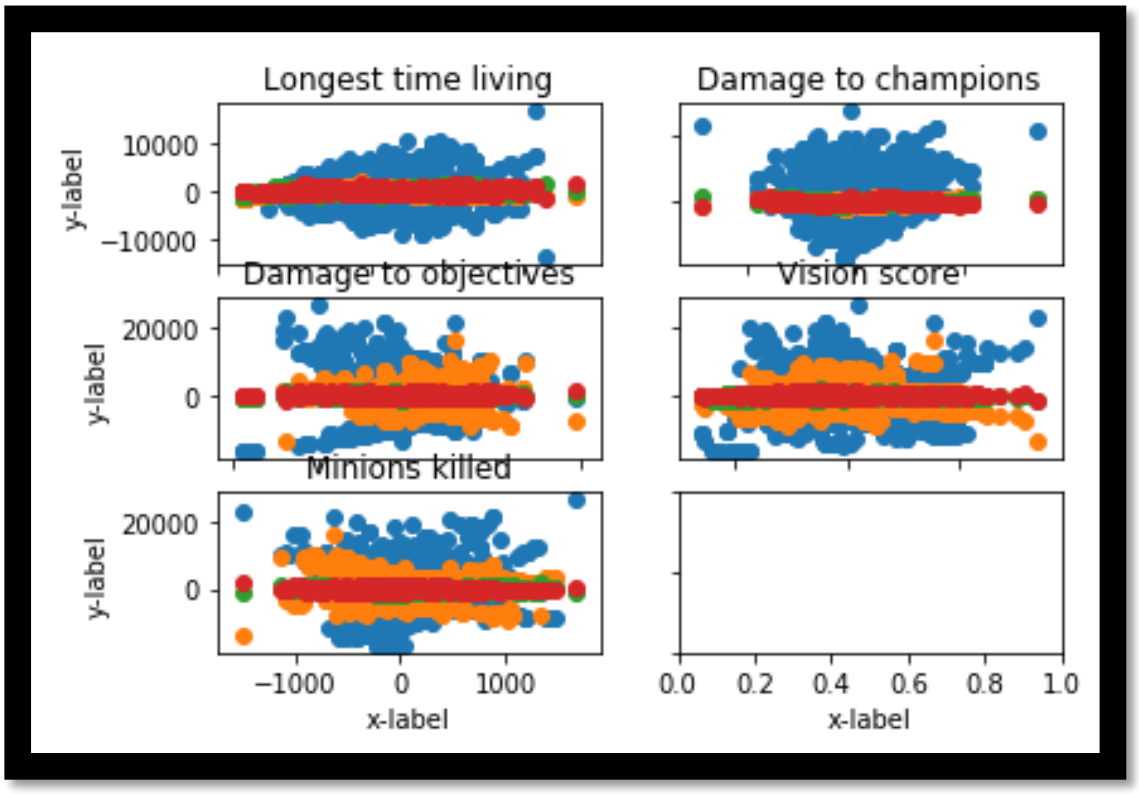
- **Isomap**

For the last dimensionality reduce we used Isomap. The objective: "This is implemented by linking the points X into the graph of geodesic distances of the training data. First the `n_neighbors` nearest neighbors of X are found in the training data, and from these the shortest geodesic distances from each point in X to each point in the training data are computed in order to construct the kernel. The embedding of X is the projection of this kernel onto the embedding vectors of the training set".

To realize the Isomap for our dataset, each variable was displayed according to the others and then we can conclude on the correlation. We had 1000 lines, so we take 31 neighbors. However, we did this on data and not images (e.g.: faces) so we cannot recover very conclusive results.

At least we can see some relations between variables (Those who have the same shape):

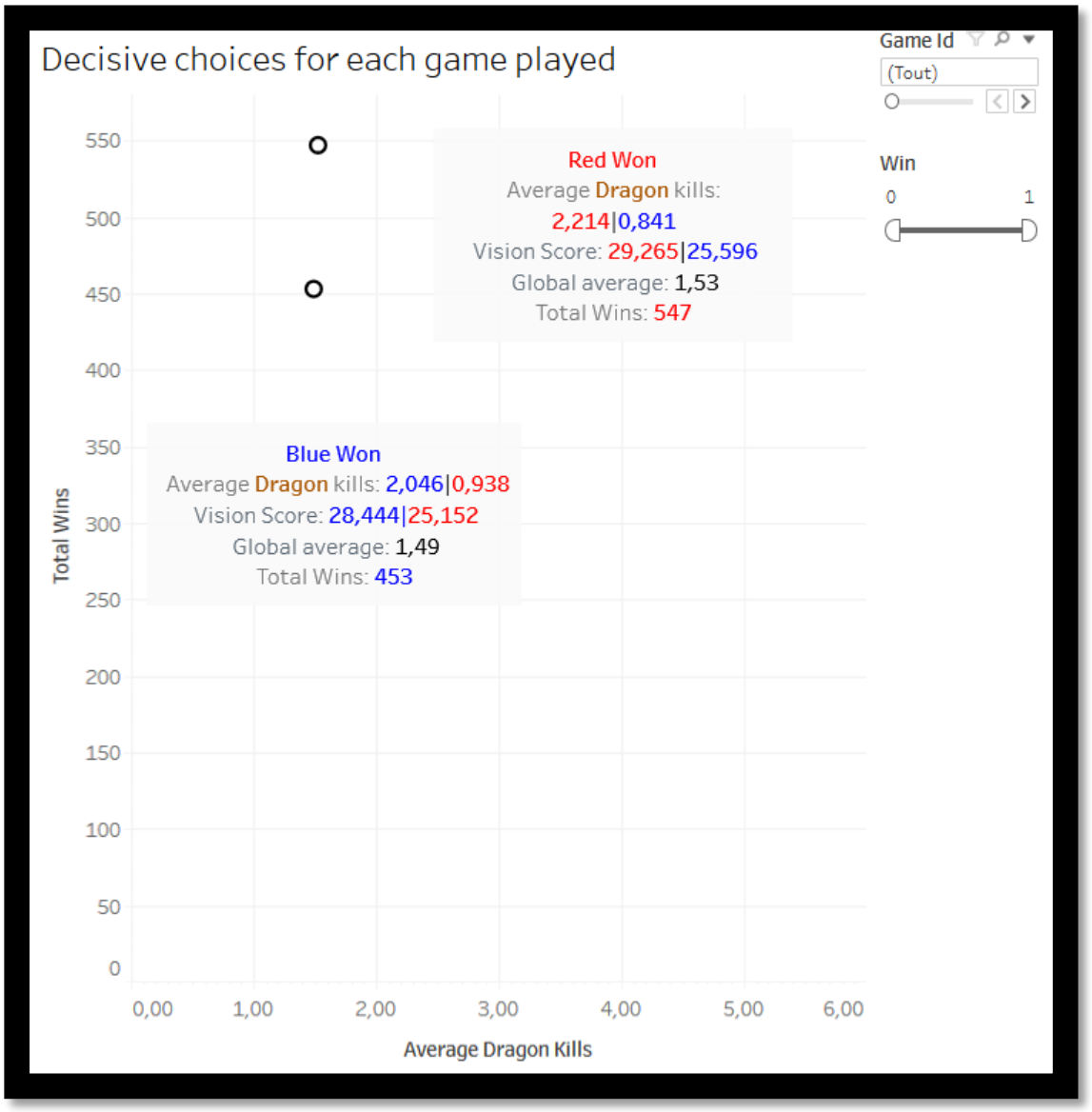
**Eliot Babet**
**Tancrède Martin**
**Alexis Ribat**



**Eliot Babet**
**Tancrède Martin**
**Alexis Ribat**

# III. The tableau software dashboards

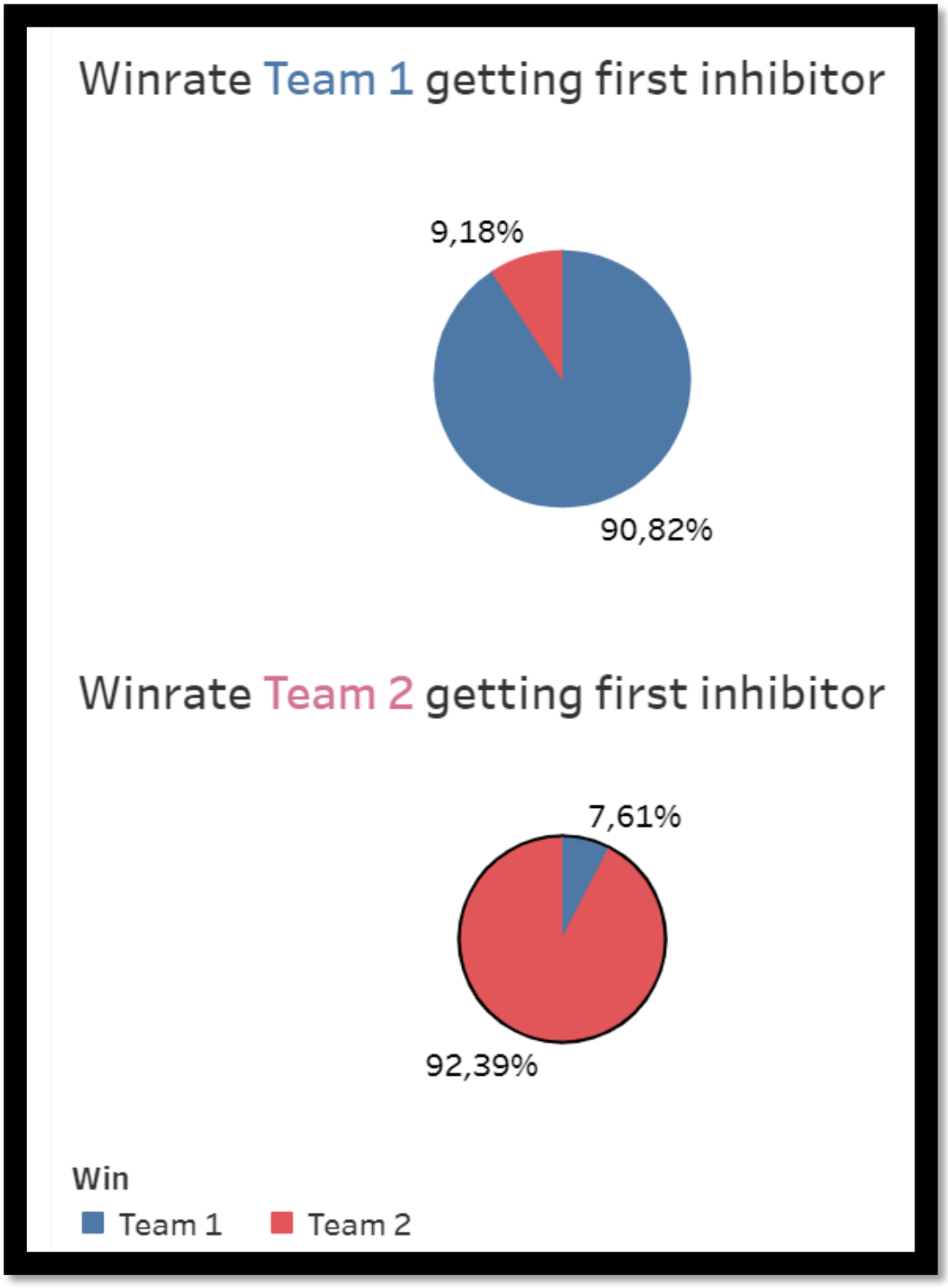
1. Decision choices for each game played




Link of the Graph:

https://public.tableau.com/profile/babet.eliott#!/vizhome/Project_BDPVTP/Tableaudebord1?publish=yes

2. Win rate regarding first inhibitor

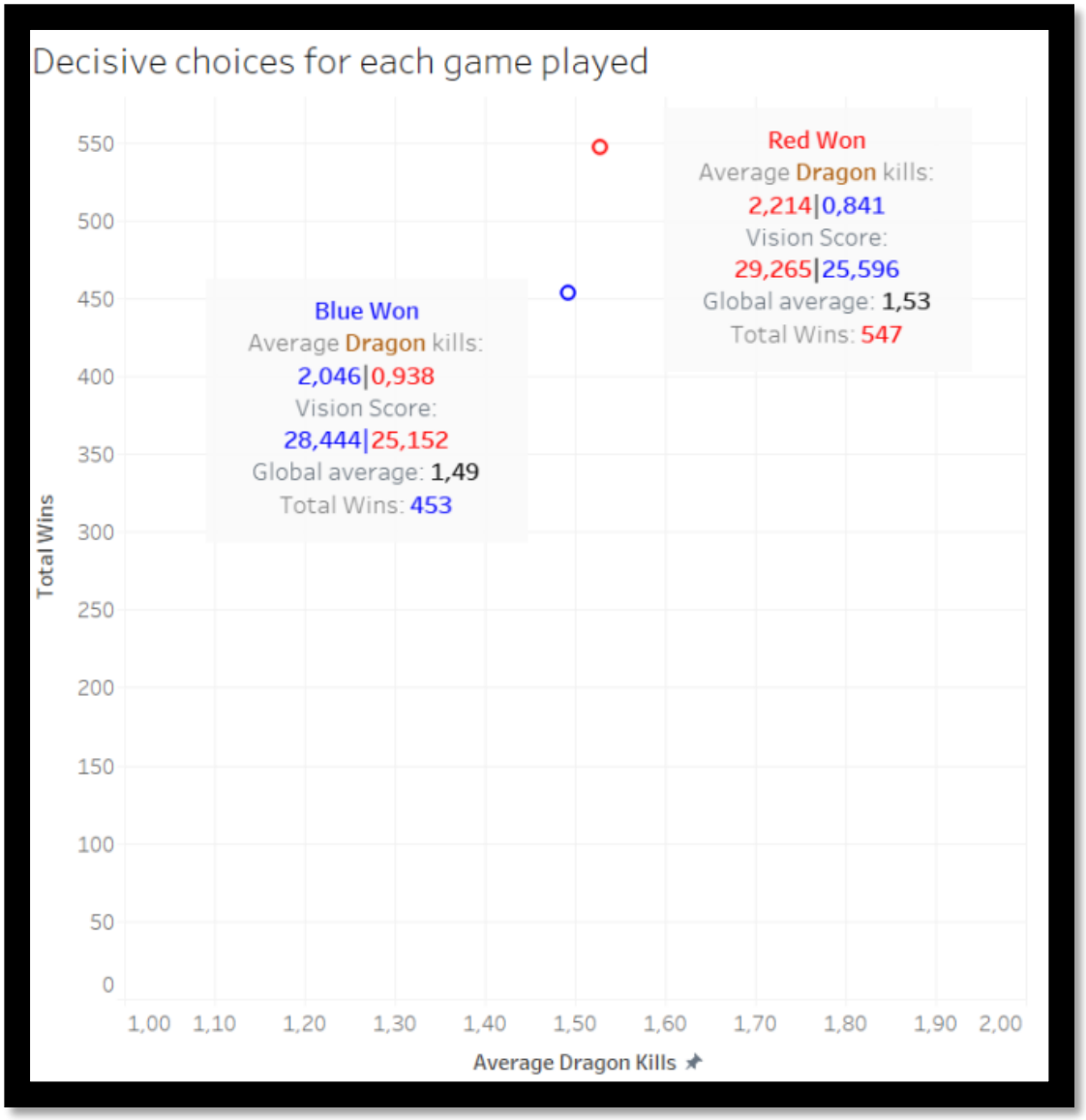



Link of the Graph:

https://public.tableau.com/profile/ribat?fbclid=IwAR0L44IyvTsEvvQVDYBN8jRkT2SP4IVp_yGU0WHX0ssR9UOJBIgd3XWBHZY#!/vizhome/Projet_BD_Virtualization_Alexis_Ribat/Tableaudebord1?publish=yes

**Eliot Babet**
**Tancrède Martin**
**Alexis Ribat**

# IV. Final visual with Tableau software




With Tableau, we created this dashboard to show the importance of wards and drake. We saw that the winner takes a lead with objectives and not with kills. We chose to represent warding and drake as two factors of win. The team that won is basically the team that place more wards and that kills more drakes.

After doing all that analyses about gaming statistics for best players in Brazilian server, we wanted to compare them with the best data analysis site for League Of Legends: League Of Legends Graphs (https://www.leagueofgraphs.com/fr/ )

And the thing that was interesting is that, when we compared our data and our results with this site, we had almost the same results at the end (which is reassuring)!

# V. Conclusion

It was an interesting project where we learned how to use **Tableau Software** and **Excel**. Even if we have a lot to learn, we explored a large part of Tableau Software and we saw the great capacities of this software, using only the free version of it. It is a great tool in order to analyse data and it could be useful in our projects in the future.

We also learned more about Excel. Same as Tableau Software, we have a lot to learn but we have seen a different way to use it. For example, in order to do a better analyse the data, it was sometimes easier to change the data in the excel file than in Tableau Software. Furthermore, we learned how to use functions in Excel (IF, AND RECHERCHEV, VLOOKUP etc...) in order to do interesting graphs.

Thanks for reading this report and we hope you liked it.