EPITA - Promo 2025

Compte rendu
04/04/24

Exploratory Data Analysis

Aupest Benjamin

Escamilla Louis

Lhomme Joris

Table des matières

Introduction

Dans ce mini projet, le but est d'identifier une **thématique** qui nous intéresse et de réaliser une **analyse visuelle exploratoire** afin de mieux comprendre la forme et la structure des données. Nous avons examiné les données afin de **répondre aux questions** et aux hypothèses que nous nous sommes posées. Ce rapport contient toutes les démarches d'analyse, nos visualisations et explications. Un **notebook** est également rendu avec ce rapport pour comprendre comment nous avons travaillé. Nous avons crée un

1 Selection des données

1.1 Choix du dataset

Après une période de recherche sur différents sites web de datasets, nous en avons trouvé un sur lesquels notre intérêt commun s'est porté. Un premier **dataset de 10000 parties** du jeu League of Legends au rang diamant en 2020. Ce dataset contient des données qui ont été récupérées lors des **10 premières minutes** de jeu. Et un second contenant les informations de **250000 fin de parties solo**.

1.2 League of Legends

League of Legends est un MOBA (multiplayer online battle arena) ou **deux équipes de 5 joueurs** s'affrontent sur une carte. Il y a 3 voies et une jungle ou les joueurs peuvent s'affronter et évoluer. Le but du jeu est de détruire le nexus adverse qui se trouve au cœur de la base.

1.3 Description du dataset

Le premier dataset que nous avons choisi contient **19 features par équipes** (donc 38 au total). Cela inclut le nombre de **morts**, la différence d'**argent**, d'**expérience**, etc. Il est au format csv et contient exactement **9 879** lignes

Le second contient **59 features** mais l'on ne s'intéressera qu'à **15** d'entre elles parmi lesquels **"killed[nom de monstre épique]","lost[nom de monstre épique]"** qui correspond au nombre de **monstres épiques** tués par son équipe ou l'équipe adverse. Il contient exatement **242572 lignes**

1.4 Questions à examiner

Nous avons voulu répondre aux questions suivantes :

Q1. Peut-on prédire l'issue d'une partie à 10 minutes de jeu ?

Q2. Est-ce qu'à partir d'un certain avantage pour une équipe, on peut dire que la partie est gagnée ou perdue ?

Q3. Y a-t-il un avantage particulier sur un côté de la carte ?

Q4. Quel monstre épique donne le plus d'avantage à une équipe ?

2 Exploratory Data Analysis with vizualisation

2.1 Identification de valeurs aberrantes, manquantes et anomalies

Afin de mener à bien notre exploration, nous nous sommes assuré de la **qualité** des dataset en vérifiant s'il y avait des données manquantes, aberrantes ou d'autres anomalies.

Nous avons utilisé le code python suivant pour vérifier s'il manquait des valeurs.

```
print(data.isna().sum())
```

En l'exécutant, nous voyons qu'il n'existe **aucune valeur manquante** parmi les colonnes de nos datasets (cf notebook).

Nous avons ensuite vérifié l'**équilibre** du nombre de victoires et du premier sang obtenu par les bleus et les rouges **dans le premier dataset**.

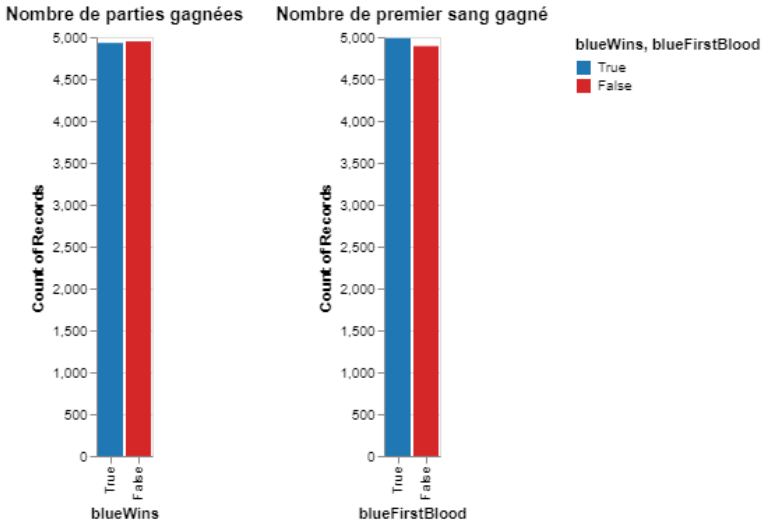

Figure 1 – *Répartition du nombre de victoires et du premier sang*

Nous remarquons que la répartition **est équilibrée**, autant pour les premiers sangs que pour les victoires (même si les bleus on l'air d'avoir plus de premier sang, sûrement dû à la taille du dataset).

Pour finir la vérification de la qualité des données, nous avons recherché des valeurs aberrantes. Pour ce faire, nous avons regardé la valeur minimale et maximale pour chaque variable du dataset en utilisant le code suivant.

```
for field in data.columns:
    print(field + " max : " + str(max(data[field])) + " min: " + str(min(data[field])))
    print(data[field].describe())
    print("")
```

Pour chaque variable, nous avons **analysé** les chiffres (grâce à nos nombreuses heures de jeu) afin de savoir s'ils sont **cohérents** avec ce que l'on peut retrouver en partie.
Voici un exemple de ce qui est affiché pour une variable :

blueWins max : True	min : False
count	9879
unique	2
top	False
freq	4949
Name : blueWins,	dtype : object

Exemple de variable

Les résultats sont très **satisfaisants**, il n'y avait aucune valeur qui sortait de sa range normale.

Nous avons ensuite souhaité regarder s'il y avait des parties **non-intéressantes** (sans premier sang avant 10 min de jeu). Pour ce faire, nous avons utilisé le code python suivant :

```
data['boring'] = ((data['blueFirstBlood'] == False) & (data['redFirstBlood'] == False))
data['boring'].describe()
```

Voici le résultat que nous avons obtenu :

count	9879
unique	1
top	False
freq	9879
Name : boring,	dtype : object

Repérage des parties ennuyantes

Comme on peut le déduire, il n'y a **aucune partie ennuyante** dans le dataset que nous avons choisi. Inutile donc de supprimer des lignes.
La dernière modification que nous avons apportée au dataset est la **transformation** de valeurs flottantes en booléennes pour les champs qui sont booléens. Pour ce faire, nous avons utilisé le code python suivant :

```
# Convertir les blueWin en booleen pour les graphiques
data['blueWins'] = data['blueWins'].astype(bool)
data['blueFirstBlood'] = data['blueFirstBlood'].astype(bool)

data["blueCSDiff"] = data["blueCSPerMin"] - data["redCSPerMin"]
data["redCSDiff"] = data["redCSPerMin"] - data["blueCSPerMin"]
```

Nous vérifions aussi l'équilibre des victoires du **second dataset** :

Figure 2 – *Répartition du nombre de victoires*

Nous allons pouvoir utiliser sereinement les données puisqu'on est maintenant sûr qu'elles sont **fiables** et représentatives.

2.2 Peut-on prédire l'issue d'une partie à 10 minutes de jeu ?

Cette question qui est de **savoir s'il faut se rendre ou non** dans le jeu fait grand débat depuis très longtemps et divise la communauté. Nous même nous la posons, c'est pour cette raison que nous avons essayé d'y répondre.

Pour commencer, nous avons voulu savoir **quelles sont les variables qui sont le plus corrélé avec la victoire**. Nous avons créé la matrice de corrélation de nos variables et ensuite gardé la colonne qui relie chaque variable à la victoire bleue. Nous avons ensuite représenté les **corrélations** sur la visualisation suivante :

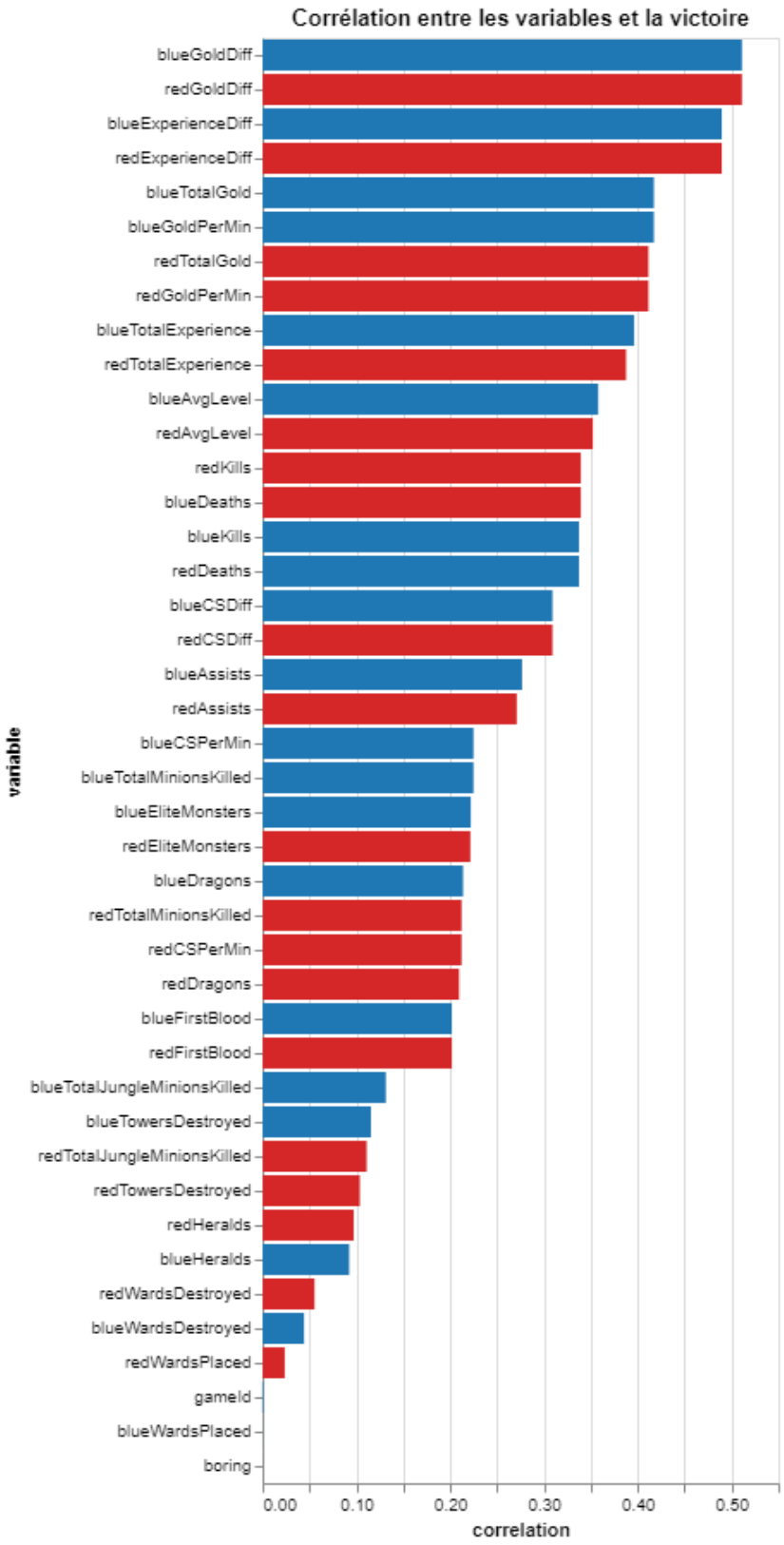

Figure 3 – *Corrélation entre les variables et la victoire*

Au vu des résultats, **on ne peut pas dire que l'issue d'une partie est sûre à 10 minutes de jeu**, car il n'y a pas de variables en corrélation forte avec la victoire.

Cependant, nous avons décidé de mener une **analyse plus approfondie**. Il s'agit d'entraîner une régression logistique avec nos données afin d'essayer de prédire la victoire. Nous observons ensuite la précision de notre modèle afin de savoir s'il est possible de **prédire de manière précise l'issue d'une partie** avec les données disponibles à 10 minutes de jeu.

Pour ce faire, nous avons entrainé une **regression logistique** en prennant successivement les **variables** qui ont une **corrélation supérieure à un certain seuil** qui vit dans la liste suivante [**0.1**, **0.2**, **0.3**, **0.4**, **0.5**]. Pour chaque modèle, nous avons divisé le dataset en deux parties, une pour l'**entraînement** et une pour **tester** la précision avec une proportion de 0.2 pour le test set.

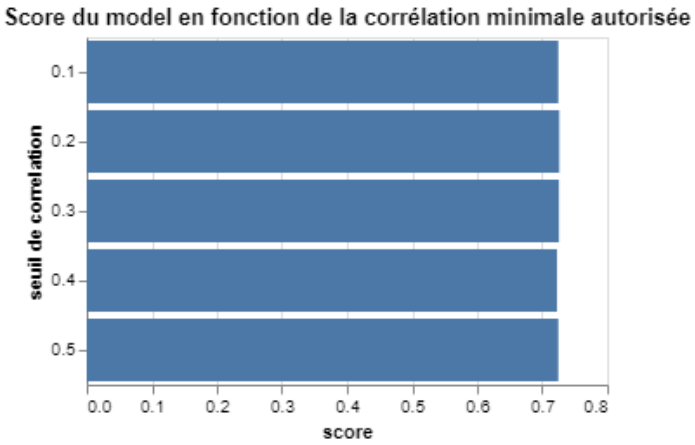

Figure 4 – *Précision de la régression en fonction des variables choisies*

Ce qu'on peut observer et déduire de cette représentation : les modèles sont **performants**, on a une **accuracy de 0.73** en moyenne. On peut dire que les **variables sont bien utiles** pour prédire l'issue d'une partie avec les données disponibles à 10 minutes de jeu, **mais pas suffisamment** pour l'affirmer avec grande certitude.

De plus, le fait que le score de précision ne varie pas significativement en retirant les variables les moins corrélées nous montre que la différence de golds a une **importance prédominante** pour le modèle.

Les **golds** sont obtenus en tuant des unités (majoritairement), ce qui **rapporte également de l'expérience** aux joueurs. Pour le vérifier, nous avons calculé une **corrélation de 0.89 entre l'avantage en golds et l'avantage en expérience**. Cela explique pourquoi en rajoutant des features, notre modèle ne performe pas mieux.
Voici un nuage de point qui nous montre cette corrélation :

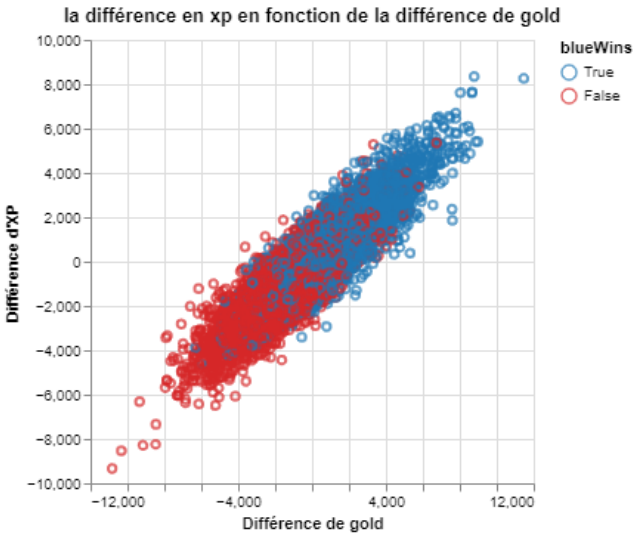

Figure 5 – *Différence d'xp en fonction de la différence de golds*

Tout ce que nous avons déduit auparavant est visible sur cette visualisation. Nous voyons l'**importance** du paramètre différence de gold et une corrélation forte entre l'avantage en **gold** et l'avantage en **expérience**.

Pour revenir à la question posée initialement, on ne peut **pas donner de réponse directe**. Mais on a pu remarquer certaines propriétés et les confirmer. Ce qu'on peut quand même affirmer, c'est qu'avec les données disponibles à 10 min de jeu, une régression logistique est capable de **prédire quasiment 3 fois sur 4 la victoire** correctement.

2.3 À partir de quel avantage de gold/expérience peut-on considérer une partie gagnée et perdue ?

Dans cette partie, nous avons voulu savoir s'il y avait des **seuils** à partir desquels il n'est **plus possible** de revenir dans la partie et si oui lesquels ?
Pour ce faire, nous avons créé les visualisations suivantes :

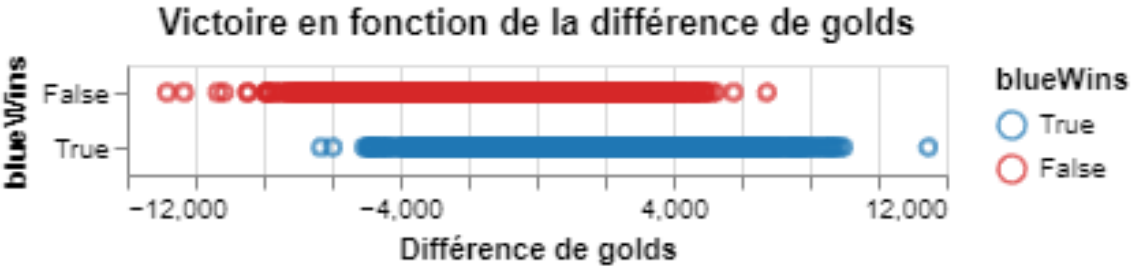

Figure 6 – *Visualisation de la victoire en fonction de la différence de golds*

Cette visualisation nous permet de voir qu'**il existe bien un seuil** de retard en golds à partir duquel il n'y a **plus de victoires** observées, et inversement pour l'avance en golds, on voit un seuil à partir duquel il n'est plus observé de défaites.

Cependant, nous pouvons quand même critiquer cette visualisation, car on ne perçoit **pas** la densité de points qui se superposent. Il nous faut donc une deuxième manière de la représenter qui nous permettra d' **apprécier** cette répartition et ainsi d'identifier les seuils.

Pour ce faire, nous avons pris comme seuils les bornes d'un **intervalle de confiance de 95%**, c'est-à-dire les **quantiles** associés.

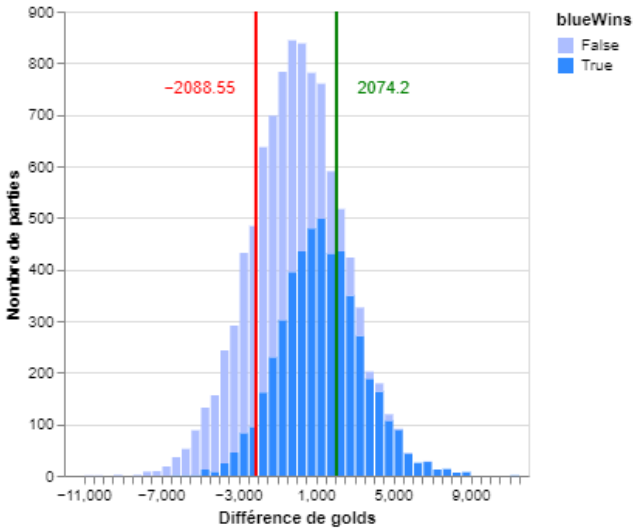

Figure 7 – *Visualisation de la répartition de l'avance en golds*

Sur cette visualisation, on peut maintenant voir qu'il existe un seuil grâce à la répartition et on peut même l'**identifier précisément** grâce à une droite et à un indicateur.

On voit donc qu'**au-delà d'environ 2100 golds de différence a 10 min, on peut considérer a 95% que la partie est perdue ou gagnée**.

Par principe, nous avons fait de même avec l'avance en **expérience** même si on a vu précédemment que ces deux paramètres étaient corrélés.

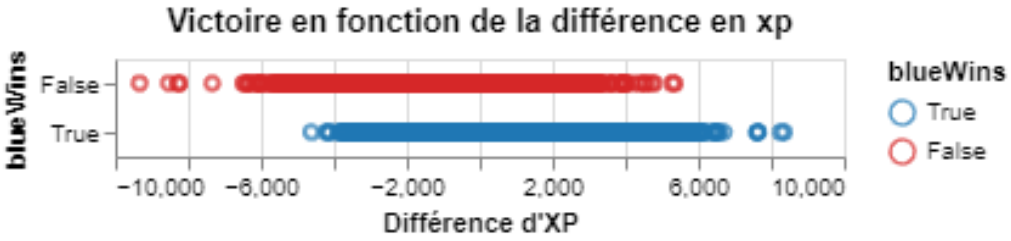

Figure 8 – *Visualisation de la victoire en fonction de la différence d'expérience*

Comme on le voit, l'avance en expérience a un effet **similaire** sur la victoire ou la défaite que l'avance en golds.

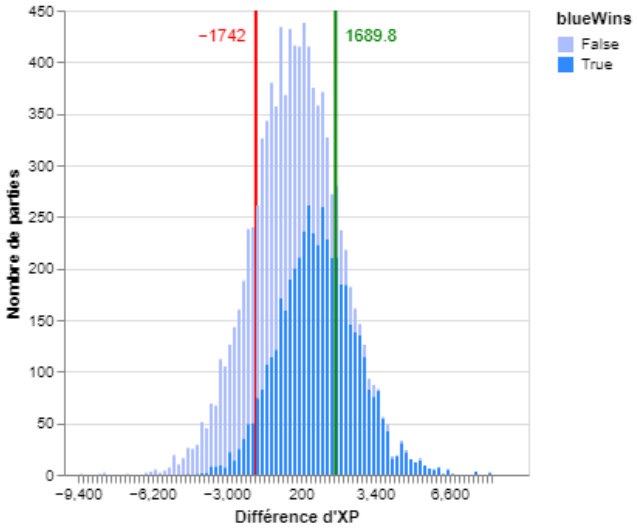

Figure 9 – *Visualisation de la répartition de l'avance en expérience*

Même remarque pour cette visualisation, nous avons pu identifier les **deux seuils**. On peut donc répondre de façon certaine à la question en affirmant qu'on peut considérer qu'**une partie est gagnée ou perdue si l'avance ou le retard en golds et en expérience dépasse un certain seuil**.

Par contre, il faut garder en tête que cette analyse est basée sur notre dataset, certes assez important mais pas représentatif de la réalité future. Le dataset que nous avons utilisé est basé sur des parties au rang diamant, ces joueurs sont beaucoup plus capable d'utiliser leurs avantages. Nous n'aurions sans doute pas eu ce genre de résultats avec des parties en fer.

2.4 Est ce qu'un coté de la carte a des avantages par rapport au coté opposé ?

Dans League of Legends, les **deux équipes s'affrontent et se font face** à chacune d'un côté de la carte (l'une dans le coins inférieur gauche et l'autre dans le coin supérieur droit). Certains joueurs peuvent

être **gênés** par l'affichage du menu des sorts et des objets ainsi que de la mini-carte qui prennent une certaine taille sur l'écran.

Il y a également des **emplacements prédéfinis** pour l'apparition de certaines créatures d'élite qui peuvent **favoriser l'accès** à une équipe plutôt qu'une autre. On peut donc imaginer qu'il y ait un léger avantage pour un côté.

Nous avons donc décidé de créer une visualisation pour voir les **écarts** sur certains paramètres intéressants en fonction du **côté** duquel se situe l'équipe.

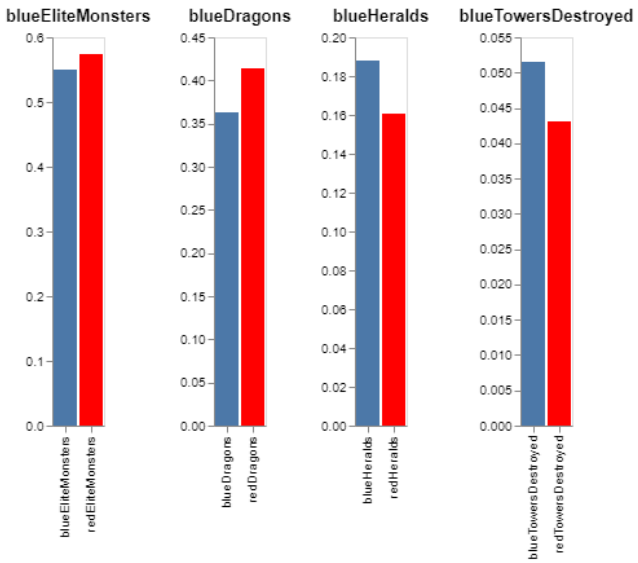

Figure 10 – *Visualisation de l'avantage sur certains paramètres en fonction du côté*

Comme on peut facilement le remarquer, il y a bien des **différences** observables en fonction du côté. L'équipe **bleue** parvient plus facilement à tuer le **héraut de la faille**, car il est plus facile d'accès du cote équipe bleue. Il permet aussi de **détruire plus facilement des tourelles** et cet **avantage se reporte** donc aussi sur le **nombre de tourelles détruites** à 10 minutes de jeu.

Cependant, il y a aussi un avantage pour l'équipe **rouge** qui est observable, car il leur est **plus simple de tuer le dragon** en étant du côté rouge. De plus, les dragons leur donnent des **bonus de statistiques** de façon permanente.

Le jeu est donc **équilibré** de cette manière : les **bleus** ont un meilleur accès aux **hérauts** et les **rouges** ont un meilleur accès aux **dragons**. En effet, nous avions vu précédemment qu'**aucun** côté de la carte n'apportait plus de possibilités de gagner.

2.5 Quel dragon donne le plus d'avantage à une équipe ?

On va maintenant se pencher sur les monstres épiques qui apparaissent plus tard dans la partie. Le but est de déterminer lequel donne le plus de chance de mener l'équipe a la victoire.

On commence par regarder les corrélations entre l'obtention de monstres épiques et la victoire :

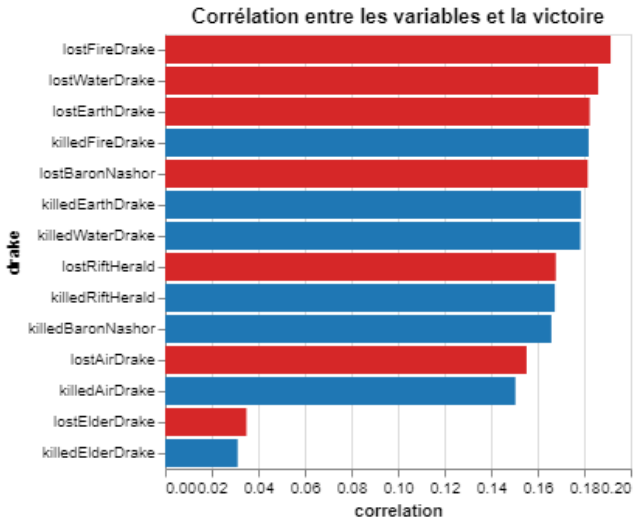

Figure 11 – *Impact des monstres épiques*

À première vue, c'est le dragon infernal qui est le plus avantageux. Mais au vu des données, on va essayer d'ajuster le dataset.

Pour cela **grouper** les "lost" et les "killed" en une seule variable **"Diff"** = "killed" - "lost" est une bonne solution. De cette manière, on distingue bien toutes les situations possibles :

En effet lorsque que les valeurs de killed ou de lost valent **0** on ne peux pas savoir si il s'agit d'une partie qui ne contient **pas** ce monstre ou si il s'agit de l'**adversaire qui l'a obtenue** ce qui **limite** la valeur de 0 au yeux de la corrélation. Avec diff lorsque l'adversaire obtient le bonus la valeur vaut donc **-1**. Cependant avec cette nouvelle valeur on peut **confondre** l'obtention d'un même bonus pour les deux équipe ou l'**absence** de celui-ci. Dans les deux cas on peut considérer que ce bonus ne sera **pas décisif** ce qui est cohérent.

Voici les corrélations obtenues en utilisant les diffs :

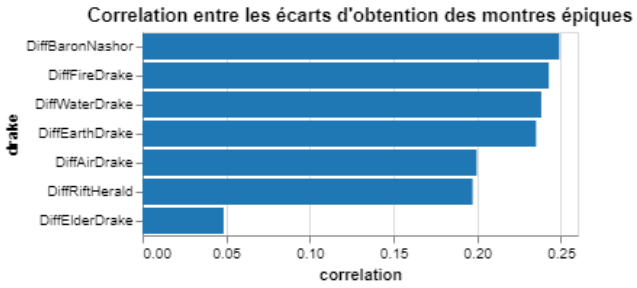

Figure 12 –

On remarque sur ce graphique que le **baron Nashor** passe en première position. Tout joueur de League of Legend pouvait s'attendre à ce résultat. En effet le baron Nashor apparaît **plus tard** dans la partie mais peut avoir de gros impact voir **mettre fin** à une partie. Il est possible d'expliquer son absence de la première place dans le précédent résultat puisqu'en étant disponible plus tard, il est **absent** de plus de parties. Il subissait donc encore plus les problèmes précédents.

Cependant un autre monstre épique sort du lot par sa corrélation très faible avec la victoire, l'**Elder Drake** ou Dragon Ancestrale en français. Est-t-il vraiment inutile ? En réalité c'est plus **compliqué**, ce monstre n'apparaît que dans très peu de parties il est donc très sous représenté.

Refaisons les calculs en ne prenant **que** les parties où celui-ci est atteint :

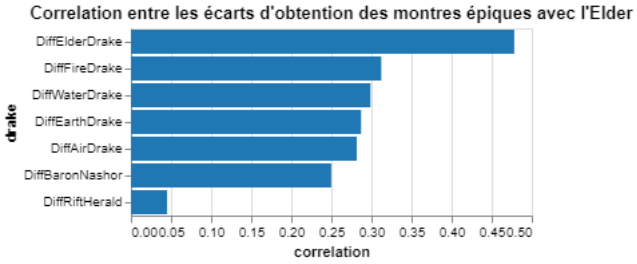

Figure 13 –

On remarque alors l'**importance** de ce monstre épique, il ne s'agit pas d'une victoire assurée mais il reste le plus **gros coefficient** que l'on a obtenu. on remarque aussi que dans ce type de partie l'importance du Baron Nashor **diminue**, cela peut l'expliquer par le fait que la puissance de l'Elder Drake **dépend** du nombre de dragon précédemment obtenue par l'équipe.

Pour conclure cette partie, le dragon le plus intéressant parmi les quatre élémentaires est le **dragon infernal**, il reste tout de même moins important que le **baron Nashor** dans la plupart des parties. le Nashor est donc dans la plupart des cas le **meilleur monstre épique** à obtenir mais si vous avez le choix entre le Baron Nashor et le Dragon Ancestral c'est le **Dragon Ancestral** qu'il faut privilégier.

3 Conclusion

Après nous être assurés de la qualité et réalisés quelques transformations pour faciliter la suite, nous nous sommes posés quatre questions. Pour autant, nous avons répondu à plus de quatre questions, car nous avons pu aller plus profond en découvrant ce qui se cachait derrière certaines et en explorant nos hypothèses.

Ce mini projet nous a donné l'occasion de trouver des réponses à des questions que nous nous posions depuis longtemps. Nous avons pris plaisir à pousser l'étude, car nous avons choisi un sujet d'exploration qui nous plaît tous et cela nous a permis de bien nous faire la main avec altair et pandas.

4 Glossaire

- **Golds** : ressource principale s'obtenant en tuant des sbires, des joueurs ennemis, des tourelles, etc. Les golds renforce la puissance du joueur en lui permettant d'acheter des objets qui donnent des statistiques.
- **Expérience** ou **XP** : points permettant d'augmenter le niveau et les statistiques d'un joueur. Elle s'obtient en tuant des sbires et des joueurs.
- **First blood** ou **premier sang** : nom donné à la première élimination d'un joueur adverse. Elle donne des golds bonus.
- **Sbires** ou **CS** : unités appartenant à une équipe et réapparaissant tout au long de la partie, ils sont la principale source de revenus de golds et d'XP pour les joueurs.
- **Monstres épiques** : cela comprend les dragons, le héraut de la faille et le baron Nashor. Ce sont des monstres neutres qui confèrent de grands bonus à l'équipe qui parvient à les tuer.
- **Tourelles** : structures qui protègent la base des équipes. Il faut les détruire pour rentrer dans la base adverse.
- **Wards** ou **totems de vision** : balise que les joueurs placent au sol pour révéler les environs pendant quelques minutes.
- Équipe **bleue** / équipe **rouge** : appellation utilisée pour différentier les équipes. La carte est sous forme de carré dans lequel l'équipe bleue a sa base en bas à gauche tandis que celle de l'équipe rouge est en haut à droite.

5 Annexe

Figure 14 – *Carte du jeu*

5.1 Dataset

Premier dataset sur les statistiques des parties à 10 min :
https://www.kaggle.com/datasets/bobbyscience/league-of-legends-diamond-ranked-games-10-min
Deuxième dataset sur les montres épiques obtenu au cour le la partie :
https://www.kaggle.com/datasets/bobbyscience/league-of-legends-soloq-ranked-games