Aabu

De l'importance des qualifications au Grand Prix de Monaco

6 juin 2019

Table des matières

Il y a quelques semaines se déroulait le Grand Prix de Monaco de Formule 1. À cette occasion, j'ai entendu dire dire que les qualifications y jouent un rôle prépondérant, car il y est difficile de doubler. Quoi de mieux que de vraies données pour tenter de vérifier cela ?

1. Trouver des données

Les données de la Formule 1 sont relativement faciles à trouver, mais pas toujours sous une forme correcte.

On notera en particulier que Wikipédia comporte quasiment tous les résultats, notamment la grille de départ (conséquence directe des qualifications modulo les pénalités) et le résultat de la course. C'est ce que j'ai utilisé au tout début pour explorer un peu les données et faire une preuve de concept avec un tableau.

Le mieux serait peut-être d'utiliser le site officiel de la discipline, qui contient une archive de tous les résultats qui m'intéressent, mais il n'est pas utilisable directement. Pour l'utiliser, il faut en extraire les données. Ce que j'ai retenu de faire.

Il existe aussi des jeux de données bien présentés, certains sites nécessitent une inscription et d'autres sont suffisamment mal référencés pour que je les trouve *après* avoir commencé à travailler les données (trop tard pour changer de plan)... J'aurai probablement utilisé ce deuxième site si je l'avais trouvé avant.

Il y a même des sites souvent commerciaux qui proposent même des API pour recevoir tout plein de données en quasi-temps réel et qui ont une granularité qui va jusqu'au tour par tour !

2. Ma boîte à outils

Comme dit plus haut, j'ai choisi d'extraire les données depuis le site officiel de la Formule 1.

J'ai procédé avec ce qui semblait le plus facile d'accès pour moi :

— Python, mon langage de prédilection ;

— `urllib` pour récupérer le contenu des pages web ;
— `lxml` pour extraire les données des pages, avec en particulier la possibilité de faire des requêtes XPath que j'avais déjà utilisées par ailleurs et qui sont très pratiques ;
— `matplotlib` pour regarder des jolis graphiques ;
— `numpy` pour calculer sur mes données le cas échéant.

3. Traitement des données

La partie la plus pénible et que je n'ai pas eu le courage d'automatiser est la récupération des URL intéressantes. Je l'ai fait à la main et me suis limité à une quarantaine de pages en tout, pour que cela reste pratiquable. Il y a sûrement moyen de le scripter en observant la structure des pages, mais ce n'était pas l'intérêt premier de mon exploration.

À partir de là, j'ai réalisé un script qui :

— télécharge les pages web ;
— extrait les données des pages web (position, nom du pilote, etc.) ;
— traite les données pour en sortir les infos qui m'intéressent (voir la partie suivante) ;
— affiche ces données pour les humains.

Il n'y a normalement là rien d'incroyable pour un développeur expérimenté, mais ce n'est pas vraiment mon cas, et j'ai appris plein de choses (en particulier tout ce qui touche au *scraping*) et ai aussi eu une piqûre de rappel sur d'autres choses (XPATH).

4. Résultats de l'analyse

Je n'ai pas passé trop de temps à réfléchir et ai opté pour l'approche qui consiste à analyser la relation entre la position de départ et d'arrivée.

4.1. Probabilité d'arriver x-ième en partant de la y-ième place

Quand on dit que les qualifications sont importantes, c'est une manière de dire qu'il sera difficile de remonter dans le classement au cours de la course. Une manière de quantifier ça, c'est de regarder pour chaque position de départ la probabilité d'arriver à une position d'arrivée donnée. Tout cela forme une matrice, où l'élément à la colonne i et la ligne j est la probabilité d'arriver à la j-ième place en partant de la i-ème position sur la grille de départ.

Le calcul est fait à l'aide des données des dix derniers Grand Prix.

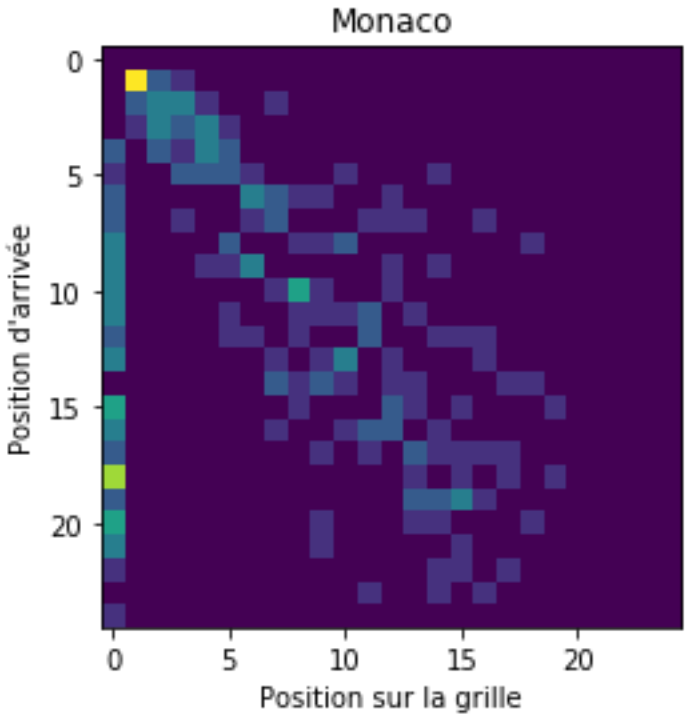

FIGURE 4. – Matrice pour Monaco.
(Dans cette matrice, la colonne de gauche correspond aux pilotes non-classés).

Avec ça tout seul, on n'ira pas très loin, aussi j'ai décidé de comparer avec une course réputée pour ses dépassements nombreux : le Grand Prix de Chine. On utilise toujours les dix derniers grands prix.

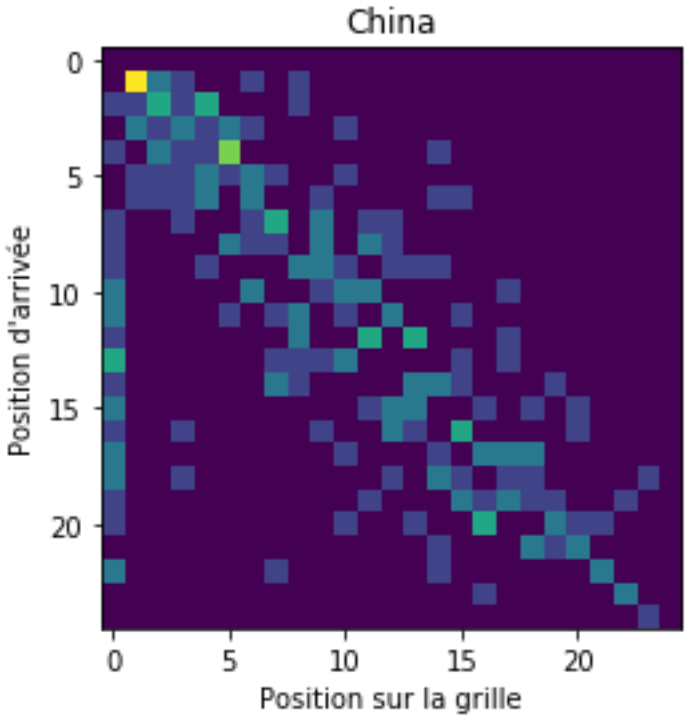

FIGURE 4. – Matrice pour la Chine.

Les deux matrices sont assez différentes, c'est intéressant ! Alors, à l'œil nu, je remarque des choses :

— dans les deux cas, celui qui part premier à plus de chance d'arriver premier (la corrélation est toute trouvée : être un bon pilote dans une bonne voiture conduit à être le plus rapide en qualification **et** en course).
— le Grand Prix de Monaco semble être ramassé à la tête (en gros la tête de course est globalement figée) et éclatée vers la fin, avec pas mal de non-classés (abandons par exemple) ;
— le Grand Prix de Chine présente une grosse bande large, qui signifie qu'on peut relativement facilement gagner et perdre quelques places quelle que soit la position de départ.

4.2. Mesure de diagonalité

Si on imagine une course sans possibilité de dépasser la matrice serait diagonale. Si on arrive à mesurer la diagonalité d'une matrice, on pourrait alors quantifier la facilité de dépassement sur une course !

J'ai eu un peu de mal à trouver les bons mots-clés pour ma recherche, mais j'ai fini par trouver une méthode intéressante . On a un score de diagonalité r. Grossièrement, plus r est proche de 1, plus la matrice est diagonale. Plus r est proche de zéro, moins c'est le cas. Statistiquement, c'est un genre de corrélation entre les différents vecteurs ligne et colonne.

Pour la matrice complète, on trouve :

— r = 0.364 pour Monaco ;
— r = 0.609 pour la Chine.

C'est l'inverse de l'attendu ! Essayons d'éliminer le nombre d'abandons qui semble grand à Monaco de l'équation en retirant la première colonne (et ligne) :

— r = 0.796 pour Monaco ;
— r = 0.818 pour la Chine.

Ce coup-ci, c'est plus proche, mais on a peut-être juste trouvé la valeur globale pour un Grand Prix de Formule 1, où la qualification est *de toute façon* importante, quel que soit le circuit. Si on s'intéresse seulement à la tête de course, on a pour le carré des cinq premiers :

— r = 0.679 pour Monaco ;
— r = 0.457 pour la Chine.

Ah ! Quelque chose qui correspond à ce qu'on voit graphiquement. Ces deux valeurs signifient que pour la tête de la course, le résultat à Monaco est plus dépendant de la qualification que pour la Chine.

Cet indicateur (comme l'espérance par exemple) est *trop* synthétique : on perd une partie de l'information, et la précaution doit être de mise dans sa manipulation.

Alors, les qualifications sont-elles plus importantes à Monaco qu'ailleurs (en Chine en l'occurrence) ? Oui, mais seulement quand on ne considère que la tête de course. Pour le reste, c'est plus incertain.

Les données de la Formule 1 (et d'autres sports) regorgent d'informations cachées (sauf peut-être des parieurs professionnels), et je pourrai encore y passer des heures. Heureusement, le reste de ma vie me rappelle à la réalité.