



DataScientest

Challenge Sorbonne

Sujet et consignes

Sujet

1.Introduction

Le sujet de ce challenge porte sur la Formule 1.

La Formule 1 (F1) est une compétition automobile internationale avec des voitures de course monoplaces régie par la Fédération Internationale de l'Automobile (FIA). Le mot « Formule » dans le nom fait référence à l'ensemble des règles auxquelles doivent se conformer les voitures de tous les participants.

Une saison de Formule 1 consiste en une série de courses, appelées Grands Prix, qui se déroulent dans le monde entier sur des circuits construits à cet effet.

| F1 UPDATED 2021 SEASON CALENDAR | | |
|--|---|--|
| R1 |  | 26-28 MAR BAHRAIN SAKHIR |
| R2 |  | 16-18 APR ITALY IMOLA |
| R3 |  | 30 APR-02 MAY PORTUGAL PORTIMAO |
| R4 |  | 07-09 MAY SPAIN BARCELONA |
| R5 |  | 20-23 MAY MONACO MONACO |
| R6 |  | 04-06 JUN AZERBAIJAN BAKU |
| R7 |  | 18-20 JUN FRANCE LE CASTELLET |
| R8 |  | 25-27 JUN AUSTRIA SPIELBERG (STYRIAN GP) |
| R9 |  | 02-04 JUL AUSTRIA SPIELBERG (AUSTRIAN GP) |
| R10 |  | 16-18 JUL UNITED KINGDOM SILVERSTONE |
| R11 |  | 30 JUL-01 AUG HUNGARY BUDAPEST |
| R12 |  | 27-29 AUG BELGIUM SPA |
| R13 |  | 03-05 SEPT NETHERLANDS ZANDVOORT |
| R14 |  | 10-12 SEPT ITALY MONZA |
| R15 |  | 24-26 SEPT RUSSIA SOCHI |
| R16 |  | 08-10 OCT TURKEY ISTANBUL |
| R17 |  | 22-24 OCT USA AUSTIN |
| R18 |  | 05-07 NOV MEXICO MEXICO CITY |
| R19 |  | 12-14 NOV BRAZIL SAO PAULO |
| R20 | | 19-21 NOV TBC |
| R21 |  | 03-05 DEC SAUDI ARABIA JEDDAH |
| R22 |  | 10-12 DEC ABU DHABI YAS ISLAND |

Note : le Grand Prix n°20 a finalement eu lieu au Qatar.

2.Format de compétition

Il existe deux championnats du monde annuels dans cette discipline :

- Le championnat du monde des pilotes (saison inaugurale en 1950).
- Le championnat du monde des constructeurs (saison inaugurale en 1958).

Chaque constructeur F1 dispose d'une écurie avec deux pilotes et les résultats de chaque course sont évalués par un système de points. Les pilotes d'une même écurie doivent travailler en équipe pour contribuer au championnat des constructeurs, tout en étant en concurrence entre eux et avec les autres pilotes de la grille pour remporter le championnat des pilotes. Pour cette raison, la Formule 1 est à la fois un sport d'équipe et un sport individuel.



Un Grand Prix de F1 se déroule sur 3 jours le week-end et se compose de 3 parties : les essais, les qualifications et la course.

- Les essais se déroulent en 3 séances : FP1, FP2 et FP3. Ce sont des sessions d'essais libres permettant aux équipes de tester leurs voitures le vendredi et le samedi.
- La séance de qualifications se déroule en 3 étapes : Q1, Q2 et Q3 (format actuel mise en place en 2006). Durant la Q1, tous les pilotes s'affrontent pour réaliser le meilleur temps au tour. Les 15 meilleurs pilotes participeront à la Q2 et les derniers pilotes seront éliminés. Même principe pour la Q2, où seuls les 10 meilleurs pilotes passeront à la Q3. C'est cette ultime session de qualification qui détermine l'ordre sur la grille de départ, le meilleur temps de la Q3 démarrera la course en pole position, le deuxième temps en seconde place, etc.

Note: Les positions sur la grille de départ peuvent être différentes de celles des qualifications, en raison de pénalités ou de problèmes mécaniques.

- La course se déroule le dimanche. Les pilotes doivent effectuer un nombre défini de tours (300km + 1 tour) en appliquant une stratégie (arrêts aux stands, choix des pneumatiques, quantité d'essence et ravitaillement, ...) pour terminer à la meilleure

place possible. A l'issue de la course, les points sont attribués aux 10 premiers pilotes et les trois premiers montent sur le podium.

GRAND PRIX DE FRANCE
CIRCUIT PAUL RICARD

TOURS : **53** | LONGUEUR : **5.842 KM** | DISTANCE : **309.690 KM**
RECORD CIRCUIT : **1:28.319** HAMILTON 2019

VENDREDI 18 JUIN

| | | |
|------------------------|---------------|---------------------|
| ESSAIS LIBRES 1 | 11:30 - 12:30 | CANAL+ SPORT |
| ESSAIS LIBRES 2 | 15:00 - 16:00 | CANAL+ SPORT |

SAMEDI 19 JUIN









| | | |
|------------------------|---------------|---------------------|
| ESSAIS LIBRES 3 | 12:00 - 13:00 | CANAL+ SPORT |
| QUALIFICATIONS | 15:00 - 16:00 | CANAL+ |

DIMANCHE 20 JUIN

| | | |
|---------------|-------|------------------|
| COURSE | 15:00 | CANAL+ C8 |
|---------------|-------|------------------|

motorsport.com

A l'issue de la saison, le pilote ainsi que l'écurie ayant cumulé le plus de points remportent respectivement le championnat du monde des pilotes et le championnat des constructeurs.

|  2021 TEAM STANDINGS AFTER BRAZILIAN GRAND PRIX | | |
|---|---------------------|--------------|
|  | MERCEDES | 521.5 |
|  | RED BULL | 510.5 |
|  | FERRARI | 287.5 |
|  | McLAREN | 256 |
|  | ALPINE | 112 |
|  | ALPHATAURI | 112 |
|  | ASTON MARTIN | 68 |
|  | WILLIAMS | 23 |
|  | ALFA ROMEO | 11 |
|  | HAAS | 0 |

3. Data et Formule 1

Le monde de la formule 1 a toujours été à la pointe de l'innovation en inventant de nombreuses technologies que l'on retrouve aujourd'hui dans les voitures de série. On peut citer notamment l'invention de l'injection directe, du frein à disque, de l'aileron, de l'ABS, de l'antipatinage, des palettes au volant...

Aujourd'hui cette innovation se retrouve dans l'utilisation des données. Lors de chaque course, 120 capteurs positionnés sur chaque voiture génèrent 3 Go de données et 1 500 relevés chaque seconde !

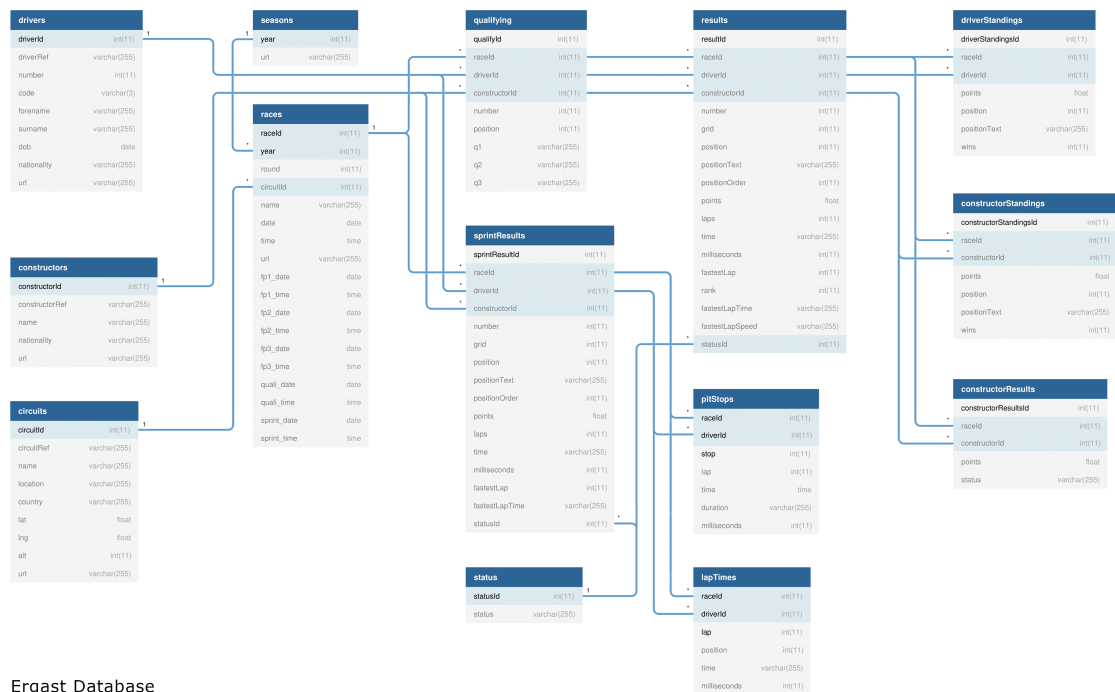
Ces données sont utilisées pour extraire des statistiques mais aussi pour établir des prédictions de course.

Une grande partie de ces données est rendue disponible au public à la fin de chaque week-end de Grand Prix par l'intermédiaire du groupement d'entreprises « Formula 1 » qui est responsable de la promotion du Championnat du monde FIA de Formule 1.

Dans le cadre de ce challenge, nous vous fournissons un extrait de la base de données [Ergast](#) qui contient des données historiques de la Formule 1 depuis le début de la discipline

en 1950.

Voici la structure des données ainsi que la description de chacune des tables :



Ergast Database

L'image en haute résolution est accessible [ici](#).

Vous trouverez ci-dessous une description des tables par ordre alphabétique :

- **circuits**: La table `circuits.csv` fournit les informations des différents circuits sur lesquels se sont déroulés les Grands Prix dans l'histoire de la Formule 1.
 - `circuitId` = ID (clé primaire)
 - `circuitRef` = ID (texte)
 - `name` = Nom du circuit
 - `location` = Ville
 - `country` = Pays
 - `lat` = Latitude
 - `lng` = Longitude
 - `alt` = Altitude
 - `url` = Page Wikipédia
- **constructor_results**: La table `constructor_results.csv` fournit les résultats des écuries à l'issue des Grands Prix.
 - `constructorResultsId` = ID (clé primaire)
 - `raceId` = ID du Grand prix
 - `constructorId` = ID écurie
 - `points` = Points gagnés par l'écurie

- status = 'D' pour disqualifié ou valeur nulle
- constructor_standings: La table `constructor_standings.csv` fournit le classement actualisé des écuries au Championnat Constructeurs à l'issue de chaque Grand Prix.
 - constructorStandingsId = ID (clé primaire)
 - raceId = ID du Grand prix
 - constructorId = ID écurie
 - points = Points gagnés par l'écurie en cumulé
 - position = Position écurie au classement constructeurs
 - positionText = Idem 'position' au format texte
 - wins = Nombre de victoires cumulées au fil d'une saison
- constructors: La table `constructors.csv` fournit les informations sur les différentes écuries ayant participé aux championnats dans l'histoire de la Formule 1.
 - constructorId = ID (clé primaire)
 - constructorRef = ID (texte)
 - name = Nom écurie
 - nationality = Nationalité
 - url = Page Wikipédia
- driver_standings: La table `driver_standings.csv` fournit le classement actualisé des pilotes au Championnat à l'issue de chaque Grand Prix.
 - driverStandingsId = ID (clé primaire)
 - raceId = ID du Grand prix
 - driverId = pilote
 - points = Points gagnés par le pilote en cumulés
 - position = Position au classement pilote
 - positionText = position au format texte
 - wins = Nombre de victoires cumulées au fil d'une saison
- drivers: La table `drivers.csv` fournit les informations sur les différents pilotes ayant participé aux championnats dans l'histoire de la Formule 1.
 - driverId = ID (clé primaire)
 - driverRef = ID (texte)
 - number = Numéro du pilote ou valeur nulle
 - code = Initiales du pilote
 - forename = Prénom
 - surname = Nom

- dob = date de naissance
 - nationality = nationalité
 - url = page Wikipédiaia
- lap_times: La table `lap_times.csv` fournit les chronomètres par tour des pilotes à chaque Grand Prix. *Note : Les données sur les chronomètres sont disponibles à partir de la saison 1996.*
 - raceId = ID du Grand Prix
 - driverId = ID du pilote
 - lap = Numéro du tour
 - position = Positon du pilote sur le tour
 - time = Chrono du pilote sur le tour
 - milliseconds = Chrono converti en millisecondes
- pit_stops: La table `pit_stops.csv` fournit les arrêts aux stands effectués par les pilotes à chaque Grand Prix. *Note : Les données sur les arrêts au stand sont disponibles à partir de la saison 2011.*
 - raceId = ID du Grand Prix
 - driverId = ID du pilote
 - stop = Numéro arrêt aux stands
 - lap = Numéro du tour
 - time = Temps début arrêt
 - duration = Durée arrêt
 - milliseconds = Durée convertie en millisecondes
- qualifying: La table `qualifying.csv` fournit les résultats des séances de qualifications des pilotes à chaque Grand Prix. *Note : Les résultats des qualifications ne sont entièrement pris en charge qu'à partir de la saison 2003.*
 - qualifyId = ID (clé primaire)
 - raceId = ID du Grand Prix
 - driverId = ID du pilote
 - constructorId = ID écurie
 - number = Numéro du pilote
 - position = Position pour la grille de départ de la course
 - q1 = Chrono en première phase de qualification
 - q2 = Chrono en deuxième phase de qualification ou nul si disqualifié Q1
 - q3 = Chrono en troisième phase de qualification ou nul si disqualifié Q2

- **racetracks**: La table `racetracks.csv` fournit les informations sur les différents Grands Prix organisés dans l'histoire de la Formule 1.
 - `trackId` = ID (clé primaire)
 - `year` = Année
 - `round` = Numéro du Grand Prix sur la saison
 - `circuitId` = ID du circuit
 - `name` = Nom du Grand prix
 - `date` = Date du Grand prix
 - `time` = Heure départ de la course
 - `url` = Page Wikipédia
 - `fp1_date` | `fp1_time` = Date et heure départ des essais libres 1 (ou nul)
 - `fp2_date` | `fp2_time` = Date et heure départ des essais libres 2 (ou nul)
 - `fp3_date` | `fp3_time` = Date et heure départ des essais libres 3 (ou nul)
 - `quali_date` | `quali_time` = Date et heure départ de la séance de qualifications (ou nul)
 - `sprint_date` | `sprint_time` = Date et heure départ de la course Sprint (ou nul)
- **results**: La table `results.csv` fournit les résultats des Grands Prix organisés dans l'histoire de la Formule 1. *Note : Les données sur le tour le plus rapide des pilotes sont disponibles à partir de la saison 2004.*
 - `resultId` = ID (clé primaire)
 - `trackId` = ID du Grand Prix
 - `driverId` = ID du pilote
 - `constructorId` = ID écurie
 - `number` = Numéro du pilote
 - `grid` = Position sur la grille de départ
 - `position` = Position à la ligne d'arrivée (ou nul)
 - `positionText` = Idem 'position' au format texte
 - `positionOrder` = Numéro pour tri
 - `points` = Points gagnés par le pilote à l'issue de la course
 - `laps` = Nombre de tours complétés
 - `time` = Temps total cumulé à l'issue de la course pour le pilote arrivé 1er / écart par rapport au 1er
 - `milliseconds` = Temps total cumulé converti en millisecondes
 - `fastestLap` = Numéro du tour le plus rapide réalisé par le pilote
 - `rank` = Classement du tour le plus rapide par rapport aux autres pilotes
 - `fastestLapTime` = Chrono du tour le plus rapide réalisé par le pilote
 - `fastestLapSpeed` = Vitesse (km/h) du tour le plus rapide
 - `statusId` = ID statut classement final (arrivé/abandon...)

- seasons: La table `seasons.csv` fournit les informations sur les Championnats organisés dans l'histoire de la Formule 1.
 - year = Année du championnat
 - url = Page Wikipédia
- sprint_results: La table `sprint_results.csv` fournit les résultats des séances "Sprint" des pilotes.
 - sprintResultId = ID (clé primaire)
 - raceId = ID du Grand Prix
 - driverId = ID du pilote
 - constructorId = ID écurie
 - number = Numéro du pilote
 - grid = Position grille de départ
 - position = Position arrivé (ou nul)
 - positionText = Idem au format texte
 - positionOrder = Numéro pour tri
 - points = Points gagnés par le pilote à l'issue du Sprint
 - laps = Nombre de tours complétés
 - time = Temps total cumulé à l'issue du Sprint pour le pilote arrivé 1er, et l'écart par rapport au 1er pour les suivants
 - milliseconds = Temps total cumulé converti en millisecondes
 - fastestLap = Numéro du tour le plus rapide réalisé par le pilote
 - fastestLapTime = Chrono du tour le plus rapide réalisé par le pilote
 - statusId = ID statut classement final (arrivé/abandon...)

La course "Sprint" est une mini-course introduite en 2021 sur 3 Grands Prix en test et reconduite pour 2022.

Le Sprint se déroule après la séance de qualifications dont le classement final constitue la grille de départ du Sprint. Le classement à l'arrivée du Sprint définit la grille de départ de la course.

- status: La table `status.csv` fournit les informations sur les différents statuts à l'arrivée de la course/sprint.
 - statusId = ID (clé primaire)
 - status = Label du statut (arrivé, abandon, disqualifié...)

4. Météo et Formule 1

La météo a toujours eu un impact en Formule 1. Pour de nombreuses raisons, certains pilotes sont plus performants que d'autres sous la pluie, certaines écuries sont plus stratégiques que d'autres et gèrent bien mieux les aléas climatiques, il y a plus d'accidents lors de gros épisodes de pluie ce qui peut bouleverser le classement...

Pour ajouter les données météo, ont été scrapées des informations sur les pages Wikipédia de chacune des courses, vous trouverez dans la table `weather.csv` les informations suivantes :

- weather:
 - `racelId` : ID du Grand Prix
 - `year` : Année du championnat
 - `weather_warm` : Le temps était-il chaud ?
 - `weather_cold` : Le temps était-il froid ?
 - `weather_dry` : Le temps était-il sec ?
 - `weather_wet` : Le temps était-il mouillé ?
 - `weather_cloudy` : Le temps était-il nuageux ?

5. Objectif et évaluation

À l'aide des données fournies ci-dessus, et éventuellement de données annexes (bibliothèque [FastF1](#), scraping...), vous aurez plusieurs objectifs.

a) Data Visualisation

Vous devrez produire une analyse graphique et visuelle avec une approche de Data Storytelling. Vous trouverez davantage de détails dans le notebook `DataVisualisation.ipynb`.

Cette partie devra être rendue à 14:00 et comptera pour 25% de la note finale.

b) Modélisation

Concernant la modélisation, votre objectif sera double, vous devrez **prédire les vainqueurs et les podiums de chaque course de l'année 2021**. Vous vous positionnerez comme un expert au départ d'un Grand Prix à qui on demande d'utiliser toutes les données en sa possession afin de prédire le vainqueur et le podium.

Vous devrez soumettre sur la plateforme un fichier de prédiction au format **.CSV** et possédant les colonnes suivantes :

- `RacelId` (en index) : ID du Grand Prix
- `driverId` (en index) : ID du pilote
- `winner` : 1 si le pilote a gagné le Grand Prix, 0 sinon

- positionOrder : 1 si le pilote est sur la plus haute marche du podium, 2 si le pilote est arrivé 2e, 3 si le pilote est arrivé 3e, 0 sinon.

L'objectif étant de prédire le plus finement possible le vainqueur et le podium, la métrique d'évaluation se basera sur le **recall**.

Pour rappel, dans le cadre d'une classification binaire, le recall se calcule ainsi :

$$Recall = \frac{TP}{TP + FN}$$

Pour la tâche de prédiction du podium, on généralisera cette métrique avec un "micro-average" sur les différentes classes comme ceci :

$$Recall_{micro-average} = \frac{TP_1 + TP_2 + \dots}{TP_1 + FN_1 + TP_2 + FN_2 + \dots}$$

Vous trouverez un exemple de travail de modélisation dans le fichier `Modelisation.ipynb`

Vous pourrez soumettre vos résultats sur la plateforme jusqu'à 16h, le classement sera lui figé à 15h. Cette partie comptera pour 75% de la note finale.

En plus de la métrique ci-dessus, nous pourrions être amenés à juger la pertinence mathématique, méthodologique et sportive de vos modèles. Faites donc bien attention à ne pas utiliser d'informations qui ne seraient pas accessibles avant le départ d'un Grand Prix, sans quoi vos résultats ne pourront pas être pris en compte.

c) Jury et Soutenance

Sur la base des résultats des sections a) et b), un classement sera établi et un premier verdict sera rendu par le jury aux alentours de 16h30.

Les 3 premières équipes du classement seront invitées à préparer une soutenance de 10min qui aura pour but de présenter les éléments suivants (liste non-exhaustive):

- présentation du sujet, du problème et des enjeux
- présentation des données (volumétrie, architecture, etc.)
- analyse des données avec figures de Data Visualization
- description et justification preprocessing effectué
- présentation des modèles entraînés et de leurs résultats
- analyse du meilleur modèle
- conclusion du projet en reliant au maximum les résultats obtenus à la problématique métier : comment prédire le vainqueur de la saison sur la base des prédictions, parier etc...
- critique et perspectives

Vous pouvez organiser vos slides de la façon dont vous le souhaitez. L'objectif est que ces éléments soient abordés lors de la présentation et que votre travail soit correctement résumé. L'utilisation d'un support visuel interactif de type [Streamlit](#) sera fortement appréciée.

Suite à chacune des présentations un temps d'échange rapide avec le jury aura lieu, à la fin des 3 soutenances le jury se réunira afin d'établir le verdict final.

Vous l'aurez compris, à la manière du choix des pneus au départ d'un Grand Prix, préparer la soutenance avant le classement intermédiaire (à 16h) est un pari risqué, mais qui pourrait s'avérer crucial pour la victoire final ! Bon courage à toutes et à tous.

