**Rapport d’Analyse de Données en Formula 1 – Méthode CRISP-DM**

1. Compréhension du Domaine

Ce projet d'analyse de données porte sur les performances des constructeurs en Formule 1 (F1) au cours des différentes saisons. L'objectif principal est de comprendre l'évolution des performances des équipes au fil des années, d'identifier les périodes de domination des équipes et d'analyser la distribution des points obtenus par les différents constructeurs.

L'analyse des résultats des courses permet d'explorer les tendances sur plusieurs décennies, de visualiser les performances des top constructeurs et d'identifier les périodes où certains constructeurs ont dominé la compétition.

**Objectifs spécifiques :**

* Identifier les équipes dominantes au fil du temps.
* Analyser la distribution des points des équipes par saison.
* Examiner les périodes de domination (équipe remportant plusieurs championnats consécutifs).

1. Compréhension des Données

Les données utilisées dans ce projet comprennent les résultats des équipes de F1 au cours des saisons, avec plusieurs attributs clés :

* **Année (Year)** : L'année de la saison de F1.
* **Constructeur (Team)** : L'équipe de F1 qui a participé à la saison.
* **Position (Pos)** : La position finale du constructeur dans le championnat de la saison (1 pour champion, 2 pour vice-champion, etc.).
* **Points (PTS)** : Le nombre de points obtenus par l'équipe pendant la saison.
* Les données proviennent de plusieurs saisons, avec des informations sur le classement de chaque équipe au fil des années.

**Points importants :**

* **Valeurs manquantes** : Une analyse des valeurs manquantes a été réalisée pour vérifier l'intégrité des données.
* **Types de données** : Les colonnes Pos et PTS ont été converties en données numériques pour permettre les analyses.

Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.

Figure (1) : Raw data

1. Préparation des Données

La préparation des données a consisté à nettoyer et organiser les informations pour rendre l'analyse plus efficace :

* **Traitement des valeurs manquantes** : Vérification des valeurs manquantes dans les colonnes et traitement des données manquantes, le cas échéant.
* **Conversion des types de données** : Les colonnes Pos et PTS ont été converties en types numériques pour éviter les erreurs lors des calculs.
* **Création de nouvelles colonnes** : Ajout de la colonne **'Era'** pour classer les années en périodes historiques, afin de mieux visualiser les performances par ère.

Une image contenant texte, capture d’écran, nombre, Police

Le contenu généré par l’IA peut être incorrect.

Figure (2) : Processed Data

1. Modélisation et Analyse

L'analyse des données a été réalisée en plusieurs étapes pour identifier les tendances et les points d'intérêt.

* 1. Visualisation des Performances des Top Constructeurs

Une des premières analyses a consisté à étudier les performances des constructeurs en fonction de leur classement dans le championnat. Nous avons identifié les équipes les plus performantes (celles qui ont remporté plusieurs championnats) et avons tracé leurs positions au fil des années.

* **Méthode** : Sélection des équipes ayant le plus grand nombre de championnats et visualisation de leurs positions chaque année. Une image contenant texte, diagramme, ligne, Tracé

  Le contenu généré par l’IA peut être incorrect.
  1. Visualisation des Clusters de Performance

Une analyse de clustering non supervisée (K-means) a été réalisée à partir des composantes principales extraites via ACP

Une image contenant texte, diagramme, capture d’écran, ligne

Le contenu généré par l’IA peut être incorrect.

Figure (3) : Clusters de performance des constructeurs (PCA)

* **Équipes d'élite** (jaune) : Dominent régulièrement le championnat.
* **Contenders solides** (rouge) : Souvent en haut du classement mais moins réguliers.
* **Midfield Teams** (bleu) : Performances moyennes.
* **Backmarkers** (vert) : Constamment en bas du classement.
  1. Distribution des Points par ére

Un graphe boxplot par période permet de visualiser l’évolution de la distribution des points.

Une image contenant capture d’écran, Rectangle, carré, Caractère coloré

Le contenu généré par l’IA peut être incorrect.

* Le volume de points augmente significativement à partir des années 2000.
* L’écart entre les équipes s’accentue dans les ères modernes.
  1. Champion vs Vice-champion : Compétitivité
* Jusqu’aux années 2000, les écarts sont relativement modérés.
* Depuis 2010, certaines saisons montrent un écart très net entre le champion et le deuxième.

Une image contenant texte, capture d’écran, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

* 1. Modélisation Prédictive : Position Finale

Une régression a été utilisée pour prédire la position finale à partir des points obtenus et d'autres métriques.

* **R² = 0.84** : Très bonne capacité de prédiction.
* **MAE = 1.06** : L'erreur moyenne est d’environ une position.
* Cela montre que les données historiques sont de bons prédicteurs de performance future.

Une image contenant texte, diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.

1. Evaluation

Les résultats du modèle prédictif sont très encourageants, avec une corrélation forte entre les positions prédites et les positions réelles. Toutefois, certaines anomalies existent, surtout dans le milieu du peloton, là où les écarts de performance sont minimes.

Les clusters identifiés via PCA confirment les intuitions historiques :

* Quelques équipes dominent sur le long terme.
* Le milieu de grille reste très compétitif.
* Les points ont tendance à se concentrer parmi les top teams dans les ères modernes.

1. Déploiement

Le projet peut être étendu avec les axes suivants :

* **Mise à jour continue des données** : Intégration des saisons futures.
* **Tableau de bord interactif** : Pour permettre l'exploration par saison, constructeur, ou ère.
* **Application prédictive** : Prévision des positions pour les saisons à venir selon les performances en début de saison.
* **Intégration avec des données pilotes et circuits** : Pour enrichir les analyses contextuelles.