Plan Hackathon :

Fichier CSV :

**Étape 1 : Exploration et Premier Diagnostic**

L'objectif ici est de "faire connaissance" avec vos données.

* Prise en main : Chargez vos données et posez-vous les premières questions : Combien de lignes et de colonnes ? Quels sont les types de données ?
* Contrôle Qualité : Vérifiez la propreté de base du jeu de données. Y a-t-il des valeurs manquantes ou des lignes en double ?
* Premières Observations : Explorez visuellement les relations entre les variables. Y a-t-il des corrélations évidentes qui sautent aux yeux ?

**Étape 2 : Nettoyage et Préparation des Données**

Ici, vous vous attaquez aux imperfections des données pour les rendre analysables.

* Détection des Anomalies : Visualisez la distribution de chaque variable. Identifiez celles qui sont très asymétriques ou qui présentent des valeurs extrêmes (outliers).
* Transformation : Appliquez des transformations mathématiques (comme le logarithme) pour normaliser les distributions et réduire l'impact des outliers.
* Ajustements : Parfois, une simple transformation ne suffit pas. Soyez prêt à utiliser des techniques plus ciblées pour traiter les valeurs extrêmes restantes.

**Étape 3 : Investigation d'un Comportement Complexe**

C'est le moment de jouer au détective avec une variable qui semble se comporter étrangement.

* Identifier le Mystère : En observant une distribution, vous pourriez remarquer un comportement inhabituel, comme la présence de plusieurs "pics" (bimodalité).
* Chercher la Cause : D'où vient ce comportement ? Est-il lié à une autre variable, comme le temps ou une catégorie spécifique ?
* Valider avec des Tests : Utilisez des tests statistiques pour confirmer que les sous-groupes que vous avez identifiés ont bien des distributions différentes.
* Appliquer une Solution Ciblée : Une fois le problème compris, traitez chaque sous-groupe de manière appropriée pour corriger sa distribution.

**Étape 4 : Construction du Jeu de Données Final**

Cette étape consiste à consolider votre travail de préparation.

* Sélection : Choisissez les colonnes les plus pertinentes pour la suite de l'analyse (les versions nettoyées et transformées).
* Finalisation : Créez un nouveau DataFrame propre, structuré et prêt à être utilisé pour l'analyse statistique et la modélisation.

**Étape 5 : Analyse Approfondie pour Extraire des Conclusions**

Maintenant que vos données sont fiables, vous pouvez enfin répondre à des questions métier.

* Analyse Temporelle : Visualisez l'évolution de vos variables clés dans le temps pour identifier des tendances de fond.
* Tests d'Hypothèses : Posez une question précise (ex: "Y a-t-il une différence significative entre deux périodes ?") et utilisez un test statistique (comme le t-test) pour y répondre de manière rigoureuse.
* Recherche de Relations Subtiles : Allez plus loin que les corrélations de base. En lissant les données (par ex. avec des moyennes mobiles), vous pourriez découvrir des relations plus profondes entre les variables.

Rendus

Vérifier si les commentaires à tonalité positive ou négative ont plus de votes positifs (score, ups) ou sont plus controversés.Lien entre la notoriété des auteurs (user\_total\_karma) et l’impact de leurs commentaires (score, ups, controversiality).Identifier si les sentiments évoluent au fil du temps (pics de catastrophes, événements politiques, COP...).Comparer la tonalité du post (post\_title, post\_self\_text) avec celle des commentaires (self\_text) pour mesurer :

* Amplification (mêmes émotions que le post)
* Divergence (ton opposé)

Analyser si les posts récompensés (post\_total\_awards\_received) sont associés à des sentiments particuliers.

100 mots les plus cités