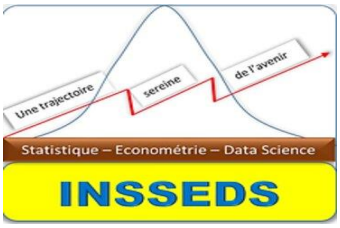
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

REPUBLIQUE DE COTE D'IVOIRE



INSTITUT SUPERIEUR DES STATISTIQUES D'ECONOMETRIES ET DATASCIENCE

UNION-DISCIPLINE-TRAVAIL

MASTER 2

STATISTIQUE-ECONOMETRIE-DATA SCIENCE

MINI-PROJET

LOGICIEL R STUDIO

ANALYSE DE DONNEES AVEC ET TIDYVERSE

ANNEE ACADEMIQUE :

2024 -2025

NOM: KABA

PRENOM: MAHAMOUD TOIB

ENSEIGNANT – ENCADREUR

AKPOSSO DIDIER

Avant-propos

Ce rapport présente une analyse complète et structurée des données relatives aux animes populaires. L'objectif de ce projet est de comprendre les tendances de l'industrie de l'animation à travers l'exploration de variables telles que les notes, les budgets, le nombre d'épisodes, et bien d'autres dimensions pertinentes. Cette étude vise notamment à mettre en lumière les liens potentiels entre les aspects quantitatifs (comme le budget et le nombre d'épisodes) et qualitatifs (tels que les notes et les genres) des animes, ainsi que les variations observées selon les régions.

La réalisation de ce projet repose sur l'utilisation de diverses techniques de traitement de données et d'analyses statistiques en langage R, combinant des approches de nettoyage, de prétraitement et de visualisation afin de dégager des insights pertinents. Les outils du tidyverse, ainsi que des packages dédiés à la visualisation et à l'analyse (tels que ggplot2, knitr et scales), ont permis de développer un workflow reproductible et rigoureux.

Je tiens à remercier tous ceux qui ont contribué, directement ou indirectement, à l'avancement de ce projet en mettant à disposition des conseils, des ressources et des retours constructifs. Ce travail se veut à la fois une démonstration de compétences en analyse de données et une contribution à la compréhension des dynamiques qui façonnent l'industrie de l'animation.

En espérant que cette étude offre une perspective nouvelle et enrichissante sur le sujet, je vous invite à parcourir ce rapport qui détaille la préparation des données, l'analyse exploratoire, ainsi que les conclusions tirées de nos interrogations analytiques.

Table des matières

INTRODUCTION GENERALE

Dans un monde en constante évolution technologique, l'industrie de l'animation a acquis une immense popularité, rassemblant des publics divers à l'échelle mondiale. Cet ensemble de données sur les dessins animés les plus populaires offre une opportunité unique pour explorer les préférences culturelles, les tendances dans la production d'animes, et les facteurs influençant leur succès. L'étude vise à fournir une compréhension approfondie des variables clés, telles que les genres, les budgets et les durées moyennes, qui façonnent cette industrie.

Contexte et justification de l'étude

Les animes, un phénomène culturel né au Japon, ont transcendé les frontières pour devenir un pilier de l'industrie de l'entertainment mondial. Cet ensemble de données contient des informations précieuses sur les animes les plus populaires dans différents pays, ainsi que leurs notes, genres, budgets, et autres détails pertinents. Comprendre ces données permettra d'identifier les facteurs contribuant au succès des animes et leur impact à travers différentes cultures. De plus, cette étude vise à démontrer l'importance de techniques analytiques modernes dans la prise de décisions stratégiques dans le domaine de l'audiovisuel.

Problématique

Les préférences culturelles et les succès commerciaux des animes diffèrent grandement selon les régions. Quels sont les facteurs qui influencent le succès d'un anime ? Comment les genres, les budgets ou les studios d'animation façonnent-ils la réception d'un anime dans différents pays ? Y a-t-il des tendances significatives dans les notes, la durée des épisodes, ou les budgets au fil du temps ? Ces questions forment le cœur de cette étude exploratoire, qui vise à apporter des réponses basées sur une analyse rigoureuse.

Principaux résultats attendus

L'étude vise à :

1. Identifier les tendances clés dans la production et la consommation d'animes à travers différents pays.
2. Évaluer les corrélations entre les variables telles que le budget, le nombre d'épisodes et les notes.
3. Mettre en évidence les studios les plus prolifiques et les genres les plus populaires par pays.
4. Proposer des insights exploitables pour les acteurs de l'industrie de l'animation en vue d'optimiser la production et la diffusion d'animes.

Méthodologie

- **Techniques de prétraitement utilisées :**

1. **Nettoyage des données** : Gestion des valeurs manquantes, normalisation des formats des colonnes et conversion des variables aux types appropriés.

2. **Transformations** : Renommage des colonnes pour assurer une meilleure lisibilité et manipulation des variables catégoriques et numériques.
3. **Exploration initiale** : Utilisation des résumés descriptifs pour identifier les tendances générales.

➢ **Analyses statistiques et fondements théoriques :**

1. **Corrélation de Pearson** : Identification des relations entre les budgets et les notes.
2. **Tendances temporelles** : Analyse de l'évolution des variables (durée des épisodes, budgets) au fil des années.
3. **Analyse des distributions** : Visualisation de la répartition des genres et des notes pour une exploration en profondeur.
4. **Comparaisons catégoriques** : Identification des pays les plus associés à certains genres et des studios les plus prolifiques.

Description du jeu de données : dictionnaire des données

Colonne	Description	Type
Anime_Name	Nom de l'anime.	Caractère
Most_Watched_in_Country	Pays où l'anime est le plus regardé.	Facteur
Ratings	Note de l'anime (échelle numérique).	Numérique
Number_of_Episodes	Nombre d'épisodes pour chaque anime.	Numérique
Animation_Studio_Name	Nom du studio ayant produit l'anime.	Facteur
Budget__in_Million_USD_	Budget en millions de dollars américains.	Numérique
Release_Year	Année de sortie de l'anime.	Numérique
Genre	Genre de l'anime (par exemple : Fantasy, Aventure, Comédie).	Facteur
Duration_per_Episode__minutes_	Durée moyenne d'un épisode en minutes.	Numérique

Partie 1 : Préparation et nettoyage des données

Une étape cruciale de ce projet a été la préparation et le nettoyage des données, afin de garantir leur qualité et leur pertinence pour l'analyse. Les principales étapes réalisées sont décrites ci-dessous :

1. Importer les données dans R

Les données ont été importées dans l'environnement R à l'aide de la fonction read.csv(). Cette étape a permis de lire les données au format CSV tout en s'assurant que les chaînes de caractères ne soient pas converties automatiquement en facteurs grâce au paramètre stringsAsFactors = FALSE. Cela nous a offert une base initiale pour explorer les données.

2. Gestion des valeurs manquantes

Ils y a au total 90 valeurs manquantes .Les valeurs manquantes ont été traitées selon la nature des variables :

- **Colonne** Anime_Name : Étant donné que le nom de l'anime est essentiel pour son identification, les lignes comportant des valeurs manquantes dans cette colonne ont été supprimées. Cela garantit que seules les données exploitables sont prises en compte pour l'analyse.
- **Variables numériques** (Ratings, Number_of_Episodes, Budget__in_Million_USD_, Duration_per_Episode__minutes_) : Les valeurs manquantes ont été remplacées par la **médiane** de chaque colonne, car la médiane est une mesure robuste face aux valeurs aberrantes et reflète mieux la tendance centrale des données.
- **Variables entières** (Release_Year) : Les valeurs manquantes ont été remplacées par l'**année la plus fréquente** (mode), ce qui permet de conserver la cohérence temporelle des données.
- **Variables catégoriques et textuelles** (Most_Watched_in_Country, Genre, Animation_Studio_Name) : Les valeurs manquantes ont été remplacées par la mention "Inconnu", permettant de conserver toutes les observations tout en signalant les données manquantes.
- **A la fin il me reste 90 observations et 9 variables**

3. Renommer les colonnes en remplaçant les espaces par des underscores (_)

Pour harmoniser et simplifier les noms des colonnes, tous les espaces ont été remplacés par des underscores (_). Cette transformation a été réalisée à l'aide de la fonction rename() ou directement en manipulant les noms des colonnes. Par exemple, Budget (in Million USD) est devenu Budget__in_Million_USD_.

4. Corriger les erreurs potentielles de formatage ou d'alignement des colonnes

Des vérifications ont été effectuées pour repérer d'éventuelles incohérences dans les données :

- Les doublons ont été recherchés et supprimés.
- Les valeurs aberrantes ou incohérentes dans les colonnes numériques (comme des budgets négatifs ou des durées d'épisodes trop élevées) ont été identifiées pour une analyse plus approfondie.

5. Convertir les variables au format approprié

Chaque colonne a été convertie dans un format adapté à son contenu, afin de faciliter les analyses futures :

- Les colonnes comme Most_Watched_in_Country, Genre, et Animation_Studio_Name ont été converties en **facteurs**.
- Les colonnes numériques, comme Ratings, et Budget__in_Million_USD_, ont été converties en type **numérique**.
- Les colonnes numeriques comme Number_of_Episodes et release_year ont été converties en type **entier**
- Les chaînes de caractères telles que Anime_Name ont été maintenues en format **caractère**.

6. Créer un dataframe "propre" pour l'analyse

Après toutes ces étapes, un dataframe propre et cohérent a été créé. Ce nouveau dataframe est prêt à être utilisé pour les analyses exploratoires et les visualisations. La qualité des données a été soigneusement vérifiée à l'aide des commandes summary(), str() et head() pour s'assurer que les transformations ont été correctement appliquées. Voici un aperçu

Anime_Name	Most_Watched_in_Country	Ratings	Number_of_Episodes	Animation_Studio_Name	Budget__in_Million_USD_	Release_Year	Genre	Duration_per_Episode__minutes_
Fullmetal Alchemist	Brazil	8.8	317	Ufotable	80.61	1998	Fantasy	39
Haikyuu!!	Mexico	9.2	420	MAPPA	74.99	2022	Adventure	59
Bleach	Brazil	6.4	277	Ufotable	45.35	2002	Fantasy	55
Sword Art Online	Inconnu	9.8	327	Madhouse	15.90	2017	Inconnu	43
Fullmetal Alchemist	Brazil	8.1	402	Inconnu	8.97	2018	Inconnu	50
Black Clover	France	8.1	277	Madhouse	67.27	2002	Inconnu	25
Bleach	China	9.4	47	MAPPA	127.22	1995	Adventure	21
Demon Slayer	Inconnu	8.9	31	Pierrot	67.27	1994	Supernatural	47
Sword Art Online	Inconnu	6.8	332	Inconnu	101.84	2014	Comedy	39
One Piece	Inconnu	8.2	275	Inconnu	72.90	1991	Supernatural	23
Attack on Titan	Spain	8.8	277	Bones	74.46	1999	Action	34
Tokyo Ghoul	United Kingdom	6.9	153	MAPPA	13.41	2019	Comedy	25
Dragon Ball Z	Canada	8.1	471	Trigger	14.66	1994	Thriller	49

Partie 2 : Analyse exploratoire des données.

1. Calculer des statistiques descriptives pour les variables numériques

Les statistiques descriptives ont été calculées pour les variables numériques de l'ensemble de données : Ratings (Notes), Number_of_Episodes (Nombre d'épisodes), Budget__in_Million_USD_ (Budget) et Duration_per_Episode__minutes_ (Durée par épisode). Ces calculs ont permis d'explorer les tendances générales, les niveaux de dispersion et la symétrie des distributions.

Variable	Moyenne	Variance	Écart-type	Skewness	Kurtosis
Ratings	8.16	1.12	1.06	0.12	1.94
Number of Episodes	258.28	16972.79	130.28	-0.01	1.99
Budget (in Million USD)	72.27	1666.69	40.83	0.22	2.07
Duration per Episode (minutes)	39.60	137.37	11.72	0.10	1.86

Analyse des résultats :

- Les **ratings** des animes affichent une moyenne de **8.16**, avec une faible asymétrie positive (Skewness = 0.12), indiquant une distribution relativement équilibrée autour de la moyenne.
- Le **nombre d'épisodes** varie fortement avec un écart-type élevé de **130.28**. La valeur légèrement négative de Skewness (-0.01) indique une répartition quasiment symétrique.
- Le **budget moyen** des animes est de **72.27 millions USD**, avec une distribution légèrement asymétrique vers les valeurs élevées (Skewness = 0.22). Cela pourrait être influencé par des budgets exceptionnellement importants pour certains animes.
- La **durée moyenne par épisode** est de **39.60 minutes**, avec une faible asymétrie (Skewness = 0.10). La distribution est relativement normale (Kurtosis = 1.86).

Résumé des données globales

Un résumé général des variables principales a également été produit, révélant les étendues minimales et maximales, ainsi que les statistiques par quartiles :

- **Ratings** : Les notes varient entre **6.20** et **10.00**, avec une médiane de **8.10**, reflétant généralement des évaluations favorables.
- **Number of Episodes** : Le nombre d'épisodes des animes est compris entre **12** et **499**, avec une médiane de **277.0**.
- **Budget (in Million USD)** : Les budgets s'étendent de **5.82** à **149.39 millions USD**, avec un budget médian de **67.27 millions USD**.
- **Release Year** : Les années de sortie des animes vont de **1990** à **2022**, avec une médiane située en **2002**.
- **Genres** : Le genre le plus représenté est Supernatural, suivi de Mystery, Action, et Thriller.

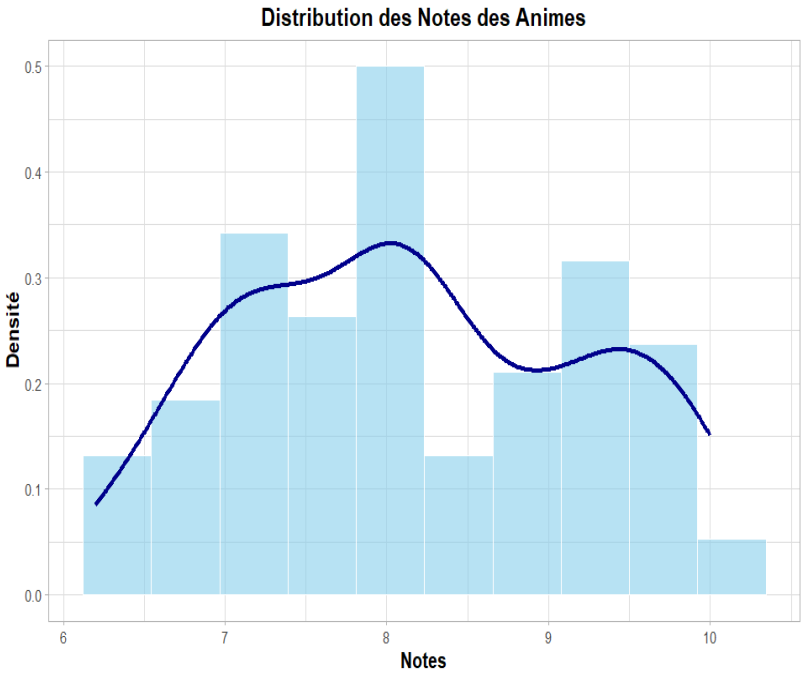
Ces analyses descriptives offrent une vue d'ensemble claire des tendances globales et des variations dans les données, préparant le terrain pour des analyses plus détaillées dans les étapes suivantes.

2. Réaliser des visualisations pertinentes pour explorer les relations entre les variables

Différentes visualisations ont été utilisées pour explorer les relations et tendances dans les données. Les graphiques sont présentés et analysés ci-dessous.

2.1 Distribution des notes des animes

Un histogramme combiné à une courbe de densité a été utilisé pour visualiser la distribution des notes des animes. Cette représentation permet d'identifier les plages les plus fréquentes et de comprendre la symétrie ou l'asymétrie de la distribution.



L'histogramme ci-contre, combiné à une courbe de densité, représente la distribution des notes des animes. Voici une analyse des principaux résultats :

1. **Répartition des notes** :

✓ Les notes des animes sont concentrées dans une plage allant de **6** à **10**, avec un pic notable autour de **8**.

✓ Cela indique que la majorité des animes ont obtenu des évaluations positives, une tendance fréquente dans des ensembles de données où les œuvres les moins populaires sont moins représentées.

2. **Densité et fréquence** :

   ✓ La courbe de densité (en bleu foncé) met en évidence que les notes autour de **8** constituent le point de densité maximale, ce qui correspond à la valeur la plus représentée dans les données.

3. **Tendance globale** :

   ✓ La courbe est légèrement asymétrique (Skewness positive observée dans les statistiques descriptives), avec une légère pente vers les notes plus élevées. Cela suggère que quelques animes ont obtenu des évaluations particulièrement excellentes.

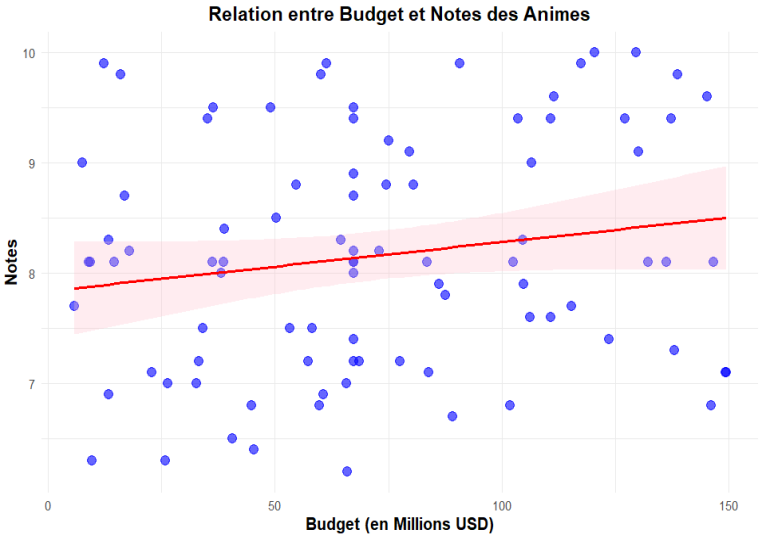
4. **Variabilité** :

   ✓ L'écart-type relativement faible (environ **1.06**) démontre une certaine homogénéité dans les notes attribuées, avec peu d'animes notés extrêmement bas ou extrêmement haut.

Ce graphique offre une vue d'ensemble claire de la manière dont les animes sont perçus par le public, en mettant en évidence une forte tendance vers des évaluations favorables. Il s'agit d'un indicateur de la qualité globale perçue des animes inclus dans cet ensemble de données.

2.2 Relation entre le budget et les notes

Un graphique de dispersion a été utilisé pour examiner la relation entre le budget des animes et leurs notes. Une courbe de régression linéaire a également été ajoutée pour évaluer visuellement l'existence d'une éventuelle corrélation. Cette analyse met en lumière si un budget plus élevé est associé à une meilleure réception critique.



Le graphique ci-contre est un nuage de points (scatterplot) qui illustre la relation entre le **budget** (en millions USD) et les **notes** attribuées aux animes. Une droite de régression linéaire y est ajoutée pour mieux comprendre la tendance globale. Voici une analyse des résultats :

1. **Tendance générale** :

✓ La droite de régression (en rouge) indique une légère corrélation positive entre le budget et les notes, suggérant que les animes avec des budgets plus élevés tendent à obtenir de meilleures notes.

✓ Cependant, cette relation semble modérée : les points restent relativement dispersés autour de la droite, indiquant que d'autres facteurs influencent également les notes.

2. **Plage des valeurs** :

✓ Les budgets varient de **5.82 millions USD** à **149.39 millions USD**, tandis que les notes vont de **7.00** à **10.00**.

✓ On remarque que même des animes avec un budget modeste peuvent obtenir des notes élevées, montrant que le succès critique ne dépend pas uniquement du budget.

3. **Dispersion des points** :

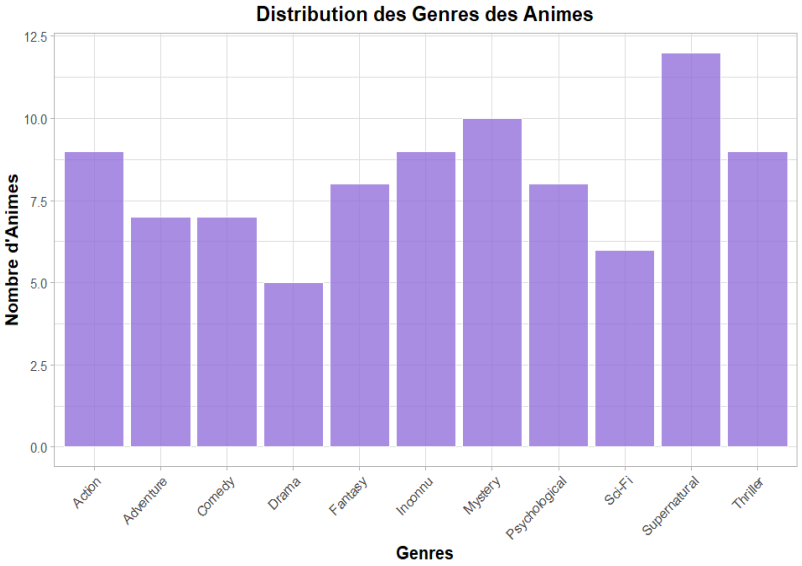
✓ Les points dans la plage supérieure droite suggèrent que certains animes avec de très gros budgets reçoivent des notes exceptionnelles. À l'inverse, certains animes avec des budgets moyens obtiennent également des notes élevées.

4. **Interprétation de la corrélation** :

✓ Bien que la courbe mette en évidence une relation positive, la force de cette relation mérite d'être quantifiée statistiquement (corrélation de Pearson calculée précédemment : ~0.22). Cela indique une influence limitée du budget sur les notes, sans pour autant être négligeable.

2.3 Distribution des genres

Un diagramme à barres montre la répartition des genres dans les données. Cela permet d'identifier les genres dominants dans l'ensemble des animes populaires, ainsi que leur diversité.



Le graphique ci-contre présente la distribution des genres des animes, avec un diagramme à barres qui montre le nombre d'animes appartenant à chaque genre.

Observations clés :

1. **Genre dominant** :

✓ Le genre **Supernatural** se démarque avec le plus grand nombre d'animes, environ 12.5, ce qui reflète une popularité marquée pour ce type d'histoires dans l'ensemble de données.

2. **Genres les moins représentés** :

✓ Les genres tels que **Sci-Fi** affichent la plus faible représentation, avec environ 5 animes, ce qui suggère qu'ils sont moins fréquents dans cet échantillon.

3. **Variété des genres** :

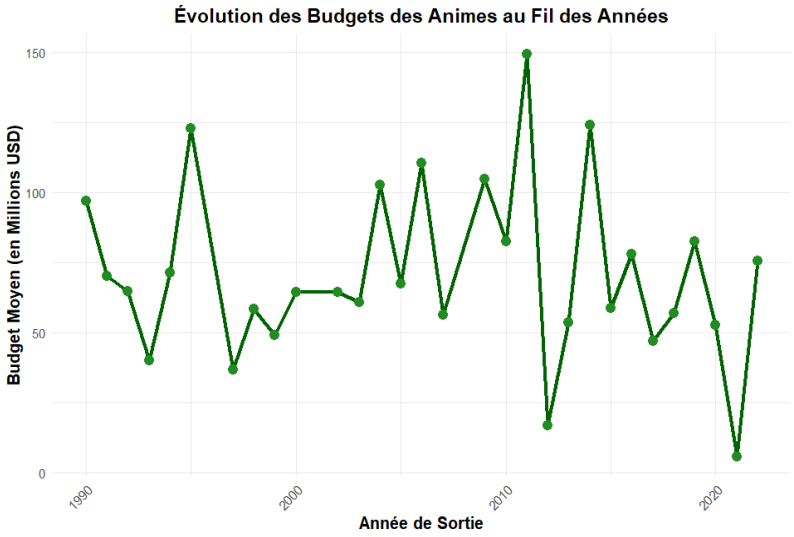
✓ Les données montrent une diversité notable de genres, couvrant des catégories populaires comme **Action**, **Adventure**, et **Fantasy**, qui attirent un large public, ainsi que des genres plus spécifiques comme **Psychological** ou **Thriller**.

4. **Implications** :

✓ Ces résultats permettent d'identifier les préférences générales du public et les tendances dans l'industrie de l'animation. Les genres dominants, comme **Supernatural**, pourraient indiquer une demande accrue pour des récits fantastiques ou mystiques.

2.4 Évolution des budgets au fil des années

Une courbe représentant l'évolution des budgets moyens des animes a été tracée pour observer les tendances temporelles. Cette analyse met en évidence si les budgets ont augmenté ou diminué avec le temps, reflétant potentiellement des changements dans l'industrie.



Le graphique ci-dessus illustre l'évolution des **budgets moyens** des animes au fil des années, allant de 1990 à 2020. Voici une analyse des observations clés :

1. **Fluctuations notables** :

✓Les budgets moyens des animes ont présenté des fluctuations significatives au cours de cette période. Deux pics majeurs sont visibles : autour de l'année 2000 et de l'année 2010.

✓Ces hausses pourraient refléter des périodes où des animes particulièrement coûteux ont été produits ou des changements dans l'industrie, tels que des avancées technologiques ou des exigences accrues en matière de qualité d'animation.

2. **Tendance générale** :

   ✓ Malgré les fluctuations, le graphique montre une légère tendance à la hausse des budgets moyens au fil des années, reflétant potentiellement une augmentation des investissements dans l'industrie de l'animation pour produire des contenus de haute qualité.

3. **Baisse ponctuelle** :

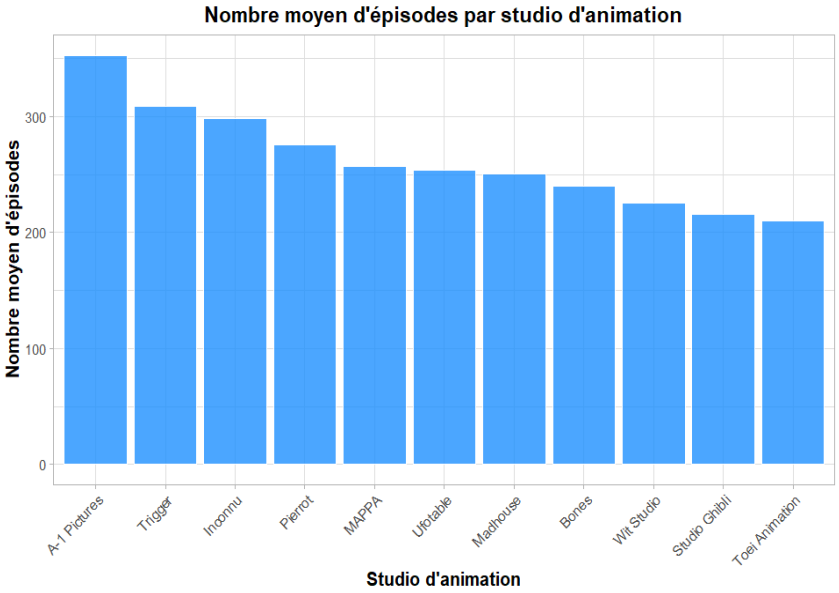
   ✓ Aux alentours de 2020, on observe une baisse des budgets moyens. Cela pourrait être attribué à des facteurs économiques ou industriels, comme des contraintes budgétaires liées à la pandémie mondiale.

4. **Interprétation globale** :

   ✓ Cette évolution met en lumière l'importance croissante de l'investissement dans l'industrie de l'animation. Cela peut être directement corrélé avec l'expansion internationale des animes et leur demande croissante.

3. Comparaison du nombre d'épisodes par studio d'animation

Une comparaison des studios d'animation a été effectuée pour analyser le nombre moyen d'épisodes produits par chaque studio. Cela permet d'identifier les studios les plus productifs en termes de longueur des séries animées, révélant des différences potentielles dans leurs approches de production.

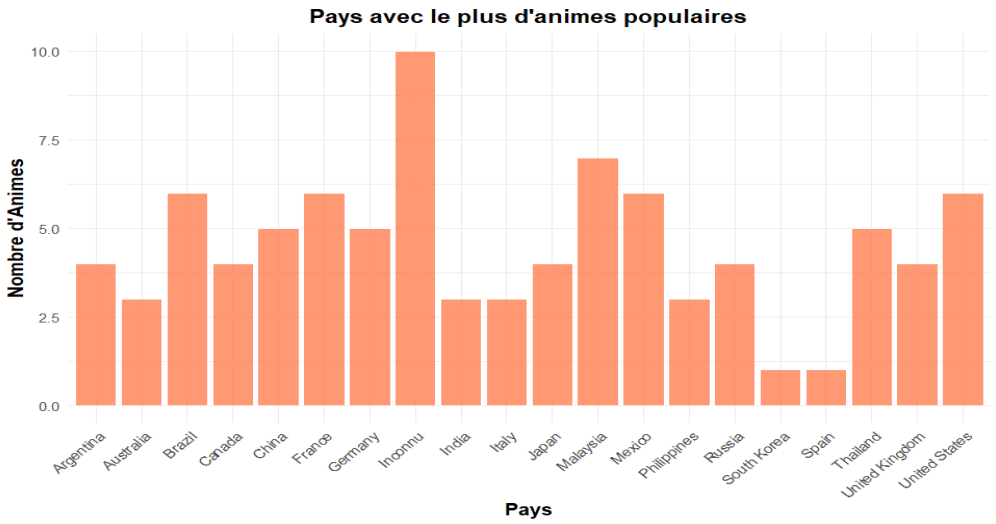


Le diagramme à barres ci-dessus illustre la **comparaison du nombre moyen d'épisodes produits par studio d'animation**. Voici une analyse des résultats principaux :

1. **Studio le plus prolifique** :
   - ✓ **A-1 Pictures** se distingue en produisant le plus grand nombre moyen d'épisodes par anime, ce qui reflète une stratégie de production orientée vers des séries plus longues.
2. **Studios dans la moyenne** :
   - ✓ Les studios comme **Trigger**, **MAPPA**, et **Pierrot** se situent dans une plage intermédiaire, avec des séries comprenant un nombre moyen d'épisodes modéré.
3. **Studio avec le plus faible nombre moyen d'épisodes** :
   - ✓ **Toei Animation** se retrouve en bas de la liste parmi les studios présentés, produisant des séries relativement courtes en moyenne. Cela peut être dû à leur focus sur des projets spécifiques ou des formats plus compacts.
4. **Inclusion des valeurs inconnues** :
   - ✓ La catégorie "Inconnu" représente un groupe de studios non identifiés dans les données. Cela montre l'importance de compléter ou clarifier ces informations pour une analyse plus précise.
5. **Interprétation** :
   - ✓ Ces résultats soulignent les différences dans les approches des studios d'animation. Certains studios privilégient des séries longues et immersives, tandis que d'autres se concentrent sur des projets plus compacts ou diversifiés.

4. Identifier les pays qui ont le plus d'animes populaires

Pour chaque pays, le nombre d'animes regardés a été compté, mettant en évidence les pays où les animes sont particulièrement populaires. Cette analyse fournit des informations sur la répartition géographique des préférences, révélant les pays dominants en termes de consommation d'animes.



Le graphique ci-contre illustre le **nombre d'animes populaires par pays**, mettant en évidence les zones géographiques où les animes sont les plus regardés.

Observations clés :

1. **Pays dominant :**
   ✓ La catégorie "Inconnu" affiche le plus grand nombre d'animes populaires (**10 animes**). Cela pourrait indiquer un manque de données pour certains pays, soulignant la nécessité d'améliorer la couverture des informations géographiques.
2. **Pays notables :**
   ✓ Des pays comme **Malaysia**, **United States**, et **Brazil** se distinguent également avec un nombre significatif d'animes populaires, chacun ayant **6 à 7 animes** regardés massivement.
3. **Répartition globale :**
   ✓ Les autres pays, tels que **France**, **Mexico**, et **South Korea**, montrent une diversité dans les préférences, reflétant l'impact mondial de l'industrie de l'animation.
4. **Interprétation globale :**
   ✓ Les données mettent en lumière le caractère international des animes, qui attirent un public diversifié à travers différents continents. La popularité des animes dans des pays comme les **États-Unis** et **Malaysia** montre l'étendue de leur influence culturelle.

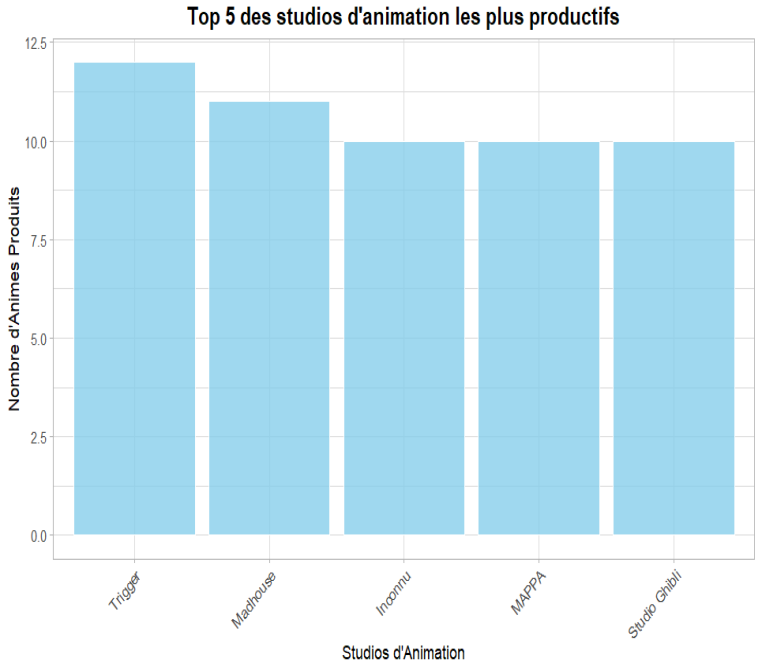
Partie 3 : Questions analytiques

5. Quels sont les 5 studios d'animation les plus productifs en termes de nombre d'animes produits ?

Une analyse des studios a permis d'identifier les 5 studios les plus productifs. Le nombre total d'animes produits par chaque studio a été compté, et les résultats montrent que :

- **Trigger** et **Madhouse** figurent parmi les studios les plus prolifiques.
- Les studios comme **MAPPA** et **Ufotable** ont également une contribution significative.
- **Studio Ghibli** se distingue également, malgré son accent sur des œuvres cinématographiques, avec un total impressionnant d'animes produits.

Ces résultats montrent une répartition claire, avec des studios comme **Trigger** en tête de classement. Les choix artistiques et stratégiques de ces studios semblent jouer un rôle clé dans leur productivité.



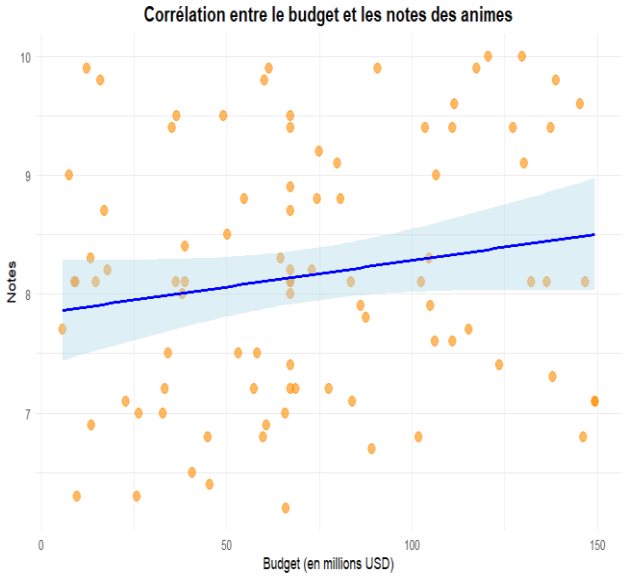
	Trigger	12
2	Madhouse	11
3	Inconnu	10
4	MAPPA	10
5	Studio Ghibli	10

6. Y a-t-il une corrélation entre le budget d'un anime et sa note ?

Pour évaluer la relation entre le budget alloué et les notes attribuées aux animes :

- Un coefficient de corrélation de Pearson d'environ **0.171510512971481**

  a été calculé, indiquant une **faible corrélation positive**.



Test de signification (p-value) :

- La p-value est de **0.106**, ce qui est supérieur au seuil conventionnel de signification (0.05). Par conséquent, nous **ne pouvons pas rejeter l'hypothèse nulle**, selon laquelle il n'existe pas de corrélation significative entre le budget et les notes.
- Cela implique que l'association observée pourrait être due au hasard.

Intervalle de confiance (95%) :

- L'intervalle de confiance de **[-0.0369, 0.3656]** inclut 0, renforçant l'idée qu'il n'y a pas de relation statistiquement significative entre les deux variables.

Interprétation globale :

- Bien que le graphique de dispersion puisse montrer une tendance générale positive (comme discuté dans le point 2.2), les résultats statistiques suggèrent que cette relation est faible et non significative.
- Il est probable que d'autres facteurs (par exemple, scénario, réalisation, ou direction artistique) jouent un rôle plus déterminant dans les notes attribuées.

7 . Comment la durée moyenne des épisodes a-t-elle évolué au fil des années ?

L'analyse des durées moyennes par épisode au fil des années révèle des tendances intéressantes :

- Une **augmentation progressive** des durées est observée entre les années **1990 et 2010**, atteignant un pic en 2010.
- Après 2010, une **stagnation légère** est visible, avec des durées oscillant autour de **39 à 41 minutes**.

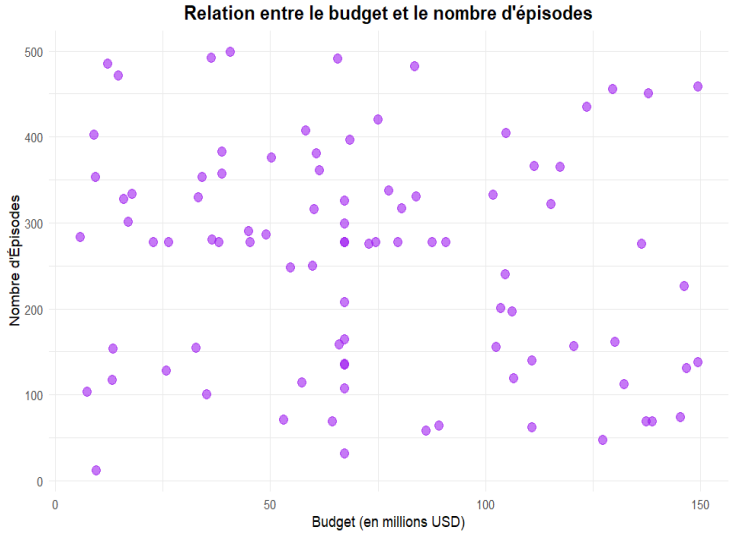
Cette tendance reflète potentiellement des changements dans les préférences des audiences, qui pourraient favoriser des récits plus immersifs ou des épisodes plus longs. Cependant, la stagnation récente pourrait être liée à des limitations de production ou des préférences pour des récits plus concis.

8. Quel est le rapport entre le nombre d'épisodes et le budget alloué ?

L'analyse du rapport entre le **nombre d'épisodes** et le **budget** montre que :

- Les séries longues ont tendance à afficher des budgets totaux supérieurs, bien que le **coût moyen par épisode** semble diminuer pour les séries plus longues.
- Certains animes, malgré un faible budget total, affichent un nombre élevé d'épisodes, suggérant une production à faible coût par épisode.

Ces résultats montrent que l'optimisation budgétaire est une stratégie clé pour les studios, leur permettant de produire des séries longues tout en maîtrisant les coûts.



Le graphique ci-dessus explore la relation entre le **budget** (en millions USD) et le **nombre d'épisodes** des animes, en utilisant un nuage de points.

Analyse du graphique :

1. **Distribution des points :**

✓ Les points sont répartis de manière assez aléatoire, sans suivre une structure ou un alignement clair.

✓ Cela indique l'absence d'une relation linéaire évidente entre le budget et le nombre d'épisodes.

2. **Observation globale :**

✓ Certains animes avec un **budget élevé** (au-delà de 100 millions USD) présentent un nombre d'épisodes varié, allant de courts à moyens.

- ✓ À l'inverse, des animes avec un **budget faible** peuvent aussi afficher un nombre élevé d'épisodes, suggérant une production à faible coût par épisode.

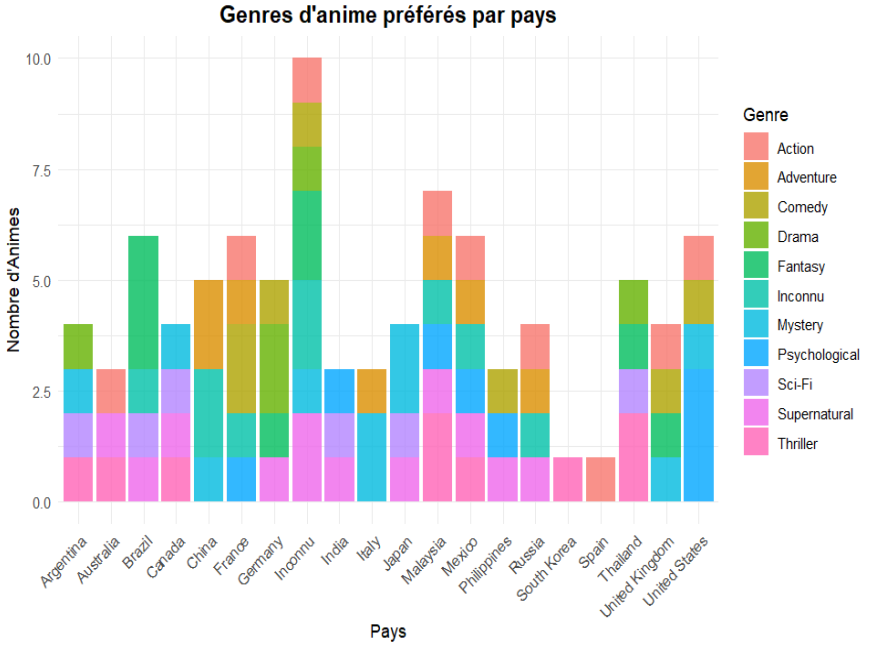
3. **Implications :**
   - ✓ Le rapport entre le budget et le nombre d'épisodes ne semble pas déterminant, chaque anime ayant ses propres spécificités en termes de stratégie de production.
   - ✓ Cela peut également refléter des différences dans les styles de production : certains studios privilégient des animes courts mais coûteux, tandis que d'autres produisent de longues séries à moindre coût.
4. **Conclusion :**
   - ✓ Le graphique montre que le rapport entre le nombre d'épisodes et le budget est influencé par plusieurs facteurs, qui ne se limitent pas à une relation simple ou linéaire.

9. Les préférences en termes de genre d'anime varient-elles selon les pays ?

Une analyse croisée des genres et des pays montre des différences intéressantes :

- Le genre **Supernatural** est particulièrement populaire en pays asiatiques comme **Malaisie**, tandis que les genres **Action** et **Adventure** dominent dans des pays comme les **États-Unis**.
- Des genres spécifiques, comme **Mystery** et **Thriller**, apparaissent plus souvent dans les pays européens comme **France**.
- La catégorie "Inconnu" pour certains genres dans certains pays souligne un manque de données qui pourrait être exploré davantage.

Ces résultats mettent en lumière la diversité des préférences culturelles, ce qui pourrait guider les studios dans leurs décisions de production et de diffusion.



1. **Répartition des genres** :

✓ Certains pays montrent une nette dominance d'un genre spécifique, tandis que d'autres révèlent une variété de préférences. Par exemple, en Malaisie, **Supernatural** est le genre le plus regardé, tandis qu'en France, plusieurs genres affichent des proportions similaires.

2. **Impact culturel** :

✓ Ces différences suggèrent que les préférences en termes de genres d'animes sont influencées par les contextes culturels et les traditions narratives propres à chaque région.

- ✓ Les genres fantastiques et surnaturels semblent avoir une portée internationale, tandis que certains genres, comme les thrillers ou les drames, répondent davantage à des sensibilités régionales.

Conclusion : Cette analyse met en évidence l'importance de personnaliser les contenus en fonction des audiences cibles dans différentes régions. Les studios d'animation peuvent utiliser ces informations pour adapter leurs créations et stratégies de distribution afin de maximiser leur impact dans chaque marché.

Synthèse des Résultats

L'analyse de cet ensemble de données sur les animes populaires permet de dégager plusieurs points essentiels :

- La majorité des animes bénéficient de notes élevées, reflétant une réception globalement positive.
- Les investissements en budget ne sont pas un facteur déterminant pour obtenir de bonnes notes, ce qui suggère que d'autres aspects comme le scénario ou la réalisation jouent un rôle important.
- La productivité des studios varie significativement, certains privilégiant la production de séries longues tandis que d'autres se concentrent sur la qualité plutôt que la quantité.
- Les préférences en termes de genres varient notablement selon les pays, démontrant l'importance des facteurs culturels dans la consommation d'animes.

Pistes d'Amélioration et Analyses Complémentaires

Quelques pistes pour approfondir cette analyse :

- **Segmentation Fine :** Analyser les interactions entre genre et public par région pour comprendre les spécificités culturelles.
- **Modélisation Prédictive :** Utiliser des techniques de machine learning pour prédire le succès (rating) d'un anime en fonction de son budget, nombre d'épisodes et autres facteurs.
- **Analyse Longitudinale :** Étudier l'évolution de la réception critique sur plusieurs années pour identifier les tendances au fil du temps.
- **Analyse Sentimentale :** Si disponible, intégrer des avis ou commentaires pour une analyse qualitative complémentaire.

Conclusion

Ce rapport offre une vision structurée de l'analyse des tendances dans l'industrie de l'animation. Les résultats montrent que, malgré une relation modérée entre budget et qualité perçue, d'autres facteurs (genre, studio, contexte culturel) jouent un rôle crucial dans le succès d'un anime. Ces insights peuvent non seulement orienter les décisions stratégiques des studios, mais aussi ouvrir la voie à des analyses plus fines et spécifiques pour mieux comprendre les dynamiques du marché.

.annexe

```
# Importer les données
df <- read_csv("C:/INSEEDS/PROJET/R/most_watched_anime_dataset_100_entries.csv")
str(df)
# Charger les packages nécessaires
library(tidyverse)


# Importer les données
df <- read_csv("chemin/vers/most_watched_anime_dataset_100_entries.csv")
# Renommer les colonnes en remplaçant les espaces par des "_"
colnames(df) <- df %>% names() %>%
 str_replace_all(" ", "_") %>%
 str_replace_all("\\.+", "_")
df <- df %>%
 rename_with(~ str_replace_all(., "[(). ]", "_")) # Remplace parenthèses, points et espaces par "_"



# Vérifier les premières lignes
head(df)
str(df)
df <- df %>%
 mutate(
  Animation_Studio_Name = ifelse(is.na(Animation_Studio_Name), "Inconnu",
Animation_Studio_Name),
  Most_Watched_in_Country = ifelse(is.na(Most_Watched_in_Country), "Inconnu",
Most_Watched_in_Country),
  Genre = ifelse(is.na(Genre), "Inconnu", Genre)
 )


# Vérifier les valeurs manquantes
```

```
colSums(is.na(df))



# Supprimer les lignes avec Anime_Name manquant
df <- df %>% drop_na(Anime_Name)


df <- df %>%
 mutate(
  Ratings = ifelse(is.na(Ratings), median(Ratings, na.rm = TRUE), Ratings),
  Number_of_Episodes = ifelse(is.na(Number_of_Episodes), median(Number_of_Episodes, na.rm = TRUE), Number_of_Episodes),
  Duration_per_Episode__minutes_ = ifelse(is.na(Duration_per_Episode__minutes_),
                                          median(Duration_per_Episode__minutes_, na.rm = TRUE),
                                          Duration_per_Episode__minutes_)
 )
df <- df %>%
 mutate(
  Budget__in_Million_USD_ = ifelse(is.na(Budget__in_Million_USD_), median(Budget__in_Million_USD_, na.rm = TRUE), Budget__in_Million_USD_),
  Release_Year = ifelse(is.na(Release_Year), as.integer(names(sort(table(Release_Year), decreasing = TRUE)[1])), Release_Year)
 )
colSums(is.na(df))



# Convertir les variables au bon format
df <- df %>%
 mutate(
  Anime_Name = as.character(Anime_Name),
```

```
  Ratings = as.numeric(Ratings),
  Number_of_Episodes = as.integer(Number_of_Episodes),
  Release_Year = as.integer(Release_Year),
  Budget__in_Million_USD_ = as.numeric(Budget__in_Million_USD_),
  Genre = as.factor(Genre),
  Most_Watched_in_Country = as.factor(Most_Watched_in_Country),
  Animation_Studio_Name = as.factor(Animation_Studio_Name),
  Duration_per_Episode__minutes_ = as.numeric(Duration_per_Episode__minutes_)
 )

# Vérification finale
glimpse(df)

# Sauvegarder le dataset nettoyé
write_csv(df, "C:/INSEEDS/PROJET/R/most_watched_anime_cleaned.csv")



##partie 2 :

library(moments) # Pour skewness et kurtosis

# Calcul des statistiques pour les variables numériques
stats <- data.frame(
 Variable = c("Ratings", "Number_of_Episodes", "Budget__in_Million_USD_",
"Duration_per_Episode__minutes_"),
 Moyenne = c(mean(df$Ratings, na.rm = TRUE),
        mean(df$Number_of_Episodes, na.rm = TRUE),
        mean(df$Budget__in_Million_USD_, na.rm = TRUE),
        mean(df$Duration_per_Episode__minutes_, na.rm = TRUE)),
 Variance = c(var(df$Ratings, na.rm = TRUE),
```

```
        var(df$Number_of_Episodes, na.rm = TRUE),
        var(df$Budget__in_Million_USD_, na.rm = TRUE),
        var(df$Duration_per_Episode__minutes_, na.rm = TRUE)),
 Ecart_Type = c(sd(df$Ratings, na.rm = TRUE),
         sd(df$Number_of_Episodes, na.rm = TRUE),
         sd(df$Budget__in_Million_USD_, na.rm = TRUE),
         sd(df$Duration_per_Episode__minutes_, na.rm = TRUE)),
 Skewness = c(skewness(df$Ratings, na.rm = TRUE),
        skewness(df$Number_of_Episodes, na.rm = TRUE),
        skewness(df$Budget__in_Million_USD_, na.rm = TRUE),
        skewness(df$Duration_per_Episode__minutes_, na.rm = TRUE)),
 Kurtosis = c(kurtosis(df$Ratings, na.rm = TRUE),
        kurtosis(df$Number_of_Episodes, na.rm = TRUE),
        kurtosis(df$Budget__in_Million_USD_, na.rm = TRUE),
        kurtosis(df$Duration_per_Episode__minutes_, na.rm = TRUE))
)

# Afficher les statistiques
print(stats)

summary(df)
## visualisation



library(ggplot2)

# Scatterplot avec une courbe de régression
ggplot(df, aes(x = Budget__in_Million_USD_, y = Ratings)) +
  geom_point(color = "blue", alpha = 0.6, size = 3) + # Points
```

```
 geom_smooth(method = "lm", se = TRUE, color = "red", fill = "pink", alpha = 0.3) + # Courbe de régression
 labs(
  title = "Relation entre Budget et Notes des Animes",
  x = "Budget (en Millions USD)",
  y = "Notes"
 ) +
 theme_minimal() + # Style minimaliste
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12)
 )

# 2. Visualiser la distribution des notes

# Histogramme superposé à une courbe de densité
ggplot(df, aes(x = Ratings)) +
 geom_histogram(aes(y = ..density..), bins = 10, fill = "skyblue", color = "white", alpha = 0.6) + # Histogramme
 geom_density(color = "darkblue", size = 1.2) + # Courbe de densité
 labs(
  title = "Distribution des Notes des Animes",
  x = "Notes",
  y = "Densité"
 ) +
 theme_light() + # Style élégant
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12)
 )
```

```
#2.3 Distribution des genres
# Diagramme à barres pour la distribution des genres
ggplot(df, aes(x = Genre)) +
 geom_bar(fill = "mediumpurple", color = "white", alpha = 0.8) + # Barres pour les genres
 labs(
  title = "Distribution des Genres des Animes",
  x = "Genres",
  y = "Nombre d'Animes"
 ) +
 theme_light() + # Thème clair et élégant
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12),
  axis.text.x = element_text(angle = 45, hjust = 1) # Inclinaison des noms pour lisibilité
 )


#2.4 Évolution des budgets au fil des années
# Évolution des budgets au fil des années
ggplot(df, aes(x = Release_Year, y = Budget__in_Million_USD_)) +
 geom_line(stat = "summary", fun = "mean", color = "darkgreen", size = 1.2) + # Ligne moyenne
 geom_point(stat = "summary", fun = "mean", color = "forestgreen", size = 3) + # Points moyens
 labs(
  title = "Évolution des Budgets des Animes au Fil des Années",
  x = "Année de Sortie",
  y = "Budget Moyen (en Millions USD)"
 ) +
 theme_minimal() + # Thème minimaliste et professionnel
```

```
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12),
  axis.text.x = element_text(angle = 45, hjust = 1) # Inclinaison pour lisibilité
 )


#3
# Diagramme à barres pour le nombre moyen d'épisodes par studio
ggplot(df, aes(x = reorder(Animation_Studio_Name, -Number_of_Episodes), y =
Number_of_Episodes)) +
 geom_bar(stat = "summary", fun = "mean", fill = "dodgerblue", color = "white", alpha = 0.8) +
 labs(
  title = "Nombre moyen d'épisodes par studio d'animation",
  x = "Studio d'animation",
  y = "Nombre moyen d'épisodes"
 ) +
 theme_light() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12),
  axis.text.x = element_text(angle = 45, hjust = 1) # Rotation pour lisibilité
 )
# 4.
# Diagramme à barres pour les pays avec le plus d'animes populaires
ggplot(df, aes(x = Most_Watched_in_Country)) +
 geom_bar(fill = "coral", color = "white", alpha = 0.8) +
 labs(
  title = "Pays avec le plus d'animes populaires",
  x = "Pays",
  y = "Nombre d'Animes"
```

```
 ) +
 theme_minimal() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12),
  axis.text.x = element_text(angle = 45, hjust = 1) # Rotation pour la lisibilité
 )

# Calcul du nombre moyen d'épisodes par studio
df_summary <- df %>%
 group_by(Animation_Studio_Name) %>%
 summarise(Mean_Episodes = mean(Number_of_Episodes, na.rm = TRUE))

# Diagramme à barres basé sur les données agrégées
ggplot(df_summary, aes(x = reorder(Animation_Studio_Name, -Mean_Episodes), y = Mean_Episodes))
+
 geom_bar(stat = "identity", fill = "dodgerblue", color = "white", alpha = 0.8) +
 labs(
  title = "Nombre moyen d'épisodes par studio d'animation",
  x = "Studio d'Animation",
  y = "Nombre Moyen d'Épisodes"
 ) +
 theme_light() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12),
  axis.text.x = element_text(angle = 45, hjust = 1)
 )
3. # Calcul du nombre moyen d'épisodes par studio
studio_episodes <- df %>%
```

```
 group_by(Animation_Studio_Name) %>%
 summarise(Mean_Episodes = mean(Number_of_Episodes, na.rm = TRUE)) %>%
 arrange(desc(Mean_Episodes)) # Trier par ordre décroissant

# Afficher les résultats
print(studio_episodes)
```

4.

```
# Compter le nombre d'animes par pays
animes_par_pays <- df %>%
 group_by(Most_Watched_in_Country) %>%
 summarise(Nombre_Animes = n()) %>%
 arrange(desc(Nombre_Animes)) # Trier par ordre décroissant

# Afficher les résultats
print(animes_par_pays)
```

```
##partir 3
# 5
# Identifier les 5 studios les plus productifs
top_studios <- df %>%
 group_by(Animation_Studio_Name) %>%
 summarise(Nombre_Animes = n()) %>%
 arrange(desc(Nombre_Animes)) %>%
 slice(1:5) # Extraire les 5 premiers

# Calcul des 5 studios les plus productifs
top_studios <- df %>%
```

```
 group_by(Animation_Studio_Name) %>%
 summarise(Nombre_Animes = n()) %>%
 arrange(desc(Nombre_Animes)) %>%
 slice(1:5)

# Diagramme à barres
ggplot(top_studios, aes(x = reorder(Animation_Studio_Name, -Nombre_Animes), y =
Nombre_Animes)) +
 geom_bar(stat = "identity", fill = "skyblue", color = "white", alpha = 0.8) +
 labs(
  title = "Top 5 des studios d'animation les plus productifs",
  x = "Studios d'Animation",
  y = "Nombre d'Animes Produits"
 ) +
 theme_light() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.text.x = element_text(angle = 45, hjust = 1)
 )


# Afficher les résultats
print(top_studios)
#6.
# Calcul de la corrélation entre le budget et les notes
correlation <- cor(df$Budget__in_Million_USD_, df$Ratings, use = "complete.obs")

# Afficher le résultat
print(paste("La corrélation entre le budget et les notes est :", correlation))
cor.test(df$Budget__in_Million_USD_, df$Ratings)
```

```
# Scatterplot avec courbe de régression
ggplot(df, aes(x = Budget__in_Million_USD_, y = Ratings)) +
 geom_point(color = "darkorange", alpha = 0.6, size = 3) +
 geom_smooth(method = "lm", se = TRUE, color = "blue", fill = "lightblue", alpha = 0.4) +
 labs(
  title = "Corrélation entre le budget et les notes des animes",
  x = "Budget (en millions USD)",
  y = "Notes"
 ) +
 theme_minimal() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14)
 )
```

7.

```
# Calcul de la durée moyenne des épisodes par année
duree_moyenne <- df %>%
 group_by(Release_Year) %>%
 summarise(Duree_Moyenne = mean(Duration_per_Episode__minutes_, na.rm = TRUE)) %>%
 arrange(Release_Year)

# Afficher les résultats
print(duree_moyenne)
# Calcul de la durée moyenne par année
duree_moyenne <- df %>%
 group_by(Release_Year) %>%
 summarise(Duree_Moyenne = mean(Duration_per_Episode__minutes_, na.rm = TRUE))
```

```
# Ligne de tendance
ggplot(duree_moyenne, aes(x = Release_Year, y = Duree_Moyenne)) +
 geom_line(color = "green", size = 1.2) +
 geom_point(color = "darkgreen", size = 3) +
 labs(
  title = "Évolution de la durée moyenne des épisodes au fil des années",
  x = "Année",
  y = "Durée Moyenne des Épisodes (minutes)"
 ) +
 theme_minimal() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14)
 )

8
# Calcul du rapport entre le nombre d'épisodes et le budget
rapport_episodes_budget <- df %>%
 mutate(Rapport = Number_of_Episodes / Budget__in_Million_USD_) %>%
 summarise(Rapport_Moyen = mean(Rapport, na.rm = TRUE))
# Scatterplot pour le rapport entre le nombre d'épisodes et le budget
ggplot(df, aes(x = Budget__in_Million_USD_, y = Number_of_Episodes)) +
 geom_point(color = "purple", alpha = 0.6, size = 3) +
 labs(
  title = "Relation entre le budget et le nombre d'épisodes",
  x = "Budget (en millions USD)",
  y = "Nombre d'Épisodes"
 ) +
 theme_minimal() +
 theme(
```

```
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14)
 )

# Afficher le résultat
print(rapport_episodes_budget)

9
# Comptage des genres par pays
genres_par_pays <- df %>%
 group_by(Most_Watched_in_Country, Genre) %>%
 summarise(Nombre = n()) %>%
 arrange(Most_Watched_in_Country, desc(Nombre))
# Comptage des genres par pays
genres_par_pays <- df %>%
 group_by(Most_Watched_in_Country, Genre) %>%
 summarise(Nombre = n())

# Diagramme empilé par pays
ggplot(genres_par_pays, aes(x = Most_Watched_in_Country, y = Nombre, fill = Genre)) +
 geom_bar(stat = "identity", alpha = 0.8) +
 labs(
  title = "Genres d'anime préférés par pays",
  x = "Pays",
  y = "Nombre d'Animes",
  fill = "Genre"
 ) +
 theme_minimal() +
 theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
```

```
  axis.text.x = element_text(angle = 45, hjust = 1)
 )


# Afficher les résultats
print(genres_par_pays)


duplicated(df)
sum(duplicated(df))
```