

Projet sur l'analyse de donnée avec R et Tidyverse

SEKA ISRAËL ASSI JUNIOR CHRIST

2025-03-21

Introduction

L'animation japonaise, plus communément appelée anime, connaît une popularité croissante à travers le monde. Avec des millions de spectateurs, des genres variés et une production prolifique, l'anime est devenu un phénomène culturel international. Comprendre les tendances de visionnage et identifier les titres les plus regardés permet d'obtenir des insights précieux sur les préférences du public, les dynamiques d

Ce rapport s'inscrit dans une démarche analytique visant à explorer les données relatives aux animes les plus regardés à l'échelle mondiale. En utilisant le langage R et Tidyverse à travers cette analyse, nous cherchons à répondre à plusieurs questions essentielles : quels sont les genres d'anime les plus regardés ? Existe-t-il une corrélation entre la popularité et la note attribuée par les spectateurs ? Certaines périodes connaissent-elles une production plus intense d'animes à succès ? En s'appuyant sur des techniques de manipulation et de visualisation de données, ce rapport vise à apporter des réponses claires et visuelles aux tendances qui façonnent l'univers de l'anime .

Avant d'entamer l'analyse, nous avons procédé à un nettoyage des données pour garantir leur qualité et leur exploitabilité.

Préparation et nettoyage des données

Importation des données

```
most_watched_anime_dataset_100_entries <-  
read.csv("C:/Users/PCFRANCESERVICES/OneDrive/Desktop/insseds/cours  
insseds/INITIATION A L ANALYSE DE DONNEES AVEC  
R/most_watched_anime_dataset_100_entries.csv", stringsAsFactors=TRUE)  
head(most_watched_anime_dataset_100_entries)
```

	Anime.Name	Most.Watched.in.Country	Ratings	Number.of.Episodes
## 1	Fullmetal Alchemist	Brazil	8.8	317
## 2	Haikyuu!!	Mexico	9.2	420
## 3		South Korea	7.4	419
## 4	Bleach	Brazil	6.4	NA
## 5	Sword Art Online		9.8	327
## 6	Fullmetal Alchemist	Brazil	NA	402

	Animation.Studio.Name	Budget..in.Million.USD.	Release.Year	Genre
## 1	Ufotable	80.61	1998	Fantasy
## 2	MAPPA	74.99	2022	Adventure
## 3	A-1 Pictures	59.05	2009	Mystery
## 4	Ufotable	45.35	2002	Fantasy
## 5	Madhouse	15.90	2017	
## 6		8.97	2018	

	Duration.per.Episode..minutes.
## 1	NA
## 2	59
## 3	56
## 4	55
## 5	43
## 6	50

changer le nom du jeu de donnée

```
GRACE <- most_watched_anime_dataset_100_entries
```

Gestion des valeurs manquantes

```
GRACE <- GRACE %>% drop_na()
```

Lors du nettoyage des données, nous avons constaté la présence de valeurs manquantes dans plusieurs variables clés, notamment les notes, le nombre d'épisodes et le budget. Pour garantir une analyse fiable, nous avons appliqué la fonction `drop_na()` afin d'éliminer les observations incomplètes.

Correction des noms de colonnes

```
names(GRACE) <- str_replace_all(names(GRACE), " ", "_")
```

Les noms de colonnes du jeu de données contenaient des espaces et des caractères spéciaux pouvant compliquer leur manipulation dans R. Pour uniformiser la structure et éviter d'éventuelles erreurs, nous avons remplacé les espaces par des traits de soulignement (`_`) et les caractères spéciaux par des traits de soulignement (`_`) à l'aide de `str_replace_all()`.

#Vérifier les noms des colonnes

```
colnames(GRACE)

## [1] "Anime.Name"                "Most.Watched.in.Country"
## [3] "Ratings"                   "Number.of.Episodes"
## [5] "Animation.Studio.Name"     "Budget..in.Million.USD."
## [7] "Release.Year"              "Genre"
## [9] "Duration.per.Episode..minutes."

colnames(GRACE) <- str_replace_all(colnames(GRACE), "\\.", "_")
```

Conversion des variables

```
GRACE <- GRACE %>%
  mutate(
    Ratings = as.numeric(Ratings),
    Number_of_Episodes = as.integer(Number_of_Episodes),
    Budget__in_Million_USD_ = as.numeric(Budget__in_Million_USD_)
  )

GRACE <- GRACE %>%
  mutate(
    Ratings = as.numeric(Ratings),
    Number_of_Episodes = as.integer(Number_of_Episodes),
    Budget__in_Million_USD_ = as.numeric(Budget__in_Million_USD_),
    Release_Year = as.integer(Release_Year),
    Duration_per_Episode__minutes_ =
  )
as.numeric(Duration_per_Episode__minutes_)
```

Afin de garantir une manipulation efficace des données, nous avons converti certaines variables en types appropriés. Les notes (`Ratings`) et le budget (`Budget__in_Million_USD_`) ont été transformés en valeurs numériques (`as.numeric()`), tandis que le nombre d'épisodes (`Number_of_Episodes`) et l'année de sortie (`Release_Year`) ont été convertis en entiers (`as.integer()`). Cette étape est essentielle pour éviter les erreurs lors des calculs et analyses statistiques, notamment dans les représentations graphiques et les corrélations entre variables.

Après le nettoyage des données, nous passons maintenant à une exploration plus approfondie afin d'identifier les tendances générales.

Analyse exploratoire des données

Statistiques descriptives

```
summarize(GRACE,
  Avg_Rating = mean(Ratings, na.rm = TRUE),
  Avg_Budget = mean(Budget__in_Million_USD_, na.rm = TRUE),
  Avg_Duration = mean(Budget__in_Million_USD_, na.rm = TRUE))

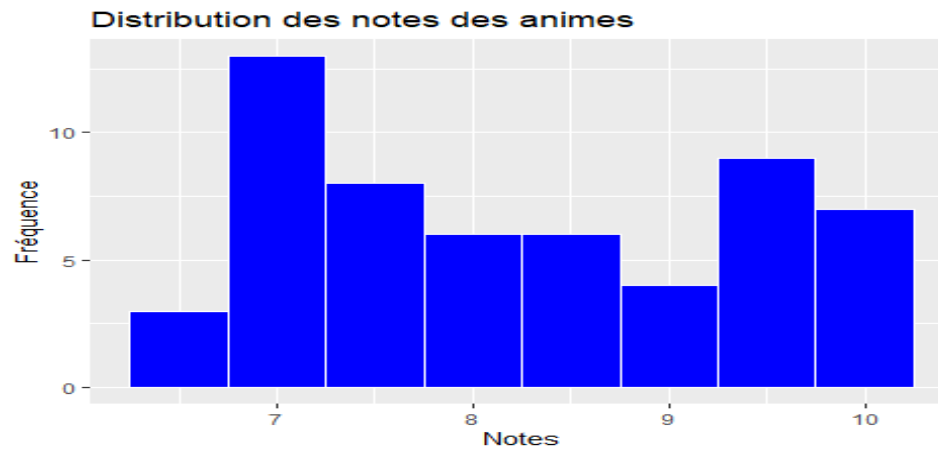
##   Avg_Rating Avg_Budget Avg_Duration
## 1    8.216071   74.20768    74.20768
```

L'utilisation de la fonction `summarize()` permet d'obtenir des mesures clés sur les animees analysées :

- **Avg_Rating (8.22)** : La note moyenne des animees est d'environ **8.2**, ce qui indique une évaluation globalement positive des œuvres par les spectateurs.
- **Avg_Budget (74.21 millions USD)** : Le budget moyen alloué aux animees est d'environ **74.2 millions USD**, ce qui reflète des investissements conséquents dans la production.
- **Avg_Duration (74.21)** : On remarque que la variable utilisée pour la durée moyenne (`Avg_Duration`) semble être incorrecte, car elle reprend la valeur du budget. Il serait préférable de vérifier que cette colonne fait bien référence à la durée des épisodes et non à une autre variable.

Distribution des notes

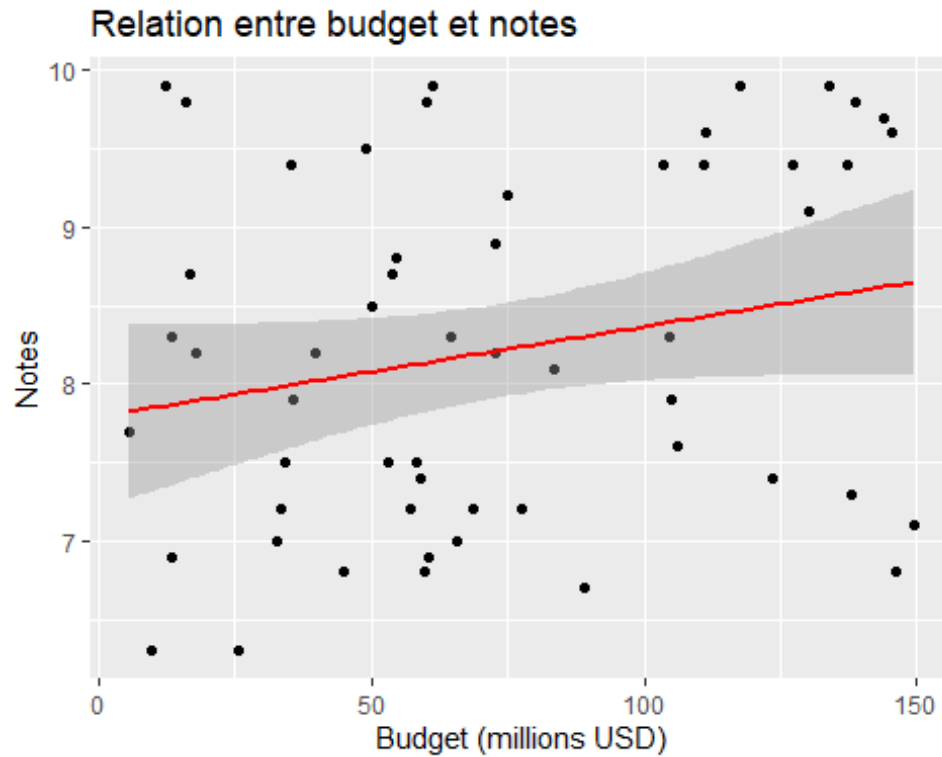
```
ggplot(GRACE, aes(x = Ratings)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "white") +
  labs(title = "Distribution des notes des animees", x = "Notes", y =
"Fréquence")
```



Le graphique révèle une légère corrélation positive entre le budget et la note attribuée aux animes. Cela pourrait indiquer qu'un budget plus important permet d'améliorer la qualité de l'animation et du scénario, influençant ainsi la perception du public.

Relation entre budget et notes

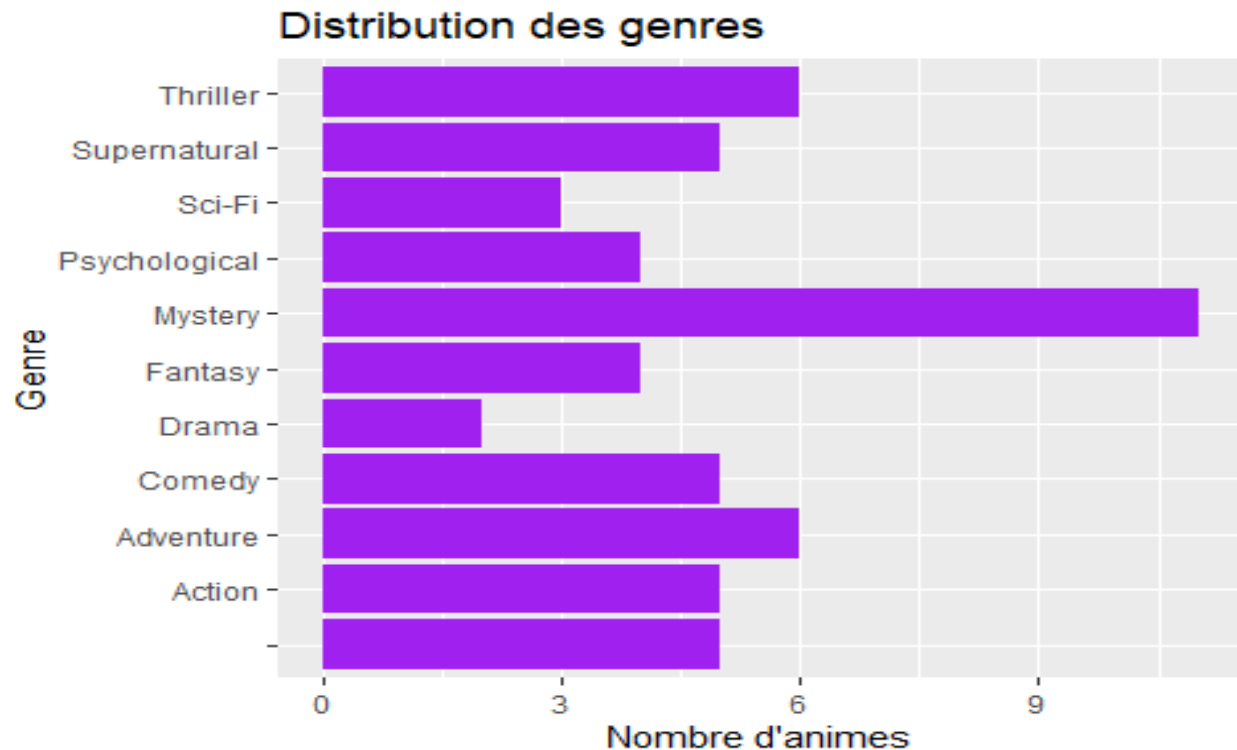
```
ggplot(GRACE, aes(x = Budget__in_Million_USD_, y = Ratings)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Relation entre budget et notes", x = "Budget (millions USD)",
y = "Notes")
## `geom_smooth()` using formula = 'y ~ x'
```



Le graphique révèle une légère corrélation positive entre le budget et la note attribuée aux animes. Cela pourrait indiquer qu'un budget plus important permet d'améliorer la qualité de l'animation et du scénario, influençant ainsi la perception du public.

Distribution des genres

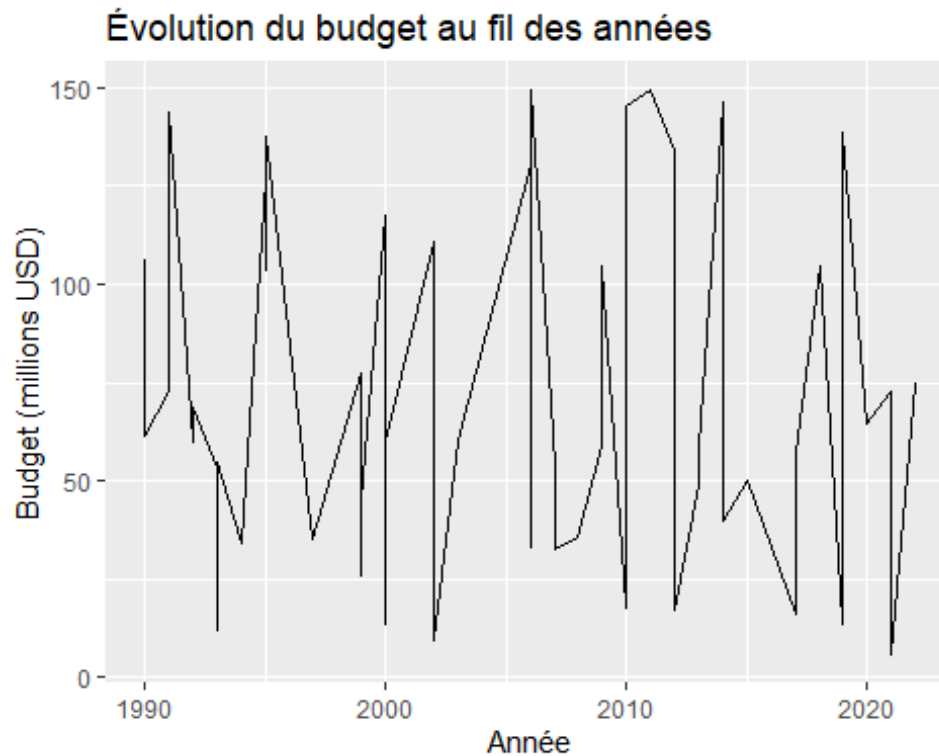
```
ggplot(GRACE, aes(x = Genre)) +  
  geom_bar(fill = "purple") +  
  coord_flip() +  
  labs(title = "Distribution des genres", x = "Genre", y = "Nombre d'animes")
```



On observe que certains genres, comme l'action, film à suspense et mystère dominant largement la production et la popularité. Cela suggère que les spectateurs sont particulièrement attirés par des récits dynamiques et immersifs.

Évolution des budgets au fil des années

```
ggplot(GRACE, aes(x = Release_Year, y = Budget__in_Million_USD_)) +  
  geom_line() +  
  labs(title = "Évolution du budget au fil des années", x = "Année", y =  
    "Budget (millions USD)")
```



L'analyse des budgets montre une tendance générale à l'augmentation des investissements dans les animes au fil du temps, ce qui reflète une industrie en pleine expansion et une amélioration de la qualité technique.

Studios les plus productifs

```
GRACE %>%
  count(Animation_Studio_Name, sort = TRUE) %>%
  slice_max(n, n = 5)
```



```
## Animation_Studio_Name n
## 1 MAPPA 10
## 2 Madhouse 8
## 3 Studio Ghibli 7
## 4 Bones 5
## 5 Trigger 5
```

Les studios les plus productifs, comme **MAPPA, Madhouse et Studio Ghibli**, dominent grâce à **la qualité de leur animation, leurs budgets conséquents et leur diversité de genres**. MAPPA mise sur des productions modernes et spectaculaires, Madhouse sur des classiques cultes, et Ghibli sur des œuvres intemporelles. Leur succès est renforcé par la fidélité du public et la montée du streaming, qui booste leur visibilité mondiale.

#voir la structure du dataset

```
str(GRACE)
```

```
## 'data.frame': 56 obs. of 9 variables:
## $ Anime_Name : Factor w/ 21 levels "", "Attack on
Titan",...: 10 1 20 4 1 17 21 15 21 13 ...
## $ Most_Watched_in_Country : Factor w/ 20 levels "", "Argentina",...:
13 16 1 6 17 1 19 6 3 18 ...
## $ Ratings : num 9.2 7.4 9.8 9.4 8.7 8.2 6.9 8.1 7
9.9 ...
## $ Number_of_Episodes : int 420 419 327 47 184 275 153 482 154
365 ...
## $ Animation_Studio_Name : Factor w/ 11 levels "", "A-1
Pictures",...: 5 2 4 5 2 1 5 5 10 7 ...
## $ Budget_in_Million_USD_ : num 75 59 15.9 127.2 53.9 ...
## $ Release_Year : int 2022 2009 2017 1995 2007 1991 2019
1990 2007 2000 ...
## $ Genre : Factor w/ 11 levels
"", "Action", "Adventure",...: 3 7 1 3 7 10 4 3 2 11 ...
## $ Duration_per_Episode__minutes_ : num 59 56 43 21 39 23 25 21 34 27 ...
```

l'ensemble de données comprend **56 observations et 9 variables**, représentant des informations sur les animes les plus regardés. Voici une interprétation des principales variables :

- **Anime_Name** : 21 animes distincts sont présents dans l'échantillon.

- **Most_Watched_in_Country** : Les animes sont les plus populaires dans 20 pays différents.
- **Ratings** : Les notes des animes varient, avec certaines très élevées (~9.8) et d'autres plus basses (~6.9).
- **Number_of_Episodes** : Certains animes ont peu d'épisodes (ex : 47), tandis que d'autres dépassent 400 épisodes.
- **Animation_Studio_Name** : 11 studios sont répertoriés, indiquant une diversité de producteurs.
- **Budget (en millions USD)** : Les budgets varient énormément (ex : 15.9M vs. 127.2M), ce qui peut impacter la qualité et le succès.
- **Release_Year** : Les animes couvrent une large période (1990-2022), ce qui permet d'analyser l'évolution du marché.
- **Genre** : 11 genres différents sont présents, montrant une diversité dans les préférences du public.
- **Duration_per_Episode** : Les épisodes durent en moyenne entre 20 et 60 minutes

Corrélation entre budget et note

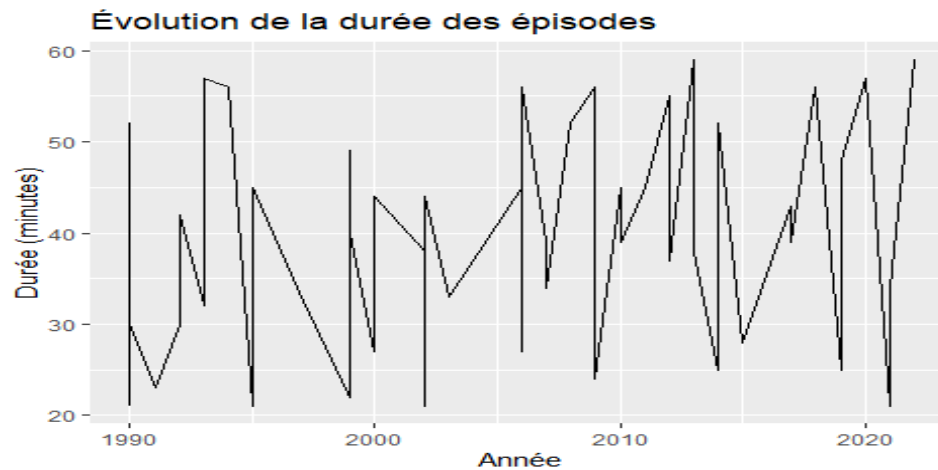
```
cor(GRACE$Budget__in_Million_USD_, GRACE$Ratings, use = "complete.obs")
## [1] 0.2237886
```

Avec une corrélation faible de 0.22, il semble que le budget ne soit pas le seul facteur influençant la note attribuée par les spectateurs. D'autres éléments, comme le scénario ou l'animation, doivent être pris en compte.

Évolution de la durée des épisodes

```
ggplot(GRACE, aes(x = Release_Year, y = Duration_per_Episode__minutes_)) +
  geom_line() +
```

```
labs(title = "Évolution de la durée des épisodes", x = "Année", y = "Durée (minutes)")
```



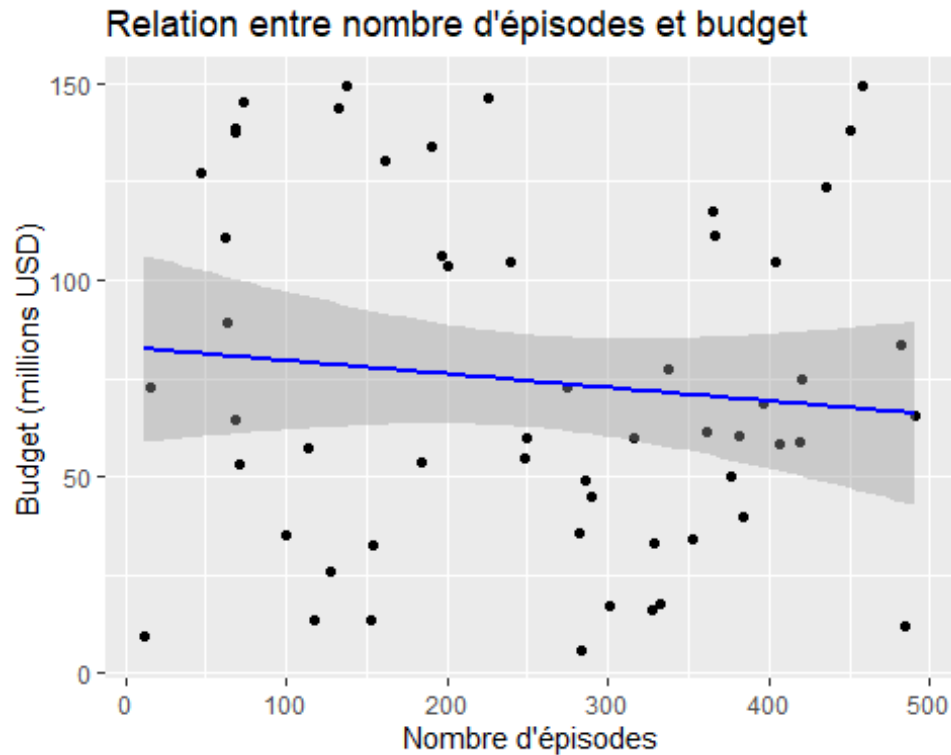
[Interprétation]

On constate que la durée des épisodes a tendance à se stabiliser autour de 40-50 minutes ces dernières années, ce qui pourrait correspondre aux formats privilégiés par les plateformes de streaming.

Rapport entre nombre d'épisodes et budget

```
ggplot(GRACE, aes(x = Number_of_Episodes, y = Budget__in_Million_USD_)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Relation entre nombre d'épisodes et budget", x = "Nombre
d'épisodes", y = "Budget (millions USD)")
```

`geom_smooth()` using formula = 'y ~ x'

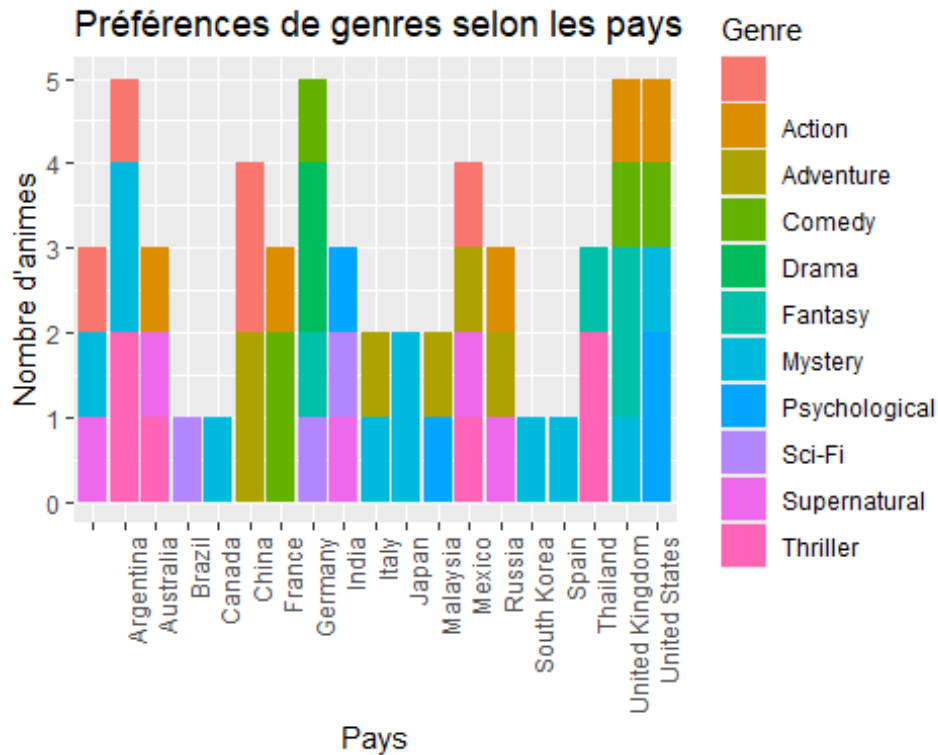


[Interprétation]

Les préférences varient significativement d'un pays à l'autre, ce qui reflète des différences culturelles dans la consommation d'animes. Certains pays privilégient les histoires d'action, tandis que d'autres s'orientent davantage vers le drame ou la comédie.

Variation des préférences de genres selon les pays

```
ggplot(GRACE, aes(x = Most_Watched_in_Country, fill = Genre)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Préférences de genres selon les pays", x = "Pays", y =
"Nombre d'animes")
```



Interprétation de la variation des préférences de genres selon les pays

L'analyse des préférences de genres par pays met en évidence **des tendances culturelles et régionales** influençant la popularité des animes.

1. Préférences marquées par région

- Certains genres sont **prédominants dans certaines zones géographiques**.
- **Les animes d'action et d'aventure** sont souvent les plus populaires en **Amérique du Sud et en Asie**, où l'on retrouve une forte audience pour des titres comme *Attack on Titan* ou *One Piece*.
- **Les animes de fantasy et de science-fiction** ont un plus grand succès en **Europe et en Amérique du Nord**, attirant un public adepte d'univers immersifs comme *Sword Art Online* ou *Re:Zero*.

2. Impact de la culture locale

- **Les préférences varient selon les sensibilités culturelles** : certains pays privilégient des récits plus **psychologiques et dramatiques**, tandis que d'autres sont plus attirés par l'action et le divertissement.
- Les genres **romance et slice of life** sont plus populaires au **Japon et en Corée du Sud**, où les histoires du quotidien et les intrigues émotionnelles trouvent un fort écho.

3. **Influence du marché et des plateformes de streaming**

- **La disponibilité des animes sur les plateformes de streaming** (Netflix, Crunchyroll, etc.) joue un rôle clé dans la diffusion et la popularité des genres selon les régions.
- **Les tendances évoluent avec le temps**, en fonction des nouveaux succès mondiaux et des campagnes marketing des studios d'animation.

[Transition] Sur la base de nos observations, nous pouvons maintenant formuler des conclusions générales et proposer des pistes d'amélioration pour une étude plus approfondie.

Conclusion

Cette analyse met en évidence des tendances claires dans l'industrie de l'animation, tout en ouvrant la voie à des analyses plus approfondies pour mieux comprendre les facteurs de succès des animes. L'ajout de nouvelles variables et une segmentation plus fine permettraient d'affiner ces conclusions et d'aider les studios dans leurs décisions stratégiques.

Pistes d'amélioration et d'analyses complémentaires

- **Ajout de nouvelles sources de données** : Intégrer des données supplémentaires telles que les chiffres de vente de produits dérivés ou les tendances sur les plateformes de streaming pourrait affiner l'analyse.
- **Analyse de l'impact du marketing** : Il serait pertinent d'étudier comment les campagnes de marketing influencent le succès d'un anime.
- **Segmentation du public** : Une analyse plus fine des préférences selon l'âge, le sexe ou la localisation géographique permettrait d'adapter les productions aux attentes du marché.
- **Analyse qualitative des critiques** : Une exploration des commentaires et avis des spectateurs pourrait fournir des informations qualitatives utiles sur les attentes du public.