Rapport de Projet

Module Statistiques

Réalisé par :

PELTIER Hugo

LIM Sundara

LI Ludovic

MACCHI Nicola

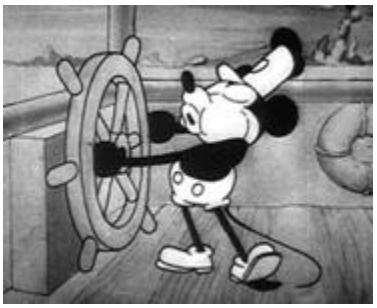
Enseignants :

ANDRIANSITOHANINA Andrimiraho / DELLA MAESTRA Laetitia

Table des matières

Préambule

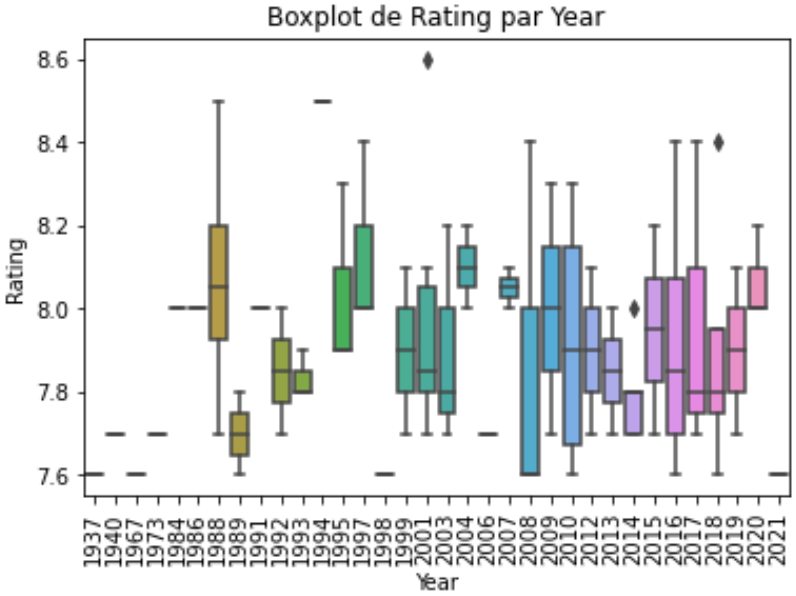
Dans le cadre de notre module 'Introduction aux statistiques', nous devions réaliser un projet de jeu de données qui vise à déduire une étude statistique du traitement d'une base de données. Dans le cas de notre groupe, nous avons choisi la base 'Films d'animations' car nous la trouvions aussi intéressante qu'originale avec certaines données difficiles à traiter, mais très enrichissant à étudier. Nous pouvons citer par exemple la donnée 'Résumé' qui est un résumé des différents films dont la taille varie énormément avec parfois plusieurs lignes de description ce qui nous a énormément compliqué la tâche pour l'analyser. Il y a aussi également la donnée Votes avec laquelle la conversion en float dans nos programmes pouvait être une source de difficulté. Cependant ce projet nous a tous tellement motivé que nous n'avons jamais cédé à la difficulté, ni à la facilité de demander à certaines intelligences artificielles de faire tous nos codes de traitement. Pour la plupart d'entre nous ce projet était une découverte de l'utilisation Python dans la Data Science, preuve que nous avons dû nous exercer des dizaines d'heures pour nous familiariser avec cette notion. Cependant nous nous sommes tous entraidés et grâce à cette solidarité que nous avons pu tous progresser afin de rendre un projet qui ne se veut pas parfait, mais que nous avons fait avec notre envie d'apprendre et notre persévérance. Ce fût une aventure incroyable, accompagné de dizaine voir de centaines d'heures de difficulté mais nous sommes fières de vous présenter notre projet, ainsi que nos différentes interprétations de ce dernier.



Afin que le projet soit le plus complet possible, nous avons créé une documentation GitHub afin que notre projet soit ouvert à tous et qu'il puisse possiblement être utiliser pour d'autres applications dans la Data Science. De ce fait n'hésitez pas à aller voir nos différents codes et à revenir vers nous en cas de questions. Nous vous souhaitons une belle découverte de notre projet.

Introduction

Notre objectif est de pouvoir proposer une analyse simple de notre base de données. Après de nombreux essaies, nous avons eu nos premiers graphes corrects cependant nous voulions proposer un projet qui était optimisé aussi bien dans la complexité de nos diagrammes, mais aussi dans la compréhension de ces derniers qui doivent servir notre étude pour mieux traiter les données. En effet, ces diagrammes sont des outils qu'il faut interpréter et non des réponses à nos questions, il faut donc les rendre optimisés pour avoir une lecture de ses informations claire et concise.
Prenons l'exemple de ce boxplot pour étude bivariée entre deux variables quantitatives :



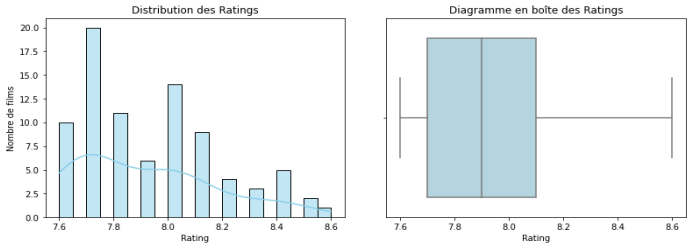
C'est un diagramme correct qui nous renvoie les bonnes informations mais un peu trop chargé donc forcément peu claire à lire. De ce fait, notre réel défi est donc de proposer des diagrammes efficaces et claires.
Nous avons donc divisé notre projet en différentes parties en fonction de l'étude statistiques des différentes variables et aussi des couples ou des ensembles de variables. Afin d'essayer de garder une certaine logique à notre projet, nous avons essayé d'orienter nos études de variables dans le but de trouver des problématiques viables propre à notre base de données. De ce fait nous avons commencé par les études univariés et bivariés que nous avons essayé de généraliser au maximum afin d'avoir des résultats qui pourront nous aiguiller sur la rechercher de nos problématiques. Ensuite, nous pourrons traiter nos différentes problématiques pour en déduire des hypothèses et répondre à ces dernières.

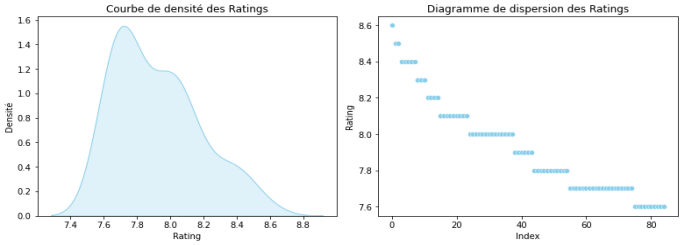
A noter que nous avons décider de traiter les études des variables avant les problématiques. Cette décision découle notre volonté de d'abord nous familiariser avec la base pour ensuite formuler des problématiques cohérentes. Dans ce cas, la pluralité de nos études nous permettent ensuite d'évaluer au mieux la résolution de nos problématiques en proposant aussi bien un test ANOVA qu'un test de Matt Whithney U pour être le plus adapté possible à notre problématique qu'une simple étude bivariée par exemple.
Dans cette méthode, tout réside donc dans la bonne interprétation de nos études, point où nous avons était particulièrement vigilent.

Etudes univariées

Nous allons commencer par l'étude univarié des variables quantitatives. Dans ce cadre nous allons vous exposer notre méthodologie suivante. Nous vous présentons nos différents graphes avec une interprétation de ces derniers afin d'extraire le plus d'informations de ces études.

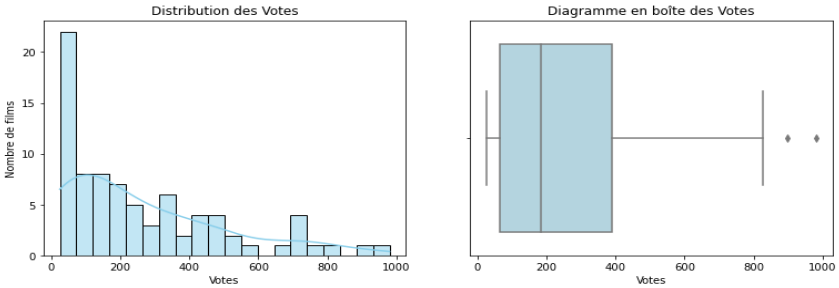
30 Avril 2024
Rapport de Projet Statistique
Jeu de données

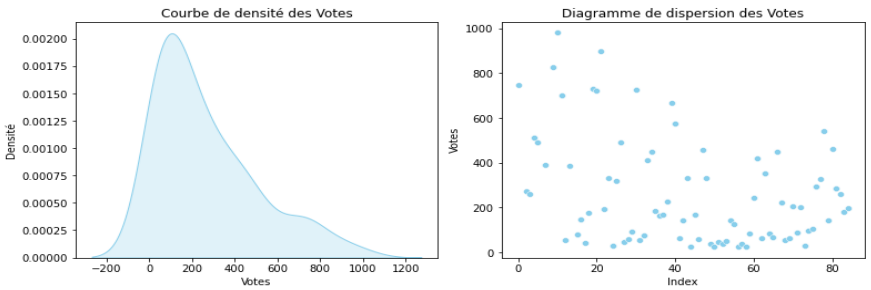




Statistique	Valeur
Moyenne	7.822359411764712
Médiane	7.9
Mode	7.7
Ecart-type	0.2579480416862359
Variance	0.06652011204481792
Minimum	7.6
Maximum	8.6
Premier quartile (Q1)	7.7
Deuxième quartile (Q2)	7.9
Troisième quartile (Q3)	8.1

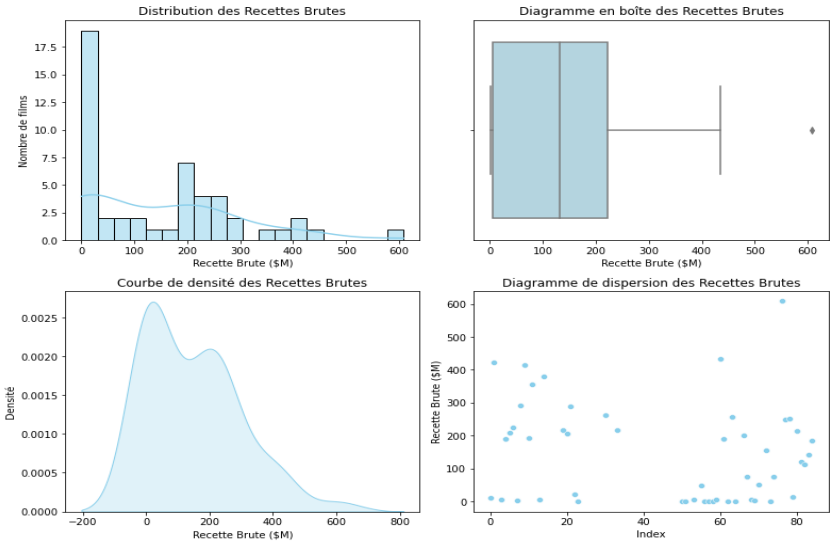
Ces quatre diagrammes nous montrent la distribution et la tendance des évaluations des films d'animation. Le code fait le calcul des statistiques descriptives telle que la moyenne, médiane, etc qui sont présentées dans le petit tableau en bas à gauche, et crée ensuite les quatre diagrammes pour comprendre les données. Le premier est un histogramme, représentant la distribution des ratings, avec le nombre de films en ordonnées et le rating en abscisse. Le deuxième, diagramme en boîte représente la même chose, en nous montrant en plus la médiane, les quartiles et les valeurs extrêmes. Le troisième, la courbe de densité, est une estimation de la distribution des ratings, finalement, le dernier, le diagramme de dispersion montre la relation entre l'index du film et son rating.





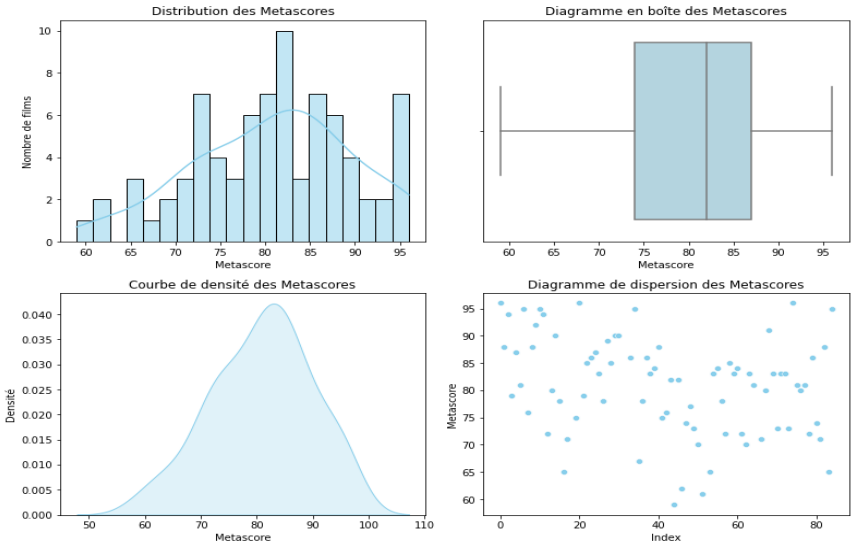
Statistique	Valeur
Moyenne	263.6555185518518 53
Médiane	184.356
Mode	25.319
Ecart-type	256.9585836465108
Variance	66048.37510205277
Minimum	25.319
Maximum	979.994
Premier quartile (Q1)	99.147
Deuxième quartile (Q2)	184.356
Troisième quartile (Q3)	388.925

Pour l'étude des Votes, même disposition et diagrammes. Nous pouvons voir qu'il y a un extremum aux alentours de 1000 (diagramme en boîte ou de dispersion), mais que la plupart des votes sont regroupés entre 70 et 400 votes.



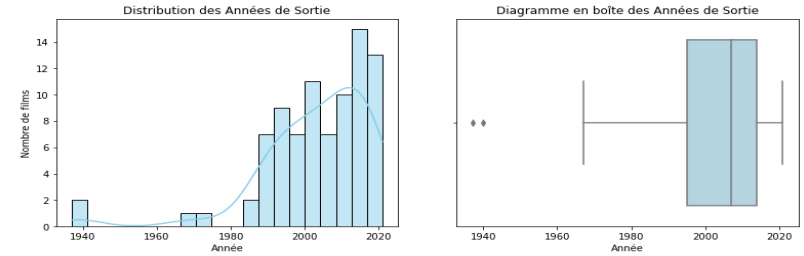
Statistique	Valeur
Moyenne	146.7416
Médiane	101.29300000000002
Mode	0.19
Ecart-type	148.82674546594478
Variance	22147.814274938776
Minimum	0.19
Maximum	809.58
Premier quartile (Q1)	5.0675
Deuxième quartile (Q2)	101.29300000000002
Troisième quartile (Q3)	222.2525

Etude univariée sur les recettes brutes des films d'animation. Les diagrammes sont les mêmes ainsi que le tableau. On peut voir d'après le diagramme de dispersion, qu'il y a de grosses différences sur les chiffres. Une grosse partie se trouve ente 0 et 20, tandis qu'il y a des plusieurs piques à plus de 300 M$. On peut voir cela dans le diagramme de distribution, 18 films dans la première partie (0~20) et 4 pour la deuxième (300~).

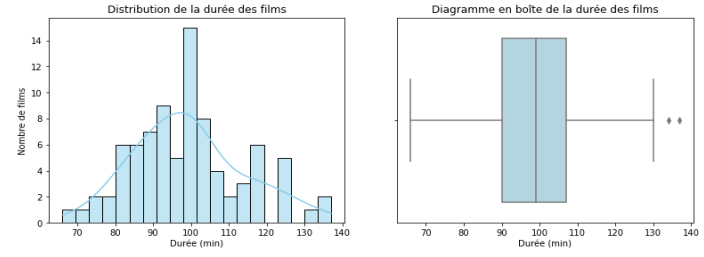


Statistique	Valeur
Moyenne	82.85
Médiane	82.0
Mode	85.0
Ecart-type	8.960616458504483
Variance	80.36227848101266
Minimum	58.0
Maximum	96.0
Premier quartile (Q1)	74.0
Deuxième quartile (Q2)	82.0
Troisième quartile (Q3)	87.0

Le Metascore représente l'appréciation générale des médias spécialisés, une note évaluée sur 100. On a les mêmes quatre diagrammes et le tableau en bas à gauche. Les notes sont bien dispersées, entre 50 et 60, car n'ayant pas de film avec un Metascore inférieur à 50. Le diagramme en boîte est assez centré et inclus les deux extrêmes.



Etude univariée des années de sortie des films d'animation : Mêmes diagrammes, années de sortie majoritairement dispersées entre 1980 et 2020. Deux films qui sont sortis avant 1960, qui ne sont pas inclus dans le diagramme en boîte.



Etude univariée sur a durée des films d'animation. La durée moyenne est de 1 heure 40, on peut le voir assez clairement sur les trois premiers diagrammes. Les durées sont bien dispersées, avec le nombre de films d'une durée inférieur et ceux supérieur à 1h40 qui sont assez proches.

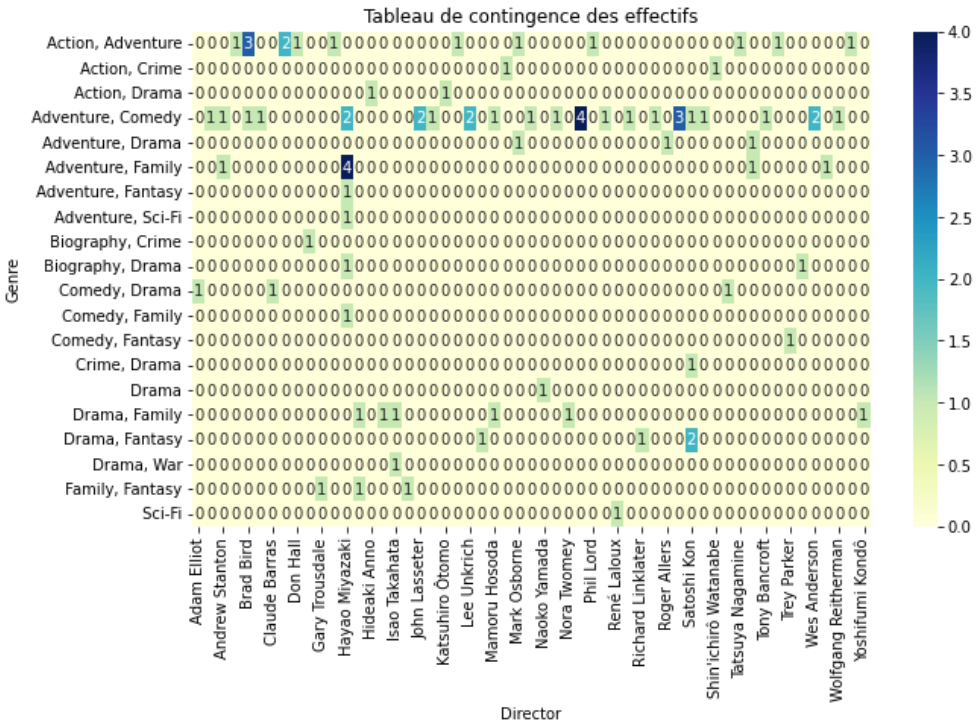
Etudes Bivariées

Nous avons réalisé deux études bivariés afin de nous familiariser avec les études à plusieurs variables. L'intérêt de ces études et de voir la relation entre deux données ce qui est très utile pour la suite et surtout pour nous aiguiller pour nos futurs problématiques.

Les études bivariées ont plusieurs utilisations et analyses selon le type de variables choisit. Par exemple, pour deux quantitatif, il y a le test du chi-deux, l'analyse de la cohérence, celle de la corrélation, les tableaux croisés..

En somme, une étude bivarié permet d'avoir une vue plus complète que des études univariés et montrer des liens, des associations entre deux variables.

Etude Bivarié Qualitative vs Qualitative :



Dans notre cas ici, nous avons un tableau de contingence des effectifs ou bien un tableau croisé entre le genre de nos films d'animations et les réalisateurs.

Ce tableau à en abscisse les réalisateurs et en ordonnées les différents genres des films. Il permet de voir par réalisateur une échelle du nombre film et de son genre.

Cette échelle a été aussi réalisé par une heatmap qui permet de grader par couleur et par numéro de 0 à 4. Ce qui nous donne un aperçu facile du genre des film par réalisateur.

De plus, dans le programme associé à ce diagramme nous avons aussi le calcul du chi-deux qui a été cité précédemment.

```
Test du chi2 pour l'indépendance entre Genre et Director :
Chi2 : 1095.5083333333334
p-value : 0.06479416559486777
```

Ici nous avons une valeur de 0,064 pour la valeur de p ce qui est un tout petit peu supérieur au seuil de 0,05. Cela signifie donc qu'il n'y a pas de preuves suffisantes pour montrer le lien entre ces deux variables.

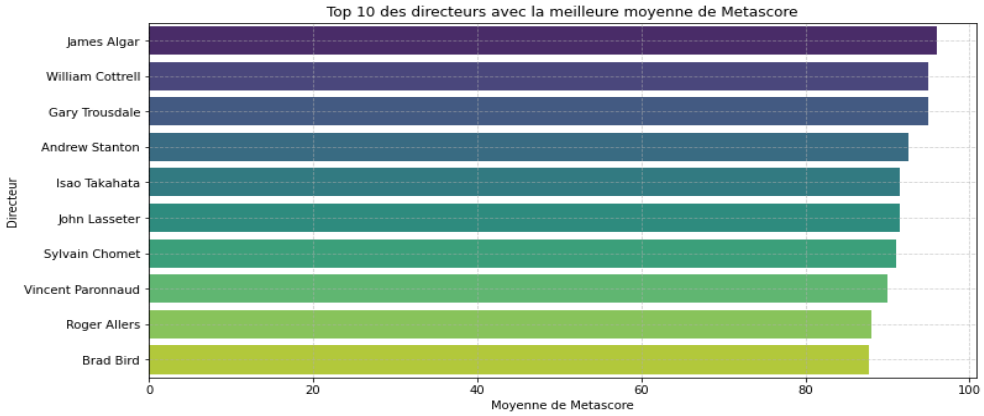
Sur ce coup si donc, le calcul du chi-deux n'a pas donné de résultat permettant de conclure sur de possible lien entre les deux variables.

Cependant, le diagramme en lui-même permet de déduire certaines informations sur les réalisateurs et son lien avec les genres de films d'animations.

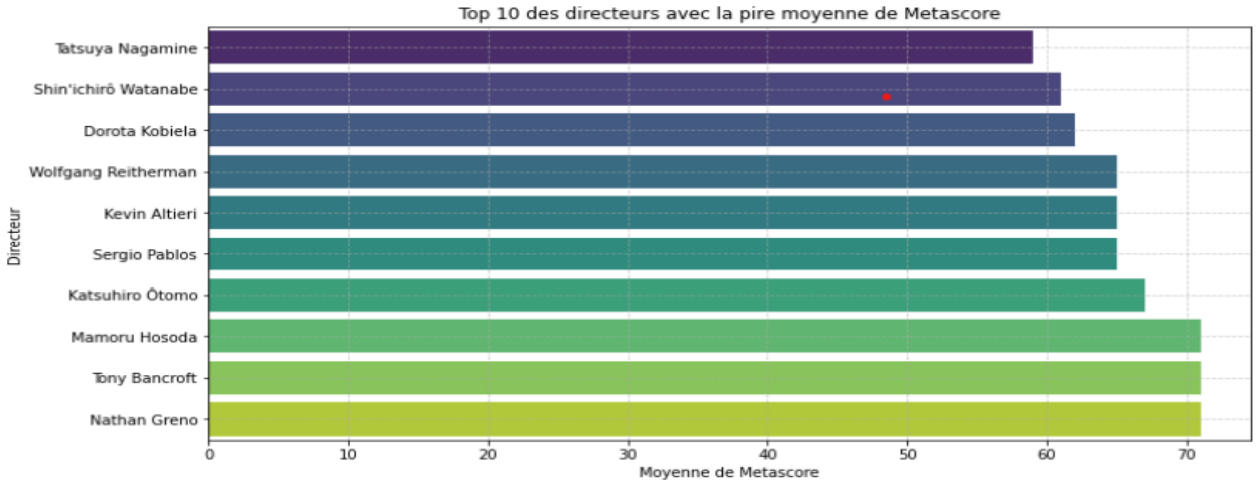
On peut donc détecter certaine tendance à travers celui-ci, comme une plus grande présence de film par réalisateur dans le genre "Adventure and Commedy" que les autres genres.
De plus, nous pouvons aussi voir la diversité que certain réalisateur propose à travers ses films avec le nombre d'effectifs par colonne.

Etude Bivarié Quantitative vs Qualitative :

Dans ce type d'étude avec une quantitative et une qualitative, il y a plusieurs utilisations principales de l'analyse. Par exemple, il y a la comparaison de moyenne ou de médiane, de test ANOVA, régression linéaire...



Ce diagramme permet l'étude bivarié d'une variable qualitative et quantitatif sous forme d'un top 10 des réalisateurs selon la Moyenne Metascore de leurs films. Nous avons donc ici une analyse de moyenne faite de manière décroissante. En abscisse, il y l'échelle du MetaScore de 0 à 100 et en ordonnée les 10 réalisateurs avec le plus grand MétaScore. Ce diagramme comme l'indique son nom permet de voir les meilleurs réalisateurs en fonction de la moyenne du score que reçoit leur film.



Ce diagramme permet une analyse complémentaire du diagramme précèdent en montrant les 10 directeurs avec la pire moyenne de MetaScore.

Nous pouvons voir que la moyenne la plus basse est d'environ 57 et que la moyenne grimpe assez rapidement étant donné que le 10-ème "pire" réalisateur trouve sa moyenne à un peu plus de 70.

Régressions linéaires

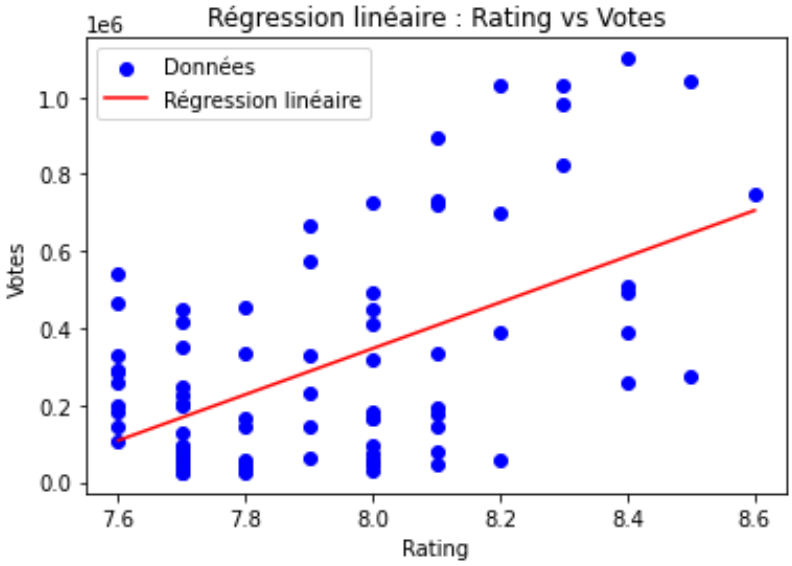
Nous avons réalisé 2 régressions linéaires : une régression linéaire simple de 2 variables quantitatives et une régression linéaire multiple de 3 variables, dont 2 quantitatives et 1 qualitative.

Les régressions linéaires permettent d'étudier la relation entre une variable dépendante et une ou plusieurs variables indépendantes.

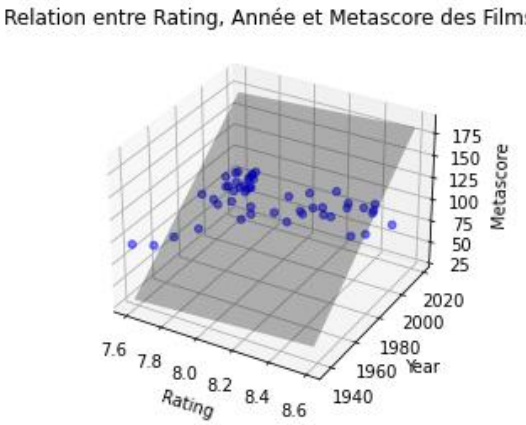
La régression linéaire sert à faire des prédictions sur les valeurs de la variables dépendantes par rapport à celles indépendantes, une analyse de la relation entre les variables, le contrôle des variables confondantes, étudier la relation entre les variables et évaluer la qualité du modèle.

Le premier graphe représente la régression linéaire de la variable dépendante 'Vote' et la variable indépendante 'Rating'.

En effet, chaque point bleu représente une paire de valeur ('Rating', 'Vote') pour un film donné. La droite rouge est la régression linéaire des données. Sa pente étant positive, on peut en conclure que lorsque le 'Rating' augmente, les 'Vote' ont tendance à augmenter aussi.



Concernant le second diagramme ci-dessous, celui-ci est la représentation de la régression linéaires multiples de 2 variables quantitatives 'Year et 'Metascore' et la variable qualitative 'Film'. Le nuage de point bleus représente pour chaque film les valeurs des variables 'Year', 'Rating', 'Metascore' La surface plane représente le plan de régression linéaire multiple.



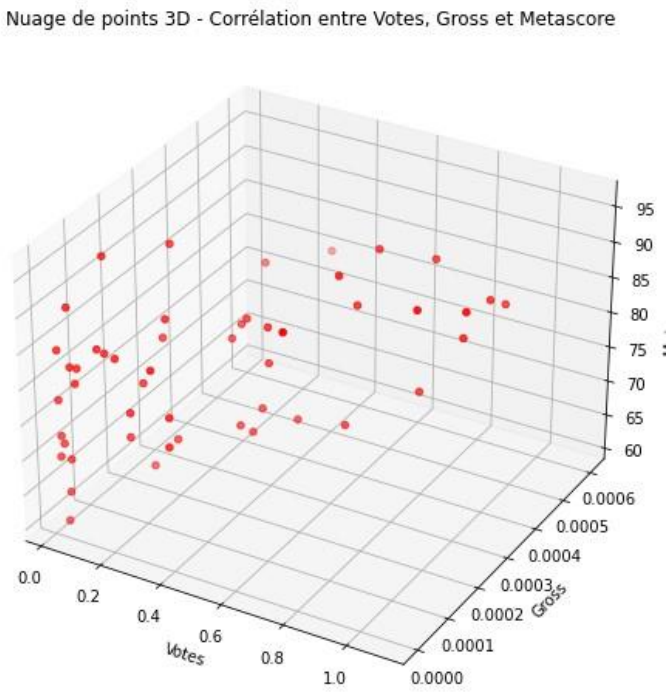
Etudes Multivariés

Les études multivariées dans notre base de données de films offrent une approche essentielle pour explorer les relations complexes entre plusieurs variables, allant des genres et des scores de critiques aux performances au box-office. Ces analyses permettent de découvrir des corrélations subtiles entre les différentes caractéristiques des films, d'identifier des tendances dans les préférences du public et de simplifier la compréhension de l'ensemble de données en réduisant sa dimensionnalité. En utilisant des méthodes telles que l'analyse en composantes principales et le clustering, nous pouvons non seulement découvrir des sous-groupes de films partageant des caractéristiques similaires, mais aussi obtenir des insights approfondis sur ce qui rend un film réussi et apprécié.

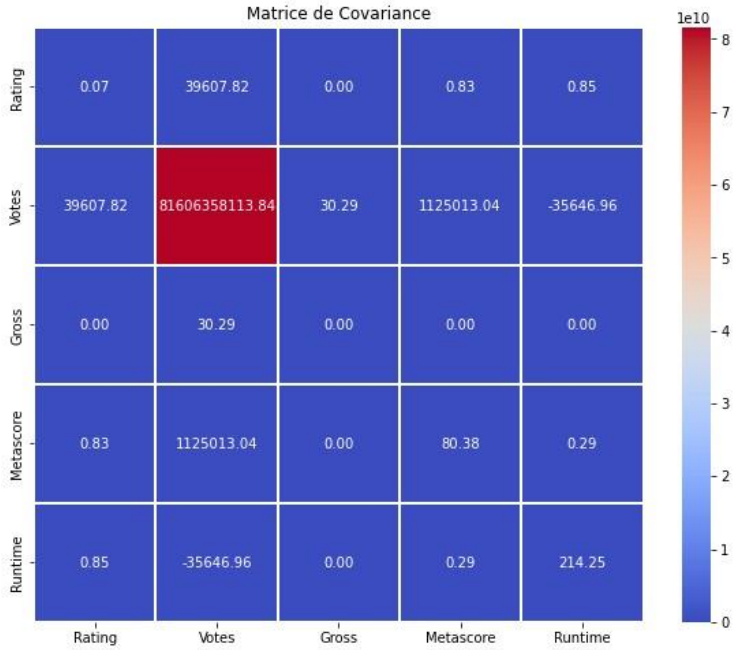
Dans notre démarche d'exploration des relations entre les variables de notre jeu de données, nous sommes résolus à explorer une gamme diversifiée de graphes multivariés. Notre objectif est d'identifier les types de visualisations les plus appropriés pour chaque combinaison de variables, afin de capturer au mieux les structures et les tendances complexes présentes dans nos données. En testant différentes approches graphiques, nous cherchons à élargir notre compréhension des interactions entre les variables et à sélectionner les outils visuels les plus pertinents pour représenter ces relations de manière

claire et informative. Cette démarche exhaustive nous permettra de choisir les techniques de visualisation les plus adaptées à chaque ensemble de variables, garantissant ainsi une interprétation précise et approfondie de notre jeu de données.

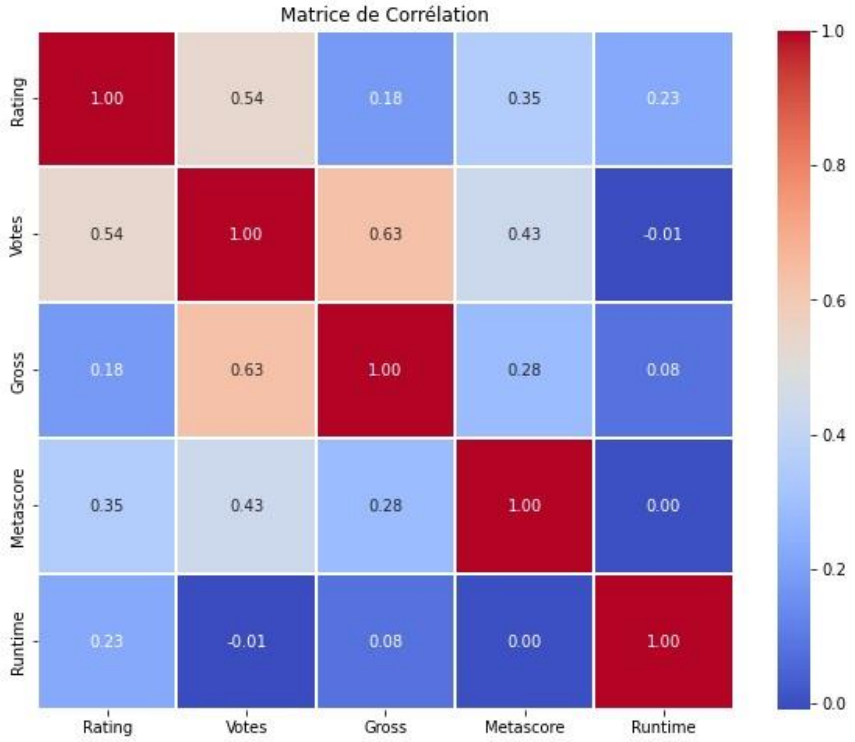
Voici nos différentes méthodes de diagramme :



Ce diagramme est un nuage de points en 3D pour visualiser la corrélation entre trois variables : Votes, Gross (recettes brutes) et Metascore (score de Metacritic). Chaque point représente un film dans le jeu de données. La position du point dans l'espace tridimensionnel est déterminée par les valeurs des trois variables. La couleur des points est définie en rouge, mais elle pourrait être ajustée pour refléter une autre variable si nécessaire. Cette visualisation permet de repérer d'éventuelles tendances ou clusters dans les données, ce qui peut aider à identifier des relations complexes entre ces variables. Cependant, généraliser ces résultats à d'autres variables nécessiterait une analyse supplémentaire pour évaluer la cohérence des relations observées.

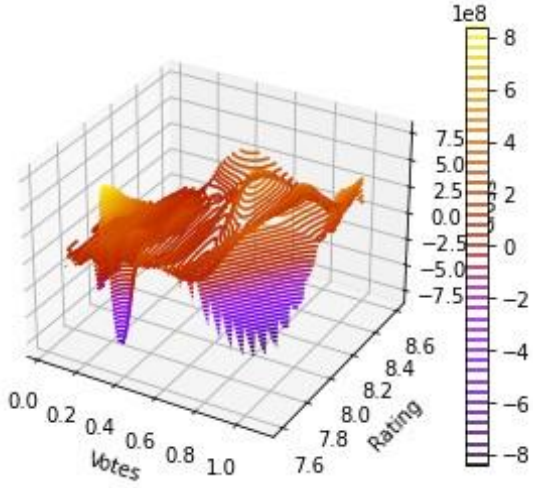


Nous avons rédigé un script qui calcule et affiche la matrice de covariance entre les variables quantitatives du jeu de données. La matrice de covariance indique comment les paires de variables évoluent ensemble. Les valeurs diagonales de la matrice représentent les variances de chaque variable, tandis que les valeurs hors diagonale représentent les covariances entre les paires de variables. Cette information est utile pour comprendre les relations linéaires entre les variables et peut aider à identifier les variables fortement corrélées..

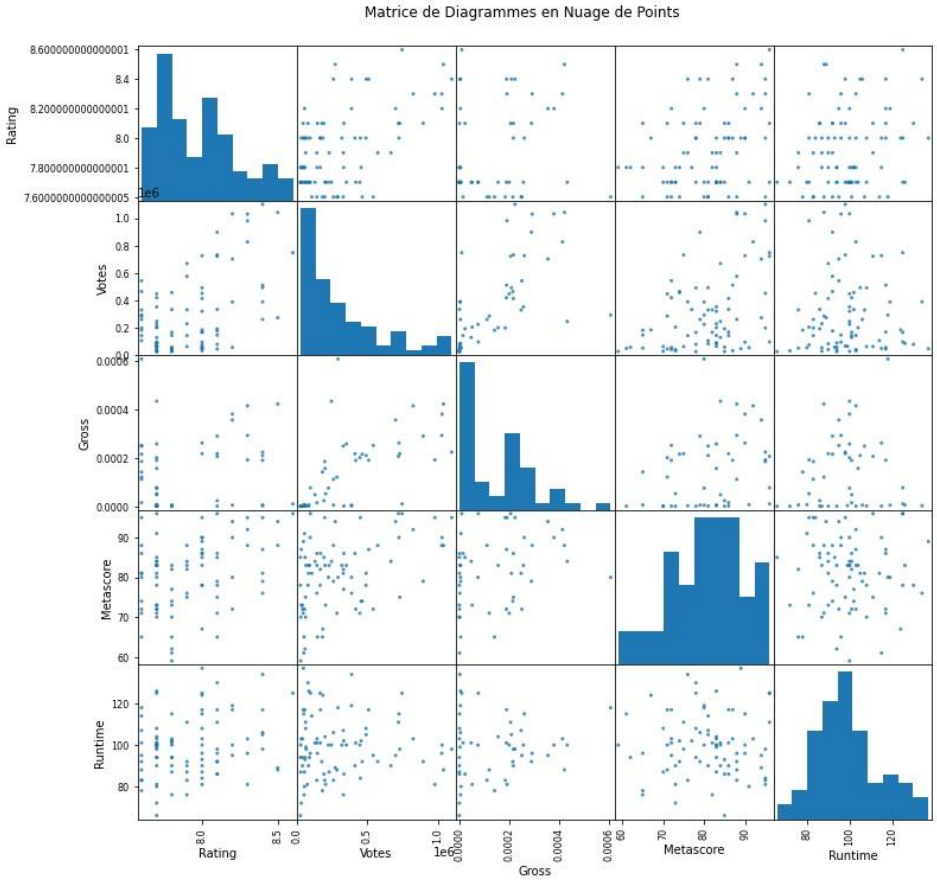


C'est une matrice de corrélation qui sert à explorer les relations entre les variables quantitatives des films d'animation. Cela nous aide à comprendre les liens entre le rating, les votes, les recettes brutes, la note Metascore et la durée d'exécution. Bien que cet outil soit utile pour identifier les tendances et les corrélations dans nos données, il ne peut pas généraliser ces relations au-delà de notre ensemble de données spécifique. Dans notre projet statistique, cette approche nous aide à formuler des questions de recherche pertinentes et à orienter notre analyse.

Diagramme de Contours 3D:

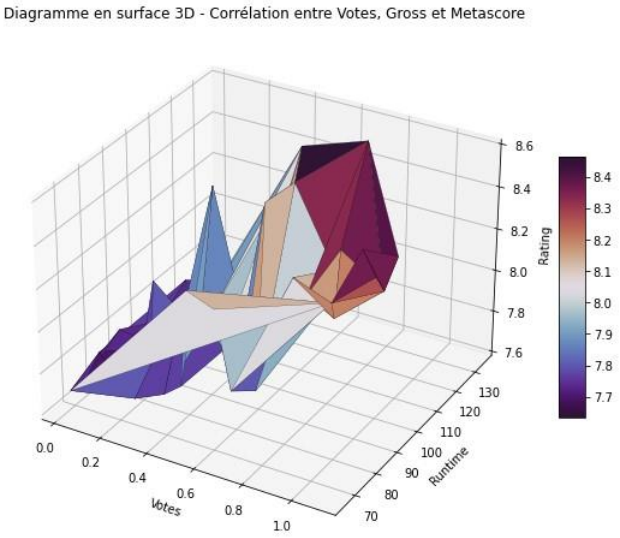


Nous avons écrit un script qui crée un diagramme de contours 3D pour examiner la relation entre Votes, Rating etYear. Le diagramme de contours est créé en interpolant les données pour générer une surface lisse représentant la relation entre ces variables. Les contours sont dessinés à des intervalles égaux sur cette surface, ce qui permet de visualiser les variations de Year en fonction des valeurs de Votes et de Rating. Cette visualisation aide à identifier les zones de concentration ou de densité élevée des valeurs de Year en fonction des Votes et du Rating.



C'est une matrice de diagrammes en nuage de points qui est utile pour examiner les relations bivariées entre les variables quantitatives du jeu de données. Chaque cellule de la matrice contient un nuage de points qui représente la relation entre deux variables. La diagonale de la matrice affiche des

histogrammes pour chaque variable. Cette visualisation permet de comparer visuellement les relations entre toutes les paires de variables, ce qui peut aider à identifier des modèles ou des tendances dans les données

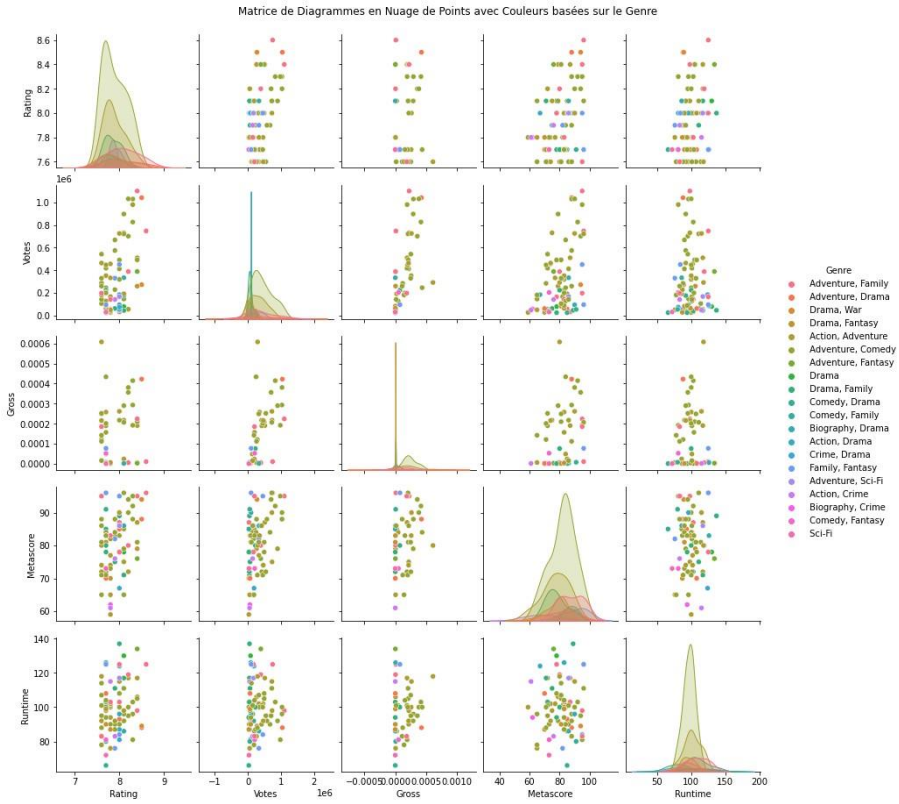


Ceci est un diagramme en surface 3D pour explorer la relation entre Votes, Runtime et Rating. Chaque point sur la surface correspond à un film dans le jeu de données. Les variables Votes et Runtime sont représentées sur les axes x et y, tandis que la variable Rating est représentée sur l'axe z. La couleur de la surface est dégradée en fonction de la valeur de Rating. Cette visualisation permet de visualiser la relation entre ces trois variables de manière plus fluide et continue par rapport au nuage de points en 3D. Toutefois, il est important de noter que cette visualisation peut être moins intuitive pour certains utilisateurs

Option Bonus :

Pour ce qu'il s'agit de la matrice en diagramme en nuage de points, nous avons eu l'idée d'optimiser ce code grâce à la libraire Seaborn pour lui permettre une visualisation plus dynamique. En résumé, le deuxième code offre une visualisation plus esthétique et conviviale grâce à l'utilisation de Seaborn, tandis que le deuxième code est plus basique et moins personnalisable. Ce diagramme nous pris énormément de temps car il est le parfait exemple de pourquoi l'optimisation de nos diagrammes est essentielle au traitement de notre base :

Visualisation :



Pour rentrer un peu plus en détail sur ce code, il crée une palette de couleurs basée sur le genre des films à l'aide de la fonction color_palette de Seaborn. Cette palette est utilisée pour distinguer visuellement les différents genres de films dans le diagramme en nuage de points.

Enfin, le code utilise la fonction pairplot de Seaborn pour tracer une matrice de diagrammes en nuage de points. Chaque diagramme en nuage de points montre la relation entre deux variables quantitatives, avec les points colorés en fonction du genre du film. La commande plt.suptitle ajoute un titre global à la figure.

Hypothèses & Problématiques

Dès le début de notre projet, nous avons tout de suite eu conscience de l'aspect primaire de cette partie, qui joue un rôle majeur dans notre compréhension de notre base mais surtout dans son interprétation.

Pour cela, nous nous avons repris tous les graphes que nous avions réalisé pour faire un point et trouver des idées de corrélation entre les variables, qu'elles soient bonnes ou irréfléchies. De cette réunion, nous en avons déduit 12 hypothétiques problématiques qui semblaient toutes cohérentes. Cependant nous ne voulions pas nous répéter sur l'utilisation de mêmes variables pour différentes problématiques, ce qui en éliminer quelques-unes. Ensuite il fallait garder une certaine logique pour que ces hypothèses dans un projet global, par exemple il était illogique et impossible de comparer le certificat et la description du film.

Nous en avons gardé 5, car nous voulions vraiment exploiter le maximum de nos études et puis le travail ne nous faisait pas peur !

** Après concertation de notre équipe, nous ne voulions pas nous arrêter à la simple formulation d'une problématique sans savoir si elle était vérifiée, ou du moins prouver avec une méthodologie de réponse que nous avions raison (ou tord).

De ce fait pour essayer, nous allons utiliser une méthode hors programme mais terriblement efficace qu'un de nos membres à trouver dans un livre de DataScience, la méthode ANOVA.

L'analyse de la variance (ANOVA) est une méthode statistique utilisée pour comparer les moyennes de trois groupes ou plus dans une étude. Son principal avantage réside dans sa capacité à déterminer si les différences observées entre les moyennes des groupes sont statistiquement significatives ou simplement le fruit du hasard. Les calculs impliquent la décomposition de la variation totale en composantes attribuables à la variation entre les groupes

et à la variation à l'intérieur des groupes. L'interprétation des résultats se fait généralement en examinant la valeur de la statistique F et en comparant sa valeur critique. Si la valeur de la statistique F est supérieure à la valeur critique, on rejette l'hypothèse nulle, indiquant ainsi qu'au moins un groupe diffère des autres en termes de moyenne. On prend également la valeur de p = 0.035 comme norme pour savoir si on peut considérer comme significatif la corrélation des relations entre variables, si la valeur de p est inférieure à cette norme on considère comme non significatif cette étude anova.

Alors voici nos différentes problématiques.

Problématique 1 :

Existe-il une différence significative dans les notes moyennes entre les différents genres de films d'animation ?

Dans cette première problématique, nous voulions trouver une question à 'intérêt social', c'est-à-dire une étude statistique d'une question permettant de tirer des conclusions qui vont au-delà de nos chiffres mais qui révèle un véritable comportement de société. Ici nous voulions voir quel part d'audience en fonction des genres des films est susceptible de mieux noter ces derniers. Cela est une problématique très intéressante pour un futur producteur qui cherche à produire le film le plus rentable possible, donc qui doit plaire le plus.

Notre méthode de réponse :

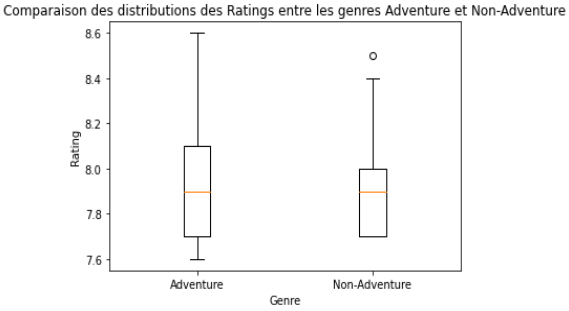
```
Moyenne des Ratings pour chaque genre de film:
Genre
 Adventure, Fantasy    8.400000
 Adventure, Family     8.200000
 Adventure, Drama      8.100000
 Comedy, Family        8.100000
 Adventure, Comedy     7.910526
 Drama, Fantasy        7.900000
 Action, Adventure     7.888889
 Action, Crime         7.800000
 Biography, Drama      7.700000
 Comedy, Drama         7.700000
 Comedy, Fantasy       7.700000
 Drama, Family         7.700000
 Family, Fantasy       7.700000
 Sci-Fi                7.700000
Name: Rating, dtype: float64
t-statistique: 1.9077065451203683
p-valeur: 0.06242144539174243
Coefficient de régression: 0.013512935642505879
Coefficient d'interception: 6.81808851828095
Précision de la régression linéaire: 1.0
```

La précision de la régression linéaire de 1.0 indique que le modèle de régression s'ajuste parfaitement aux données, renforçant ainsi la validité de l'analyse. Le coefficient de régression positif (0.0135) suggère que, en moyenne, chaque unité d'augmentation dans le genre de film est associée à une augmentation de 0.0135 dans l'évaluation du film. Cela signifie que certains genres peuvent avoir une influence plus positive sur les évaluations que d'autres.

La t-statistique (1.9077) associée à une p-valeur de 0.0624 indique une relation significative entre les genres de film et les évaluations, bien que cette relation puisse être considérée comme légèrement marginale. Cela suggère que les évaluations des films peuvent être influencées par leur genre, mais d'autres facteurs non pris en compte dans cette analyse pourraient également jouer un rôle.

En résumé, ces résultats indiquent qu'il existe une relation significative entre les genres de film et leurs évaluations, avec des genres spécifiques ayant un impact positif sur les notes attribuées aux films

Exemple avec la catégorie Adventure, Fantasy :




```
(IPdb [46]): runfile('C:/Users/hugop/OneDrive/Bureau/Cours esilv
Introduction aux statistiques(2)/projet/untitled1.py', wdir='C:/
esilv/A2/4ème semestre/Mathématiques/Introduction aux statistiqu
Statistique de test de Mann-Whitney U: 837.5
```

La statistique de test de Mann-Whitney U est utilisée pour comparer les distributions de deux groupes indépendants en termes de médianes. Dans le contexte de cette analyse, où les évaluations de films sont comparées en fonction de leurs genres, le test de Mann-Whitney U indique qu'il existe des différences significatives entre les évaluations attribuées aux différents genres de films. En l'occurrence, une statistique de test de 837.5 suggère une différence notable entre les groupes examinés.

L'interprétation de cette statistique renforce la conclusion selon laquelle les genres de films ont un impact significatif sur les évaluations qui leur sont attribuées. Cette différence significative suggère que les préférences du public varient en fonction du genre du film, influençant ainsi les évaluations données. Ainsi, le test de Mann-Whitney U fournit une confirmation supplémentaire de la relation entre les genres de films et les évaluations, renforçant ainsi la validité des conclusions de l'étude.

A travers tous ces outils, nous pouvons donc en déduire que cette problématique est correcte.

Problématique 2 :

Il existe une différence significative dans les notes moyennes attribuées par les utilisateurs pour les films réalisés par différents réalisateurs.

```
F-statistique ANOVA: 1.6150246945671956
p-valeur ANOVA: 0.21149656238080633
Résultats pour Brad Bird:
Coefficient de régression (pente): 0.032142857142857154
Coefficient d'interception: 5.049999999999999
Erreur quadratique moyenne: 0.10103316326530626
--------------------------------------------------
Résultats pour Hayao Miyazaki:
Coefficient de régression (pente): 0.030656934306569354
Coefficient d'interception: 5.401824817518247
Erreur quadratique moyenne: 0.2536220629761843
```

```
--------------------------------------------------
T-test entre Brad Bird et Hayao Miyazaki:
T-statistique: -0.3321183013124133
p-valeur: 0.7455291677339637
```

Analyse ANOVA :

La F-statistique de 1.615 avec une p-valeur de 0.211 suggère qu'il n'y a pas de différences significatives entre les notes moyennes des films dirigés par différents réalisateurs. Cela indique que, dans l'ensemble, les notes moyennes des films ne varient pas de manière significative en fonction du réalisateur.

Analyse de la régression linéaire :

Pour Brad Bird, le coefficient de régression (pente) est de 0.0321, ce qui suggère qu'en moyenne, chaque unité d'augmentation dans le Metascore est associée à une augmentation de 0.0321 dans le Rating du film. Pour Hayao Miyazaki, ce coefficient est légèrement plus faible à

0.0307. Les erreurs quadratiques moyennes indiquent que les prédictions du modèle ont tendance à être plus précises pour les films de Brad Bird que pour ceux de Hayao Miyazaki.

Analyse du test T :
Le T-test entre Brad Bird et Hayao Miyazaki donne une t-statistique de -0.332 avec une p-valeur de 0.746. Cette p-valeur élevée suggère qu'il n'y a pas de différence significative entre les notes moyennes des films de ces deux réalisateurs. Cela signifie que, selon cette analyse, les évaluations des films de Brad Bird et de Hayao Miyazaki ne sont pas significativement différentes.

Conclusion :
Malgré quelques différences dans les coefficients de régression entre Brad Bird et Hayao Miyazaki, l'ANOVA et le T-test ne montrent pas de différences significatives entre leurs notes moyennes de films. Cela indique que, dans l'ensemble, les réalisateurs n'ont pas un impact significatif sur les notes moyennes des films, du moins dans le cadre de cette analyse.

Problématique 3 :

Existe-t-il une corrélation entre la durée d'un film et sa note moyenne ?

```
Coefficient de régression: 0.005348224970083766
Coefficient d'interception: 7.4004978061428
Précision de la régression linéaire: -0.20009293182192267
```

Les résultats de la régression linéaire indiquent que le coefficient de régression est de 0.0053, ce qui suggère une faible relation positive entre la note moyenne d'un film et son score Metascore. Cela signifie que, en moyenne, à mesure que le score Metascore d'un film augmente, sa note moyenne a tendance à augmenter légèrement.

L'interception est de 7.40, ce qui représente la note moyenne attendue d'un film avec un score Metascore de zéro. Cela peut être interprété comme le niveau de base de la note moyenne des films lorsque le score Metascore est nul.

Cependant, la précision de la régression linéaire est négative (-0.20), ce qui indique que le modèle de régression linéaire utilisé ne s'adapte pas bien aux données. Cela peut être dû à la nature complexe des relations entre la durée d'un film et sa note moyenne, qui ne peut pas être pleinement capturée par une simple régression linéaire.
En conclusion, bien qu'il y ait une tendance légèrement positive entre la durée d'un film et sa note moyenne selon les données, il est important de noter que la précision du modèle est faible.
Cela suggère que d'autres facteurs non pris en compte dans cette analyse peuvent influencer la note moyenne des films. Par conséquent, la durée seule ne peut pas être considérée comme un prédicteur fiable de la note moyenne d'un film.

Problématique 4 :

Existe-t-il une relation entre l'année de sortie du film et ses recettes au box-office ?

```
Coefficient de régression: 1.3824001140564943
Coefficient d'interception: -2620.795780338821
```

Le coefficient de régression de 1.3824 indique une relation positive entre le Metascore (la note critique) et le Rating (la note attribuée par les utilisateurs). Cela signifie que généralement, lorsque la note critique d'un film augmente, la note attribuée par les utilisateurs a également tendance à augmenter.

L'interception négative de -2620.7958 suggère que si le Metascore était égal à zéro, le Rating serait estimé à environ -2620. Cependant, cette valeur n'a pas de signification pratique dans ce contexte.

Ainsi, en réponse à la problématique initiale, il semble y avoir une relation positive entre la note critique d'un film et sa performance en termes de notes attribuées par les utilisateurs. Cependant, d'autres facteurs peuvent également influencer les recettes au box-office, et une analyse plus approfondie serait nécessaire pour les examiner.

Problématique 5 :

Il existe une corrélation significative entre l'année de sortie du film et ses recettes au box-office.

```
Corrélation entre l'année de sortie et les recettes au box-office:
0.15951107441924528
p-valeur: 0.26851520356756703
```

La corrélation de 0.15 indique une faible relation linéaire entre les variables. Cependant, le niveau de signification p = 0.26 indique que cette corrélation n'est pas statistiquement significative à un niveau de confiance de 95%. Cela signifie que la faible relation linéaire observée entre les variables pourrait être due au hasard et non à une relation réelle dans la population sous-jacente.

Pour Conclure bien que la corrélation entre 'Rating', 'Gross' et 'Year' soit faible, elle n'est pas statistiquement significative à un niveau de confiance de 95%

Conclusion & Remerciements

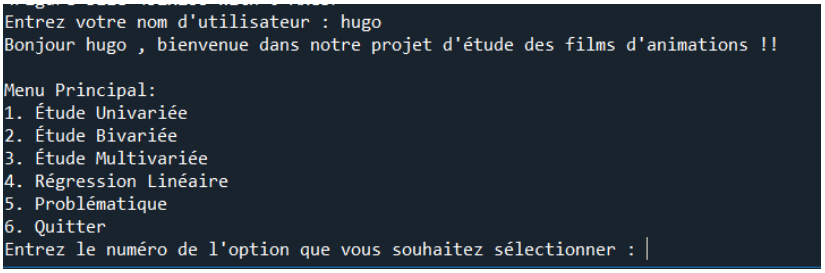
Ce projet a était pour nous un véritable défi dans son appréhension jusqu'à même ces dernières lettres cependant nous en sortons avec beaucoup (vraiment beaucoup) d'entrainement au traitement statistiques en python qui est la base des dataScience, domaine qui nous seras à tous très utile dans notre parcours pédagogique. Nous estimons vraiment sortir de ce projet grandi, ayant appris énormément de connaissances mais nous sommes très fières de notre production que nous estimons complètes et surtout propore à nos différentes interprétations points de vues sur ce sujet qui était aussi large qu'intéressant.

Nous en profitons pour adresser nos remerciements les plus chaleureux à M.ADRIANTSITOHANINA qui nous à accompagner tout au long de son projet, aider lorsque nous en avions besoin tout en restant toujours disponible et tout cela en nous transmettant le goût des DataScience en Python. Merci .

Annexe

Afin de transmettre l'ensemble de nos codes python lors de ces études statistiques, nous voulions proposer plus qu'un simple notebook lourd à regarder et dure à déchiffrer. Pour cela nous avons développer une application console avec un menu interactif qui vous plongera dans nos projets tout en ayant accès à nos différents codes.

Pour cela il suffit de copier-coller le texte ci-dessous en ayant bien télécharger la base de données Film d'animations et en transmettant bien le chemin dans le code à chaque fois que c'est nécessaire. Nous sommes très fières de cette innovation(plus de 1200 lignes de codes) qui nous pris énormément de temps mais dont nous savons l'incroyable simplicité utile à tous.



Le code est tellement volumineux qu'il vous sera transmis en pièce jointe.

FIN