Projet data

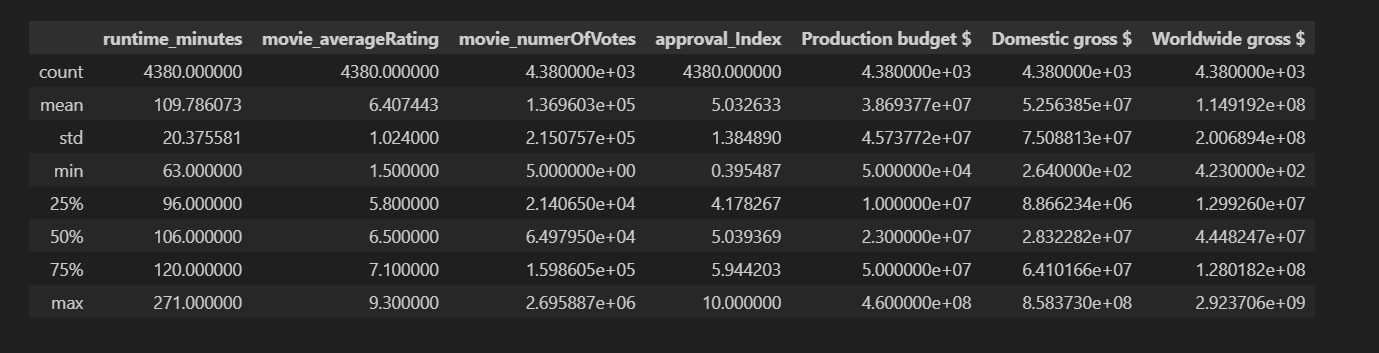
1. Analyse descriptive : (df.describe())

** runtime\_minutes (durée des films) :**

* Moyenne autour de **110 minutes**, avec un minimum de **63 min** et un maximum de **271 min**.
* L’écart-type de **20 minutes** indique une **variabilité modérée** autour de la moyenne, et les quartiles montrent une **distribution assez homogène** (la moitié des films ont une durée entre 96 et 120 minutes).
* Quelques films particulièrement longs (>180 min) peuvent être considérés comme **valeurs extrêmes**.

** movie\_averageRating (note moyenne) :**

* Moyenne de **6.4/10**, ce qui est typique pour des bases de données publiques comme IMDb.
* La note varie de **1.5 à 9.3**, avec une **dispersion relativement faible** (écart-type ≈ 1), ce qui montre une **certaine homogénéité dans la notation**.
* **movie\_numberOfVotes** (nombre de votes) :
  + Très grande dispersion : de **5 votes à plus de 2.6 millions**.
  + L’écart-type élevé et la forte différence entre la médiane (~64k) et la moyenne (~137k) traduisent une **asymétrie vers la droite**, causée par quelques films extrêmement populaires.
* **approval\_Index** :
  + Moyenne de **5.03**, avec une plage de 0.39 à 10.
  + Relativement symétrique autour de la moyenne, l'écart-type modéré (~1.38) indique que la **majorité des films sont proches de l’indice moyen**.
* **Production budget $** :
  + Grande hétérogénéité, allant de **50 000 $ à 460 millions $**, avec une moyenne autour de **38 millions $**.
  + La médiane (23M) étant bien inférieure à la moyenne, cela révèle une **distribution asymétrique vers les très gros budgets**.
* **Domestic gross $** et **Worldwide gross $** :
  + Très grandes variations, avec des **revenus mondiaux extrêmes** allant jusqu’à **2.9 milliards $**.
  + Les écarts-type élevés montrent que ces variables sont **très dispersées**, dominées par une minorité de blockbusters à très gros revenus.
  + La forte différence entre la moyenne et la médiane indique aussi une **asymétrie importante**.

2-distributions :

**Runtime (en minutes)**

* **Tendance centrale** : La durée moyenne d’un film est de **109,79 minutes**, avec une médiane de **106 minutes**, ce qui suggère une distribution légèrement asymétrique vers la droite.
* **Dispersion** : L’écart-type est modéré (**20,38**), indiquant une certaine variabilité autour de la moyenne.
* **Forme de la distribution** : La distribution est **asymétrique positive** (skewness = **1.59**) avec une **kurtosis de 5.04**, ce qui signifie qu’elle est plus concentrée autour de la moyenne mais avec une queue plus longue à droite.
* **Valeurs extrêmes** : **132 valeurs aberrantes** ont été détectées, ce qui peut correspondre à des films particulièrement longs ou courts.

**Production budget ($)**

* **Tendance centrale** : Le budget moyen est de **38,69 millions $**, avec une médiane plus basse (**23 millions $**), ce qui confirme une **asymétrie positive** importante.
* **Dispersion** : L’écart-type est très élevé (**45,7 millions $**), montrant une grande hétérogénéité dans les budgets.
* **Forme de la distribution** : Skewness = **2.44** et kurtosis = **8.17**, indiquent une forte concentration de films à petits/moyens budgets, avec quelques très gros budgets extrêmes.
* **Valeurs extrêmes** : **315 outliers** présents, représentant les grosses productions hollywoodiennes à très fort budget.

**Domestic gross ($)**

* **Tendance centrale** : Le revenu moyen domestique est de **52,56 millions $**, médiane à **28,32 millions $**, ce qui reflète encore une **asymétrie à droite** marquée.
* **Dispersion** : Écart-type très élevé (**75 millions $**), indiquant des revenus très disparates.
* **Forme de la distribution** : Skewness = **3.69** et kurtosis = **21.10**, ce qui traduit une distribution très étalée, dominée par quelques succès majeurs au box-office.
* **Outliers** : **357 valeurs aberrantes**, représentant des blockbusters ou des succès commerciaux notables.

**Worldwide gross ($)**

* **Tendance centrale** : Revenu mondial moyen de **114,9 millions $**, avec une médiane de **44,48 millions $**, très influencée par les **valeurs extrêmes**.
* **Dispersion** : Extrêmement dispersée (écart-type de **200 millions $**), preuve d’une très grande diversité dans les performances commerciales.
* **Forme de la distribution** : Très fortement asymétrique à droite (**skewness = 4.60**, **kurtosis = 34.34**) avec **433 outliers**, indiquant que seule une minorité de films génère des revenus mondiaux exceptionnels.

**movie\_averageRating**

* **Tendance centrale** : La note moyenne est de **6.41**, médiane légèrement supérieure à **6.5**, ce qui indique une **légère asymétrie négative** (skewness = **-0.66**).
* **Dispersion** : Relativement faible (**écart-type de 1.02**), donc les notes sont assez concentrées.
* **Forme de la distribution** : Distribution modérément aplatie (**kurtosis = 1.11**), et **peu d’outliers** (**82**), ce qui suggère une répartition équilibrée autour de la moyenne.