Data Mining Application Play Store

Canete Valentin, Perrin Thibaud, Tisserand Jean-philippe, Bruyat Julian, Dailler Arnaud

**Introduction :**

Le but de ce rapport est d’étudier les caractéristiques des applications du Play Store pour connaître les types d’applications qui réussissent le mieux sur le Play Store et de faire une estimation des revenues des applications de ces différents types pour comparer leurs rendements. Les résultats seront vérifiés par une méthode de clustering (K-mean/ DBScan), L’étude sera donc portée sur les corrélations entre les différentes caractéristiques d’une application. L’utilisation d’une base de données de commentaire de certaines applications permettra de comparer l’avis d’une personne sur une application et sa réussite sur le Play store.

Mots clés: Data mining, Clustering, Play-Store, Correlation

**Données :**

Les données sont récupérées du fichier Google-Play Store-Full.csv. Ce fichier contient 267 053 applications et leurs caractéristiques (App Name : nom de l’application, Catégories, Reviews : nombres de notes, Installs : nombre de téléchargement, Size : taille de l’application en Mo, Price : prix en dollars, Content Rating : public visé, Last Updated : date dernière Maj, Minimum version : version minimum, Latest version : dernière version) . Nous avons décidé de ne pas utiliser les versions pour l’analyse.

Traitement des données : Avant de commencer toute analyse il a fallu mettre en place une phase de traitement des données : Rassembler certaines catégories, Convertir type de donnée, Supprimer les missing value (workflow 1 : Annotation “Traitement des données”)

Beaucoup d’applications portent la mention “Varies with device”. Pour gagner un nombre d’applications suffisantes, nous avons décidés de ne pas convertir cette colonne. A la fin du traitement des données il reste 267 051 Applications.

Pour la répartition du nombre d’applications par catégories on obtient

Une image contenant capture d’écran

Description générée automatiquement

Nous avons rassemblé tous les types de jeu au sein d’une même catégorie “GAME”, le détail de la répartition des applications par type de jeu est le suivant :

Une image contenant capture d’écran

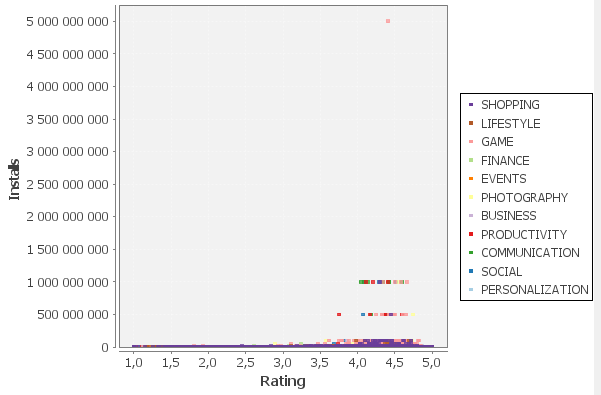
Description générée automatiquement

**Analyse des caractéristiques :**

Le nombres de téléchargements et la note d’une application sont deux paramètres importants pour indiquer la réussite sur l’App Store, dans un premier temps nous avons cherché des corrélations entre les paramètres souhaités (nombre de téléchargement) et les paramètres influent (note , taille, note, prix, public visé, catégorie) afin d’extraire les paramètres qui peuvent faire d’une application, une bonne application.

Nous présentons un scatter plot (coloré) pour montrer la répartition des applications en fonction de leurs groupes :

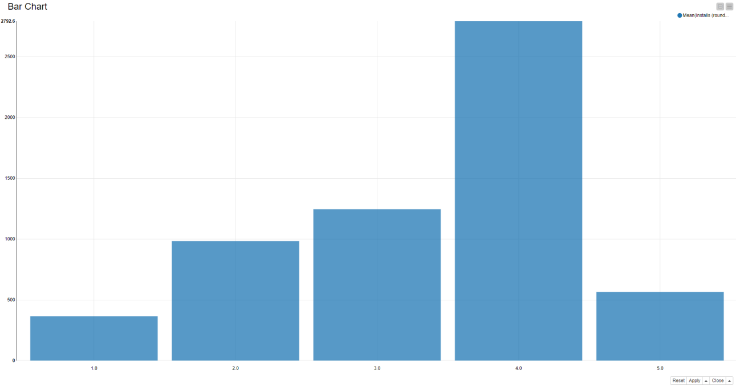
**Etude des catégories sur le nombre de téléchargements :**



Comme on peut le voir, les catégories n’ont pas réellement d’influence sur les paramètres de réussite d’une application dans le Play store, cependant on peut commencer à voir certaines tendance (jeu plus souvent téléchargé, etc..) mais on peut également voir qu’une application qui n’a pas une note inférieure à 3 ne dépassera pas les 100 000 000 de téléchargements.

**Etude de la note sur le nombre de téléchargement**

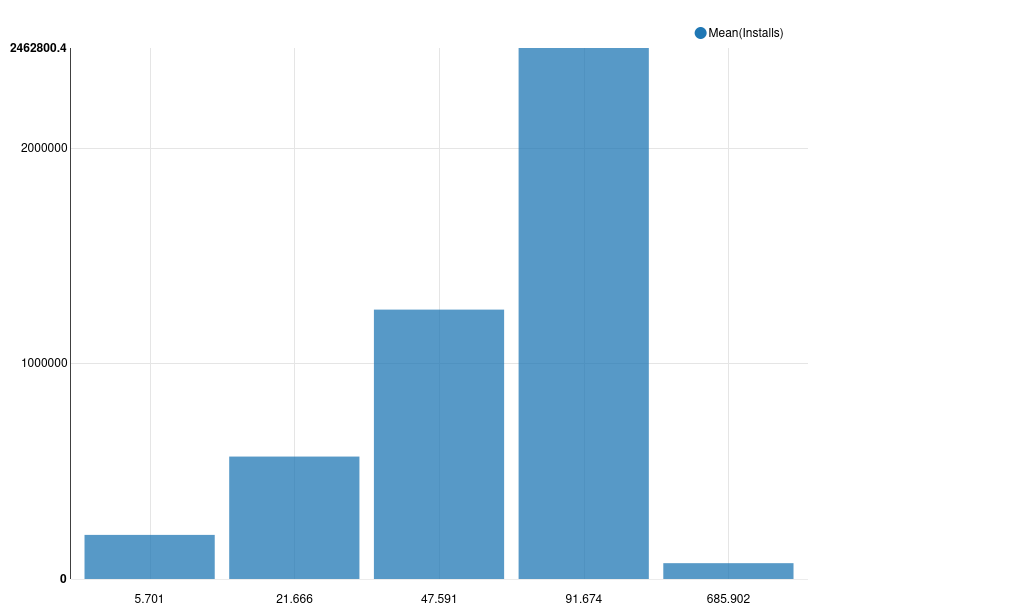
La notes et le nombre de téléchargement sont donc deux caractéristiques fortement corrélées



Il est donc conseillé de garder une note de l’application supérieure à 3 pour atteindre un nombre de téléchargement suffisamment élevé

**Etude du nombre de téléchargement selon le poids de l’application :**

Le poids est un facteur qui peut paraître important si on considère le fait qu’un utilisateur peut être réticent à l’idée d’installer une application lourde sur son téléphone. C’est donc un paramètre qui doit être étudié, pour cela on rassemble les poids par clusters et on regarde la moyenne des téléchargements par cluster



On voici ici que pour obtenir un nombre d’installation maximum, il faut veiller à créer des jeux aux alentours 90 Mo par application, une taille trop petite ou trop grande fera diminuer le nombre de téléchargement

**Etude du nombre de téléchargement selon le public visé :**

Le public visé est un facteur qui va forcément influencer le nombre de téléchargement, il existe 5 catégories de public visé, on observe la moyenne de téléchargement

Une image contenant texte

Description générée automatiquement

Selon le graphique obtenu, la catégorie 10 ans et + est la plus téléchargée. Et viser un public supérieur à 18 ans empêche un grand nombre de personne de télécharger l’application et donc baisse le nombre de téléchargement.

**Etude du nombre de téléchargement selon le prix de l’application :**

Pour le prix nous avons également regroupe par cluster pour faire des intervalles de prix et afficher le nombre de téléchargement :

Une image contenant capture d’écran

Description générée automatiquement

Comme on pouvait l’espérer, moins l’application est chère, plus le nombre de téléchargement est élevé.

Si on fait une conclusion sur les observations précédentes, une application type qui engendrerait un maximum de téléchargement serait une application de n’importe quelle catégorie, gratuite, avec un poids inférieur à 100 Mo, qui s’adresse à un public de 10 ans et plus. Il est important de conserver une note moyenne supérieur à 3.

**Recherche de corrélation entre le nombre de téléchargement et les autres colonnes :**

Afin de comprendre ce qui peut entraîner un fort nombre d’installation d’une application nous avons cherché les corrélations avec différents paramètres. Nous avons procédé en plusieurs étapes, la première chose que nous avons fait sur l’ensemble de notre jeu de données, a été de rassembler les différentes catégories des jeux (GAME\_ACTION, GAME\_CASUAL, …) sous la catégorie GAME, afin que les résultats ne soient pas déséquilibré par rapport aux autres catégories.

Ensuite nous clustérisé à l’aide de K-Means nos données en fonction du nombre d’applications. Ainsi nous obtenons 4 clusters en fonction du nombre d’installations, dans un premier temps nous avons décidé de travailler sur le cluster comportant le plus de téléchargement, car celui-ci va contenir les applications les plus populaires tel que Netflix, Facebook, TripAdvisor…

Une fois ce pré-processing fait nous avons cherché à observer des corrélations entre le nombre d’installation de ces applications et les différents autres paramètres qui nous étaient accessibles.

Nous avons d’abord regardé le nombre d’installation par catégories, par notes et par type de contenu.

*Remarque : Nous n’avons pas regardé le nombre d’installation par prix car toutes les applications de ce cluster étaient gratuites.*

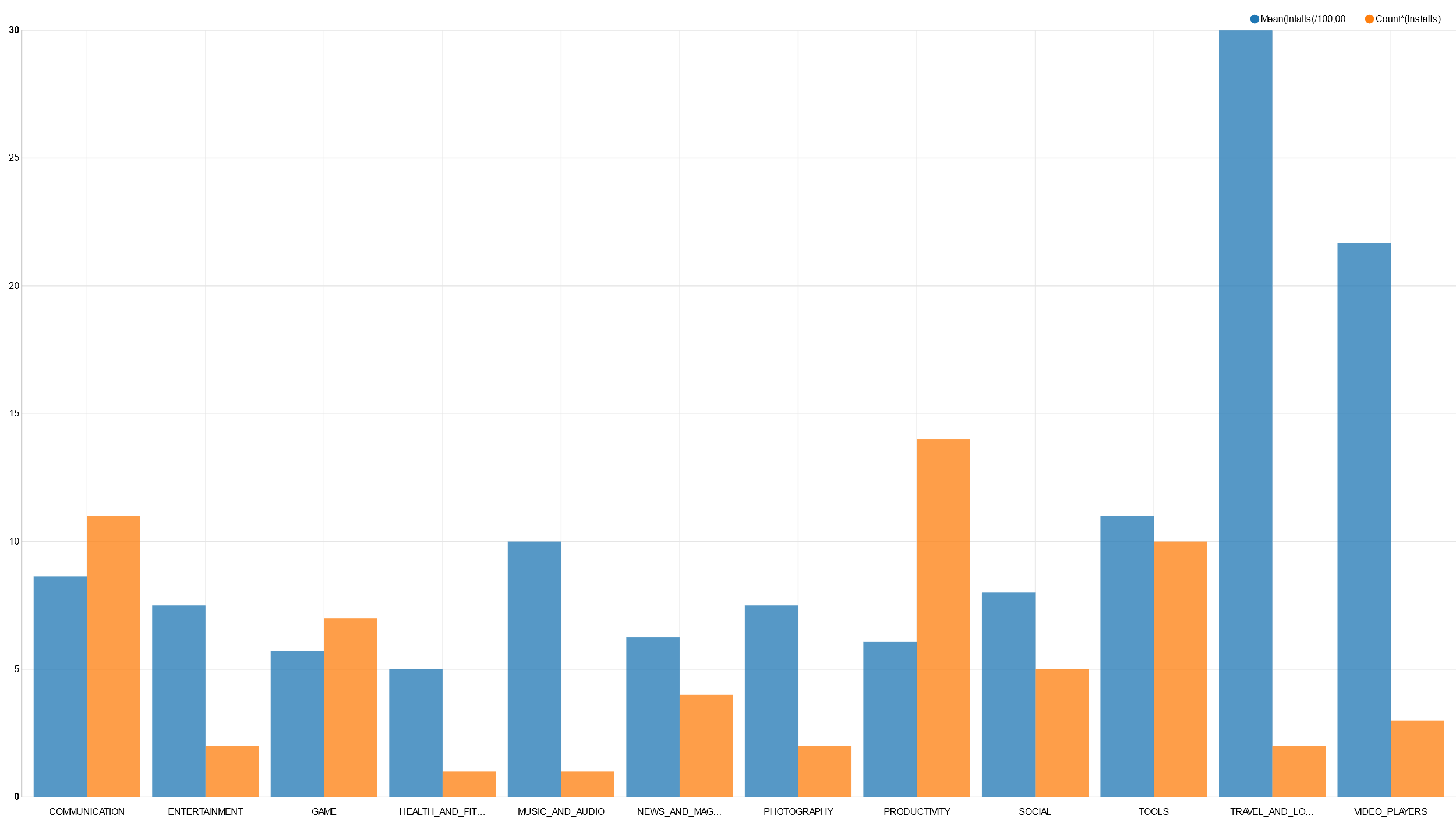


Figure : Nombre d'applications (bleu) et moyenne du nombre d'installations /100,000,000 (orange) par catégories.

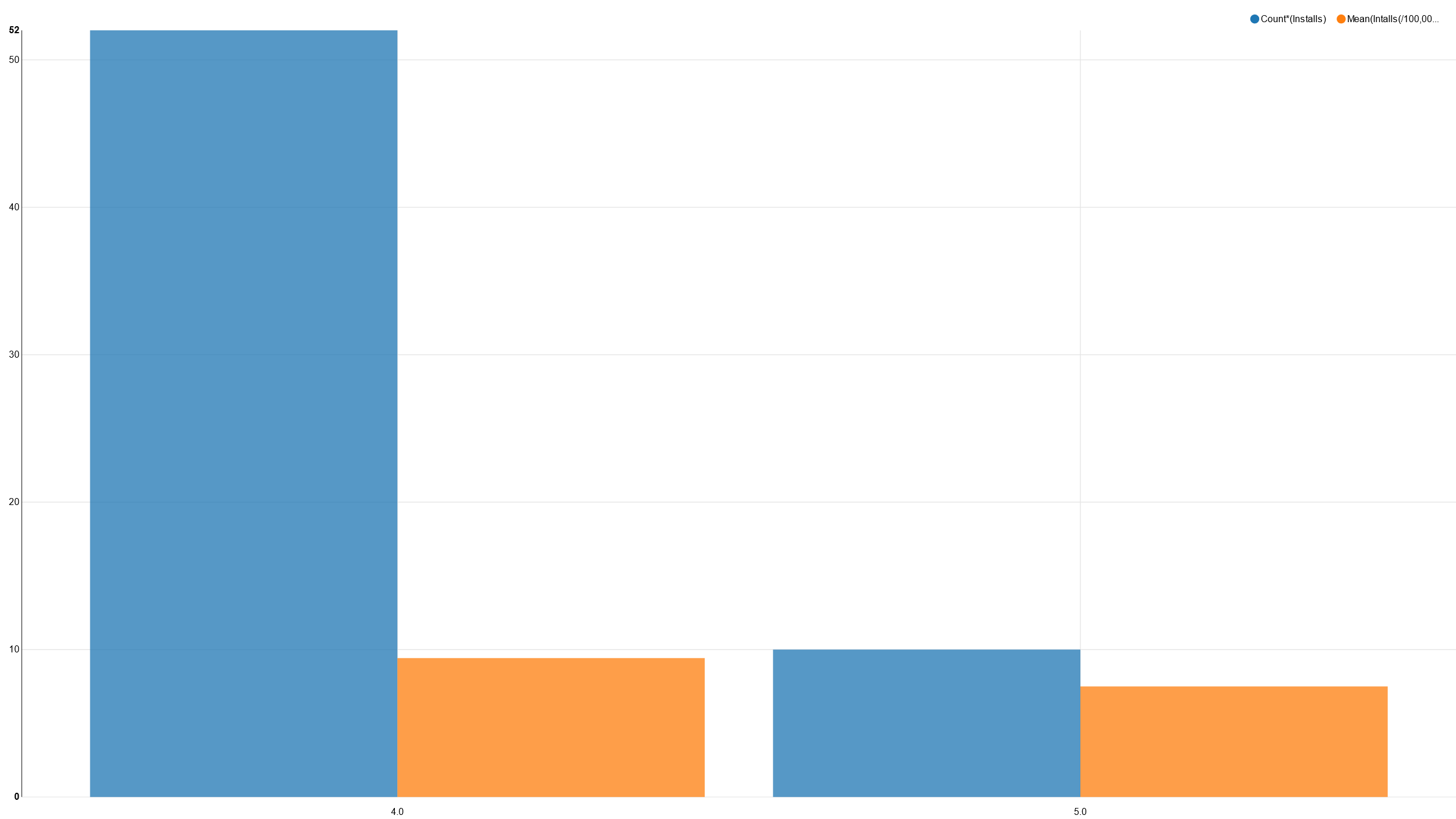


Figure : Nombre d'applications (bleu) et moyenne du nombre d'installations /100,000,000 (orange) par notes.

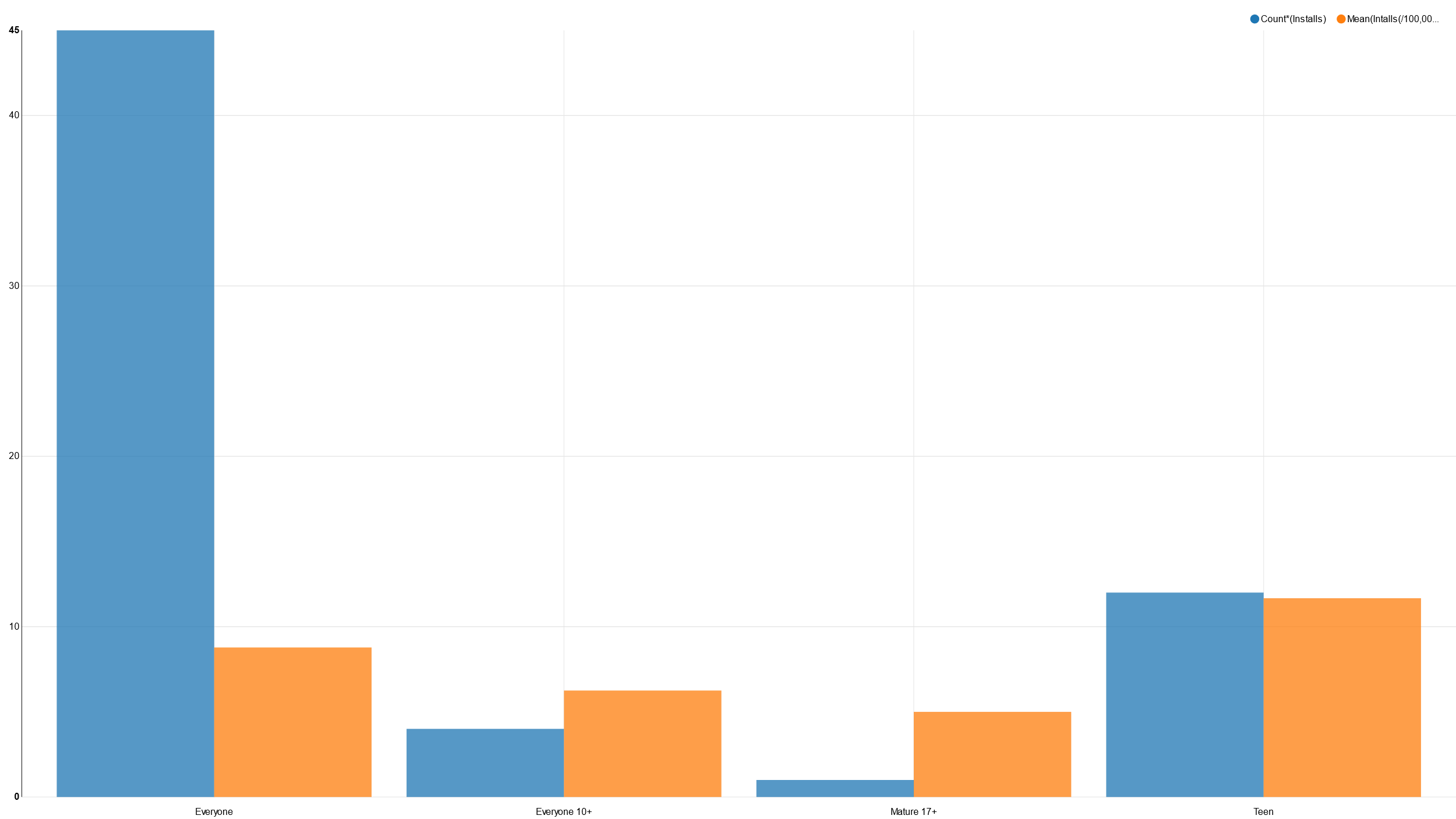


Figure : Nombre d'applications (bleu) et moyenne du nombre d'installations /100,000,000 (orange) par type de contenu.

A l’aide de ces trois graphiques on observe que les applications les plus téléchargées :

* Parmi les catégories, sont de type *TRAVEL\_AND\_LOCAL*.
* Parmi les notes, elles obtiennent en moyenne plus souvent 4/5 que 5/5.
* Parmi le type de contenu, *Teen* est légèrement majoritaire.

Une fois que nous avons déterminé séparément ces paramètres nous avons souhaité voir si le nombre d’installations étaient corrélé aux notes.

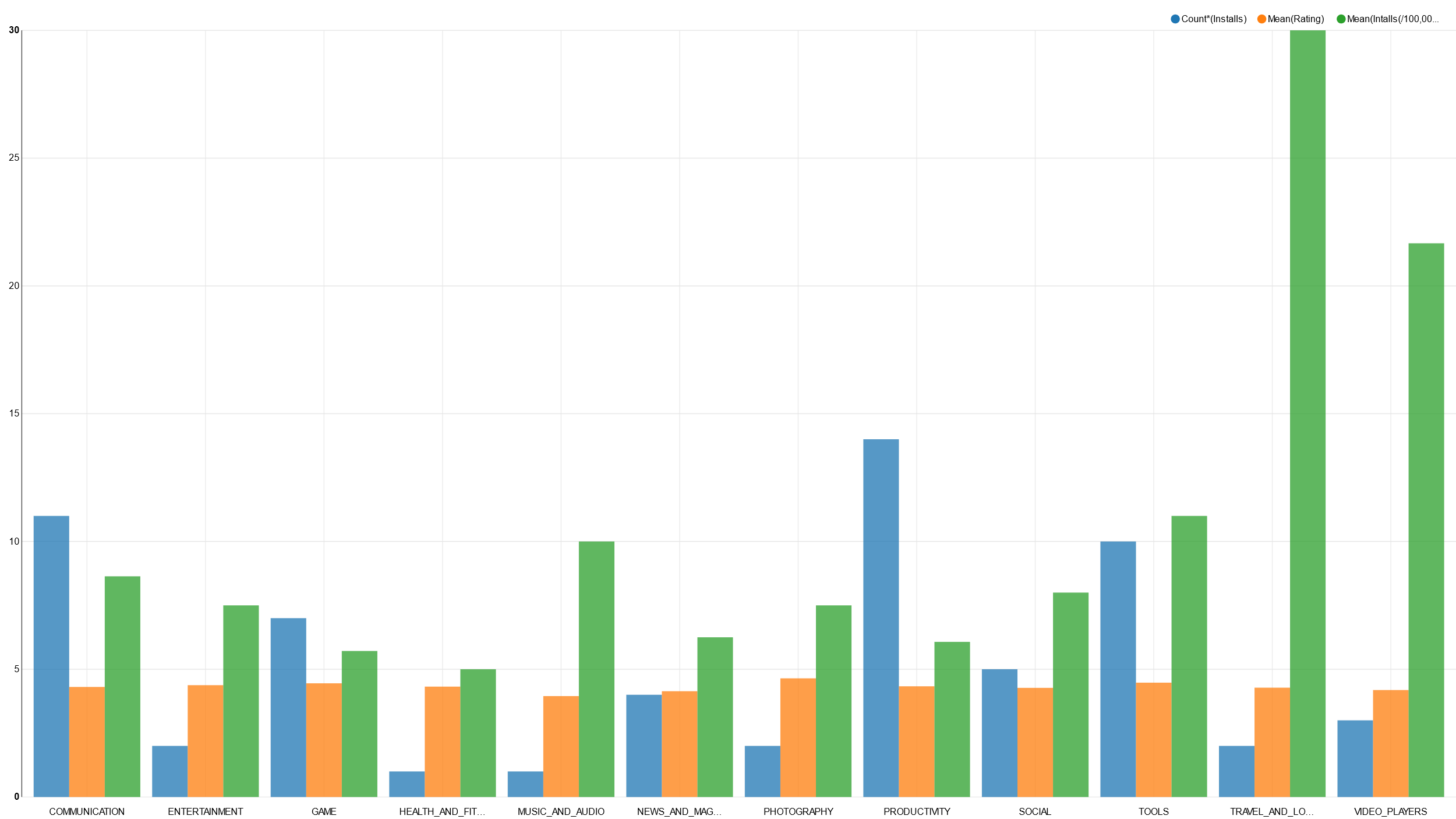


Figure : Nombre d'applications (bleu) moyenne de la note (orange) et moyenne du nombre d'installations /100,000,000 (vert) par catégories.

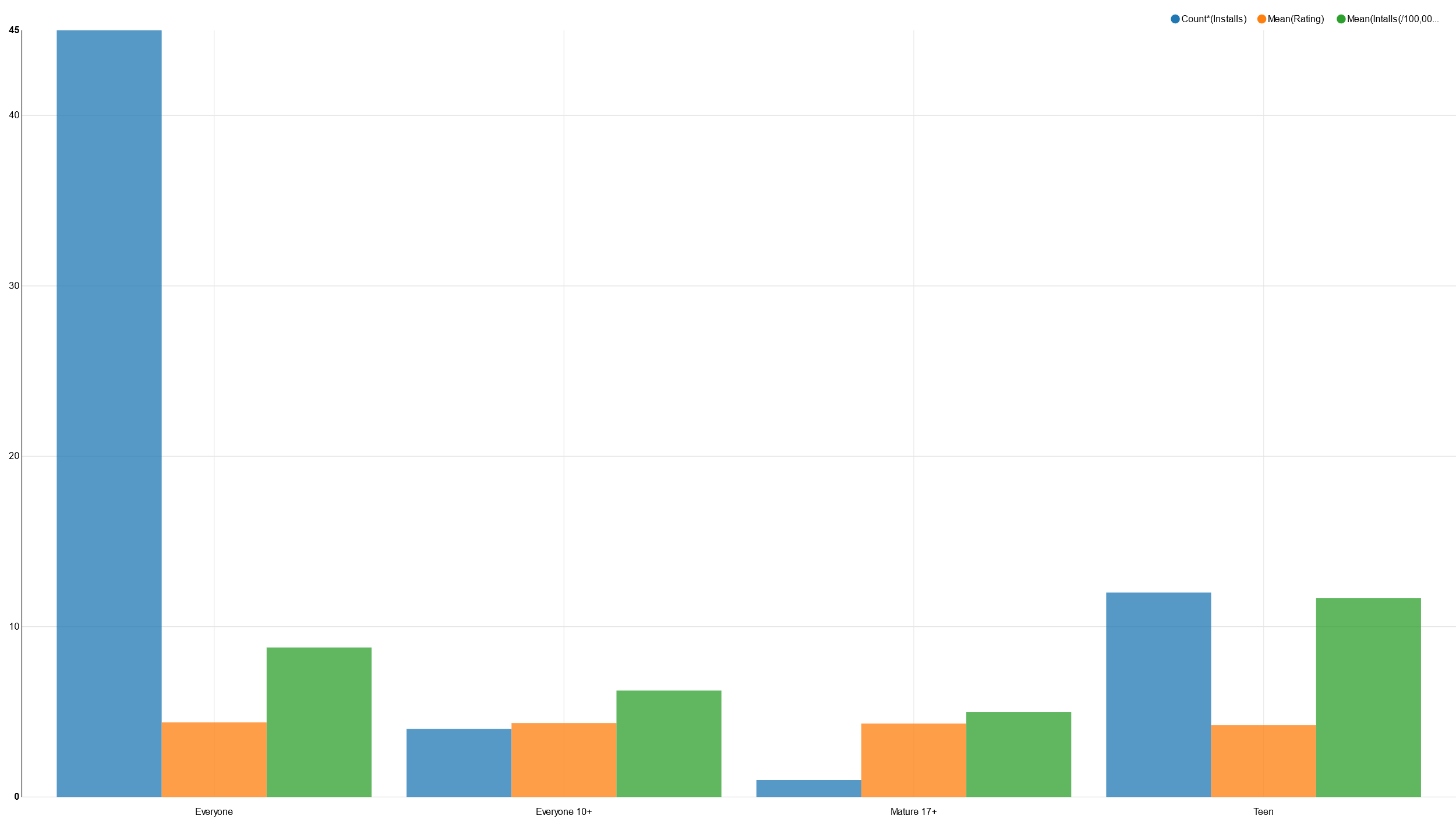


Figure : Nombre d'applications (bleu) moyenne de la note (orange) et moyenne du nombre d'installations /100,000,000 (vert) par type de contenu.

On observe que les note sont relativement similaire et donc que peu importe la catégorie ou le type de contenu et peu importe le nombre de téléchargement, en moyenne la note reste proche de 4/5.

On à fait de même mais cette fois ci avec le nombre de commentaires.

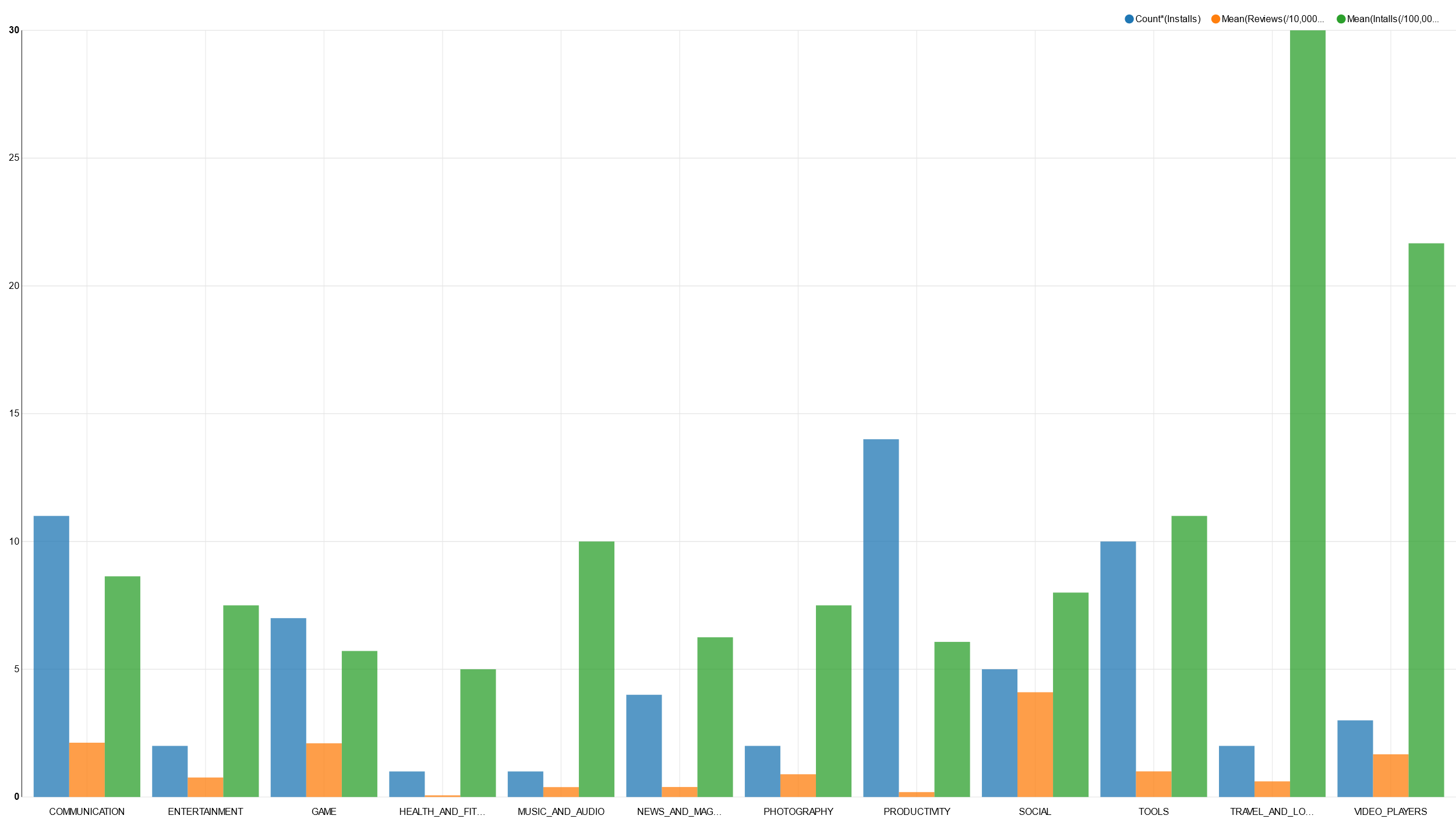


Figure : Nombre d'applications (bleu) le nombre de review /10,000,000 (orange) et moyenne du nombre d'installations /100,000,000 (vert) par catégories.

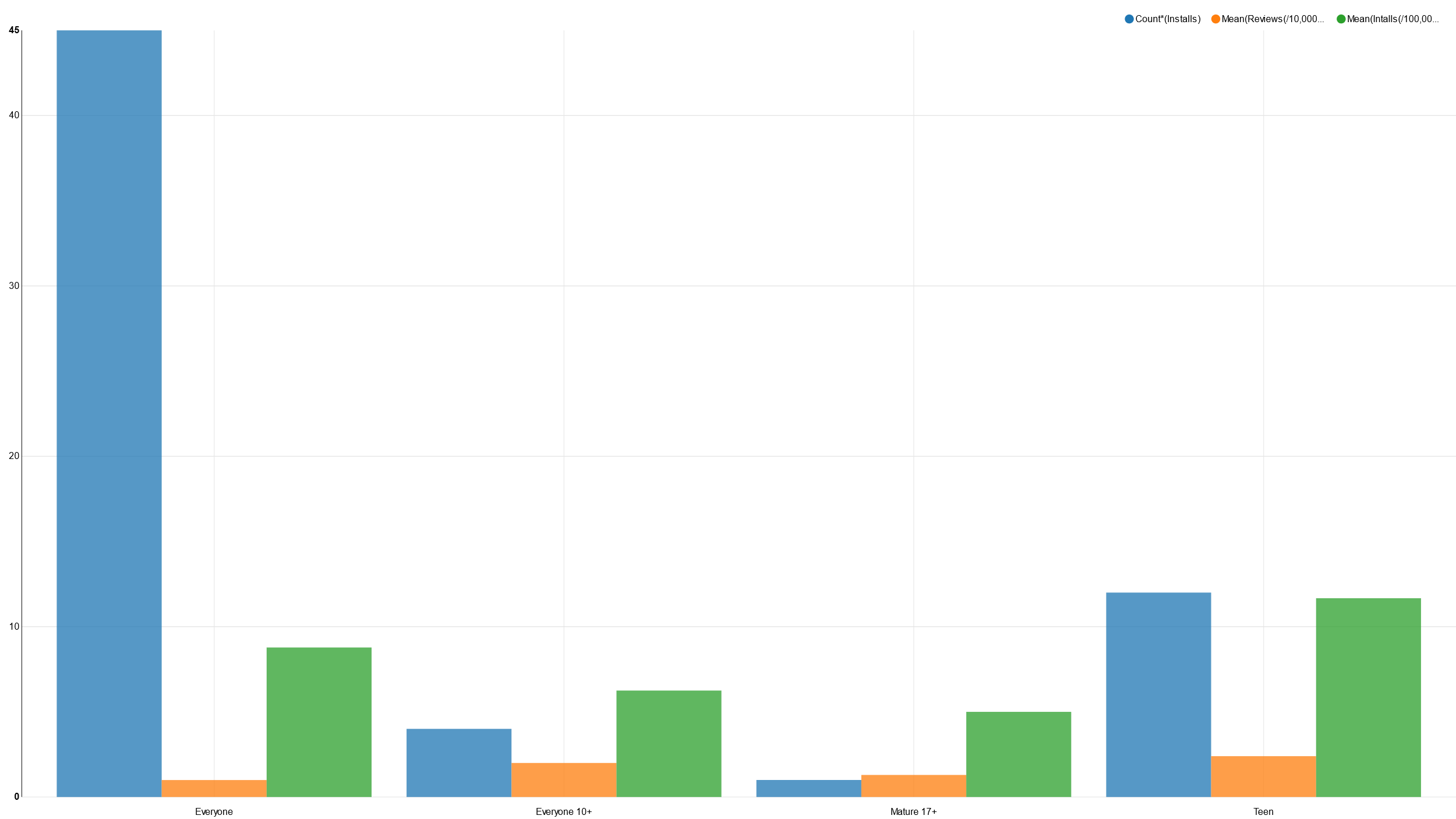


Figure : Nombre d'applications (bleu) le nombre de reviews /10,000,000 (orange) et moyenne du nombre d'installations /100,000,000 (vert) par type de contenu.

A l’issue de ces deux graphes on observe que le nombre de commentaires dans la catégorie *SOCIAL* est largement au-dessus des autres catégories ce qui montre qu’il n’y a aucune corrélation entre le nombre moyen de téléchargement et le nombre de commentaires. En ce qui concerne les types de contenu, *Teen* qui dispose en moyenne du plus de téléchargement dispose également du plus de commentaires, cependant cette différence n’est pas aussi marquée que pour la catégorie *SOCIAL*. Il est donc difficile d’établir un lien entre le nombre de téléchargement par type de contenu et le nombre de reviews.

Après Observations de tous ces graphes nous avons établie que faire ces mesures sur le cluster comportant les applications les plus populaire biaisais peut-être nos résultats. Nous avons donc décidé de recommencer les 2 dernières mesures (*Figure 6 et Figure 7)* mais en choisi les clusters centraux afin de retirer les extrêmes pour déterminer une tendance plus commune aux applications du Play store.

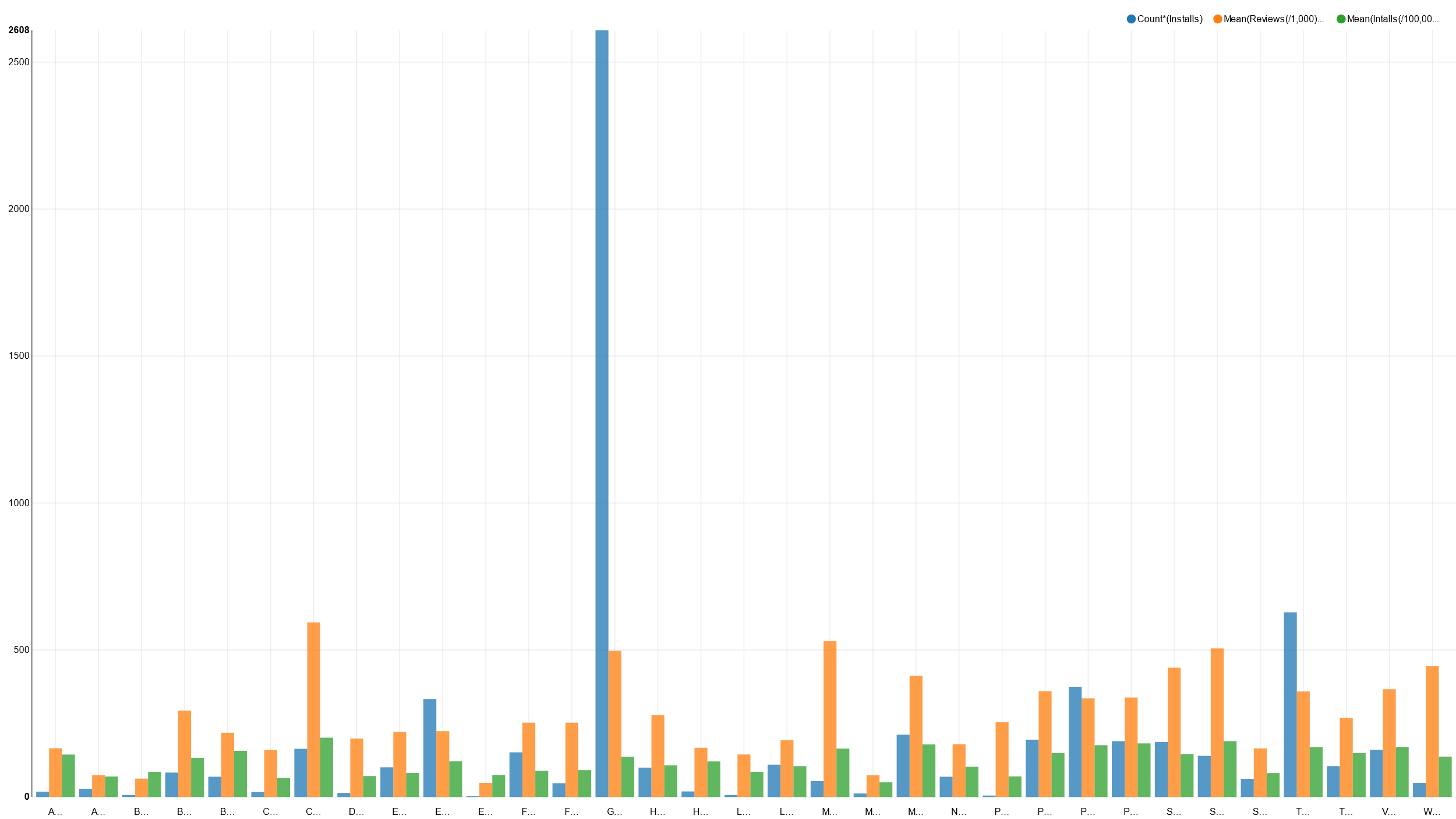


Figure : Nombre d'applications (bleu) le nombre de reviews /1,000 (orange) et moyenne du nombre d'installations /100,000 (vert) par catégories.

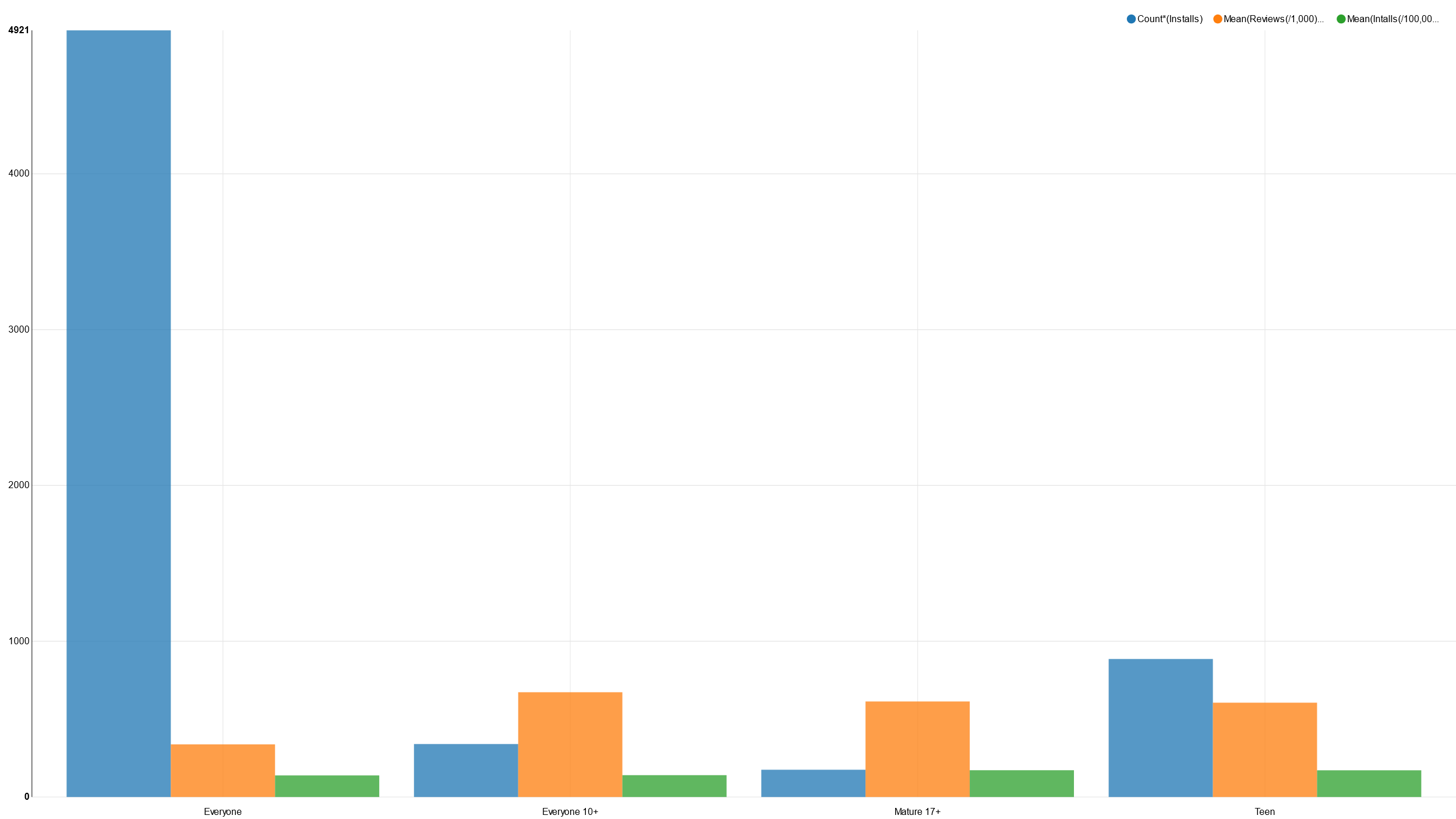


Figure : Nombre d'applications (bleu) le nombre de reviews /1,000 (orange) et moyenne du nombre d'installations /100,000 (vert) par type de contenu.

Ces données ne contiennent plus les extrêmes (application avec trop ou trop peu de téléchargement), et malgré tout nous ne voyons par de rapport de dessiner entre le nombre de téléchargement moyen et son type de contenu ou ça catégorie. Cependant on peut remarquer, au niveau des catégories que celles qui contiennent le plus de commentaires sont les catégories *SOCIAL* dans le premier test et *COMMUNICATION* dans le second test, qui sont deux catégories ou les applications ce bases souvent sur l’interaction entre les utilisateurs, ce qui semble donc correspondre.