TD sur le logiciel SPAD (Niveau 1)

Tutoriel de référence :

http://tic-recherche.crifpe.ca/docs/guides/fr/SPAD7_guide.pdf

Données : Le fichier « » recense les ventes de jeux vidéos depuis 2012. L'objectif est de réaliser l'une des premières démarches du data scientist, nettoyer les données en vue de réaliser les premières statistiques descriptives. Deux fichiers sont à votre disposition pour réaliser le TD. Ils sont disponibles via le lien ci-dessous :

https://drive.google.com/open?id=0B_3cHtrgDKSPVEc3dEw2NUcxSUU

Détail des variables:

Name	Nom du jeu
Platform	Nom de la plateforme
Year_of_Release	Date de sortie
Genre	Type de jeu
Publisher	Producteur du jeu
NA_Sales	Volume vente Amérique du Nord (en millions)
EU_Sales	Volume vente Europe (en millions)
JP_Sales	Volume vente Japon (en millions)
Other_Sales	Volume vente Autres (en millions)
Global_Sales	Volume Totale (en millions)
Critic_Score	Score agrégé de Metacritic
Critic_Count	Nombre de votant Metacritic
User_Score	Score agrégé des Utilisateurs
User_Count	Nombre de vote des utilisateurs
Developer	Développeur du jeu
Rating	PEGI (jeu déconseillé pour certain public)

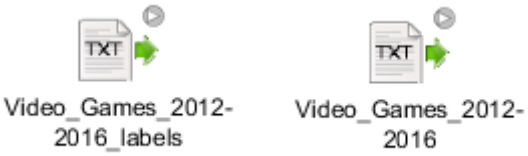
A faire :

1. Préparation des données

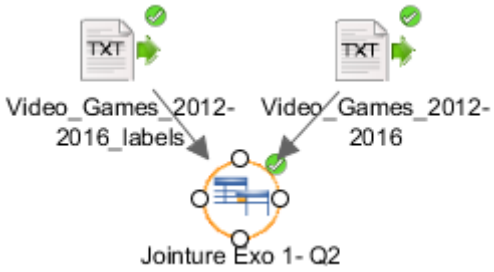
Démarrez SPAD et créez un nouveau projet

Exercice 1 : Data management

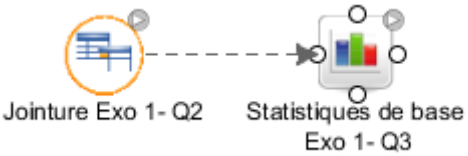
Question 1 : Importer les 2 fichiers de données à l'aide des méthodes d'import de SPAD. Vérifier que les variables ont bien été typées (nominales, continues, identifiant, **délimiteur de texte**) dans les métadonnées



Question 2 : Joindre les deux fichiers en utilisant la méthode Jointure (cf. Data management)



Question 3 : Combien y a-t-il d'observations et de variables dans le fichier. Vous pouvez utiliser la méthode Statistiques de base dans SPAD.

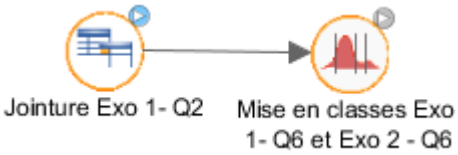


Question 4 : Utiliser la méthode "Validation des données"(sans l'executer) pour avoir une vue globale des variables quantitatives et qualitatives. Y-a-t-il des variables présentant des données manquantes ?

Question 5 : Vérifier que la variable Plateform possède 11 modalités. Quels sont les différentes plateformes répertoriées dans le Dataset (les méthodes " Sélection - Ordre" + "Distinct Dédoublonnage" pourraient vous aider) ? Que constatez-vous ? (vérifier du côté des paramètres d'importation)

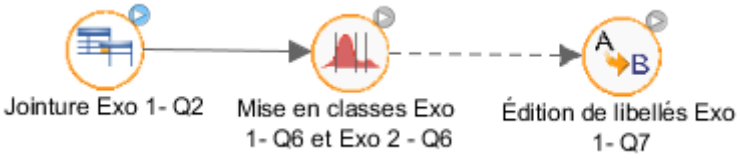
Question 6 : La variable Platform présente beaucoup de modalités la rendant inexploitable pour l'analyse. On souhaite donc effectuer un regroupement par éditeur de plateforme selon la table de correspondance ci-dessous (Mise en classes - Regroupement de modalités) :

PC	PC
XOne	XBOX
X360	
PS4	PLAYSTATION
PS3	
PSV	
PSP	
WiiU	NINTENDO
Wii	
3DS	
DS	



Question 7 : Pour étudier la variable Rating dans la partie 2, on souhaite éditer des libellés plus explicites. Modifier les libellés de cette variable d'après cette table de correspondance (https://www.esrb.org/ratings/ratings_guide.aspx) :

E	Everyone
E10+	Everyone 10+
T	Teen
M	Mature 17+

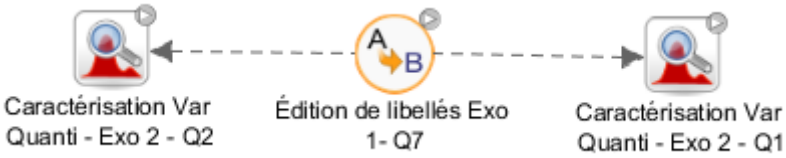


Question 8 : Les variables Critic Score et User Score ont-elles le même ordre de grandeur ? Si oui, jetez de nouveau un œil à la question 3. Si non, mettre les deux scores à la même échelle en divisant le Critic Score par 10 (Générateur de nouvelle variable)

Exercice 2 : Statistique exploratoire / Dataviz

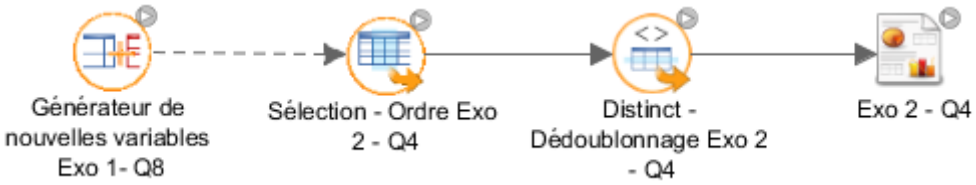
Question 1 : On souhaite caractériser les notes des utilisateurs (User_Score) selon le type de jeu (Genre) et l'âge conseillé (Rating). La méthode "Caractérisation d'une variable quantitative" pourrait vous servir. Afficher les résultats avec la sortie graphique de SPAD pour les modalités des variables nominales. Est-ce pertinent au vue de la structure du jeu de données ? Essayer de nouveau avec une variable de pondération. Quel est le genre de jeu préféré des utilsateurs ?

Question 2 : Effectuer la même analyse pour les notes des journalistes (Critic_Score). Sont-ils plus ou moins cohérents avec les résultats issus de la question précédente ?



Question 3 : On souhaite déterminer graphiquement l'évolution du volume des ventes depuis 2012 selon les zones géographiques (NA_sales,UE_sales,JP_sales, Global_sales). Utilisez le générateur de graphique de SPAD (indicateur de variable continue). Que constatez-vous ?

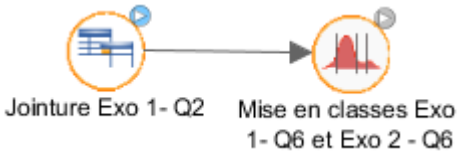
Question 4 : Cette forte baisse pourrait s'expliquer par la diminution du nombre de jeu sortie chaque année. Pour y vérifier, on cherche à dénombrer le nombre de jeux distincts (un même jeu peut sortir sur plusieurs plateformes) sortis chaque année. Utilisez le générateur de graphique de SPAD (Histogramme avancée de variable nominale). Que constatez-vous ? Les données 2016 ont été arrêtées le 22 décembre 2016. On suppose donc que les ventes réalisées durant les dernières semaines de l'année ne sont pas comptabilisées. De ce fait, on va exclure les données de 2016 pour la suite de nos analyses. Appliquer le filtre logique sur le "Générateur de nouvelle variable" utilisé auparavant.



Question 5 : Quel type de jeux ont été le plus vendu en 2015 ?

Question 6 : On souhaite regrouper les plateformes par type. Effectuez un nouveau regroupement sur la méthode déjà présente dans le diagramme SPAD selon la table de correspondance ci-dessous :

PC	Fixe
XOne	
X360	
PS4	
PS3	
WiiU	
Wii	
3DS	Portable
DS	
PSV	
PSP	



Question 7 : Avec l'avènement des smartphones, nous pensons que les ventes de jeux sur console portable sont en baisse depuis 2012. Pouvez-vous confirmer cette hypothèse graphiquement pour les ventes globales ? Utiliser le générateur de graphique de SPAD (indicateur de variable continue), attention un filtre logique pourrait-être nécessaire au préalable. Observe-t-on le même phénomène en Amérique, Europe et Japon ?

Question 8 : Constatez-vous une évolution des leaders sur le marché des jeux vidéos depuis 2012 ? Pour cela, on souhaite visualiser graphiquement le volume des ventes par année et par éditeur (cf Exercice 1 Q6). Utilisez le générateur de graphique (Secteur pour variable nominales). L'utilisation d'une variable de pondération peut aider.

Question 9 : On veut confronter graphiquement par un nuage de point les notes des journalistes (Critic_Score) et des utilisateurs (User_Score). Pour ajouter de l'information au nuage, on peut insérer une variable de poids et une variable nominale comme le type de plateforme (Cf : Exo 2 Q6) pour différencier les groupes. Quel est le jeu le moins bien noté et mieux noté ?

Exercice 3 : Analyse de données

Question 1 : Y-a-t'il une corrélation entre le vote des utilisateurs (User Score) et le volume des ventes de jeu vidéo ? Vous pouvez utiliser la méthode de statistique de base. Obtient-on le même constat pour les notes des journalistes (Critic_Score).

Question 2 : Est-ce que les scores sont les mêmes chez les Critic_Score et les User_Score (Test de comparaison de moyenne)?

Question 3 : Testez la normalité de l'échantillon par rapport aux ventes globales par année (Tests statistiques, Shapiro Wilk). Que remarquez-vous?

Question 4 : Existe-t-il un lien entre le type de plateforme et le Rating (Test du Khi-Deux) ? Si oui, mesurez la force de la relation (V de Cramer). Et entre Rating et l'Éditeur (Cf : Exo 1 Q6) ?