Découverte d'un outil d'interrogation et d'exploration visuelle de données « Tableau Desktop »

Objectifs

Lors de ce TP, vous allez découvrir un outil d'interrogation et d'exploration visuelle de données « Tableau Desktop ». Cet outil, disponible gratuitement pour tout étudiant et enseignant, est multiplateforme depuis le site https://www.tableau.com/fr-fr/academic pour télécharger le logiciel.

« Tableau » est un outil permettant principalement :

- de se connecter à différentes sources de données, locales ou distantes,
- d'intégrer et préparer des données provenant de différentes sources,
- d'explorer et visualiser les données intégrées,
- d'analyser les données, par exemple par des techniques de clustering,
- de créer des tableaux de bord et de les partager.

Un tel outil est destiné à des analystes et/ou décideurs ne possédant pas de fortes compétences en informatique. Une connaissance de langages d'interrogation et manipulation de données tels que SQL ou MDX est non nécessaire, les requêtes et visualisations se construisant visuellement et de manière interactive.

Partie I – Démarrage et connexion à une source de données

1) Après avoir lancé l'outil « Tableau», l'outil vous propose soit d'ouvrir un classeur (ou workbook) déjà existant, soit d'établir une connexion à une nouvelle source de données. Cette source de données peut aussi bien être un simple fichier (de type texte, PDF, Excel, etc.) qu'une base de données (de type SQL Server, MySQL, Oracle, etc.).
2) Choisissez de vous connecter à un fichier texte, le fichier « Sales-UTF8.csv » que vous aurez préalablement chargé depuis « Moodle ». Un aperçu des données contenues dans le fichier texte est directement affiché dans la partie centrale de l'écran. Ces données représentent des commandes (Order ID) de produits (Item ID) réalisés par des clients (Customer ID). Vous noterez le nombre d'attributs identifiés dans le fichier, ainsi que le nombre d'attributs par type distinct (les types d'attributs sont représentés par de petits icône – abc pour le type texte, # pour le type nombre, une mappemonde pour un lieu ou rôle géographique, etc. ; en cliquant sur un de ces icônes, il est possible de voir son intitulé)
3) Cliquez ensuite sur l'onglet « Feuille 1 » pour faire apparaître l'espace de travail principal de « Tableau ». Cet espace comprend notamment :

    - sur la gauche, deux encadrés « Données » et « Analyse ». Dans l'encadré « Données », vous verrez apparaître les attributs de votre source de données répartis en deux catégories : les dimensions (par défaut, sont placés dans cette catégorie, les attributs de type textuels, date et géographique) et les mesures (par défaut, sont placés dans cette catégorie, les attributs de type numérique). Quant à l'onglet « Analyse », il permettra de réaliser des analyses avancées sur les données, telles que des opérations de segmentation de données (clustering).
    - un ensemble de zones « Colonnes », « Lignes », « Filtres », « Repères » dans lesquelles pourront être placés par glisser/déposer des attributs de la source de données.

- une partie centrale où sera construit interactivement une visualisation des données à explorer.
- en haut à droite, un encadré « Montre-moi » permettant d'obtenir rapidement des recommandations de visualisation d'ensemble de données définis par un simple sous-ensemble d'attributs. Cet encadré peut être fermé ou ouvert en cliquant sur l'intitulé « Montre-moi ».
- en haut et en bas, de petites icônes permettent de lancer rapidement des actions. Par exemple, en bas, à côté de l'onglet « Feuille 1 », des icônes permettent l'ajout d'une nouvelle feuille de calcul (similaire à la « Feuille 1 » actuellement visualisée), d'un nouveau tableau de bord (construit principalement par regroupement de plusieurs visualisations de feuilles de calcul) ou d'une nouvelle histoire (concept permettant de construire une présentation scénarisée intégrant plusieurs tableaux de bord).

Partie II – Création d'une première visualisation avec des histogrammes

1) Les concepts de mesure et dimension sont des concepts fondamentaux pour l'exploration et la visualisation des données sous « Tableau ». Alors que les mesures représentent des valeurs pouvant être agrégées (par des sommes, moyennes, etc.), les dimensions vont déterminer à quel niveau de granularité les données vont être visualisées (par exemple, par année, mois ou jour, par région ou département, etc.).
2) Sélectionnez dans l'encadré « Données » la mesure « Sales », puis glisser-déposer cette mesure dans le champ « Colonnes ». Intuitivement, qu'observez-vous dans la première visualisation construite ? Dans le champ « Colonnes », vous noterez la fonction d'agrégation « SOMME » appliquée automatiquement à la mesure « Sales ». Cette fonction d'agrégation a été choisie par défaut. Elle peut être modifiée en cliquant sur « SOMME(Sales) » et en ouvrant le menu contextuel associé.
3) Sélectionnez maintenant dans l'encadré « Données » la dimension « Department », puis glisser/déposer cette dimension dans le champ « Lignes ». Quelle nouvelle visualisation obtenez-vous ?
   Notez que « Tableau » a choisi automatiquement une représentation des données sous la forme d'histogramme. Ce choix peut être remise en question en sélectionnant une autre forme de visualisation qu'« automatique » dans le menu de la zone « Repères ». Testez d'autres formes de visualisation avant de revenir à une visualisation sous la forme de « Barre».
4) Notez que pour obtenir la visualisation précédente, aucune requête SQL n'a dû être écrite directement. Quelle est la requête SQL que « Tableau » a dû construire automatiquement pour obtenir le résultat affiché.
5) Dans le champ « Lignes », ajoutez maintenant à gauche de département la dimension «Region ». Quelle nouvelle visualisation obtenez-vous ?
   Afin de comparer plus aisément les niveaux de vente par département, cliquez sur le bouton « Trier » (12ième bouton de la barre de bouton en haut de l'interface). Quel département réalise le plus grand volume de vente quel que soit la région considérée ?
6) Notez que la même information peut être visualisée de différentes manières. Par exemple, déplacez la dimension « Région » du champ « Lignes » vers le bouton « Couleur » de la zone « Repères » (par simple « Glisser-déposer »). Examinez la nouvelle visualisation obtenue. Cliquez finalement sur le bouton « Permuter » de la barre d'outils en haut de l'interface. En

double-cliquant sur le nom de l'onglet « Feuille 1 », renommez finalement votre première feuille en « Ventes par région et département ».

7) Sauvegardez votre classeur en cliquant sur le bouton adéquat de la barre d'outils. Notez que « Tableau » vous permet soit de sauvegarder votre tableau (sans les données), soit votre tableau complet (avec les données extraites des sources) avec l'option « Exporter le tableau complet … ».

Partie III – Création d'une visualisation d'évolution des ventes

1) Commencez par créer une nouvelle feuille de calcul intitulée « Evolution des ventes ».
2) Placer la mesure « Sales » dans le champ « Lignes », puis la dimension « Order Date » dans le champ « Colonnes ». Quelle nouvelle visualisation obtenez-vous ?
   Vous notez que les ventes ont été agrégées par « Année », ce niveau d'agrégation ayant été choisi par défaut par « Tableau ». Cette visualisation permet d'observer une augmentation des niveaux des ventes au cours du temps.
3) « Tableau » dispose de hiérarchies prédéfinies sur les dates. L'existence d'une telle hiérarchie est visible par la présence d'un symbole « + » devant la dimension « Année(Order Date) » dans le champ « Colonnes ». Ouvrir le menu contextuel associé à la dimension « Année (Order Date) » et comparez les résultats obtenus en choisissant d'abord le niveau « Trimestre T2 », puis le niveau « Trimestre T2 2015 ». Vous expliquerez dans chacun des cas le résultat obtenu
   Notez que la visualisation obtenue après la sélection du niveau « Trimestre T2 2015 » permet d'observer une variation saisonnière des ventes, avec toujours un niveau de vente plus importante au quatrième trimestre (T4).
4) Afin d'examiner si cette saisonnalité des ventes est la même dans toutes les régions, glisser/déposer la dimension « Région » dans la partie « Couleur » de la zone « Repères ». Qu'en concluez-vous ?
   Notez qu'un encadré « Région » est maintenant apparu en haut à droite de la visualisation. Cet encadré permet de surligner certaines courbes par rapport à d'autres en cliquant sur une légende donnée. Testez cette possibilité, ce principe étant particulièrement intéressant lorsqu'un très grand nombre de courbes sont superposées.
5) Sauvegardez l'état actuel de votre classeur.

Partie IV – Création d'une visualisation de la répartition géographique des ventes

1) Commencez par créer une nouvelle feuille de calcul intitulée « Répartition par état des ventes ».
2) Glisser-déposer la dimension « State » (tout en bas de la liste des dimensions) dans la zone blanche de visualisation « Déposer champ ici ». Notez les noms des mesures intégrées automatiquement dans les champ « Lignes » et « Colonnes ».
   Le terme « généré » indique que les mesures longitudes et latitudes ont été générées automatiquement par « Tableau » (à partir des noms d'état retrouvés dans la dimension « State »).
3) En bas de la carte affichée, si vous observez « X inconnu(e) », double-cliquez sur ce champ, puis sélectionnez « Modifier les emplacements ». Vous noterez que les valeurs dans la colonne « Vos données » n'ont pas trouvé automatiquement de correspondant. Pour ce faire, sélectionnez dans le menu déroulant « Etats-Unis ». Les correspondances nécessaires ont maintenant été réalisées et vous devrez voir apparaître un point dans chaque état des Etats-Unis.

4) Glisser-déposer la mesure « Sales » dans la partie « Couleur » de la zone « Repères », et interprétez la visualisation obtenue.
5) Plutôt que de construire des cartes par remplissage de zone, « Tableau » permet également d'associer des symboles à des positions géographiques. Créer une nouvelle feuille de calcul intitulée « Répartition par code postal des ventes et profits », puis glisser-déposer dans cette feuille la dimension « Postal Code ».
6) Glisser-déposer la mesure « Sales » dans la boite « Taille » de la zone « Repères ». Qu'obtenez-vous comme visualisation ?
   Notez que la taille des cercles peut être paramétré en cliquant sur la boite « Taille » de la zone « Repères ».
7) Glisser-déposer maintenant la mesure « Profit » dans la boite « Couleur » de la zone « Repères », puis en cliquant sur « Couleur », choisissez la palette de couleur « Rouge-bleu divergent », choisissez d'ajouter une bordure noire aux cercles et un niveau d'opacité de 75%. Comment interprétez-vous la nouvelle visualisation obtenue ?
   Notez qu'une telle visualisation peut permettre d'identifier des zones avec des niveaux de ventes importants alors que les profils sont assez faibles. Quels sont les symboles permettant d'identifier de telles zones géographiques ?
8) Sauvegardez l'état actuel de votre classeur.

Partie V – Création d'un tableau de bord

1) Dans l'outil « Tableau », un tableau de bord permet de présenter dans un même document tout un ensemble de visualisations, tout en rendant interactive ces visualisations.
2) Créer un nouveau tableau de bord en cliquant sur le bouton correspondant (dans la barre d'onglets en bas de l'interface). Dans l'encadré à gauche, différentes options vous sont proposées. Il est par exemple possible de préciser :
   a. - sur quel dispositif le tableau de bord construit sera principalement visualisé. Vous pourrez choisir une taille « Automatique » (qui redimensionnera automatiquement le tableau de bord en fonction de l'espace disponible dans votre fenêtre).
   b. - quelles feuilles vous souhaitez intégrer dans votre tableau de bord. Vous choisirez les feuilles « Ventes par région et département », « Evolution des ventes » et « Répartition par code postal ».

   Pour ce faire, il suffit de double-cliquer sur les feuilles correspondantes. Examinez avec quelle facilité, il est possible de redéfinir la taille et la disposition des différentes tuiles.

3) Quand vous cliquez sur une tuile, quatre boutons apparaissent en haut à droite de cette dernière. Ces quatre boutons que vous pouvez survoler permettent respectivement :
   - de supprimer une tuile / visualisation d'un tableau de bord,
   - de revenir à la feuille de calcul correspondante,
   - d'utiliser la tuile pour effectuer des filtrages dynamique (ce point est détaillé un peu plus loin),
   - d'ouvrir un menu textuel offrant une palette plus riche d'actions.
4) A titre d'exemple, effectuez les modifications suivantes :

   - Revenez à la feuille « Ventes par région et département » et supprimez la dimension d'analyse « Par région », puis indiquez (au niveau du tableau de bord) que cette vue pourra être utilisée comme filtre.
   - Revenez à la feuille « Evolution des ventes » pour supprimer la dimension d'analyse « Par région », puis indiquez comme précédemment que cette vue pourra être utilisée

comme filtre. La tuile qui contenait la légende pour les régions pourra alors être supprimée.
- Revenez à la tuile « Répartition par code postal des ventes et profits » et indiquez enfin que cette vue pourra être utilisée comme filtre.

5) Nous allons maintenant observez les liens de filtrage créés automatiquement entre visualisation. Par exemple, cliquez sur la barre « Furniture » de la tuile « Ventes par département ». Quelles évolutions constatez-vous dans les autres vues ?
6) Testez maintenant les possibilités offertes de filtrage en sélectionnant différents points des autres vues, sachant que plusieurs éléments peuvent être sélectionnés conjointement (en utilisant la touche « Ctrl » sur PC ou « Command » sur MAC). Notez que dans chaque tuile, un menu contextuel permet de tout re-sélectionner.

Partie VI – Manipulation avancées de données

Après avoir donné un premier aperçu des fonctionnalités apportées par un outil comme « Tableau », nous allons maintenant en examiner quelques éléments plus avancés. Pour cette partie du TP, créer un deuxième classeur de nom « Join TP1 ».

1) Dans ce nouveau classeur, commencez par vous connecter au fichier texte « Sales-UTF8.csv», et dans une feuille construisez une visualisation représentant les ventes par département (comme au tout début du TP).
2) Nous allons maintenant examiner comment construire une hiérarchie sous « Tableau ». Dans notre jeu de données, un département offre différentes catégories de produit, une catégorie de produit étant offerte par un seul département. Par conséquent, il est possible d'indiquer à « Tableau » que « Category » et « Department » appartiennent à une même hiérarchie, « Category » représentant un niveau d'analyse plus fin des ventes par « Department ». Pour ce faire, dans l'encadré « Données », glisser-déposer la dimension « Category » sur la dimension « Department » Vous noterez « Product » la hiérarchie construite. Ajouter également à cette hiérarchie, au-dessous du niveau « Category » son niveau le plus fin « Item ».
3) Dans le champ « Lignes » (ou « Colonnes ») apparaît maintenant le symbole « + » devant le nom de la dimension « Department ». Cliquez sur ce symbole pour naviguer dans la hiérarchie créée, jusqu'au niveau le plus fin « Item », avant de revenir au niveau « Category».
4) L'outil « Tableau » permet également d'effectuer des sélections sur des ensembles de données en créant des filtres. Par exemple, glisser-déposer la dimension « Order Date » dans la zone « Filtres » et sélectionnez l'année 2016. Notez que vous pouvez faire apparaître le filtre créé en cliquant sur l'intitulé du filtre (ce qui permet d'ouvrir un menu contextuel) et en choisissant « Afficher ». Vous pouvez maintenant sélectionner d'autres années que l'année 2016 ou un sous-ensemble particulier de données
5) Sur le même principe, créer maintenant un filtre sur la somme des ventes par plage de valeurs, en précisant dans un deuxième temps que vous souhaitez voir apparaître ce filtre. En testant ce filtre, si la visualisation devient transparente, cliquez sur le bouton « Exécutez mise à jour » de la barre d'outils (7ième bouton en partant de la gauche).
6) Supprimer finalement les deux filtres créés (via les menus contextuels associés à chaque filtre).
7) Nous allons maintenant examiner comment créer une jointure entre deux tables avec un outil comme « Tableau ». Dans l'onglet « Source de données », cliquez sur « Ajouter » (une nouvelle connexion) et sélectionnez le fichier texte « States-UTF8.csv ». « Tableau » cherche immédiatement à joindre les tables des deux fichiers, mais échoue car les deux tables ne possèdent pas d'attribut de nom commun. En cliquant sur l'item « Ajouter une nouvelle

clause join », sélectionnez l'attribut « States » de votre source données « Sales-UTF8.csv » puis l'attribut « State or territory » de la source « States-UTF8.csv ». Notez que la jointure est immédiatement visualisée en dessous, une jointure interne étant réalisée par défaut.

8) Dans une nouvelle feuille « Population par état », placer en lignes la mesure « Census Population (2010) » et en colonne la dimension « State or territory ». En survolant la barre de l'état de Californie, quel nombre d'habitant vous est indiqué ? En utilisant le menu contextuel associée au champ « SOMME(Census Population (2010) », changez la mesure d'agrégation en choisissant de calculer un MIN ou un MAX. Quel nombre d'habitants est maintenant indiqué pour l'état de Californie ? Selon vous, comment expliquer les différents résultats obtenus ? Vous pourrez vous aider en spécifiant la requête que l'outil « Tableau » doit exécuter.
   Cet exemple montre avec quelle précaution il faut utiliser un outil comme « Tableau » générant automatiquement les requêtes sur des sources de données. En effet, le résultat fournit par « Tableau » n'est pas nécessairement le résultat souhaité.

Partie VII – Utilisation de sources de données séparées

Nous allons maintenant examiner comment éviter les problèmes posés par l'exemple de la question 8 de la partie précédente. Pour ce faire, nous allons étudier comment créer un classeur avec deux sources de données initialement séparées, puis comment lier ces deux sources au niveau d'agrégation souhaité pour réaliser certaines vues.

1) Commencer par créer un nouveau classeur intitulé « Blending TP1 ».
2) Ajouter à ce classeur une première source de données « Sales-UTF8.csv ».
3) Via le menu général « Données » de l'application « Tableau », ajouter maintenant une deuxième source de données « States-UTF8.csv ».
4) Pour identifier ce qui a changé par rapport à la partie VI précédent (où nous avions ajouté une nouvelle connexion et non une source de données », ouvrez maintenant la première feuille du classeur. En haut de l'encadré « Données », vous devez maintenant voir apparaître deux sources de données distinctes. Vérifiez que vous pouvez simplement passer d'une source à une autre en cliquant sur l'icône correspondant.
5) Dans une telle configuration, il est possible de créer des visualisations utilisant des données d'une seule des deux sources. Créez ainsi dans des feuilles séparées « Ventes par état » et « Population par état » des visualisations des sommes de vente et population par état.
6) En effectuant une opération dite de « blending » sous « Tableau », nous allons maintenant créer une visualisation combinant des données des deux sources de données.
7) Pour ce faire, dans une nouvelle feuille, commencez par glisser-déposer la mesure « Sales » de la source de données « Sales-UTF8 » dans le champ « Colonnes ». Vous noterez l'apparition d'un petit disque bleu coché sur la source de données « Sales-UTF8 ». Cet icône indique que la source « Sales-UTF8 » sera considérée pour cette feuille comme la source de données primaire.
8) Glisser-déposer maintenant la mesure « Census Population (2010) » de la source de données « States-UTF8 » dans le champ « Colonne ». Vous verrez apparaître une boite de dialogue indiquant qu'il est nécessaire qu'une relation soit créée avec « Sales-UTF8 » via l'item « Modifier les relations » du menu « Données » de « Tableau ». Notez également le petit disque orange coché à côté de la source « States-UTF8 ». Une icône de cette couleur indique que la source de données en question servira de source de données secondaire
9) Après avoir ouvert la fenêtre « Modifier les relations », vous noterez que « Tableau » essaie par défaut de construire une mise en relation automatique. Une telle relation fonctionne

quand les deux sources de données (primaire et secondaire) possèdent un attribut de même nom, ce qui n'est pas le cas ici. Choisissez d'ajouter une relation « Personnalisée » en sélectionnant l'attribut « States » de la source primaire et l'attribut « State or territory » de la source secondaire. Après validation de la relation, vous noterez l'apparition d'un trombone à côté de la dimension « State or territory » de la source « States-UTF8 ».

10) Ajouter maintenant au champ « Lignes » la dimension « States » de la source de données « SalesUTF8 ». Quelle est la nature de la visualisation obtenue ?
11) Nous souhaitons maintenant visualiser les montants des ventes (en millier de dollars) par état, mais par habitant. Dans le champ « Colonnes », cliquez sur « SOMME(Sales) » pour ouvrir le menu contextuel associé et sélectionner l'item « Modifier dans l'étagère ». Après « SUM([Sales]) », commencez par taper « / Cens » et laisser « Tableau » complété par le nom complet de la mesure « Census Population (2010) » (vous noterez que « Tableau » ajoute automatiquement la fonction d'agrégation par défaut et le lien avec la source de données secondaire), puis tapez « * 1000 » et « Entrée ». Ajouter enfin au champ « Colonnes » (entre les deux mesures déjà existantes) la mesure « Sales » de la source de données « Sales-UTF8 ». Une telle visualisation fait apparaître que si les niveaux de vente sont dans certains états très élevés en absolu (cas de la Californie), ils le sont moins par habitant (ce qui n'est par contre pas le cas pour le Massachusetts où les ventes sont élevées en absolu et en relatif). Pour finir, écrivez la requête SQL qui permettrait d'obtenir comme sur votre visualisation les ventes en millier de dollar par état et par habitant.
12) Nous allons maintenant ajouter la dimension « Region » de la source « Sales-UTF8 » au début du champ « Lignes ». Notez que les agrégations sont réalisées correctement que ce soit pour les niveaux de vente ou taille de population. Enfin, créer une hiérarchie « Location » en plaçant la dimension « State » sous la dimension « Region », supprimer la mesure « State » du champ « Lignes » et naviguer dans la hiérarchie créée.

Partie VIII – Analyse exploratoire de données

Un outil comme « Tableau » intègre également des fonctions élémentaires de fouille de données. En particulier, il est possible de segmenter des jeux de données (à l'aide de la technique k-means introduite en cours).

1) Ajouter à votre classeur une nouvelle feuille intitulée « Clustering ».
2) Glisser-déposer dans le champ « Lignes » la mesure « Sales » de la source de données « SalesUTF8 », puis dans le champ « Colonne » la mesure « Profit » de la même source de données.
3) Placez ensuite dans la boite « Détail » de la zone « Repères » la dimension « States ». Comment interprétez-vous cette visualisation ?
4) Pour réaliser un premier clustering de l'ensemble de points visualisés, glisser-déposer le modèle « Cluster » de l'encadré « Analyse » dans la zone d'affichage de l'ensemble des points. Quelles sont les variables que « Tableau » a automatiquement sélectionnées pour réaliser la segmentation et combien de classes a-t-il distingué ?
5) Un des clusters contient un nombre faible d'états, les états pour lesquels des montants de ventes et de profits élevés ont été réalisés. Avec un clique-droit sur PC (ou Command-clique sur MAC), sélectionnez l'ensemble de ces points, et choisissez de les exclure. Vous noterez que le clustering est automatiquement mis à jour (avec toujours deux classes) et que la zone « Filtres » de la feuille permet de voir quels sont les états du jeu de données qui ont été exclus (il est ainsi possible de réaliser des sélections visuelles d'ensemble de données, un des intérêt d'un outil comme « Tableau » en terme d'IHM).

6) Modifiez le nombre de clusters à trois. Le nouvel attribut ainsi créé peut maintenant être ajouté aux dimensions d'analyse dans l'encadré « Données » (par simple Glisser-déposer). Réalisez cette action en nommant « Cluster (par ventes et profits) » cette nouvelle dimension.
7) Examinez finalement dans une nouvelle feuille comment utiliser la segmentation créée en visualisant la répartition des niveaux de ventes par cluster. Vous noterez qu'une valeur « Pas dans un cluster » a été introduite où sont placés les états exclus du processus de clustering.

Pour conclure, notez enfin que ce TP offre seulement un aperçu rapide de quelques fonctionnalités apportées par « Tableau ». Tout un ensemble de vidéos en ligne peuvent vous permettre de découvrir plus en profondeur cet outil (voir la page https://www.tableau.com/fr-fr/learn/training pour les curieux et curieuses).

Remarque : les fonctions de fouille de données d'un outil comme « Tableau » restent assez élémentaires. Néanmoins, il est possible d'intégrer un outil de fouille plus puissant comme « R » à « Tableau », sachant qu'en master BDMA vous serez familiarisés avec l'outil « R » en première année de master.