Data Wrangling

Ce document présente les modifications apportées sur les jeux de données pour chaque visualisation.

Il ne contient pas l'historique précis, mais une description du nettoyage effectué.

Pour chaque jeu de données, nous avons fait le choix par précaution, de retirer certaines, mais pas toutes les colonnes à priori inutiles à la réalisation de notre analyse.

Toute récupération de données manuelle a été effectuée avec les sites :

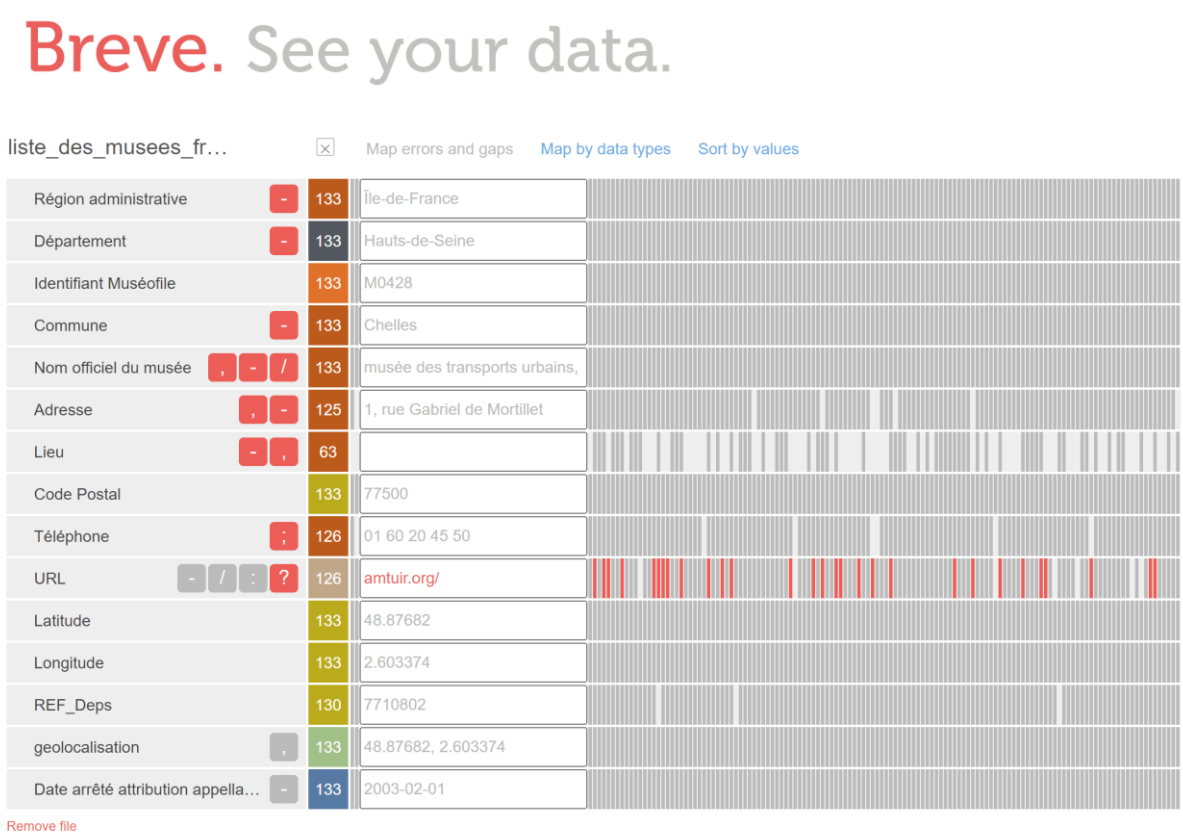
- Wikidata,
- Wikimedia,
- Wikipedia,
- la Plateforme Ouverte du Patrimoine (Ministère de la Culture),
- Google Maps,
- les sites officiels des musées.

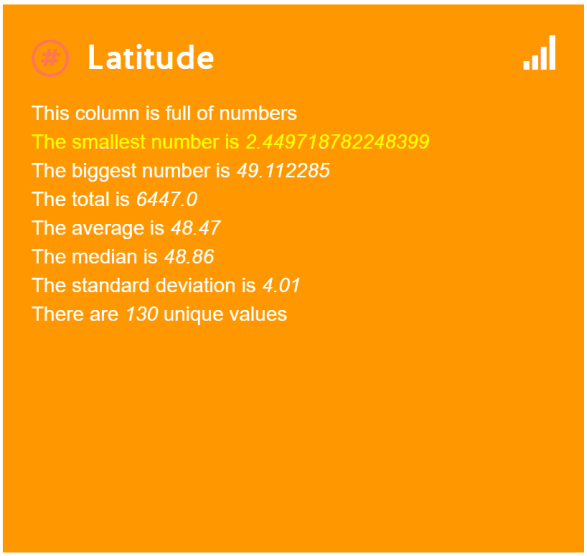
Table des matières

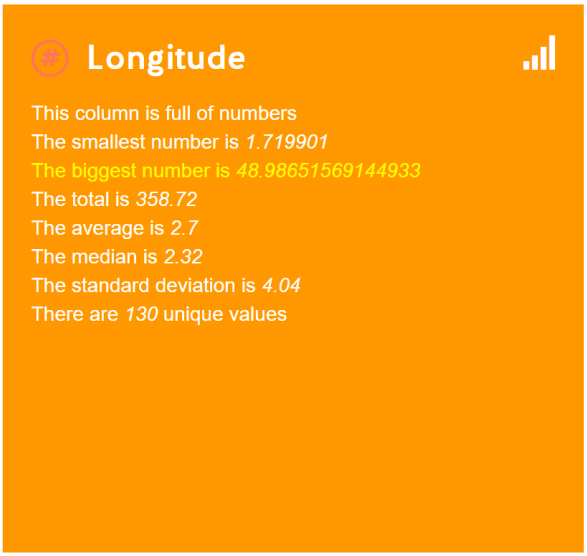
Prévisualisation des données

Prévisualisation de contenu de jeu de données avec BREVE See your data tables (stanford.edu) et WTFcsv (databasic.io) :

Fichier : *liste-des-musees-franciliens.csv*:







Les prévisualisations ont été utiles pour repérer certaines erreurs et incohérences (comme les données de latitude et de longitude dans les mauvaises colonnes), mais globalement, ce n'était qu'en allant dans les données, en les explorant et en travaillant dessus que nous avons découvert les plus graves erreurs (musées fermés, données de géolocalisation erronées, doublons, etc.).

Fichier : 1_liste-des-musees-franciliens-avec-images.csv

Ce jeu de données reflète la liste officielle des institutions dotées de l'appellation "Musée de France" au sens du Code du patrimoine. Cette liste est administrée par le Service des musées de France (direction générale des patrimoines et de l'architecture).
Elle est actualisée au moins une fois par an, en fonction des avis rendus par le Haut Conseil des musées de France, sur l'attribution, le changement ou le retrait de l'appellation "Musée de France".

Pour en savoir plus sur l'appellation "Musée de France".

Ce jeu de données remplace l'ancien jeu de données dont vous retrouverez une partie des informations dans "Fréquentation des musées de France".

- **Visualisation – Les "Musées de France" en Île-de-France :** nous souhaitons afficher une carte de l'Île-de-France et de ses Musées de France.
- **Visualisation – Répartitions des musées en Île-de-France :** nous voulons voir le nombre de musées en Île-de-France par département et par commune.
- **Visualisation – Musées situés à Paris labélisé "Musées de France" :** nous souhaitons afficher, dans un tableau, la liste des musées parisiens ayant le label Musées de France.
- **Visualisation – Date de création des musées parisiens labélisés "Musées de France" :** nous souhaitons sur une frise chronologique des dates d'ouverture officielle des musées parisiens ayant le label Musées de France.

1. Nous vérifions qu'il n'y a pas de doublons de musées avec une facette par doublons sur la colonne « Identifiant Muséofile ».

2. Nous vérifions et corrigeons les valeurs des colonnes contenant les données de géolocalisation.

3. Nous effectuons une réconciliation sur la colonne « Nom officiel du musée » à l'aide de la propriété P539 sur la colonne « Identifiant Muséofile ». Le Musée d'histoire et de société de la ville Gonesse n'a pas pu être réconcilié puisqu'il s'agit d'un « futur musée [qui] sera accueilli à Gonesse dans l'ancien hôpital hospice de 1841 ». (Source : Le Projet Musée | Ville Gonesse (ville-gonesse.fr))

4. Nous décidons de mettre les initiales en majuscule dans la colonne « Nom officiel du musée ».

5. Nous uniformisons la colonne « Adresse » (on remplit manuellement les données manquantes et on enlève les virgules avec l'expression GREL « value.replace(',', '') »)

6. Le musée municipal de Crécy-la-Chapelle et le musée des travaux publics de Courbevoie n'ont pas d'adresse définie, mais puisqu'ils ont des coordonnées géographiques et qu'ils ne sont pas définitivement fermés, nous décidons de les laisser.

7. Une facette par doublons sur les colonnes « geolocalisation » et « Adresse » nous permet de vérifier que les données ne sont pas dupliquées et de les corriger si elles le sont, en retrouvant les valeurs sur Google Maps.

8. Nous supprimons le musée Hébert de la région parisienne, qui est fermé.

9. Nous ajoutons une colonne « Adresse_2 » à partir de la colonne « Adresse » avec l'expression « value + ", " + cells["Code Postal"].value ».

10. Nous isolons les musées parisiens ave une facette sur la colonne « Commune ». Nous ajoutons ensuite une colonne « date of inception » à partir des valeurs réconciliées avec la propriété P571.

11. Nous remplissons manuellement les valeurs manquantes à la d'une facette par valeur nulle.

12. À partir de la colonne « Nom officiel du musée », nous créons une colonne récupérant les identifiants Q.

13. Une transformation sur la colonne crée avec l'expression GREL « (wd:" + value + " " + (row.index + 1) + ") » nous permet de préparé notre requête Wikidata qui nous permettra de récupéré les photos des « Musées de France » dans l'ordre.

14. Après avoir exporté ces données sous format Excel, nous utilisons la formule « =CONCAT(F2:F50;"") » pour récupérer tous les Q sur une seule ligne.

15. On utilise ce fichier pour obtenir le fichier : 6_liste-des-musees-franciliens.geojson.

Fichier intermédiaire : query-images.csv

Il s'agit d'une requête Wikidata qui nous sert à récupérer les photos des « Musées de France » situés à Paris dans une colonne « imageLink » que nous allons copier/coller dans le fichier *1_liste-des-musees-franciliens.csv*.

1. Nous vérifions que les données correspondent aux bons identifiants.

2. Nous remplissons manuellement les images manquantes avec un lien vers le logo « Musées de France ».

Fichier : 4_query-liste-des-musees-parisiens.csv

Visualisation – Musées situés à Paris labélisé "Musées de France" : nous souhaitons afficher tous les musées situés à Paris. Cette liste nous servira lors de la création d'une carte montrant les musées parisiens avec et sans le label « Musées de France ».

Les résultats de la requête contiennent des données manquantes.

1. Nous effectuons donc une réconciliation dans OpenRefine sur la colonne « musee » pour récupérer les données des pages Wikidata correspondants aux identifiants des items.

2. À partir des données réconciliées, nous créons une colonne nous permettant de vérifier que les musées ne sont pas définitivement fermés, et les supprimer le cas échéant.

3. Une facette par valeur nulle sur la colonne « lat_long » nous permet de voir les musées n'ayant pas de coordonnées. Nous recherchons ces données sur Google Maps et nous supprimons les musées qui ont été remplacés, déplacés ou qui n'existent plus.

4. À partir de la colonne « musee », nous ajoutons les colonnes « located on street » et « postal code » avec les propriétés P669 et P281 (nous n'utilisons pas les propriétés P670 et P6375 car il manque beaucoup trop de données).

5. Nous ajoutons une colonne « Adresse_2 » à partir de la colonne « located on street » avec l'expression GREL « value + ", " + cells["postal code"].value ».

6. Nous complétons les données manquantes manuellement.

7. Nous renommons la colonne « museeLabel » : « Nom officiel du musée ».

8. Nous téléchargeons le fichier sous format Excel.

9. On utilise ce fichier pour obtenir le fichier : 5_query-wikidata-liste-des-musees-parisiens.GEOJSON.

Fichier intermédiaire : liste-des-musees-verification-doublons.xsl

Fichier de comparaison : Pour cela, nous allons comparer les données des deux listes et supprimer les doublons.

1. Dans le fichier 1_*liste-des-musees-franciliens* sur OpenRefine, une facette textuelle sur la colonne « Commune » nous permet d'isoler les lignes de la commune de Paris et de les télécharger sous format Excel.

2. Nous renommons le fichier téléchargé *liste-des-musees-verification-doublons.xsl*.

3. Dans ce fichier, nous ne gardons que les colonnes « Nom officiel du musée » et « geolocalisation »

4. Nous copions depuis le fichier *4_query-wikidata-liste-des-musees-parisiens.csv*, les colonnes « Nom officiel du musée » et « lat_long » pour les coller sous les colonnes « Nom officiel du musée » et « geolocalisation », respectivement.

5. La mise en forme conditionnelle sur les valeurs en doubles nous permet de détecter 2 mêmes musées présents dans les listes.

4. Nous les retirons donc directement sur OpenRefine, dans la liste obtenue avec Wikidata Query Service.

Fichier : 2_frequentation-des-musees-de-France.csv

Visualisation – Top 10 des musées parisiens labélisés "Musées de France" les plus fréquentés : nous souhaitons afficher sur un graphique, les 10 musées parisiens ayant le label « Musées de France » les plus fréquentés chaque année, de 2012 à 2021.

Fréquentation totale, payante et gratuite dans les Musées de France de 2001 à 2021.

2001 : Les musées n'avaient pas encore l'appellation. Ils avaient le statut de musée classé ou contrôlé.
2002 : Loi musée de janvier 2002. Les musées nationaux et les musées classés sont devenus automatiquement des musées de France.
2003 : Les musées contrôlés sont devenus des musées de France - Février 2003

La colonne "Note": signale : "F" : le musée est fermé ; "NC" : la fréquentation n'a pas été communiquée ; "R" : retrait de l'appellation "Musées de France".

Attention : Certains sites communiquent une fréquentation cumulée pour l'ensemble des musées du site. La colonne « Observations » contient les références des musées dont la fréquentation est cumulée (cf. colonne « Réf du musée »).

1. À l'aide d'une facette textuelle sur la colonne « VILLE », nous supprimons les lignes des musées n'étant pas situés à Paris.

2. Une autre facette dans la colonne « NOTE » nous permet de retirer de la liste :

- les musées qui ont été fermés (NC),

- les musées pour lesquels l'appellation "Musées de France" à était retirée (R),

- les lignes pour lesquelles la fréquentation n'a pas été communiquée (F).

3. Encore une facette, dans la colonne « ANNEE » nous permet de ne garder que la période qui nous intéresse : 2011 à 2021.

4. En appliquant une facette textuelle sur la colonne « ANNEE » et un trie décroissant sur la colonne « TOTAL », nous marquons les 10 musées les plus fréquentés de chaque année par des étoiles.

5. Nous notons que selon la description du jeu de données, certains musées présentent une fréquentation cumulée pour l'ensemble des musées du site. Nous décidons tout de même de les garder.

6. Enfin, en appliquant une facette sur toutes les colonnes pour afficher les lignes étoilées, nous obtenons nos Top 10 des musées les plus fréquentés de 2011 à 2021. Nous exportons ces données sous formats csv : 2*_frequentation-des-musees-de-France.csv*.

Fichier : 3_frequentation-totale-mdf-2001-a-2016.csv

Visualisation – Fréquentation des musées parisiens labélisés "Musées de France" : Nous souhaitons afficher l'évolution de la fréquentation des « Musées de France » parisiens de 2001 à 2011 et voir lesquels sont les le plus et lesquels sont le moins fréquentés.

1. À l'aide d'une facette textuelle sur la colonne « VILLE », nous supprimons les lignes des musées n'étant pas situés à Paris.

2. Nous sélectionnons et supprimons toutes les lignes ayant des données manquantes.