Data/IA Simplon

Data Visualisation

Transparence Santé

ligne horizontale

# Exemple d'image

# Introduction

DATA VISUALISATION

02-05 Mars 2020

Objectif: Déterminer la valeur ajouté d’une base de données à partir des statistiques descriptive.

Tuteur: Ulysse Nangbe, Noura Benhajji

Intervenant: Samuel Querichon, Atilla Kaya, Simon Lamontagne

Outil mis en œuvre: Python, Git, VSCode/Jupyter, Dash, MatplotLib, Pygal, LibreOffice.

Compétences visées:

• Acquérir des données, les combiner et les structuré en donnée propre

• Exploiter des données

• Interpréter les données grâce à des outils de visualisation de données

Problématique:

L’inauguration de l’école de l’IA Microsoft de Nancy a eu un grand impact médiatique.

La coloration santé donnée à cette école a suscité l’intérêt d’un groupe de journalistes territoriaux qui veulent passer au crible les données publiées par le ministère de la santé dans le cadre de la transparence entre les entreprises et les acteurs du domaine de la santé.

Les données ne manquent pas, mais le groupe de journalistes n’a pas les armes pour s’attaquer à ce gros volume de données et le faire parler pour faire une synthèse et aller à l’essentiel.

Ils vous sollicitent pour votre expertise en statistiques pour faire une première étude et leur faire une synthèse qui sera la base d’une étude plus approfondie par la suite.

Les données rassemblées par Transparence Santé sont volumineuses et particulièrement intéressantes. Elles permettent en théorie de savoir s'il n'y a pas d'accointances entre certaines entreprises pharmaceutiques et des chirurgiens en hôpitaux par exemple. La base permet de connaître chaque transaction financière à partir de 10 euros, un repas, un voyage ou encore un cadeau.

Le site permet d'effectuer des recherches par entreprise comme par bénéficiaire.

Notre objectif sera de produire des données qui pourront ensuite être représentées sous forme hiérarchique mais l’on a pas vérifié l'exactitude de ces données au travers de tests statistiques.

Pour ce faire on utilise Visual Studio Code avec un sous module de jupyter afin de traiter plus aisément les différents modules pour ne pas avoir à recharger les fichiers CSV à chaque itération.

Pour attaquer la base de donnée et la traiter on prendra Pandas, Plot.ly pour créer nos différents graphiques ainsi que Pygal pour le mapping et enfin Dash pour la visualisations de données.

On a choisi d’utiliser Github pour le versionning et la centralisation de notre travail collaboratif.

# Résultat des stats

# Graphiques

# Interprétations

Pour conclure,

ce projet nous a permis d’apprendre à récupérer des données volumineuse, les traiter avec un langage adapté afin de construire une représentation qui servira à une visualisation clair.

On a su mettre en œuvre nos compétences au service d’un groupe de travail et développé nos accointances personnelle et ainsi entraîner une dynamique efficace.

Nous avons tenu compte de la difficulté a déterminer si les corrélations effectuées sont véritablement fiable étant donnée l’absence de test statistique.

Par ailleurs la Base de Donnée étant elle même sujet à une vérification arbitraire et rempli à la discrétion des laboratoires. Elle ne peut être à elle seul un éléments pertinent à la lecture de la transparence du lien d’intérêt entretenue par les entreprises privé vis à vis du secteur de la santé.

Il nous reste encore beaucoup à faire avec ces trois fichiers contenant plus de 20 millions de lignes. Notamment obtenir des test statistique probant ainsi qu’un “nettoyage” plus précis et ainsi envisager une intégration sous forme de machine learning.