Charles DELEMAZURE
Hugo DEBES
Aurélien DELICOURT

ESILV

Rapport projet Python (TD9-10) :

Nettoyage des données :

Pour commencer le projet, nous avons choisi Google Colab comme plateforme de travail. Elle permet d'avoir accès à de nombreux outils simplifiant le travail d'équipe et le partage d'idée. Nous avons également choisi de rédiger un document Word d'idées, de tâches à faire et d'état de leur avancement. Celui-ci nous permettait de nous coordonner et évitait la répétition du travail.

Le premier objectif était de parvenir à charger le dataset et de le prétraiter. Cela consiste à nettoyer les valeurs indésirables nuisant la qualité de nos graphiques. Plusieurs techniques ont été mises en place comme la suppression des valeurs nulles ou de colonnes répétitives. Dans cette partie nous avons aussi enrichi le dataset avec des informations combinés de plusieurs colonnes. Nous avons calculé le prix au mètre carré. La création de cette nouvelle donnée, qui est rapidement calculable pour chaque région géographique, nous permet d'obtenir de nouvelles conclusions et visuels intéressants ce qui rend notre étude plus pertinente. Une fois cette partie achevé, notre dataset était utilisable pour notre étude. Parmi les nombreux problèmes mineurs rencontrés pendant le nettoyage, on peut noter la présence de colonnes vides ou quasi-vide ou ayant une information redondante qui sont inutilisables. Il a fallu trouver un moyen de cibler ces colonnes et d'entreprendre des mesures de modification ou de suppression. Un exemple de modification dans le but de supprimer des informations redondantes est l'implémentation du code Rivoli ou encore le code des communes.

Un problème que nous avons rencontré était l'impossibilité de classer les valeurs par date. En effet les dates étaient représentées par des String de la forme JJ/MM/AA. Lors d'un tri nous n'obtenions pas l'ordre chronologique. Pour remédier à ce problème nous avions eu une première approche où nous transformions les dates au format américain. La solution était viable mais demandait un certain temps de calcul. Nous avons finalement opté pour un changement de donnée de String en Date Time de la colonne. Le format était donc plus intuitif et rapide à mettre en œuvre.

Un second problème rencontré au fur et à mesure du nettoyage fut la présence de données incohérentes qui faussaient nos graphiques. Ces valeurs étaient anormalement élevées et souvent en fin d'année civile. Autour du 23 décembre nous avions 1000 lignes avec la valeur de 160 000 000 d'euros de valeur foncière. Pour pallier ce problème nous avons cherché quelles pouvaient être les raisons de ces anomalies. Nous avons estimé que cela correspondait à des complexes immobiliers ou commerciaux s'étalant sur plusieurs communes. Nous avons décidé de regrouper les entrées uniques dans un nouveau dataframe mais cela n'en éliminait qu'une partie. Finalement nous n'avons pas eu d'autre choix que de supprimer ces valeurs de notre dataframe afin d'avoir des graphiques non parasités par des valeurs immensément grandes.

Enfin, nous avons aussi dû faire face à de nombreuses données manquantes ou nulles. C'est un problème majeur car cela peut rendre les colonnes associées non

pertinentes dans leur ensemble. Nous avons dû faire des choix stratégiques tel que supprimer certaines données, attribuer des valeurs nulles à certaines lignes pour garder les colonnes essentielles.

Nous avons choisi d'écrire le nouveau dataset dans un fichier csv à la fin de la procédure de nettoyage afin de ne pas avoir à la réaliser à chaque lancement du notebook et ce choix nous a permis de tous pouvoir télécharger ce dataset propre pour pouvoir derrière travailler avec un même fichier.

Interprétation des données:

Pour l'interprétation des données, nous avons réfléchi à de nombreuses idées de visuels afin de mettre en évidence les informations du dataset. Un gros challenge de cette interprétation était d'associer nos idées aux bons graphiques pour rendre l'étude la plus lisible possible. Les valeurs foncières et leur analyse sont un sujet qui nous étaient étrangers au début de ce projet. La première étape consistait à nous documenter afin de savoir quels étaient les chiffres clés et quelles sont les informations pertinentes à mettre en évidence lors d'une analyse foncière nationale globale.

Un deuxième challenge était de produire les visuelles pour illustrer nos conclusions. Notre seule référence en la matière était le TD 8 de langage Python et ses nombreux visuels sur l'épidémie. Il s'agissait d'une excellent base de départ mais nous voulions aussi des visuels originaux. Cette recherche nous a permis d'explorer d'autres pistes de conclusion, c'est pourquoi nous avons entrepris de longues recherches sur le net pour trouver des Template de visuels. La recherche de ces visuels était assez compliquée car les codes et notions de python correspondants étaient, quelquefois, particulièrement denses notamment lorsque nous voulions des visuels dynamiques.

Lors de la création des cartes un problème majeur était le sens de lecture des données qui ne permettait pas d'afficher les valeurs du département sur sa position géographique. Nous avons donc implémenté une fonction pour classer notre dataset dans l'ordre de lecture de la carte afin de pouvoir obtenir une correspondance entre les valeurs des départements et leur position.

Nous sommes fiers du résultat final de notre qui nous a demandé beaucoup d'investissement de temps. Les rendus visuels nous semblent appropriés et nous sommes parvenus à créer plusieurs cartes différentes. Enfin nous avons implémentés un visuel dynamique permettant une vision claire de l'évolutions au cours du temps.

Charles DELEMAZURE
Hugo DEBES
Aurélien DELICOURT

Django:

La partie Django fut assez compliquée car nous étions très limités dans le temps pour comprendre et implémenter les notions. Nous avons passé beaucoup de temps en recherche bibliographique sur le net afin de bien comprendre l'utilisation de Django. Nous n'avons pas réellement trouvé des explications que nous étions en mesure de comprendre et d'implémenter dans un temps réduit ce qui est assez frustrant pour notre équipe. Django est un Framework très performant mais nous n'avons pu l'utiliser comme nous le souhaitions. Nous avons, finalement, opté pour une implémentation en html pour l'affichage dynamique qui était plus simple à mettre en place dans l'intervalle de temps qui restait à notre disposition.

Contributions:

Notre méthode de travail sur Google Colab nous a permis de pouvoir faire des sessions de travail ou nous étions tous les trois présent et en appel sur Zoom simultanément. Une grosse partie de l'avancement du projet s'est effectué de la sorte. Finalement nous n'avons pas vraiment défini de tache allouée à chaque membre donc il est assez délicat d'associer un pourcentage à la contribution de chaque membre mais si nous devions être honnêtes vis à vis du travail effectué par chacun nous serions d'accord pour estimer nos participations comme-ci :

-Nettoyage des données : Hugo 35% Charles 33% Aurélien 32%
-Interprétation des données : Hugo 33% Charles 33% Aurélien 33%
-Django : Hugo 33% Charles 35% Aurélien 32%
-Rapport : Hugo 33% Charles 33% Aurélien 34%

Perspective :

Si nous devions poursuivre ce projet, nous nous concentrerions sur deux points :

1) Un approfondissement sur Django pour rendre plus dynamique le projet
2) Une étude corrélées avec les données Covid pour mesurer plus précisément son impact sur le secteur immobilier