Jérome PIGEON

Outils de programmation pour la science des données

8PRO408

MINIPROJET

Travail présenté à

M. Habba'S Ngodjou Doukaga

UQAC Université du Québec à Chicoutimi

Le 10 décembre 2025

1. SYNTHÈSE

Dans le cadre de mon travail de fin d'année, j'ai décidé de faire l'analyse d'un fichier csv contenant 55 500 données à l'aide de Jupiter notebook. Ces données fictives, générées sur ordinateur, représentent des données médicales de patients ayant fréquenté des hôpitaux.

Dans un premier temps, il est intéressant de mentionner que les données contenues dans le fichier étaient assez ''salles''. En effet, la colonne des noms était sous une mauvaise forme. De plus, il y avait des cellules qui contenaient des montants en négatif. Enfin, les noms d'hôpitaux dans le fichier contenaient des doublons et il y avait plusieurs lignes de données qui étaient en double. Les données du fichier csv ont dû, donc, être nettoyées et uniformisées pour permettre une bonne analyse des données.

Suite à l'analyse détaillée des données, voici les constats auxquels j'arrive. Tout d'abord, il est très clair que les données ne sont pas réelles et sont le fruit d'une génération automatique aléatoire par ordinateur. En effet, suite à mes analyses, j'ai remarqué que presque toutes les fréquences et que les proportions de chaque colonne sont toujours uniformément divisées entre les valeurs possibles. Un exemple, par exemple, le sexe des patients qui est de 50% d'hommes et de 50% de femmes ou le type sanguin des patients qui est d'environ 12.5% peut importe le type sanguin (A+, A-, etc.) Ceci est très irréaliste considérant que dans la vraie vie certains groupes sanguins comme ''O-'' sont très rares. De plus, encore une fois, les données contenues dans le fichier csv sont tellement trop bien réparties peu importe la colonne observée que, par exemple, il y a 33% de cas normaux, 33%

anormaux et 33% non-conclusifs. Et ces cas sont toujours uniformément répartis peu importe la deuxième dimension analysée. Par exemple, les proportions de cas anormaux pour les hommes et les femmes sont exactement les mêmes. On constate, par exemple, que la durée moyenne d'hospitalisation selon la condition médicale est toujours d'environ 15.5 jours, c'est-à-dire la même que la quantité moyenne de jours d'hospitalisation sur le total des données. Un autre exemple serait de dire que la proportion de patients ayant reçu l'une des 5 médications mentionnées dans les données est de 20% par médicament, ce qui est, encore une fois, trop parfait. Il en allait de même pour les proportions de patients selon le fournisseur d'assurance.

Ainsi, on constate que les proportions sont toujours trop parfaites peu importe la catégorie et qu'il est, donc, impossible de faire de réelle conclusion entre deux ou plusieurs dimensions qui pourraient avoir une corrélation sur une donnée. Malgré cela, effectuons un résumé synthèse rapide des données.

Tout d'abord, au niveau de l'analyse du profil des patients, on constate que ceux-ci sont âgés d'entre 20 et 79 ans avec une moyenne de 51.54 ans environ et qu'il y a environ autant d'hommes que de femmes. De plus, les types sanguins ‘’A+’’ et ‘’A-‘’ sont légèrement plus fréquents dans les données recueillies (mais de façon très négligeable). Enfin, on ne constate pas de lien entre l'une de ces données et une autre (par exemple le sexe et l'âge).

Au niveau de l'analyse hospitalière des données, on constate que, de façon négligeable encore une fois, le diabète et l'arthrite semblent être les deux conditions médicales les plus fréquentes. De plus, l'arthrite semble être un peu plus fréquente chez les femmes et l'asthme un peu plus chez les hommes. De plus, comme mentionné plus tôt, la durée moyenne

d’hospitalisation d’un patient est de 15.5 jours et la condition médicale du patient ne semble pas avoir de lien avec la durée.

Il est possible de faire un constat. En effet, selon les données, on constate que les clients semblent fréquenter en grande partie les hôpitaux Smith. Cela semble se refléter considérant que le docteur Michael Smith est le docteur qui a soigné le plus de clients.

Au niveau de l’analyse financière de chaque client, on constate que le cout moyen par client est de 25 546.24$ avec la majorité des données se situant entre environ 12 243.72$ et 37 819.86$. On constate que le Medicare semble être l’assureur qui, en moyenne, génèrerait des dépenses un peu plus élevées que les autres fournisseurs (mais ceci est négligeable). De plus, en moyenne, les hommes engendrent 2 000$ de plus que les femmes.

Enfin, au niveau de l’analyse des cas anormaux, mes analyses ne permettent pas de conclure un lien entre les cas déclarés comme anormaux et n’importe quelle donnée. En effet, l’âge moyenne des cas anormaux reste autour de celle de l’ensemble des données, la médication prescrite aux cas anormaux reste de 20% pour chaque médicament, il ne semble pas y avoir de groupe sanguin qui est fréquemment déclaré comme anormal et, enfin, il ne semble pas y avoir de condition médicale créant une situation anormale de façon plus fréquente.

En conclusion, nous pouvons dire, encore une fois, qu’il est clair que les données fournies dans le csv sont clairement générées par ordinateur et les données sont clairement uniformes, ce qui ne permet pas de faire de lien clair entre une colonne de donnée et une autre dimension.

2. LIENS

Voici le lien vers la page GitHub de mon travail : https://github.com/jpigeonUQAC/MiniProjetData

Voici le lien vers la page Streamlit hébergée : https://miniprojetdata-8gghdulowcjbhvdteugpdu.streamlit.app