Introduction

Le Dr Ignaz Semmelweis était un médecin hongrois né en 1818, qui travaillait à l'Hôpital Général de Vienne. Autrefois, on pensait que les maladies étaient causées par des "mauvaises odeurs" ou des esprits malins. Mais au 19e siècle, les médecins ont commencé à s'intéresser davantage à l'anatomie, à pratiquer des autopsies et à formuler des arguments basés sur des données. Le Dr Semmelweis soupçonnait que quelque chose ne se passait pas bien dans les procédures de l'Hôpital Général de Vienne. Il voulait comprendre pourquoi tant de femmes dans les services de maternité mouraient de fièvre puerpérale (une infection bactérienne qui survient généralement après l'accouchement).

Aujourd'hui, vous allez devenir le Dr Semmelweis. Vous allez analyser les mêmes données collectées entre 1841 et 1849.

1. Lire les données (monthly_deaths.csv, annual_deaths_by_clinic.csv)
2. Data Exploration:
    a. Quelle est la forme des DataFrames df_yearly et df_monthly ?
    b. Combien y a-t-il de lignes et de colonnes ?
    c. Quels sont les noms des colonnes ?
    d. Quelles années sont incluses dans le jeu de données ?
    e. Y a-t-il des valeurs NaN ou des doublons ?
    f. Quelle était la moyenne du nombre de naissances qui avaient lieu par mois ?
    g. Quelle était la moyenne du nombre de décès qui avaient lieu par mois ?
3. Visualiser le nombre total de naissances et de décès au fil du temps:
    a. Tracer les données mensuelles sur des axes jumeaux:
        i. Formater l'axe des abscisses en utilisant des localisateurs (locators) pour les années et les mois (Indice : nous l'avons fait dans le notebook Google Trends).
        ii. Définir l'intervalle sur l'axe des abscisses de manière à ce que les lignes du graphique touchent les axes des ordonnées.
        iii. Ajouter des lignes de grille.
        iv. Utiliser le bleu ciel et le cramoisi pour les couleurs des lignes.
        v. Utiliser un style de ligne en pointillés pour le nombre de décès.
        vi. Changer l'épaisseur des lignes à 3 pour les naissances et 2 pour les décès respectivement.
        vii. Remarquez-vous quelque chose vers la fin des années 1840 ?
4. Les données annuelles réparties par clinique:
    a. Utilisez Plotly pour créer des graphiques en ligne des naissances et des décès des deux cliniques différentes de l'Hôpital Général de Vienne.

        i. Quelle clinique est plus grande ou plus occupée en se basant sur le nombre de naissances ?
        ii. L'hôpital a-t-il eu plus de patients au fil du temps ?
        iii. Quel a été le nombre le plus élevé de décès enregistrés dans la clinique 1 et la clinique 2 ?
    b. Calculer la proportion de décès dans chaque clinique.
    c. Tracer la proportion des décès annuels par clinique.
        i. Quelle clinique a une proportion de décès plus élevée ?
        ii. Quel est le taux de décès mensuel le plus élevé dans la clinique 1 par rapport à la clinique 2 ?
5. L'effet du lavage des mains (The Effect of Handwashing):
   "Le Dr Semmelweis a rendu le lavage des mains obligatoire à l'été 1847 ('1847-06-01'). En fait, il a ordonné aux gens de se laver les mains avec du chlore (au lieu de l'eau)."
    a. Ajouter une colonne appelée "pct_deaths" à df_monthly qui contient le pourcentage de décès par rapport aux naissances pour chaque ligne.
    b. Créer deux sous-ensembles à partir des données df_monthly : avant et après que le Dr Semmelweis ait ordonné le lavage des mains.
    c. Calculer le taux de mortalité moyen avant juin 1847.
    d. Calculer le taux de mortalité moyen après juin 1847.
    e. Créer un DataFrame qui contient le taux de mortalité moyen mobile (rolling average)sur 6 mois avant l'obligation du lavage des mains.
    f. Ajouter 3 lignes distinctes au graphique : le taux de mortalité avant le lavage des mains, après le lavage des mains, et la moyenne mobile sur 6 mois avant le lavage des mains.
    g. Afficher le taux de mortalité mensuel avant le lavage des mains sous forme de ligne noire fine en pointillés.
    h. Afficher la moyenne mobile comme une ligne cramoisie plus épaisse.
    i. Afficher le taux après le lavage des mains comme une ligne bleu ciel avec des marqueurs ronds.
    j. Quelle était la moyenne du pourcentage de décès mensuels avant le lavage des mains ?
    k. Quel était le pourcentage moyen de décès mensuels après que le lavage des mains ait été rendu obligatoire ?
    l. De combien le lavage des mains a-t-il réduit la chance moyenne de mourir lors de l'accouchement en termes de pourcentage ?
    m. Comment ces chiffres se comparent-ils à la moyenne pour toute la décennie des années 1840 que nous avons calculée plus tôt ?
    n. Combien de fois les chances de mourir après le lavage des mains sont-elles plus faibles qu'avant ?
    o. Utiliser des diagrammes en boîte (box plots) pour montrer comment le taux de mortalité a changé avant et après le lavage des mains:

    i. Utiliser la fonction .where() de NumPy pour ajouter une colonne à df_monthly qui indique si une date particulière est avant ou après le début du lavage des mains.
    ii. Ensuite, utiliser Plotly pour créer un diagramme en boîte (box plot) des données avant et après le lavage des mains.
    iii. Comment les statistiques clés telles que la moyenne, le maximum, le minimum, le 1er et le 3e quartile ont-elles changé à la suite de la nouvelle politique ?
p. Utiliser une estimation de densité par noyau (KDE) pour visualiser une distribution lissée :
    i. Utiliser la fonction **.kdeplot()** de **Seaborn** pour créer deux estimations de densité par noyau (KDE) du **pct_deaths**, une pour avant le lavage des mains et une pour après.
    ii. Utiliser le paramètre **shade** pour donner à vos deux distributions des couleurs différentes.
    iii. Quelle faiblesse remarquez-vous dans le graphique lorsque vous utilisez simplement les paramètres par défaut ?
    iv. Utiliser le paramètre **clip** pour résoudre le problème.
q. Utiliser un **T-test** pour montrer la signification statistique :
    i. Utiliser un t-test pour déterminer si les différences dans les moyennes sont statistiquement significatives ou simplement dues au hasard.
    ii. Si la valeur **p** est inférieure à 1 %, alors nous pouvons être sûrs à 99 % que le lavage des mains a eu un impact sur le taux de mortalité mensuel moyen.
    iii. Importer stats de **scipy**.
    iv. Utiliser la fonction **.ttest_ind()** pour calculer la statistique t et la valeur p.
    v. La différence dans la proportion moyenne de décès mensuels est-elle statistiquement significative au niveau de 99 % ?