Lab 1: Line and Scatter Plots

Objectifs du TP

Dans ce TP, vous utiliserez vos nouvelles connaissances pour proposer des solutions à des scénarios réels. Pour réussir, vous devrez importer des données dans Python, répondre à des questions en utilisant ces données et effectuer des tâches de visualisation de données, telles que la génération de graphiques linéaires pour comprendre les tendances dans les données.

1 Background

Dans ce TP, vous apprendrez à créer de nombreux types de graphiques avancés. Voici quelques commandes rapides que vous pouvez utiliser pour réaliser ces graphiques :

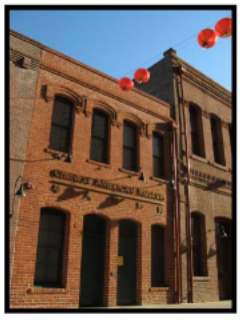
- **sns.lineplot** - Les graphiques linéaires sont les meilleurs pour montrer les tendances sur une période de temps, et plusieurs lignes peuvent être utilisées pour montrer les tendances de plusieurs groupes.
- **sns.scatterplot** - Les nuages de points montrent la relation entre deux variables continues; s'ils sont colorés, nous pouvons également montrer la relation avec une troisième variable catégorielle.
- **sns.regplot** - Inclure une ligne de régression dans le nuage de points facilite la visualisation de toute relation linéaire entre deux variables.
- **sns.histplot** - Les histogrammes montrent la distribution d'une seule variable numérique.
- **sns.countplot** - Affiche les décomptes des observations dans chaque catégorie à l'aide de barres.
- **sns.barplot** - Les diagrammes en barres sont utiles pour comparer des quantités correspondant à différents groupes.

2 Premier Scénario

Vous avez récemment été embauché pour gérer les musées de la ville de Los Angeles. Votre premier projet se concentre sur les quatre musées illustrés dans les images ci-dessous.

Avila Adobe

Firehouse Museum



Chinese American Museum

America Tropical Inventive Center

Vous utiliserez les données du portail de données de Los Angeles, qui suivent mensuellement le nombre de visiteurs de chaque musée.

2.1 Charger les données

- Votre première tâche consiste à importer et configurer les bibliothèques Python dont vous avez besoin pour compléter cet exercice.
- Ensuite, lisez le fichier de données des visiteurs des musées de Los Angeles ('museum_visitors.csv') dans un data-frame `museum_data`. De plus, le nom de la colonne à utiliser comme index est "Date".

2.2 Examiner les données

Dans cette partie, nous allons examiner le jeu de données, pour comprendre ce que vous avez "entre les mains".

- Quelle est la taille de la forme du jeu de données `museum_data` ?
- Affichez les 7 premières et dernières lignes du data frame `museum_data`. Indice : Vous pouvez utiliser df.head(), df.tail()
- Listez les noms des caractéristiques incluses dans ce jeu de données et leur type de données.
- Vérifiez et supprimez toutes les lignes contenant des valeurs NaN, ou manquantes, dans le Data-Frame `museum_data`.
- Vérifiez et supprimez toutes les lignes en double dans le Data-Frame `museum_data`. Indice : cherchez les méthodes duplicated() et drop_duplicates() de la bibliothèque pandas.

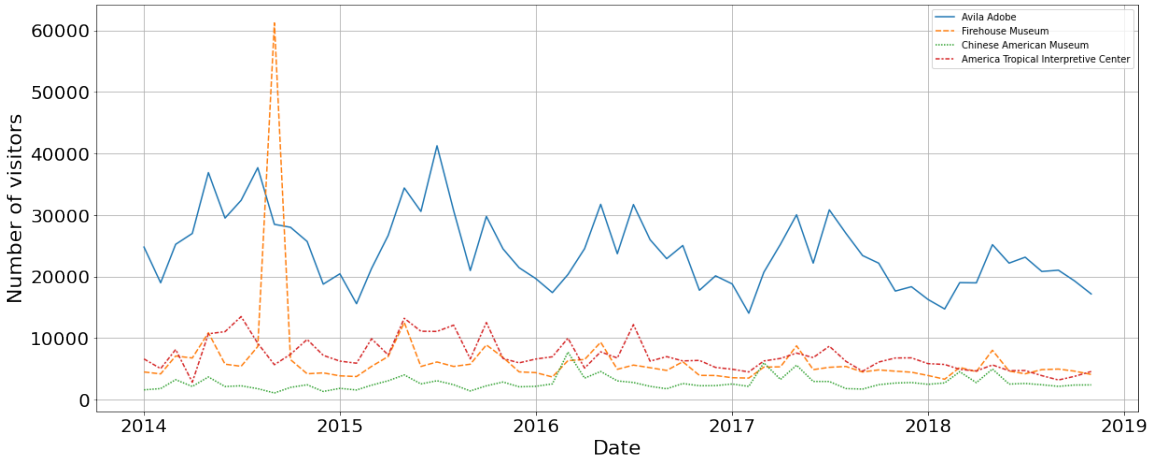
3 Résumer les données

- Fournissez une statistique résumée, qui inclut la valeur moyenne, la valeur minimale, et la valeur maximale pour chaque caractéristique, ainsi que son écart-type, et les percentiles de 25%, 50% (c'est-à-dire la médiane), et 75%.

Convaincre le conseil du musée

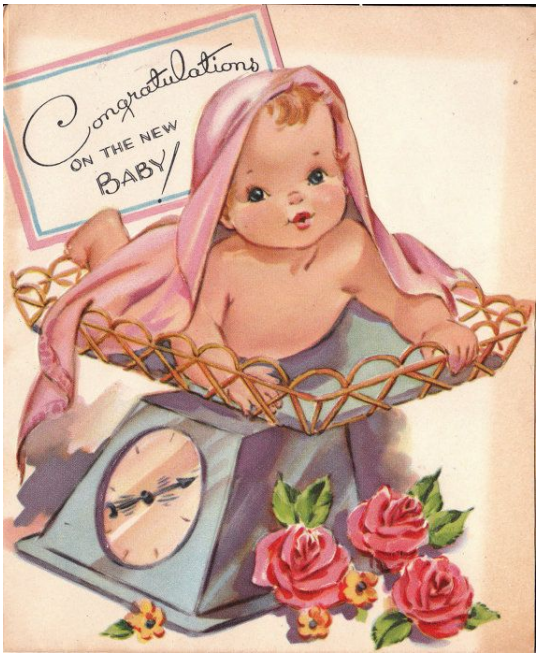
Le Firehouse Museum affirme avoir organisé un événement en 2014 qui a attiré un nombre incroyable de visiteurs, et qu'ils devraient obtenir un budget supplémentaire pour organiser à nouveau un événement similaire. Les autres musées pensent que ces types d'événements ne sont pas si importants, et que les budgets devraient être répartis uniquement en fonction des visiteurs récents lors d'une journée moyenne.

Par conséquent, pour montrer au conseil du musée comment l'événement se compare au trafic régulier de chaque musée, créez un graphique linéaire qui montre comment le nombre de visiteurs de chaque musée a évolué au fil du temps. Votre figure doit comporter quatre lignes (une pour chaque musée) comme indiqué ci-dessous.



Deuxième Scénario

Nous allons en fait pratiquer la visualisation de données en utilisant les données sur les naissances de l'État de Caroline du Nord. Le jeu de données ("ncbirth.csv") qui apparaît dans votre environnement est un grand data frame. Chaque observation ou cas correspond à la naissance d'un seul enfant. De plus, ce jeu de données contient 800 observations (lignes ou cas) et 13 variables (colonnes).



Les variables :

- "*weeks*" représente la durée de la grossesse.
- "*weight*" représente le poids du bébé à la naissance en livres.
- "*premie*" indique si une naissance était prématurée (*premie*) ou à terme.
- "*mature*": le statut de maturité de la mère.

- "*gained*": poids gagné par la mère pendant la grossesse en livres.
- "*gender*": sexe du bébé, féminin ou masculin.
- "*habit*": statut de la mère en tant que non-fumeuse ou fumeuse.

Tâches

1. Écrire le code nécessaire pour charger les données.
2. Analyser la relation entre "*weeks*" et "*weight*" en utilisant la fonction requise de la bibliothèque "seaborn". Incluez les étiquettes des axes avec les unités de mesure et un titre. Y a-t-il une relation positive ou négative entre ces variables ?
3. Faites un graphique montrant les semaines à nouveau sur l'axe des x et la variable "gained" sur l'axe des y (le poids qu'une mère a pris pendant la grossesse). Incluez les étiquettes des axes avec les unités de mesure et un titre.
4. Évaluer visuellement l'association entre le poids des bébés (*weight*) et le gain de poids de la mère (*gained*) ?
5. Coloriez les points du graphique précédent en fonction de la variable "*premie*". Combien de variables sont maintenant affichées sur ce graphique ? Quelle est votre conclusion ?
6. Créez un nouveau nuage de points qui montre l'âge de la mère sur l'axe des x (variable appelée *mage*) et le poids à la naissance des nouveau-nés sur l'axe des y (*weight*). Colorez les points du graphique en fonction du sexe du bébé (variable appelée *gender*). Y a-t-il une relation forte apparente entre l'âge de la mère et le poids de son nouveau-né ?

Questions sur les histogrammes

1. Tracez la distribution de la variable 'habit' (la mère est non-fumeuse ou fumeuse) en fonction du statut de maturité des mères (mature). Que pouvez-vous conclure ?
2. Tracez la distribution de la durée de grossesse (variable appelée *weeks*) en fonction de la classification prématurée (premie ou full-term). Ajoutez un titre et des étiquettes d'axes. L'axe des y est étiqueté "count". Que pouvez-vous conclure ?
3. Faites une distribution des naissances de nouveau-nés par sexe de l'enfant en fonction du statut de maturité des mères (mature). Que pouvez-vous conclure ?