



Data Visualisation Statistiques descriptives

Exercices

Visualisation d'un nuage de points et exploration des variables

Pour cette partie, nous allons utiliser le dataset suivante : *house_pricing.csv* que vous pouvez trouver dans le dossier *Data Sources*. Nous allons tenter de comprendre un peu mieux ce dataset par la visualisation

- 1) Importez les librairies qu'il nous faut (pandas, seaborn, matplotlib, numpy)
- 2) Donnez la liste des variables présentes dans ce dataset ainsi que leur nature (sont elles qualitatives, quantitatives, discrète etc...) et leur type (float, int, str etc...)
- 3) En créant un nuage de points, regardez comment se comporte la colonne *LotArea* par rapport au *SalesPrices*
- 4) Affinez votre visualisation en ne gardant uniquement les maisons qui ont un *LotArea* inférieur à 20 000 pieds carrés et un prix inférieur à 500 000\$
- 5) En créant un nuage de points, regardez la relation entre le *LotFrontage* et le *LotArea*
- 6) De la même manière, affinez votre visualisation en ne gardant uniquement les maisons qui ont un *LotFrontage* inférieur à 200 pieds carrés et un *LotArea* inférieur à 100000 pieds carré

Visualisation de relations continues

Pour cette partie, nous utiliserons le dataset suivant : *sales_predictions.csv* que vous pouvez trouver dans le dossier *Data Sources*.

- 1) Importez les librairies qu'il vous faut
- 2) En utilisant *relplot()*, construisez un graphique qui va vous permettre de voir l'évolution des prix par rapport au temps. Que pouvez vous voir ?
- 3) Corrigeons le problème de visualisation, en utilisant la fonction *.sample()* de Pandas, prenez un échantillon de 50 éléments dans votre dataset
- 4) Retentez de faire votre visualisation, créez une figure de taille (20,6). Que voyez vous ?
- 5) En utilisant la fonction *pd.to_datetime()*, convertissez votre colonne *date* en *datetime*
- 6) Retentez une dernière fois votre visualisation.



Visualisation de variables catégoriques

Pour cette partie, nous utiliserons le dataset suivant : *ibm_hr_attrition.csv* que vous pouvez trouver dans le dossier *Data Sources*

- 1) Importez les librairies qu'il vous faut
- 2) En utilisant *catplot()*, construisez un graphique qui vous permette de voir la distribution des personnes qui ont quitté l'entreprise par rapport à leur *JobSatisfaction*
- 3) Changez de graphique et utilisez plutôt un boxplot, que pouvez vous conclure ?
- 4) Peut-on dire grâce aux boxplot si une variable présente des valeurs aberrantes ?
- 5) Isolez les observations présentant des valeurs aberrantes pour la variable *JobSatisfaction* en fonction de la variable indiquant si ils ont quitté l'entreprise ou non
- 6) Utilisez la méthode *.describe* de pandas pour obtenir un tableau récapitulatif des statistiques descriptives du dataset.
- 7) Comparez les individus présentant des valeurs aberrantes pour la variable *JobSatisfaction* avec les statistiques descriptives du dataset, remarquez vous des différences notables ?
- 8) Faites de même avec le *TotalWorkingYears*. Que pouvez vous conclure ?
- 9) Remplacer les valeurs de la colonne *Attrition* par 1 pour *Yes* et 0 pour *No*
- 10) En utilisant un Histogramme, regardez la répartition du taux de départ par *EducationField*

Visualisation d'une distribution

Continuons sur *ibm_hr_attrition.csv*

- 1) Importez les librairies qu'il vous faut
- 2) En utilisant *distplot()*, construisez un graphique qui permette de voir la distribution des revenus par mois chez IBM (*MonthlyIncome*)
- 3) On peut voir que les très haut salaires biaisent notre distribution, essayons de voir la distribution des salaires entre 0 et 5000\$ / mois

Visualisation d'une relation linéaire

Toujours sur *ibm_hr_attrition.csv*

- 1) Importez les librairies qu'il vous faut



- 2) On voudrait connaître la probabilité de partir de la l'entreprise par rapport à la distance par rapport à la maison. Utilisez *lmpplot()* pour visualiser cela. Que pouvez vous conclure ?
- 3) Tentez le cette fois avec le nombre d'années passées avec le Manager (*YearsWithCurrManager*), que pouvez vous conclure ?

Visualisation d'une heatmap

Revenons sur *house_pricing.csv*, et tentons de faire un peu de ce qu'on appelle communément : *feature_engineering*

- 1) Importez les librairies habituelles
- 2) Découpez votre dataset pour ne garder que les 15 dernières colonnes
- 3) Créez une matrice de corrélation avec toutes les variables du Dataset
- 4) Créez une heatmap avec les différentes valeur de corrélation
- 5) Quel est le top 3 des features qu'on devrait garder pour prédire notre prix ?