Travaux Pratiques Analyse Statistique et R 01

Généralités sur R

Alison PATOU
Alison.patou@gmail.com
1/02/2021

Introduction

Ce premier TP sur le langage R va vous permettre de manipuler les données sous R et de créer des visualisations que vous devrez bien évidemment interpréter.

Exercice 1

Info: Nous travaillons sur le dataset **House Pricing** qui contient 1460 lignes et 81 colonnes.

Question 1

- Charger le jeu de données House Princing (fichier house pricing.csv)
  Décrivez les variables de ce dataset (qualitatives, quantitatives, discrète etc…, float, int, str etc…)

Question 2

- Combien y-a-t-il de valeurs différentes dans la colonne *LotConfig?*

Question 3

- En créant un nuage de points, regardez comment se comporte la colonne *LotArea* par rapport au *SalePrices*

Question 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$.
  Ajouter le titre, remplacer les points par des croix, nommer les axes.
  Tracer sur ce graphique une droite horizontale bleue correspondant à la moyenne du *LotArea.*

Question 5

- En créant un nuage de points, regardez la relation entre le *LotFrontage* et le *LotArea.*
  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

Exercice 2

Info: Nous travaillons sur le dataset **IBM HR Attrition** qui contient 1370 lignes et 35 colonnes.

Question 1

- Charger le jeu de données IBM HR Attrition (fichier IBM_HR_Attrition.csv)
  Quel est le ratio des personnes ayant quitté l'entreprise (Attrition = Yes) vs. celles qui sont restées (Attrition = No). Visualisez les résultats.

Question 2

- En utilisant un barplot empilé, regardez la répartition du taux de départ par *EducationField*

Question 3

- Visualiser la répartition des attritions par département. Ajouter un titre, des libellés d'axes et des barres en bleu.
  Comparer avec la répartition des personnes n'ayant pas quitté l'entreprise.
  Les proportions sont-elles conservées ?

Question 4

- Tracer l'histogramme de la répartition des âges du dataset. Quelle tranche d'âge est la plus représentée?

Question 5

- Tracer l'histogramme de la répartition des âges avec les classes d'âges suivantes : [0,20], [20,40], [40,60]

Question 6

- Tracer l'histogramme de densité de la distribution des revenues par mois chez IBM (*MonthlyIncome).*
  Avec le fonction lines(), ajoutez sur ce graphique la courbe de densité en rouge.

Question 7

- On peut voir que les très haut salaires biaisent notre distribution, essayons de voir la distribution des salaires entre 0 et 5000$ / mois

Exercice 3

Info: Nous travaillons sur le dataset **rp2012** qui contient 5170 lignes et 60 colonnes et contient les données du recensement en 2012 sur plusieurs départements.

Question 1

- Importer le package **questionr** et **ggplot2**
  Charger le jeu de données rp2012 avec la fonction *data()*

Question 2

- Visualiser avec la fonction *ggplot()* associée, le nombre de commune par département du dataset (vous devriez avoir besoin de la fonction *data.table()* en amont)

Question 3

- Créer un dataset rp_rhone qui contient seulement les données de rp2012 filtrées sur le Département du Rhône

Question 4

- Avec la fonction *ggplot()* visualiser la répartition des étudiants de ce dataset avec un histogramme

Question 5

- Ajouter un titre à votre graphique ainsi que les titres sur les axes