LMAFY1101 - Exercices - Série 2

Statistiques descriptives

Exercice 1

Pour cet exercice, nous allons utiliser le jeu de de données `iris` disponible ici. Ce sont les mesures en centimètres des variables suivantes : longueur du sépale (`Sepal.Length`), largeur du sépale (`Sepal.Width`), longueur du pétale (`Petal.Length`) et largeur du pétale (`Petal.Width`) pour trois espèces (`Species`) d'iris: `setosa`, `versicolor` et `virginica`.

1. Sauvegardez le fichier "iris.txt" puis importez ces données dans R. [Tuyau]: vous pouvez utiliser le menue RStudio "Import Dataset".
2. Faites le nécessaire pour que lorsque vous examinez la structure des données vous obteniez la sortie suivant.

```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "versicolor","virginica",..: 3 3 3 3 3 3 3 3 3 3
```

3. Faites un résumé de `iris` montrant les statistiques de base pour chaque variable. Voici la sortie que vous devez obtenir.

```
  Sepal.Length   Sepal.Width    Petal.Length   Petal.Width          Species
 Min.   :4.30   Min.   :2.00   Min.   :1.00   Min.   :0.1   versicolor:50
 1st Qu.:5.10   1st Qu.:2.80   1st Qu.:1.60   1st Qu.:0.3   virginica :50
 Median :5.80   Median :3.00   Median :4.35   Median :1.3   setosa    :50
 Mean   :5.84   Mean   :3.06   Mean   :3.76   Mean   :1.2
 3rd Qu.:6.40   3rd Qu.:3.30   3rd Qu.:5.10   3rd Qu.:1.8
 Max.   :7.90   Max.   :4.40   Max.   :6.90   Max.   :2.5
```

4. Donnez un tableau récapitulatif qui montre les moyennes des quatre variables numériques (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`) pour chacune des trois espèces. Voici la sortie que vous devez obtenir.

```
     Species Sepal.Length Sepal.Width Petal.Length Petal.Width
1 versicolor         5.94        2.77         4.26       1.326
2  virginica         6.59        2.97         5.55       2.026
3     setosa         5.01        3.43         1.46       0.246
```

5. Créez un histogramme qui montre la distribution `Sepal.Width`.

6. Faites deux graphiques différents, de votre choix, qui permettent de visualiser la distribution de `Sepal.Width`. Faites la même analyse, mais cette fois-ci séparément pour chaque espèce.
7. Créez un graphique qui montre `Petal.Width` en fonction de `Sepal.Width`. Faites la même analyse, mais cette fois-ci séparément pour chaque espèce.
8. Sur base de `Petal.Width` créez une nouvelle variable `Petal.WidthC` qui divise les données en trois parts (presque) égales (de même effectif). [Tuyau]: vous pouvez utiliser les fonctions `quantile` et `cut`.
9. Faites la même chose avec `Sepal.Width` puis étudiez le lien entre les deux variables `Petal.WidthC` et `Sepal.WidthC`.

Exercice 2

Cet exercice a pour prérequis d'avoir résolu le dernier exercice de la séance 1 qui porte sur la base de données `diamonds` disponible dans le package `ggplot2`. Si ce n'est pas déjà fait, commencez par charger ce dernier.

— ***Variables qualitatives***

1. Complétez le tableur de contingence suivant qui fournit les effectifs des différentes qualités de découpage (`cut`).

Fair	Good	Very Good	Premium	Ideal
1610				

Transformez ces chiffres en des pourcentages et représentez ces derniers graphiquement.

2. Complétez le tableur suivant qui fournit les pourcentages des différentes couleur (`color`) pour chaque découpage (`cut`).

color/cut	Fair	Good	Very Good	Premium	Ideal
D	10.12				
E	13.91				
F	19.38				
G	19.50				
H	18.82				
I	10.87				
J	7.39				
Sum	100	100	100	100	100

Parmi les diamants ayant une découpe "Ideal", quelle est la proportion de ceux qui ont la meilleure couleur possible (couleur D)?

3. Représentez les pourcentages du tableau précédent à l'aide d'un graphique de votre choix.

— ***Variables quantitatives***

Nous allons, dans cette partie, nous intéresser principalement aux variables `price` et `carat`.

4. Calculez quelques résumés numériques pour les variables `price` et `carat`.
5. À l'aide de la fonction `aggregate`, obtenez les résumés numériques de la variable `price` (1) en fonction de la couleur et (2) en fonction de la découpe.
6. À l'aide d'un diagramme à points, faites-vous une première idée des valeurs prises par les variables `price` et `carat`.
7. Étudiez comment le prix change (1) en fonction de la couleur, (2) en fonction de la découpe, et (3) en fonction des deux (couleur et découpe). Utilisez des boxplots pour répondre à cette question.
8. Affiner vos conclusions faites au point (2) précédent, en utilisant, à présent, des courbes de densité ?

— ***Relation entre deux variables numériques***

9. Étudiez graphiquement le lien entre les variables `carat` (en abscisse) et `price` (en ordonnée) ?
10. Étudiez graphiquement le lien entre les variables "$log(carat)$" et "$log(price)$", pour les découpes "Fair" et "Ideal" séparément ? Que pouvez-vous dire comme remarques ?