المدرسة الوطنية للعلوم التطبيقية - بني ملال

ⵜⴰⵙⴷⴰⵡⵜ ⵜⴰⵏⴰⵎⵓⵔⵜ ⵏ ⵜⵙⵙⵏⴰⵡⵉⵏ ⵜⵉⵜⵜⴰⵡⵉⵏ - ⴱⵏⵉ ⵎⵍⴰⵍ

Ecole Nationale des Sciences Appliquées - Béni Mellal

ANALYSE DES DONNÉES ET BIG DATA

1

Filière : Transformation Digitale Industrielle (TDI)

Pr. ESSWIDI AYOUB

A.U. 2024-2025

1

PLAN PRÉVISIONNEL ET OBJECTIFS

- Analyse de données
  - Prétraitement et nettoyage des données
  - Visualisation des données (EDA).
  - Feature Selection, Feature engineering, …
  - Analyse des données multimédia, Analyse du Web et des réseaux sociaux
- Science de données et Data Mining
  - Introduction et types d'apprentissage
  - Algorithmes de ML
  - Algorithmes de Deep Learning
  - Exemple d'application en IA générative
- Big Data
  - Introduction au Big Data: définition, caractéristiques, Historique, domaine d'application.
  - Architectures du Big Data
  - Ecosystèmes : Hadoop et Map-Reduce, Apache spark, kafka, HBASE, …

https://www.youtube.com/watch?v=UYseSZJR-wo&list=PLirv9XJLkgiULbuDa0GpAG65gHY1I-ohU&index=3

2

2

DATA PREPROCESSING (PRÉPARATION DES DONNÉES)

"The purpose of computing is insight, not numbers"

-Richard Wesley Hamming

Education: BS, mathematics, University of Chicago, 1937; MA, mathematics, University of Nebraska, 1938; PhD, mathematics, University of Illinois, 1942.

Professional Experience: Bell Telephone Laboratories; Naval Postgraduate School.

Honors and Awards: IEEE Computer Society Pioneer Award, 1980; fellow, IEEE.
[IEEE named their medal for exceptional contributions to the information sciences and systems after Richard Hamming.]

Richard W. Hamming's invention of error-correcting codes for computers was the result of fortune favoring the prepared mind-and of frustration.

3

3

POURQUOI L'ANALYSE EXPLORATOIRE DE DONNEES (EDA)

- ✓ Mieux comprendre les données
- ✓ Découvrir la distribution des données
- ✓ Détecter les anomalies et les aberrations
- ✓ Vérifier les relations
- ✓ Vérifier la qualité des données
- ✓ Tester ou vérifier des hypothèses
- ✓ Découvrir des structures/patterns cachées dans les données

- ➢ l'analyse des données consiste à fournir des insights et des graphiques aux décideurs

4

4

STATISTIQUES

- Les statistiques descriptives
  - permettent de résumer les informations relatives les observations
  - organiser et de simplifier les données pour mieux les comprendre

- Les statistiques inférentielles
  - utilisent les observations (données d'un échantillon) pour faire des déductions des inférences sur l'ensemble de la population
  - généraliser d'un échantillon à une population

5

STATISTIQUES

Les statistiques descriptives

- Nombre / fréquence
- Moyenne,
- Médiane,
- Mode
- Écart Type,
- Variance,
- Quartile
- Corrélation,
- KPIs (key performance indicator)
- ……

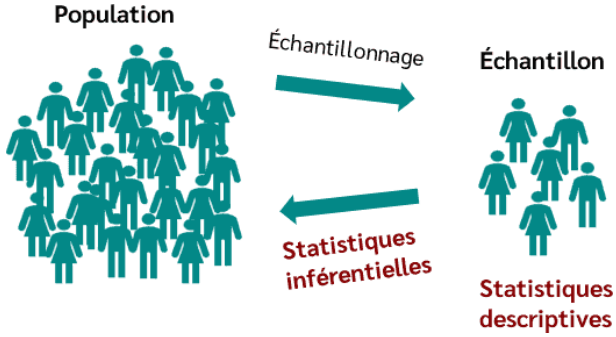
- + Exploration à travers des graphiques de visualisation

6

STATISTIQUES

Statistique inférentielle

- Elle utilise la théorie des probabilités et des modèles statistiques pour estimer les paramètres de la population et tester des hypothèses sur la population à partir de données d'échantillonnage.



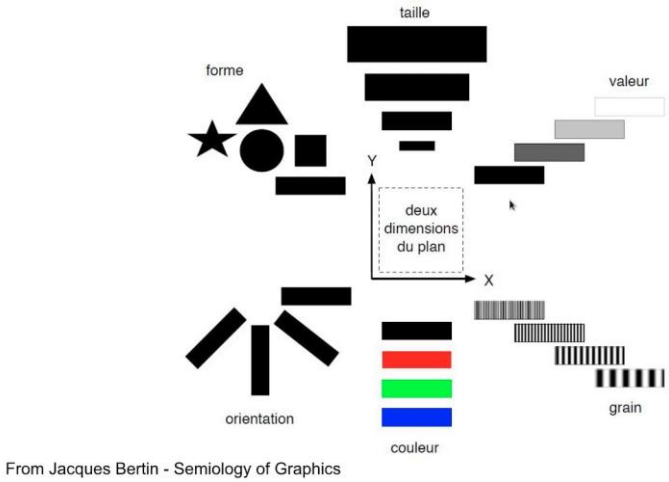
NB: Inutile pour une analyse exploratoire

7

7

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

canal et mark



From Jacques Bertin - Semiology of Graphics

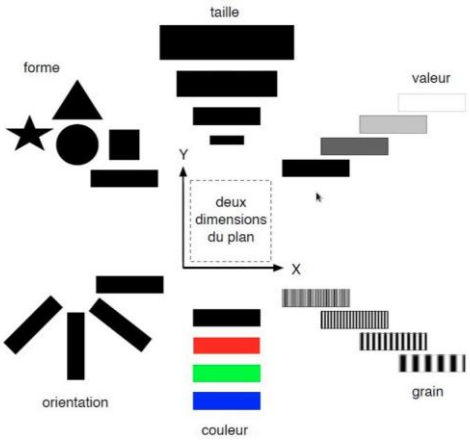
8

8

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

canal

- Un canal visuel est un moyen de contrôler l'apparence graphique des marques, proportionnellement à leurs dimensions.
- Les canaux changent l'apparence en fonction de l'attribut
- Un canal est une variable visuelle
  - Couleur
  - Longueur
  - Position
  - Angle



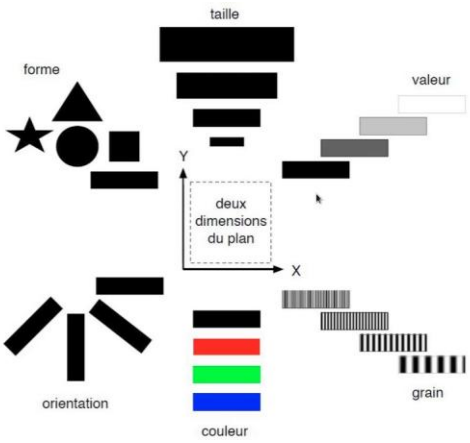
9

9

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

marques

- Une marque est un élément graphique de base dans une image.
- Les marques sont des objets géométriques primitifs utilisés pour présenter un dataset
  - Points
  - Lignes
  - Régions
  - …

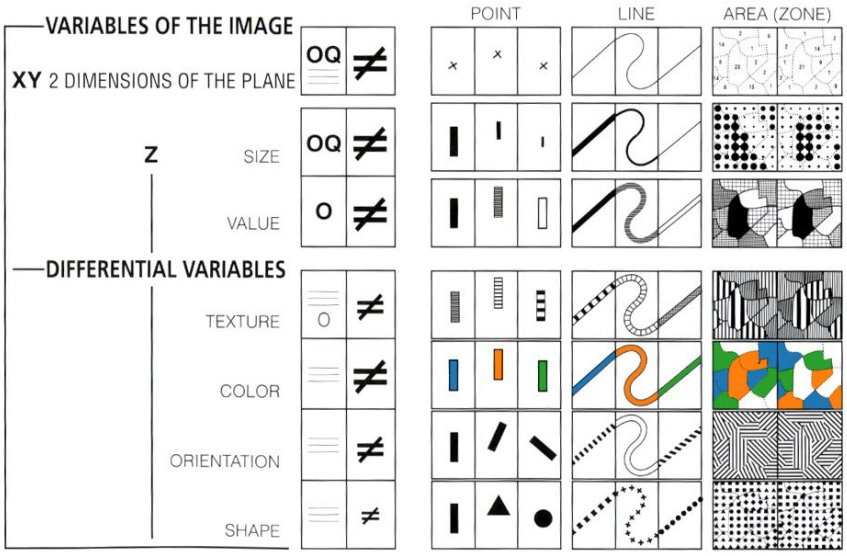


10

10

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

canal et mark



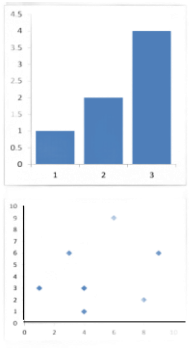
11

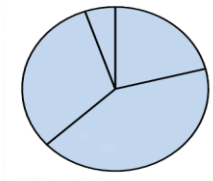
11

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

canal et mark - Exemples

- Les figures sont une combinaison de marques et de canaux





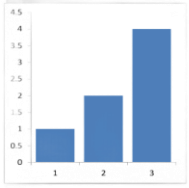
12

12

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

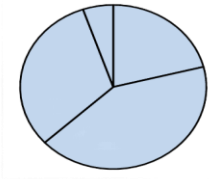
canal et mark - Exemples

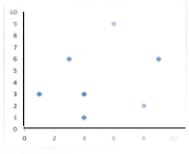
❑ Les figures sont une combinaison de marques et de canaux



1 Mark = Rectangle
1 Channel = Length of longest side

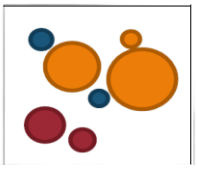
1 Mark = Circle segment
1 Channel = Angle





1 Mark = Diamond shape
2 Channels = X position, Y position

1 Mark = Circle
4 Channels:
X position
Y position
Area
Colour



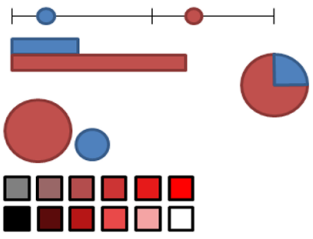
13

13

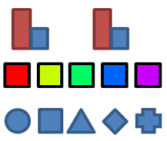
ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

types de canaux

- Quantitative
  - Position on scale
  - Length
  - Angle
  - Area
  - Colour (saturation)
  - Colour (lightness)



- Qualitative
  - Spatial Grouping
  - Colour (hue)
  - Shape

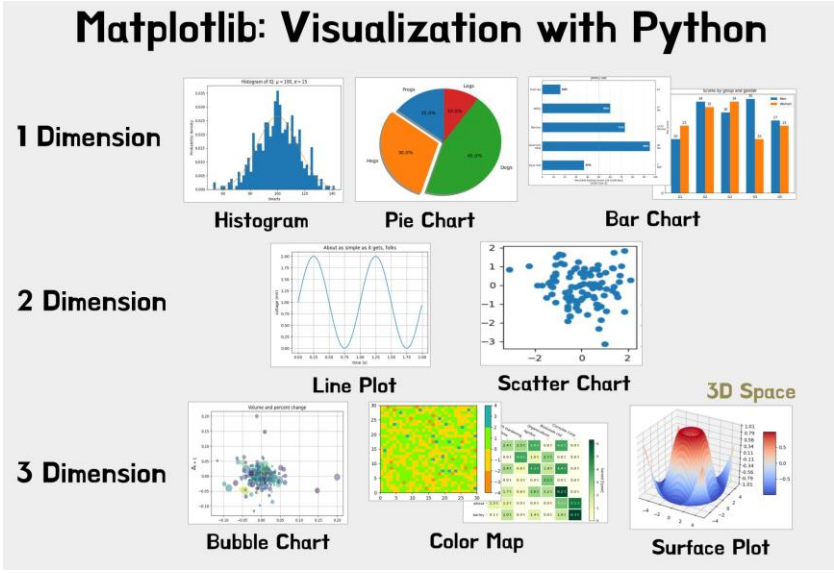


14

14

ANALYSE EXPLORATOIRE DES DONNÉES (EDA)

Visualisation



15

15

OUTILS ET BIBLIOTHÈQUES

- **Bibliothèques python :**
  Pandas, Matplotlib, Seaborn, Plotly, Altair, ….
- **Logiciels :**
  Power Bi, Tableau, Excel, Google Charts, Qlik Sense
- **Autres :**
  Rattle (R Package), Rapidminer, IBM Cognos Analytics, Polymer Search, …

16

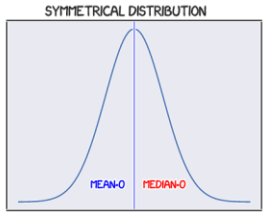
16

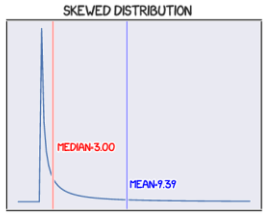
HISTOGRAMME

- L'histogramme est un graphique qui montre la distribution des données.
- Il regroupe les données numériques dans des barres, affichant les barres sous forme de colonnes segmentées.

- Ils sont utilisés pour décrire la distribution d'un ensemble de données.
  - Montre le centre,
  - la variabilité, l'asymétrie, la modalité,
  - valeurs aberrantes ou modèles étranges.

- La largeur et la position des barres est impotente

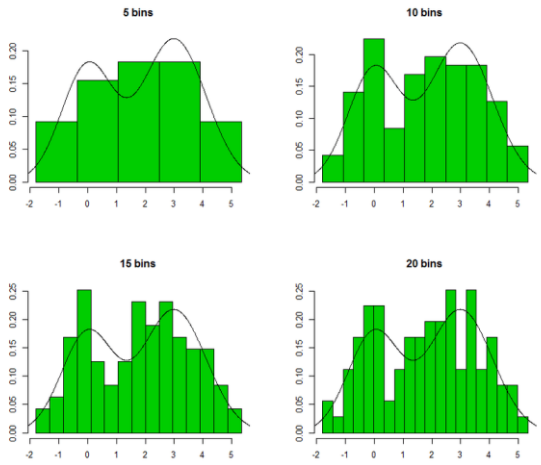


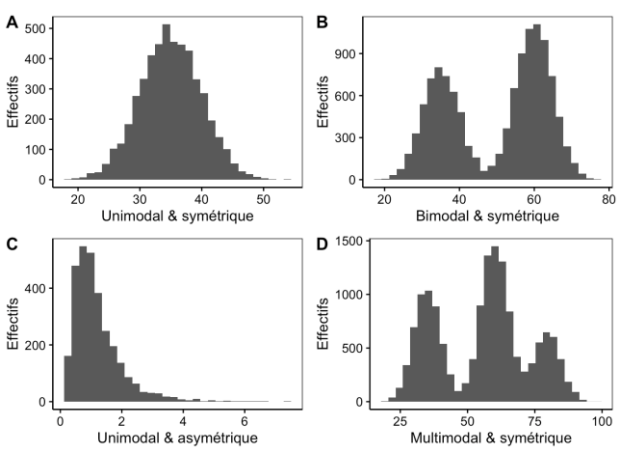


17

17

HISTOGRAMME





18

18

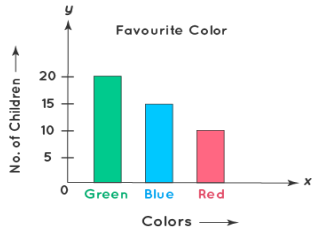
BAR PLOT (DIAGRAMME EN BÂTONS)

- Montrer la fréquence ou la proportion des catégories.
- Un graphique à barres utilise (diagramme en bâtons) des barres rectangulaires pour représenter des catégories de données, où la longueur ou la hauteur des barres est proportionnelle à leurs valeurs.
- Les graphiques à barres peuvent être affichés horizontalement ou verticalement.

19

19

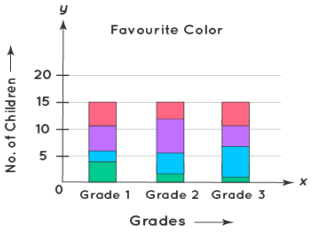
BAR PLOT



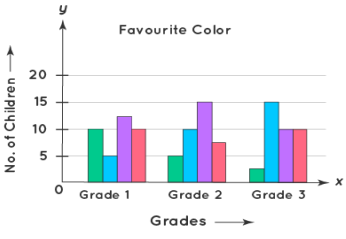
Vertical Bar Graph

Horizontal Bar Graph

Graphiques à barres empilées



Stacked Bar Graph



Grouped Bar Graph

Graphique à barres groupées

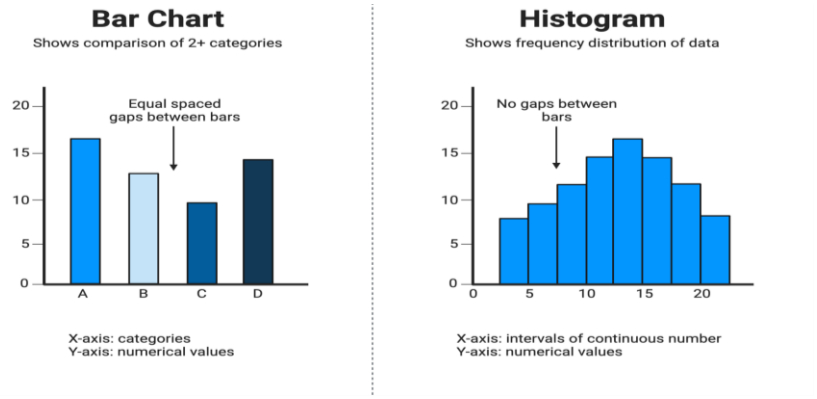
20

20

BAR PLOT VS HISTOGRAMME

- **barcharts :** Les structures sont une étiquette qui représente une variable catégorique. La hauteur de la colonne indique la taille du groupe défini par les catégories.

- **L'histogramme :** Les structures sont une étiquette qui représente une variable quantitative. L'étiquette de colonne peut être une valeur unique ou une plage de valeurs.



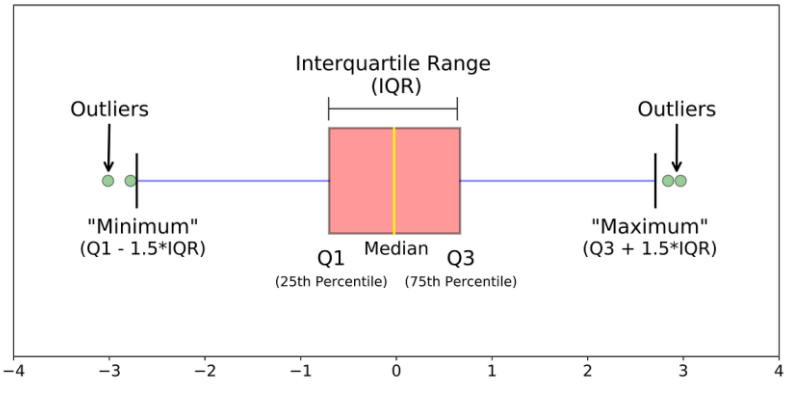
21

21

BOXPLOT (BOÎTE À MOUSTACHES)

- Utilisable pout identifier les valeurs aberrantes, et la dispersion des données.

- Affiche de nombreuses informations sur une variable dans une parcelle : Min, Max, Médiane, quartiles (Q_1, Q_2, Q_3) et IQR , Asymétrie

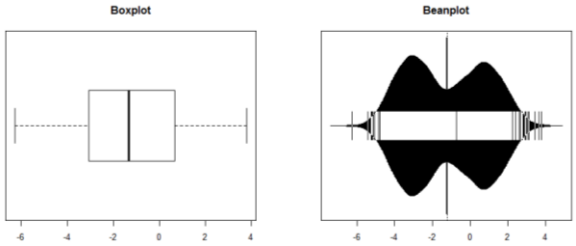


22

22

BEAN PLOTS

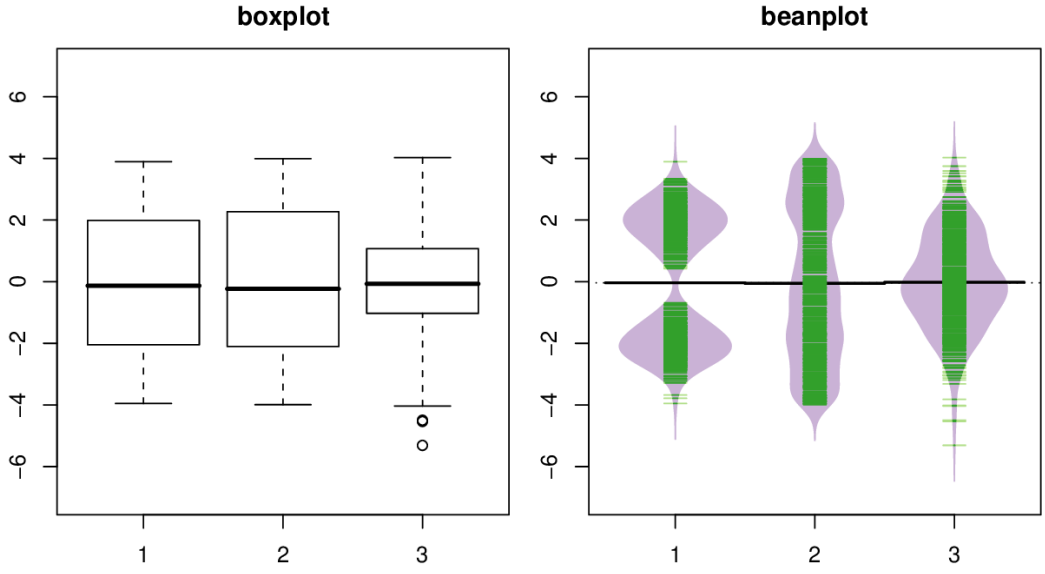
- Similaire au **boxplot** mais montre aussi la densité des données. Faciles à expliquer aux non-mathématiciens
- Le **beanplot** est une alternative au **boxplot** pour la comparaison visuelle des données univariées entre groupes.
- Contrairement au boxplot, qui résume les données en utilisant des statistiques comme la médiane et les quartiles, le beanplot montre les observations individuelles, ce qui permet de visualiser la distribution des données de manière plus détaillée.



23

23

BEAN PLOTS

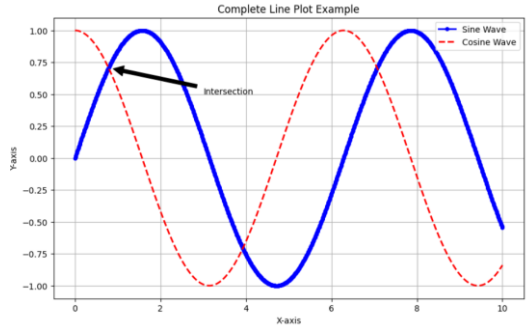


24

24

LINE CHART(COURBE DE DENSITÉ)

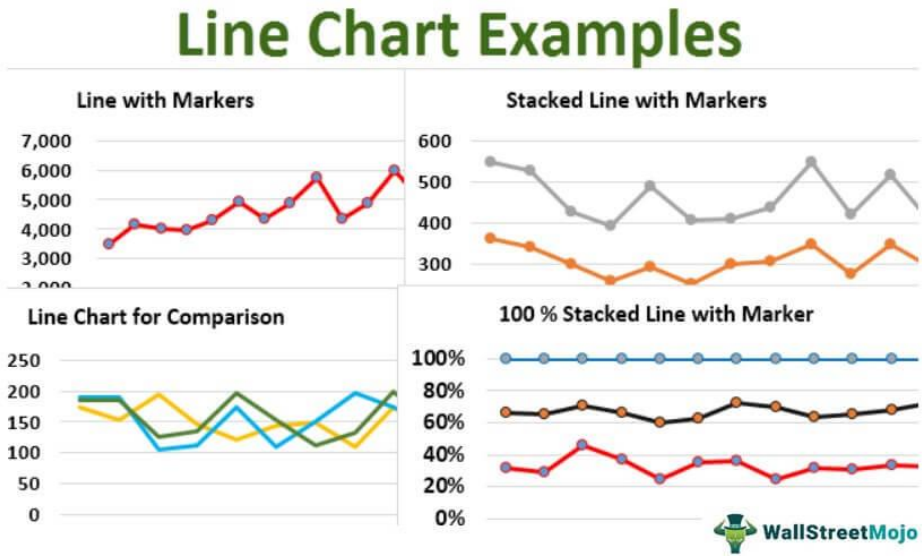
- Un graphique linéaire, est une représentation graphique utilisée pour afficher des points de données reliés par des lignes droites, des fonctions mathématiques, des séries temporelles, …
- Il est souvent utilisé pour montrer des tendances ou la comparaison de deux ensembles de données.



25

25

LINE CHART(COURBE DE DENSITÉ)

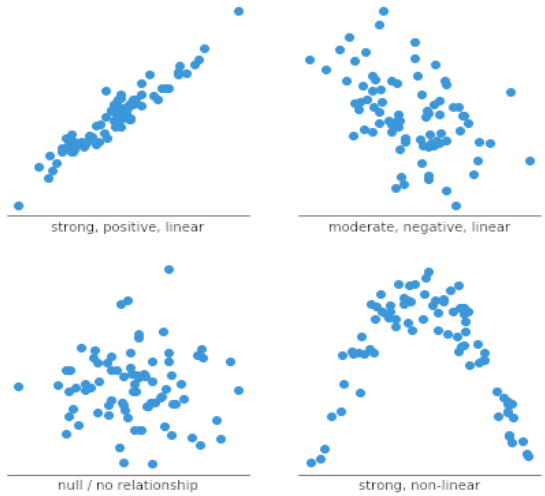


26

26

SCATTER PLOT(NUAGE DE POINTS)

- Un diagramme de dispersion, ou scatter plot, est un graphique où chaque valeur d'un ensemble de données est représentée par un point.
- Il est utilisé pour afficher la relation entre deux variables sur un graphique bidimensionnel connu sous le nom de plan cartésien.



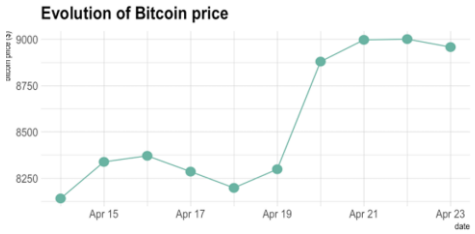
27

27

SCATTER PLOT(NUAGE DE POINTS)

Connected scatterplots

- Un nuage de points connecté affiche l'évolution d'une variable numérique. Les points de données sont représentés par un point et reliés par des segments de droite.
- Il montre souvent une tendance dans les données sur des intervalles de temps: une série chronologique.
- Fondamentalement, il s'agit de la même chose qu'un graphique linéaire (line chart) dans la plupart des cas, sauf que les observations individuelles sont mises en évidence.

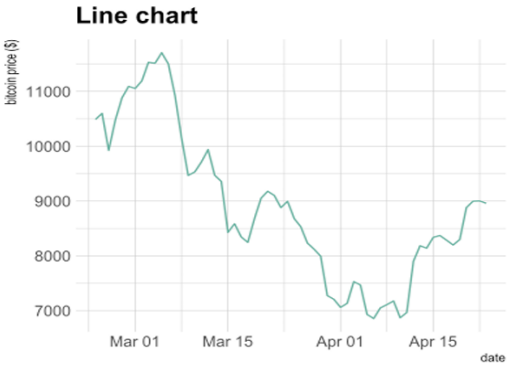


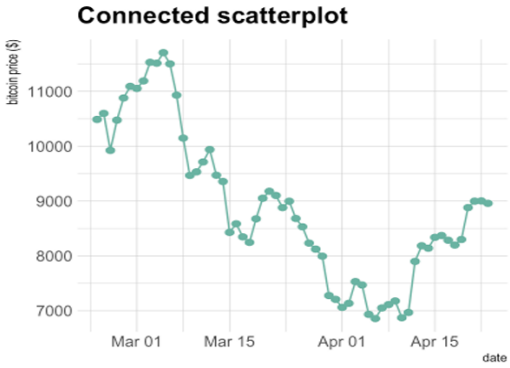
28

28

SCATTER PLOT(NUAGE DE POINTS)

Connected scatterplots





Lorsque votre axe X est ordonné, vous devez relier les points pour obtenir un diagramme de nuage de point connecté. En effet, le motif est difficile à lire si les points ne sont pas connectés

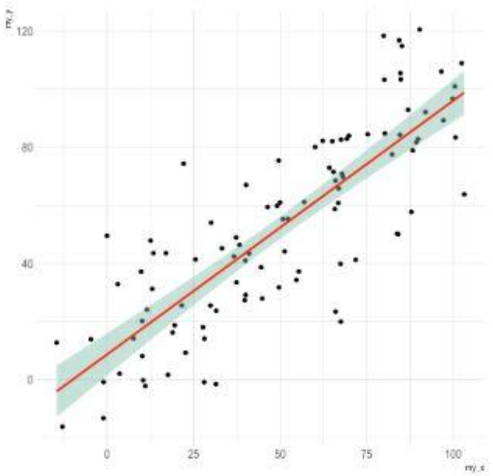
29

29

SCATTER PLOT(NUAGE DE POINTS)

Fited Scatterplots

- Les lignes de l'intervalle de confiance indiquent les limites de la variation à laquelle on peut s'attendre pour la fonction de régression ajustée.
- La largeur de l'intervalle de confiance donne une indication de la qualité de la fonction de régression ajustée.

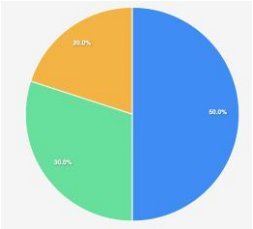


30

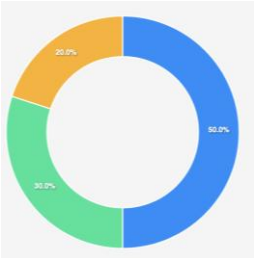
30

PIE CHART (GRAPHIQUE DE SECTEUR)

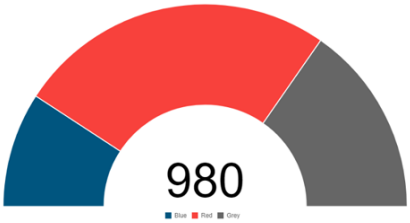
- Le **pie chart** est un graphique circulaire utilisé pour représenter des proportions ou des pourcentages, chaque segment représentant une catégorie.



- **Donut chart**



- **Half donut charts**

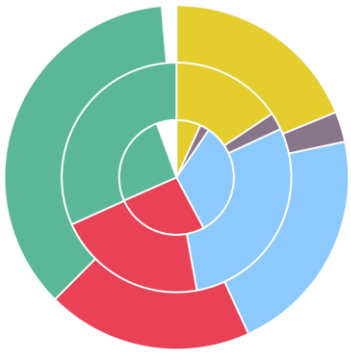


31

31

MULTI LAYER CHART

- Groupement de pie chart and donut charts.
- Il permet de représenter des données complexes en représentations visuelles
- Ce type de visualisation de données n'est pas aussi facile à créer que les autres ; il faut une certaine stratégie pour que toutes les catégories s'emboîtent et soient faciles à comprendre.
- En termes techniques, cette visualisation est constituée de trois pie charts superposés.

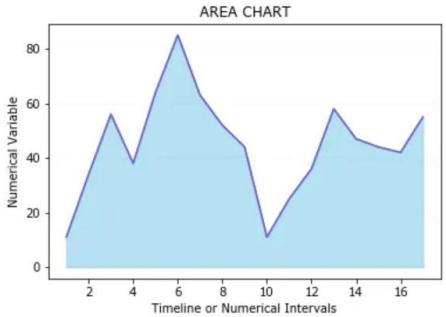


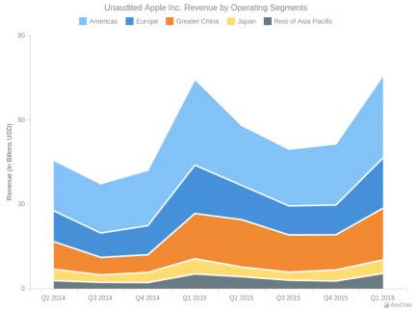
32

32

AREA CHART AND STACKED AREA CHART

- La région colorée nous montre l'évolution d'une variable dans le temps.
- Les graphiques en aires sont idéaux pour illustrer clairement l'ampleur du changement entre deux ou plusieurs points de données.
- **Exemples:**

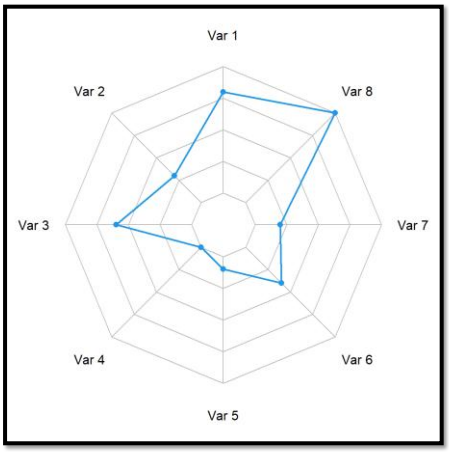


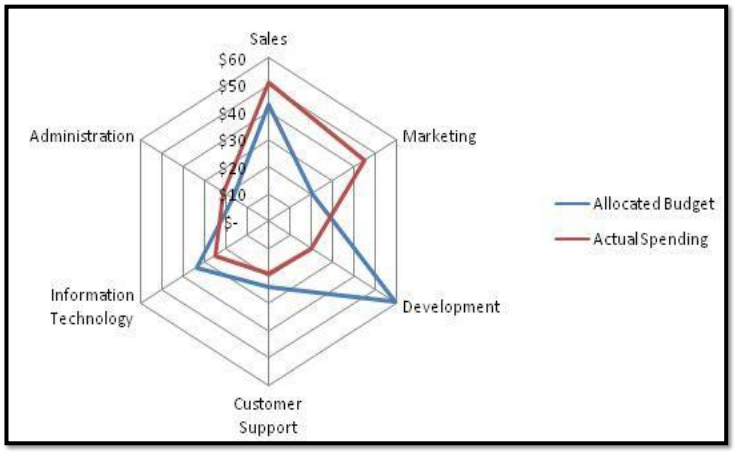


33

33

RADAR PLOT (SPIDER PLOT)



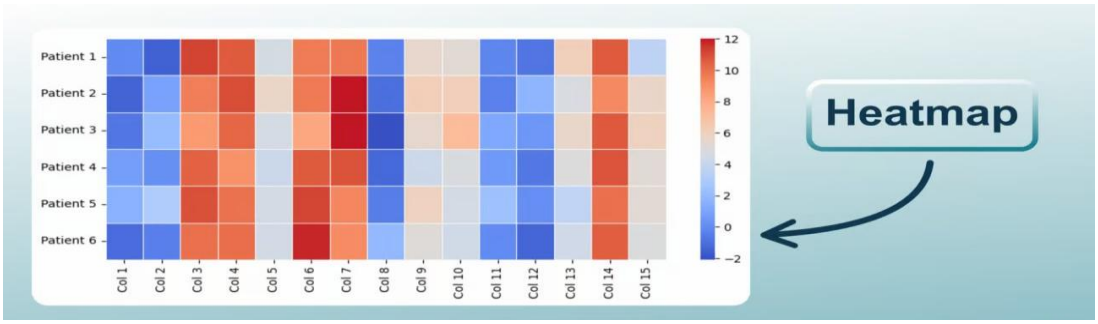


34

34

HEATMAP

- Chaque cellule est entièrement occupée par une marque de zone codant un attribut de valeur quantitative unique avec une couleur.
- Les cartes heatmap sont souvent utilisées avec des jeux de données bioinformatiques ou pour chercher la corrélation en les dimensions d'une dataset

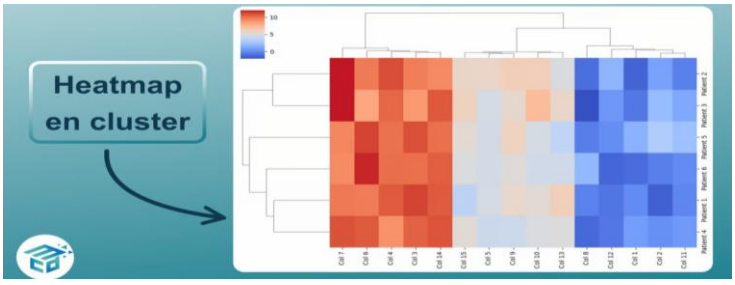


35

35

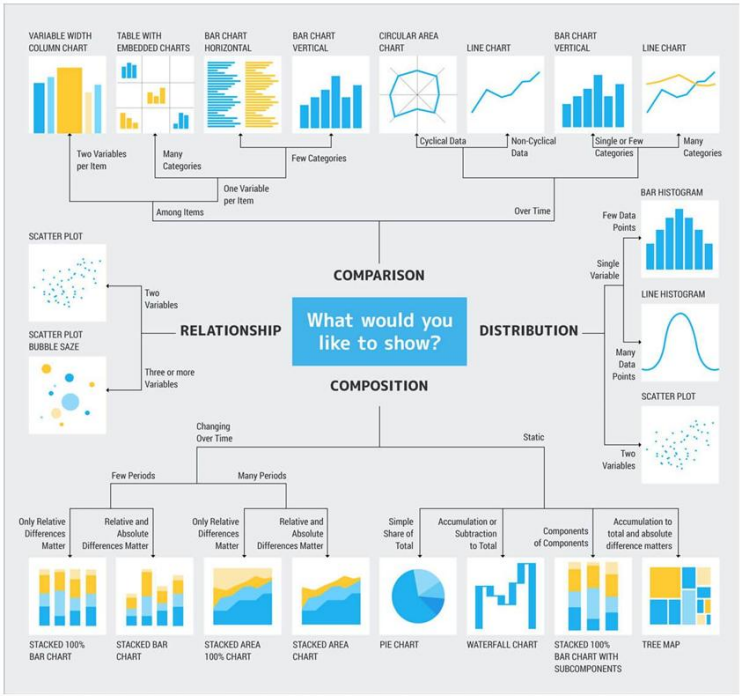
CLUSTER HEATMAP

- **Cluster heatmap** combine la carte heatmap de base avec le ré-ordonnancement de matrice où deux attributs sont combinés, l'objectif étant de regrouper des cellules similaires afin de rechercher des motifs à grande échelle entre les deux attributs et de voir les tendances sur un seul.
- Un cluster heatmap est la combinaison d'un heatmap et de deux dendrogrammes montrant les données dérivées des hiérarchies de cluster utilisées dans le réordonnancement.



36

36



37

37

VISUALISATION AVEC PYTHON

Notebook :

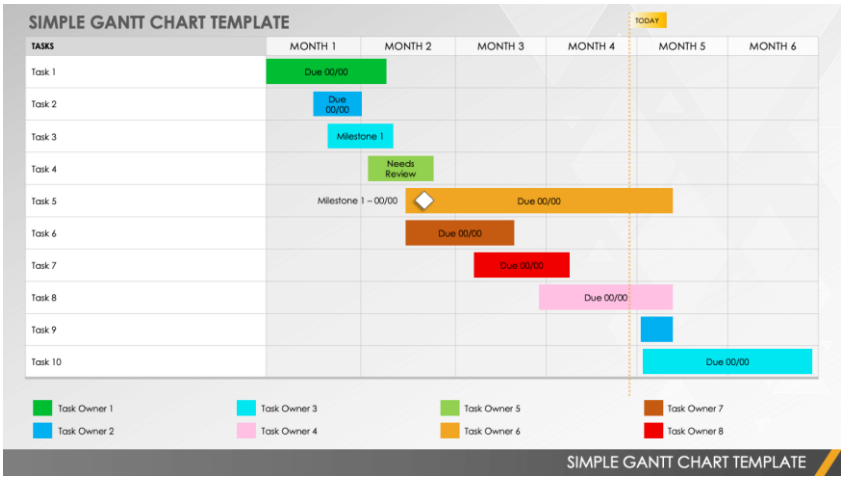
Visualisation avec matplotlib.pdf

38

38

GANTT CHARTS

- Le diagramme de Gantt est un diagramme à barres horizontales (planification d'un projet)



39