



Visualisation de données

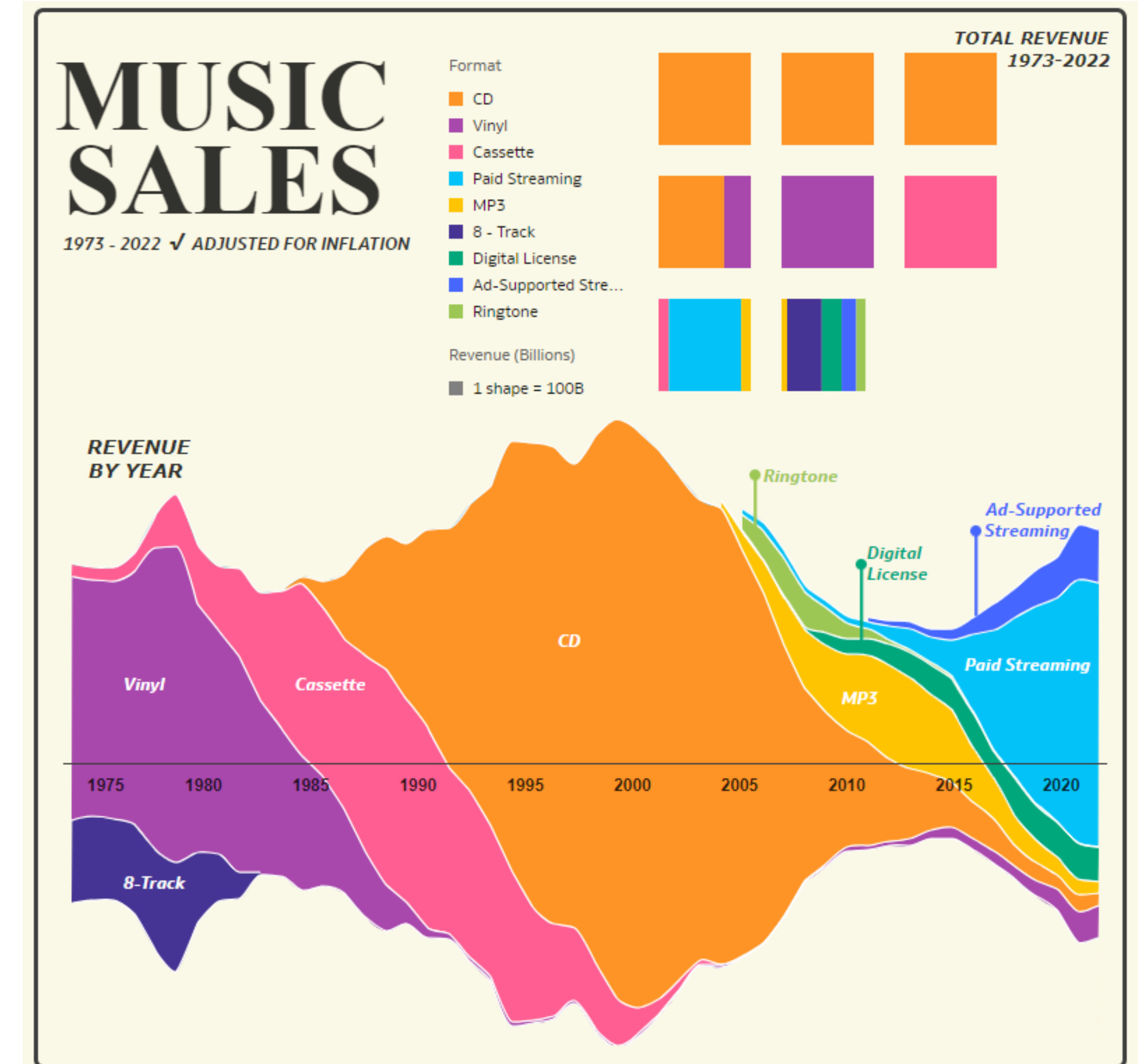
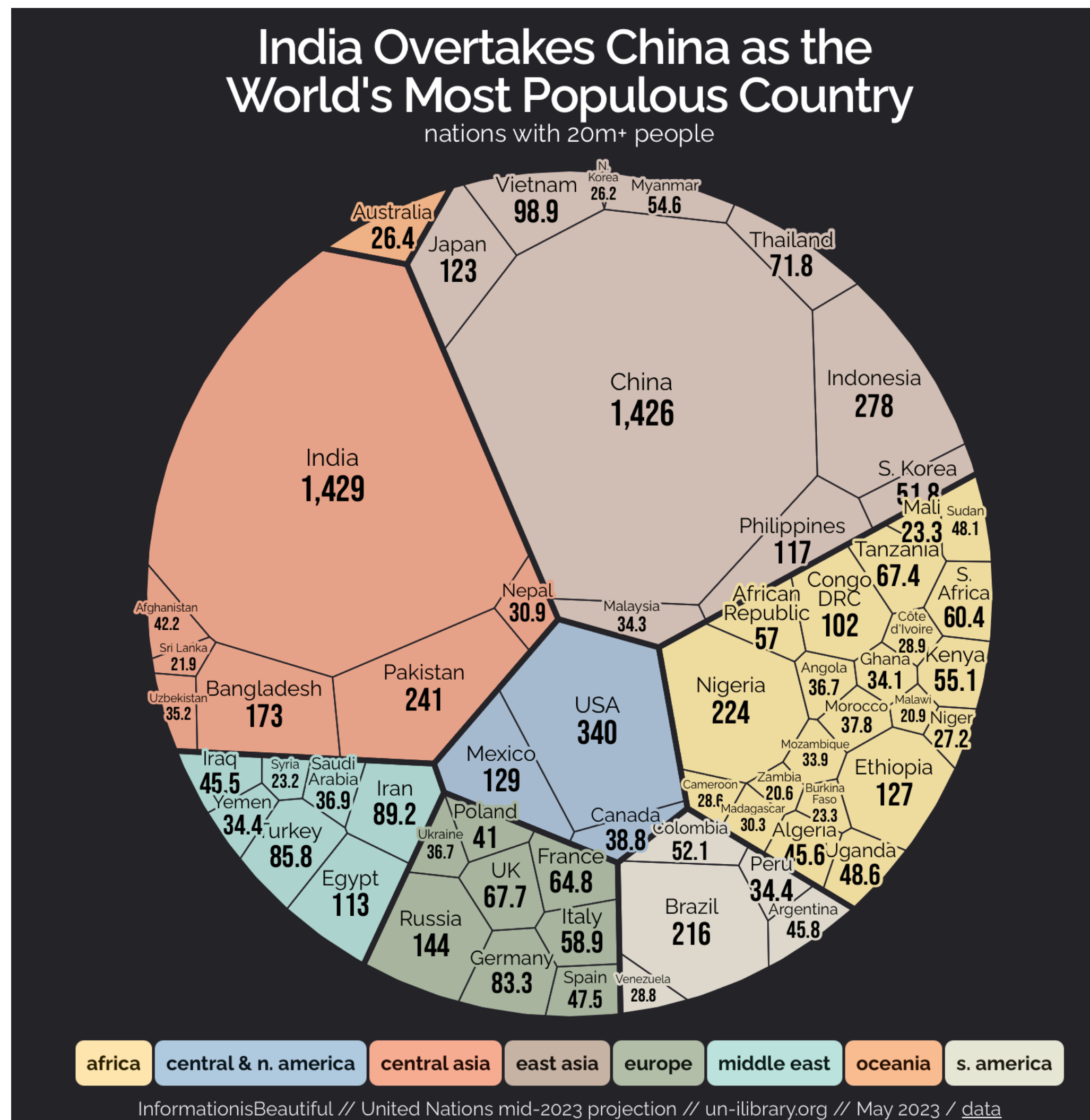
Visualisation des données

La visualisation des données consiste à transformer des données complexes en représentations visuelles (graphiques, cartes, diagrammes) afin de faciliter la compréhension, l'analyse et la communication des informations.

Visualiser des données ne consiste pas seulement à créer des graphiques, mais à **raconter une histoire avec les données**, en choisissant les formes visuelles les plus adaptées au message à transmettre.

Une bonne visualisation permet de :

- repérer rapidement des tendances, des motifs et des anomalies
- rendre les données accessibles à un public non spécialiste
- soutenir une prise de décision éclairée



Une bonne visualisation des données se suffit à elle-même : elle rend les résultats compréhensibles immédiatement et met en évidence les principaux enseignements sans nécessiter de longues explications.



Checklist pour une visualisation efficace des données

1. Contexte et interprétation

- Indiquer clairement période, source, unités.

2. Couleur

- Utiliser les couleurs **pour mettre en évidence**, pas juste pour décorer.
- Utiliser des palettes de couleurs prédéfinies (ex. ColorBrewer) pour assurer cohérence et lisibilité.
- Palette uniforme et cohérente entre graphiques.
- Couleurs distinguables et accessibles (colorblind-friendly).
- Limiter le nombre de couleurs.

Checklist pour une visualisation efficace des données

3. Titres et étiquettes

- Titre clair et descriptif.
- Axes et unités bien indiqués.
- Légende lisible et positionnée correctement.
- Ajouter annotations si nécessaire pour précision.

4. Lisibilité

- Texte et éléments graphiques lisibles.
- Couleurs et formes cohérentes.
- Grille légère pour faciliter la lecture



Checklist pour une visualisation efficace des données

5. Simplicité

- Éviter surcharge d'information.
- Supprimer éléments inutiles ("chartjunk").

6. Axes et échelles

- Axes proportionnés et cohérents (linéaire / logarithmique).
- Éviter distorsions visuelles.
- Ticks et graduations clairs.

Types de visualisation

Distribution
Comparaison
Composition
Relations entre variables
Évolution dans le temps
Visualisations complexes / Hiérarchie et flux
Visualisation géographique

Distribution

Les graphiques de distribution servent à représenter la manière dont une **variable quantitative se répartit au sein d'une population ou d'un échantillon**. Ils permettent de visualiser la fréquence des valeurs, leur dispersion et la forme globale des données.

Ils sont particulièrement utiles pour :

- observer la répartition des individus selon une variable (âge, revenus, notes, etc.)
- repérer des valeurs atypiques (outliers) ou des déséquilibres
- comparer la distribution entre plusieurs groupes

Ce sont généralement les **premières visualisations à réaliser** lorsqu'on analyse une nouvelle base de données, car elles donnent une vision d'ensemble et aident à repérer rapidement la structure des données.

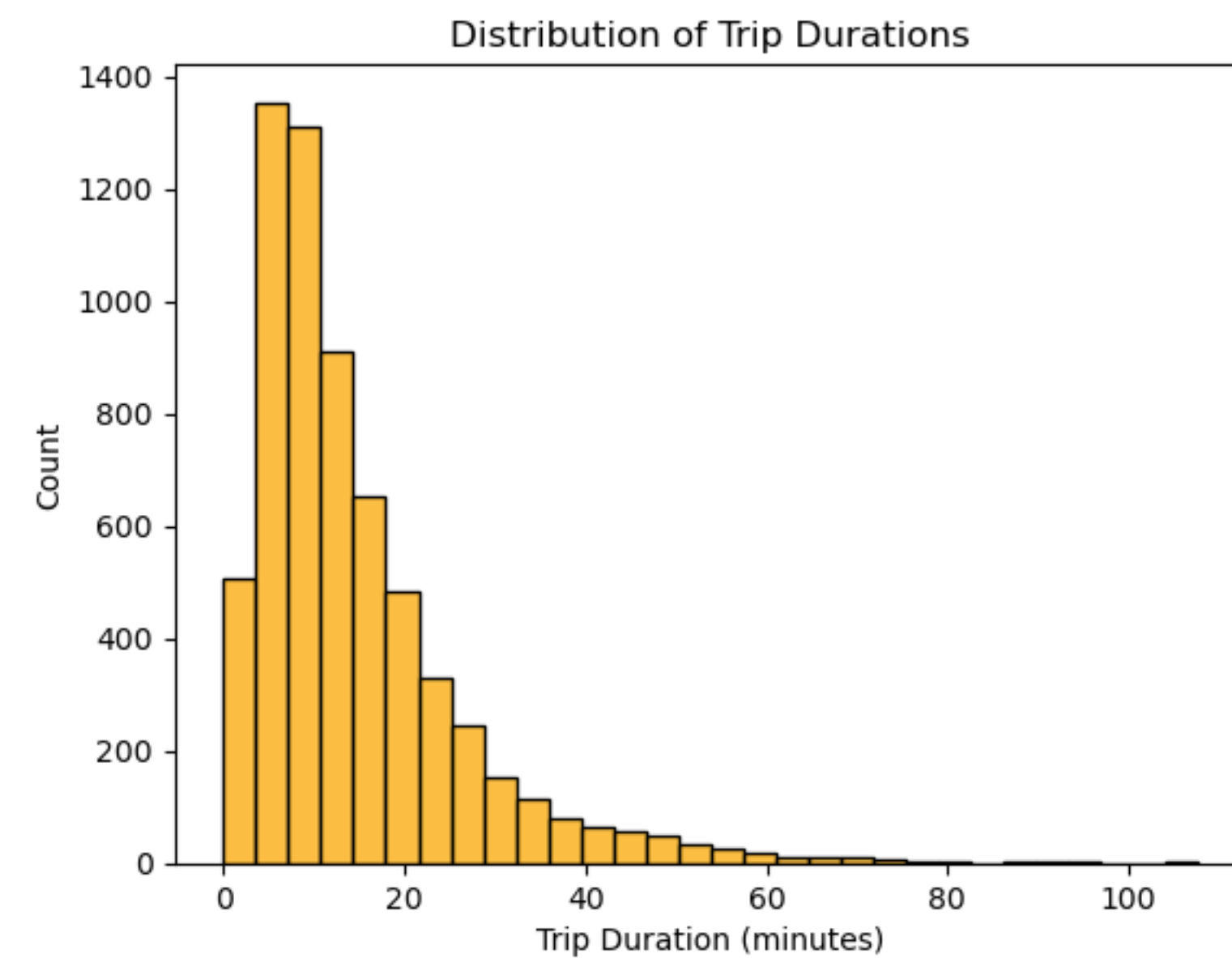
Distribution

L'histogramme est un graphique de distribution qui permet de représenter la fréquence des valeurs d'une variable quantitative en les regroupant en classes (intervalles).

Chaque barre correspond à un intervalle de valeurs, et sa hauteur indique le nombre d'observations (ou leur proportion) dans cet intervalle.

N'utiliser pas avec moins de 50 observations!

Nb observations
par classe, ou
fréquence
relative



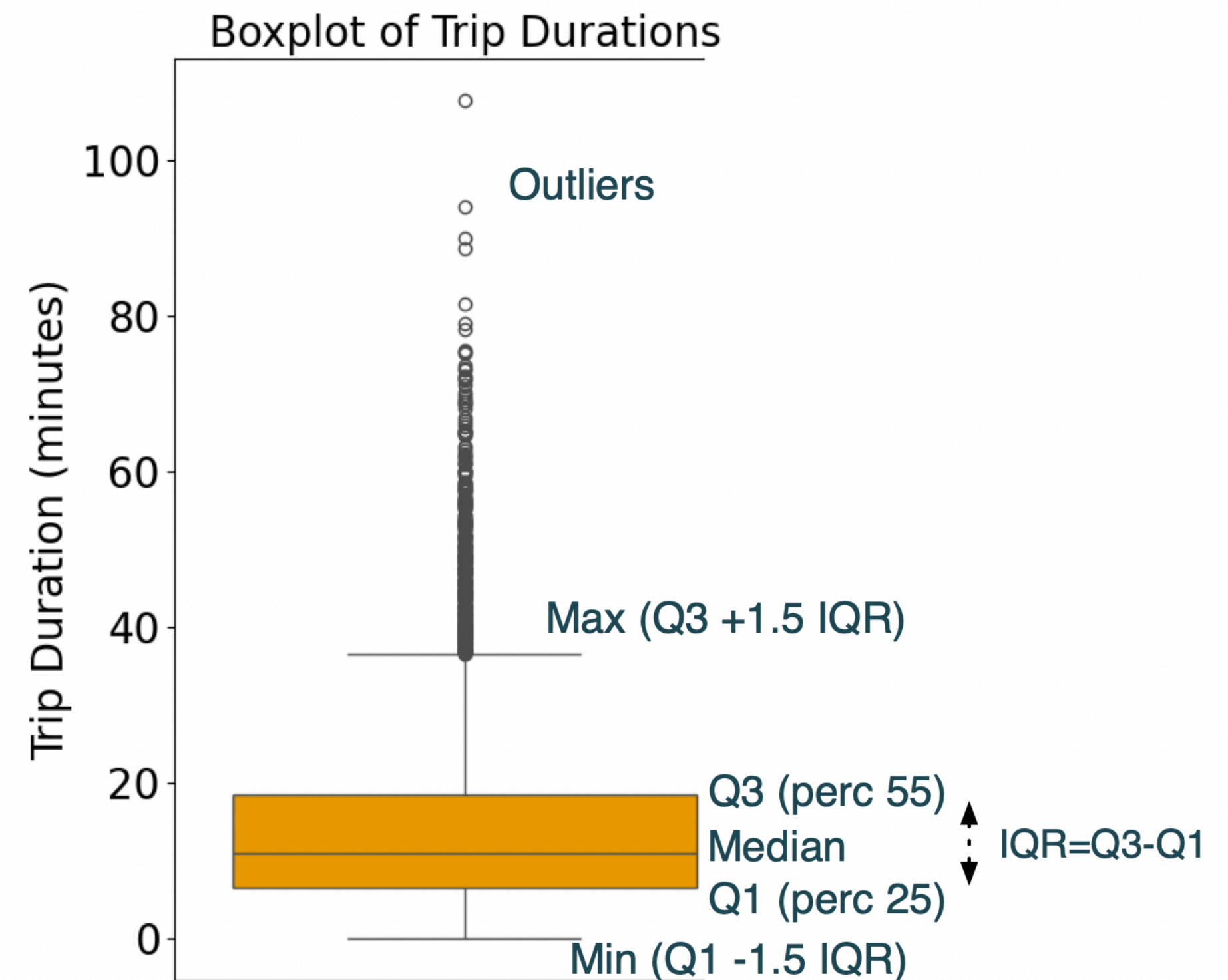
Classes

Distribution

La boîte à moustaches (box plot) donne une idée de la distribution d'une variable même quand le nombre de données est faible. Il permet de repérer des valeurs aberrantes.

Elle représente :

- la médiane (valeur centrale)
- les quartiles (Q1 et Q3), qui délimitent la moitié centrale des données
- les moustaches, qui s'étendent vers les valeurs les plus basses et les plus élevées (hors valeurs atypiques)
- les valeurs atypiques (outliers), affichées individuellement



Distribution

Un KDE plot (Kernel Density Estimate) est un graphique de distribution qui permet de visualiser la densité d'une variable continue plutôt que de compter les occurrences comme dans un histogramme.

Comment ça marche

Il estime la "forme" de la distribution en lissant les données.

Chaque point contribue à la densité autour de lui selon une fonction de noyau (kernel), souvent gaussienne.

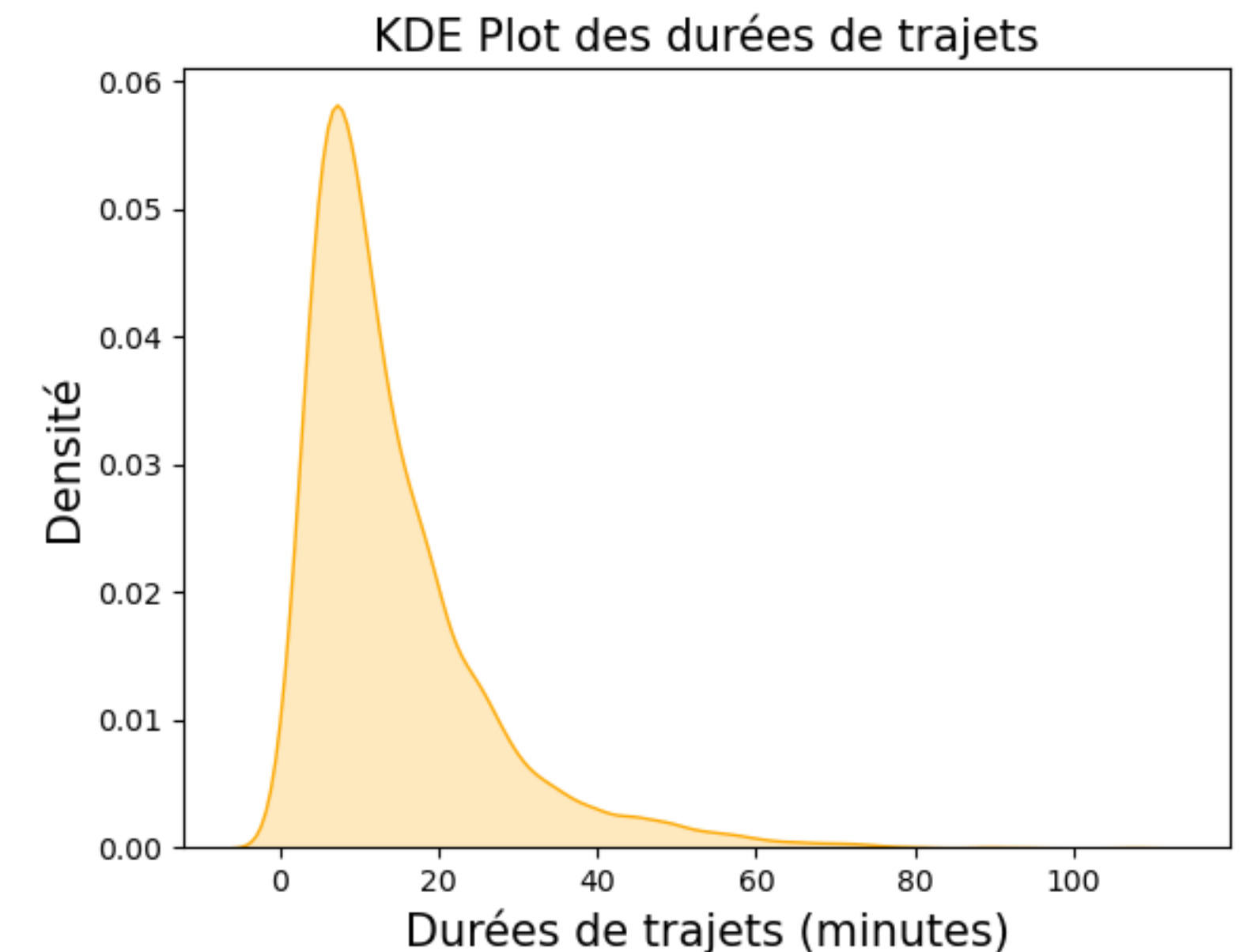
Le résultat est une courbe lisse qui montre où les valeurs sont les plus concentrées.

À quoi ça sert

- Voir la distribution sous forme continue sans dépendre des intervalles comme pour l'histogramme.
- Identifier les pics (modes) et les zones où les données sont plus ou moins concentrées.

Différence avec l'histogramme

- Histogramme : barres discrètes, dépend des intervalles choisis.
- KDE : courbe lisse, meilleure pour visualiser la tendance générale de la distribution.



Comparaison

Les graphiques de comparaison sont très utilisés dans le traitement des données d'enquête, notamment en sciences sociales, car ils permettent de **mettre en regard plusieurs catégories ou groupes** (par exemple selon le genre, l'âge, le niveau d'études, etc.).

Ils aident à repérer rapidement des différences, des écarts ou des similitudes entre sous-populations. Ils répondent à des questions comme : quel groupe est le plus élevé ?, les valeurs sont-elles proches ou très différentes ?, comment une variable varie-t-elle selon les catégories ?

Ils sont particulièrement utilisés pour :

- Comparer des sous-populations (par exemple selon le genre, l'âge, la catégorie socioprofessionnelle, le niveau d'études, etc.)
- Mettre en évidence des écarts ou des inégalités entre groupes
- Suivre des évolutions dans le temps en comparant plusieurs vagues d'enquête
- Comparer plusieurs variables au sein d'une même population (par exemple opinions, pratiques, revenus)
- Visualiser la distribution d'une variable selon différentes catégories (dispersion, médiane, valeurs extrêmes)

Comparaison

Diagramme à barres : permet de comparer des valeurs agrégées (moyenne, total, proportion) entre différentes catégories. Il est simple à lire et très efficace pour des comparaisons directes.

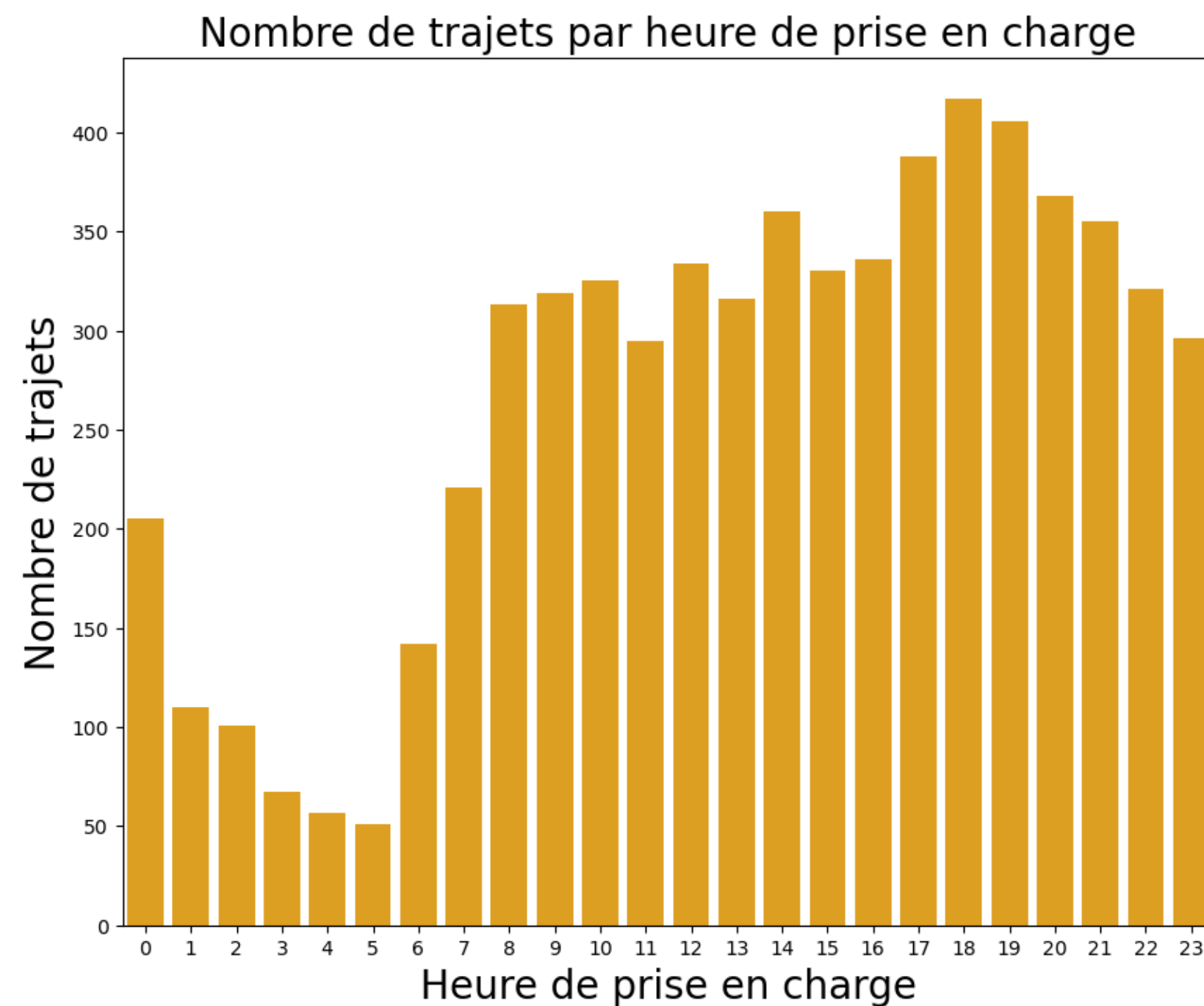
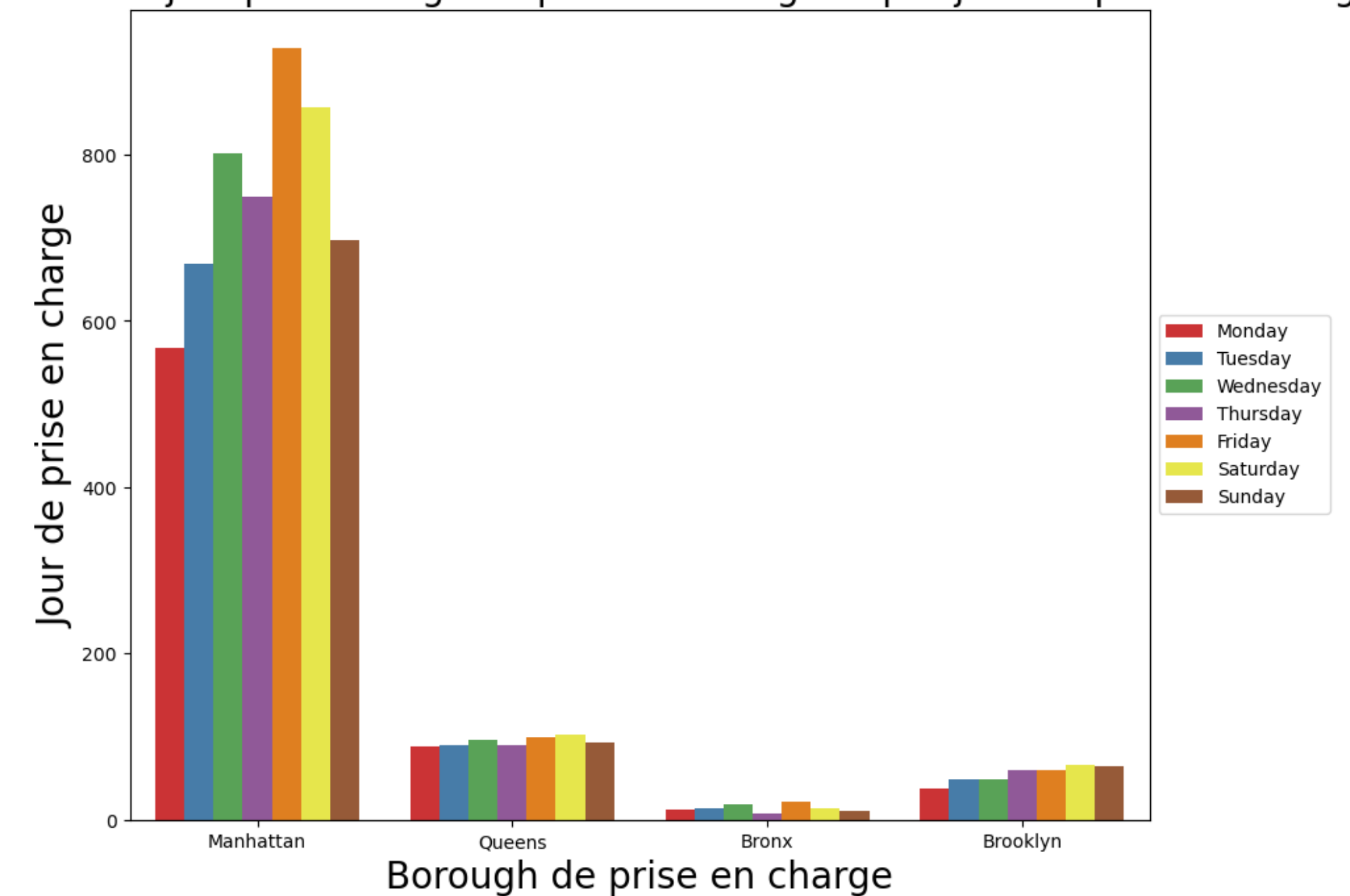


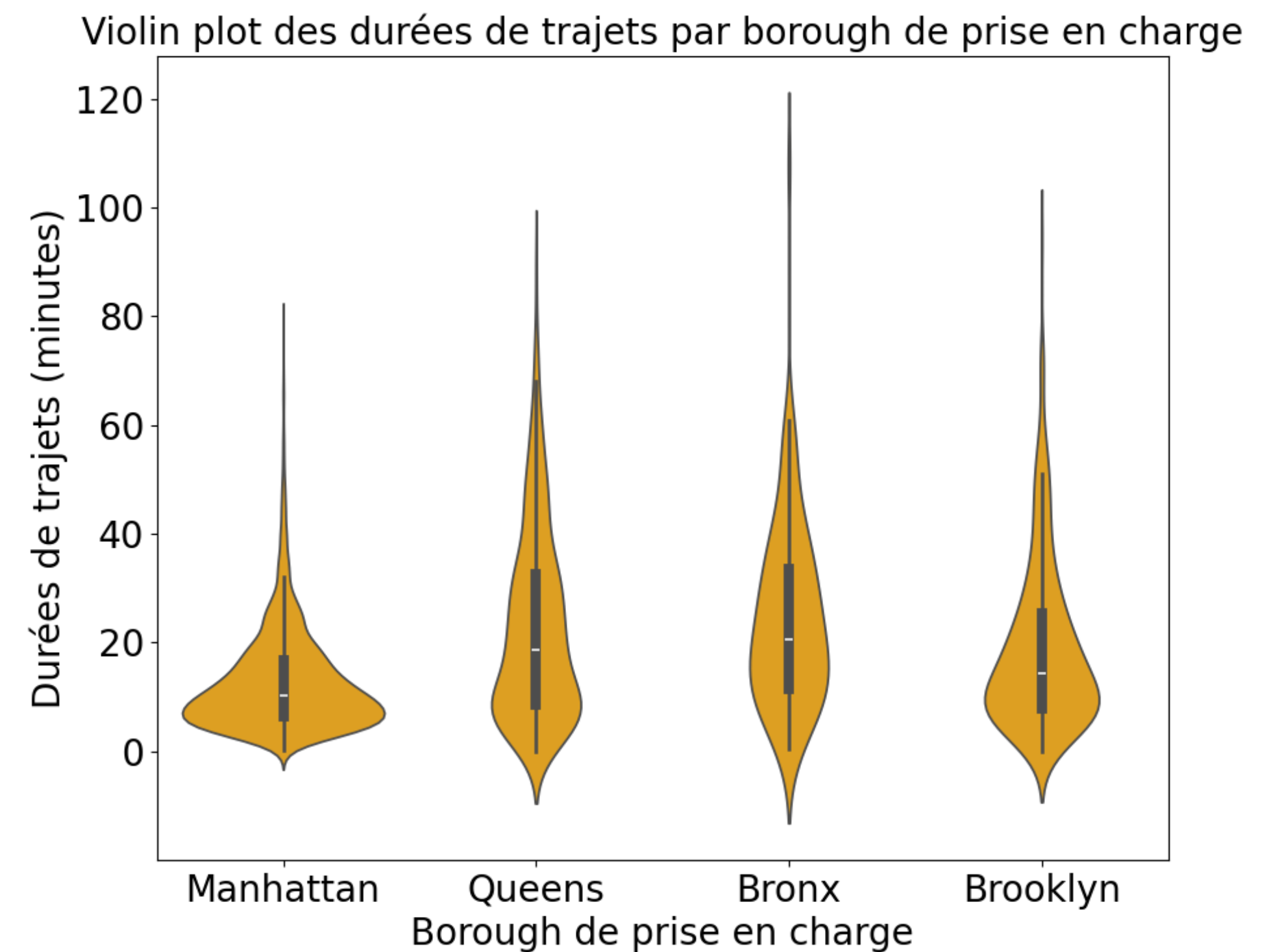
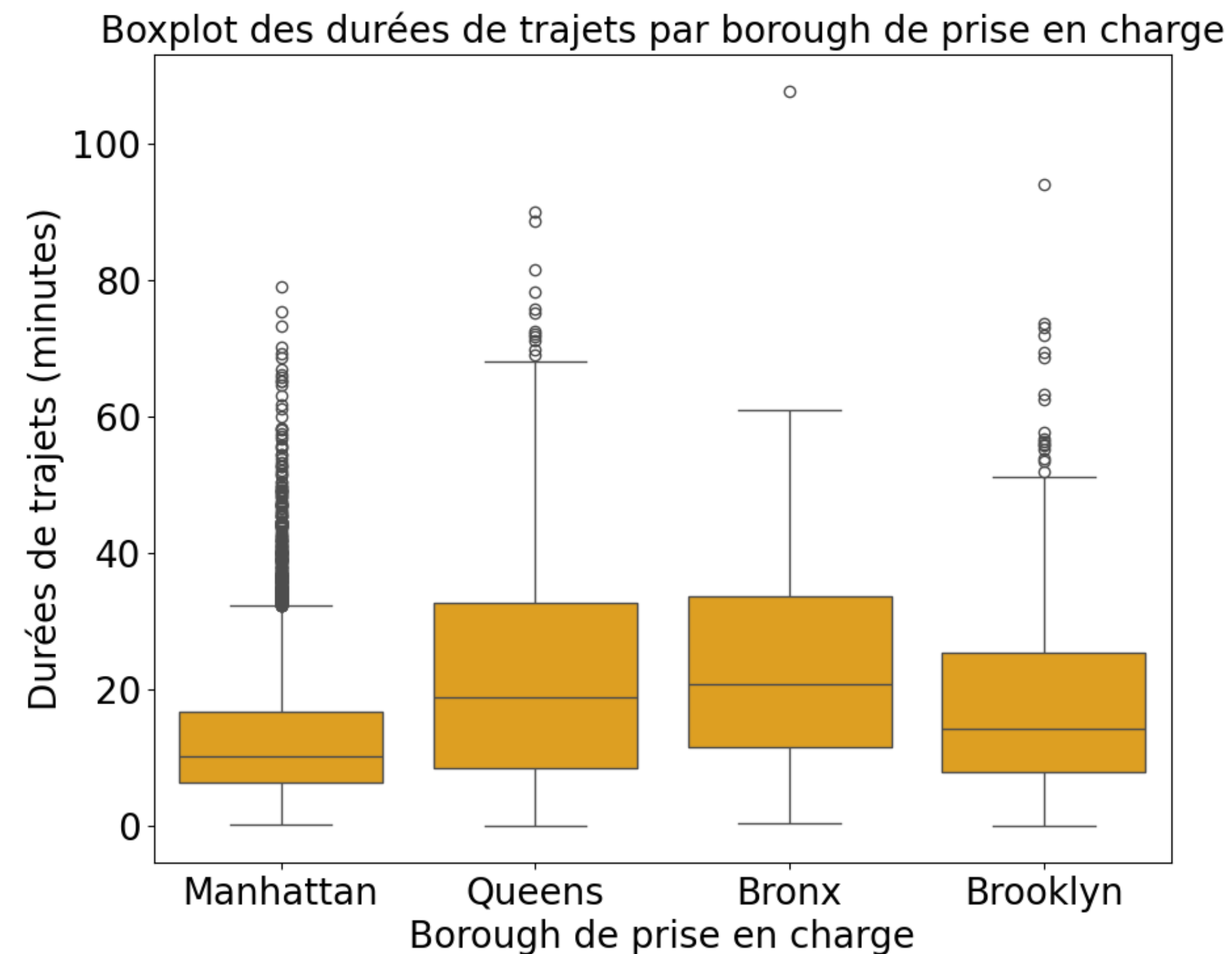
Diagramme à barres groupées : utile lorsque l'on souhaite comparer plusieurs variables ou sous-groupes au sein de chaque catégorie (par exemple hommes/femmes, années différentes, etc.).

Nombre de trajets par borough de prise en charge et par jour de prise en charge



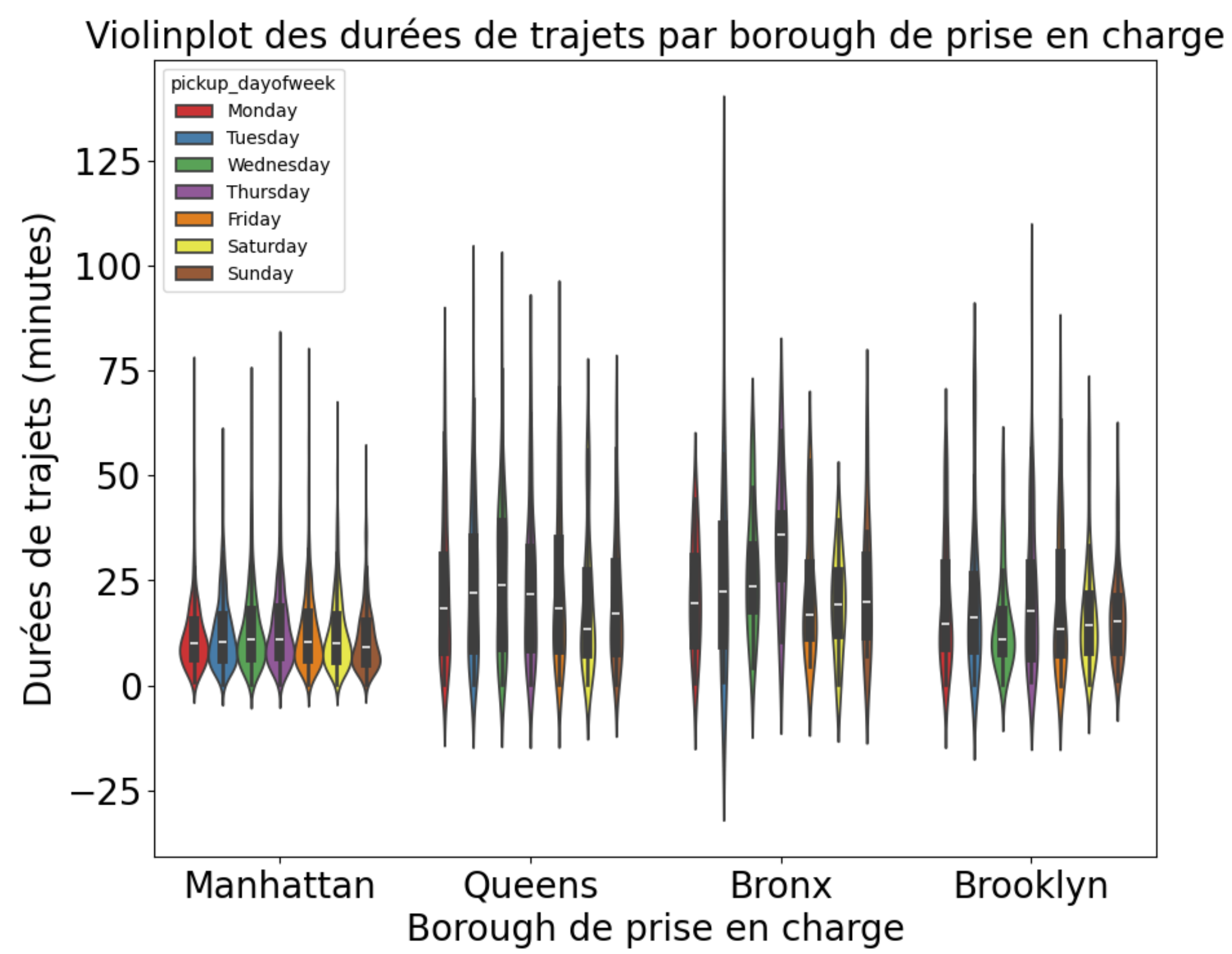
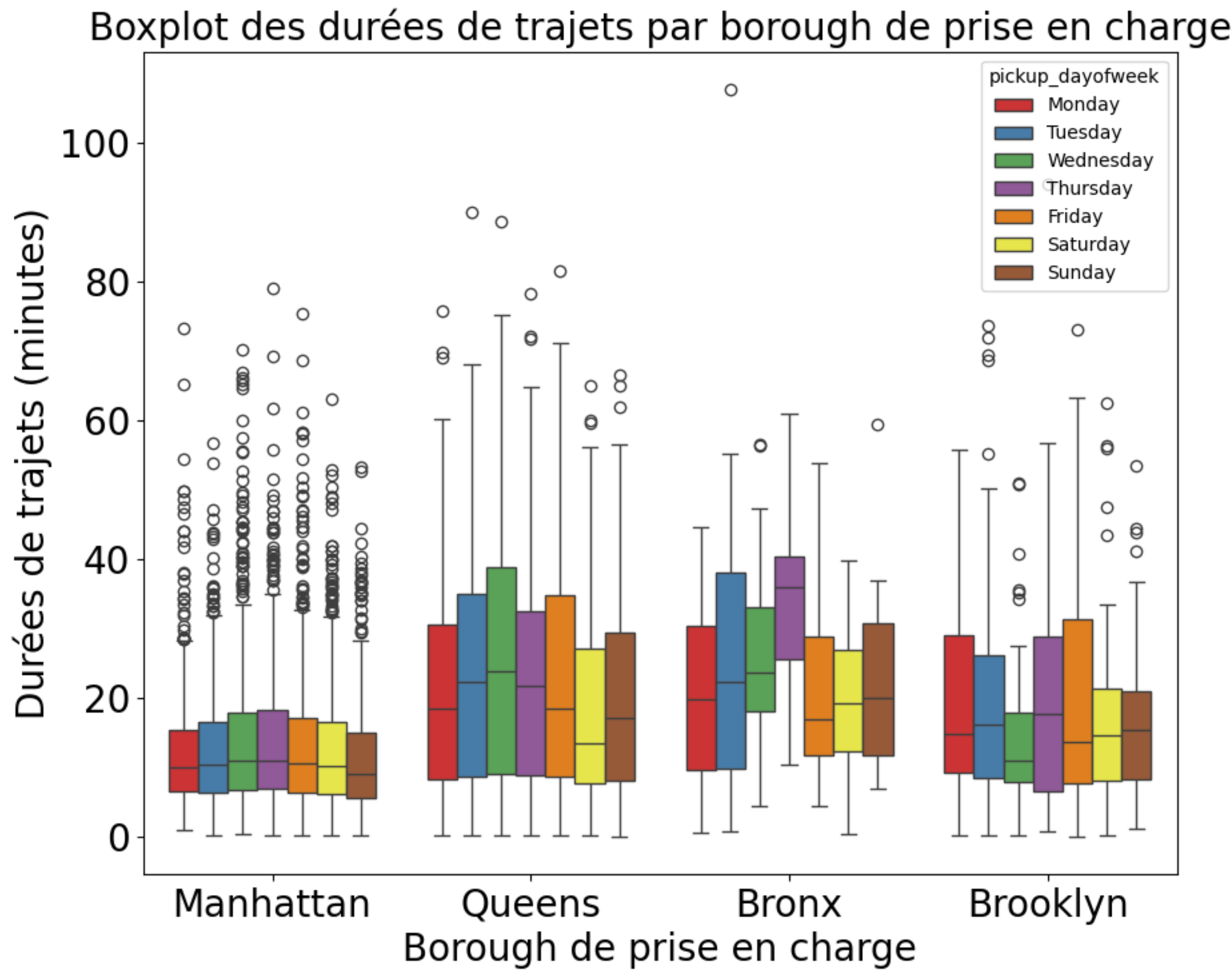
Comparaison

Boxplots / Violinplots multiples : permettent de comparer la distribution d'une variable entre plusieurs groupes (médiane, dispersion, valeurs extrêmes), et pas seulement une valeur moyenne.



Comparaison

Boxplots / Violinplots multiples groupés: permettent de comparer la distribution d'une variable entre plusieurs groupes (médiane, dispersion, valeurs extrêmes), et pas seulement une valeur moyenne.



Composition

Les graphiques de composition servent à représenter la **structure interne d'un ensemble**, c'est-à-dire la manière dont un total se répartit entre différentes catégories. Ils sont très utilisés dans les données d'enquête pour montrer la part relative de chaque modalité (par exemple répartition des réponses, des pratiques, des opinions). Ils permettent de répondre à des questions comme : quelle est la proportion de chaque catégorie ?, comment la composition évolue-t-elle ?, quels éléments dominent le total ?

Ils sont particulièrement utilisés pour :

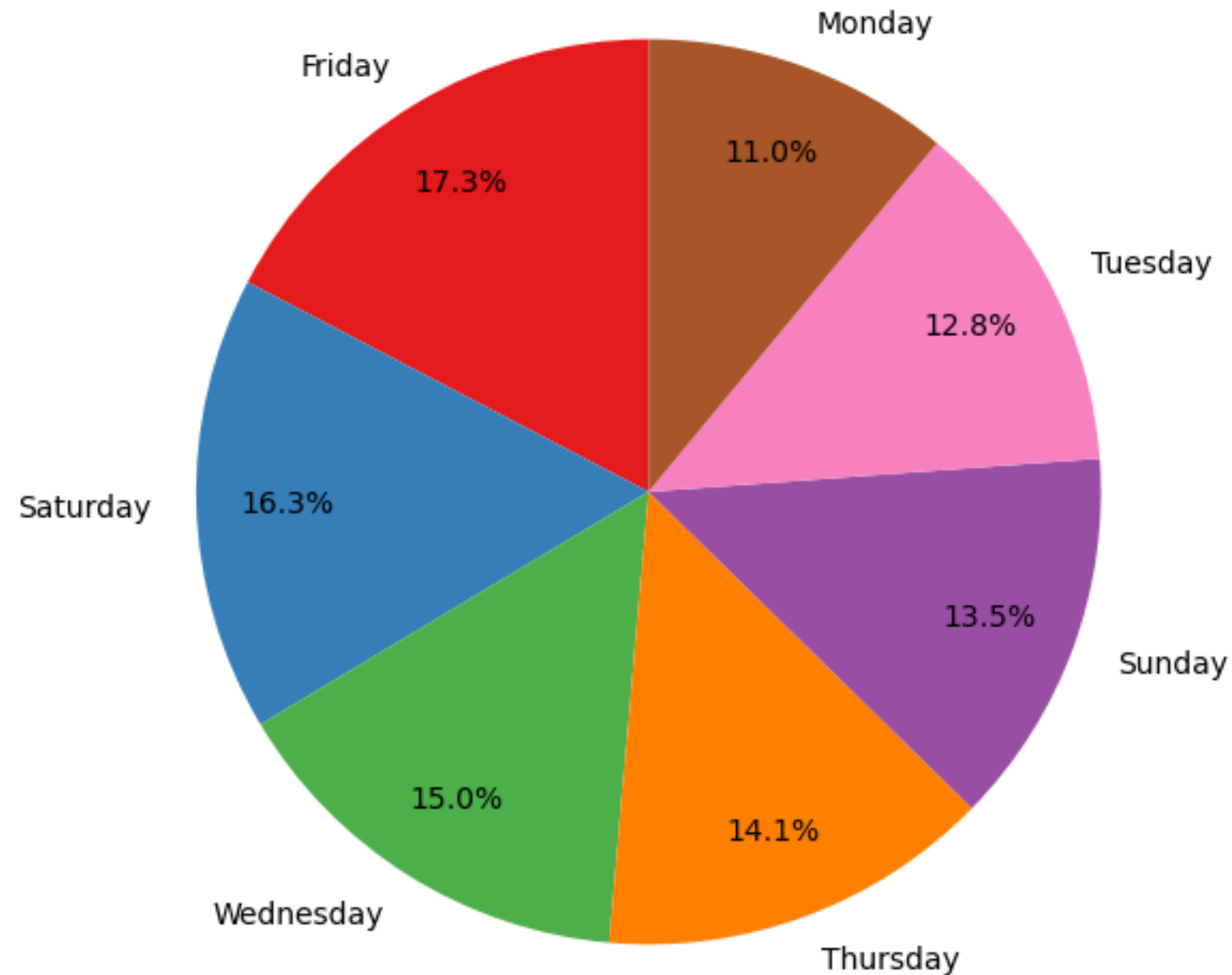
- Montrer la répartition d'un total entre différentes catégories (parts, proportions, pourcentages)
- Comparer la composition entre plusieurs groupes (par exemple selon le genre, l'âge, le territoire)
- Observer l'évolution de la structure dans le temps (changement de composition entre vagues d'enquête)
- Identifier les catégories dominantes ou minoritaires au sein d'un ensemble
- Mettre en évidence des déséquilibres ou des transformations de structure
- Résumer visuellement des distributions catégorielles issues de questions fermées d'enquête

Composition

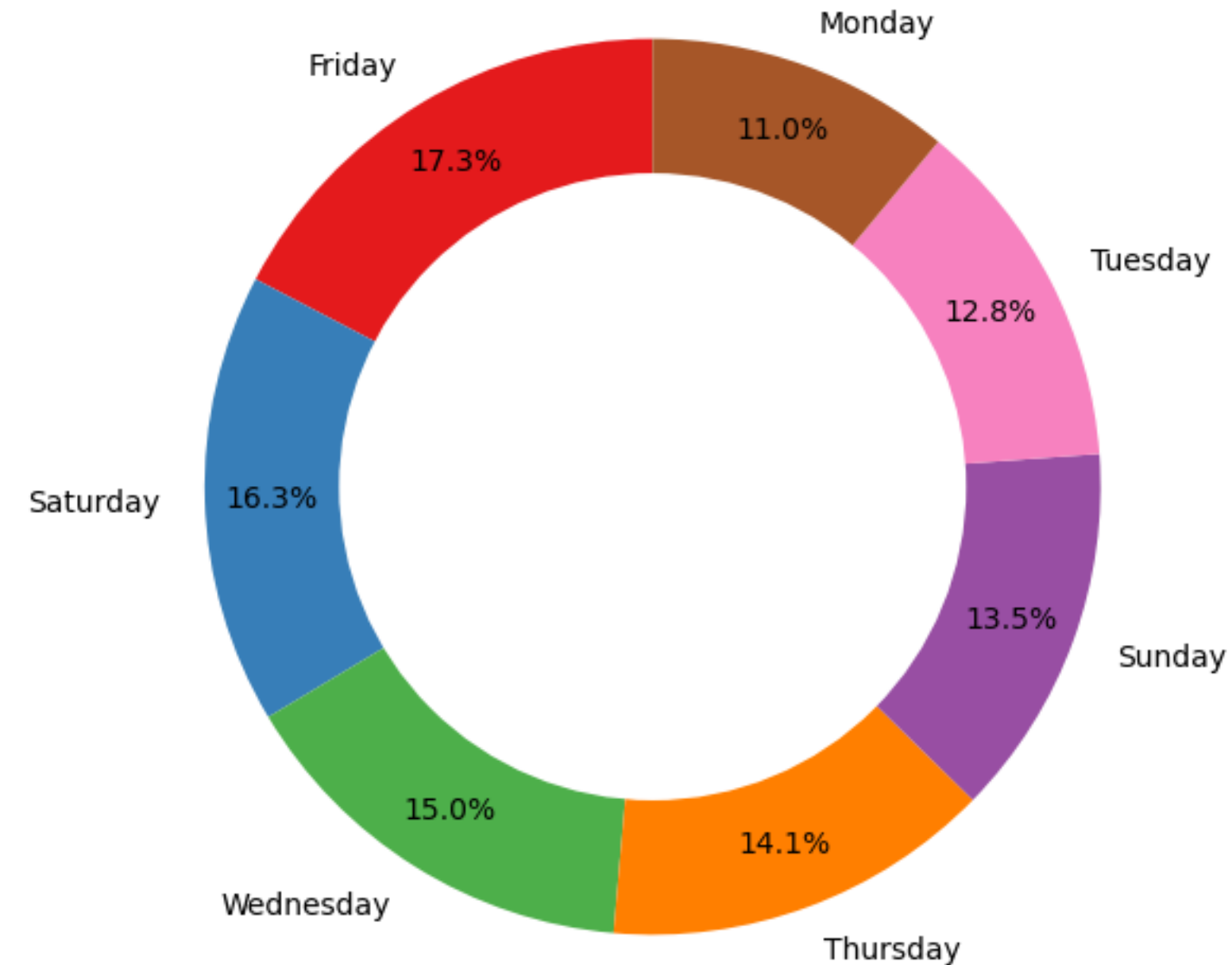
Diagramme en secteurs (camembert) ou Diagramme en anneau (donut):

permet de comparer des valeurs agrégées (moyenne, total, proportion) entre différentes catégories. Il est simple à lire et très efficace pour des comparaisons directes. L'angle de chaque secteur est proportionnel à la part qu'il représente dans l'ensemble.

Répartition des trajets par jour de prise en charge



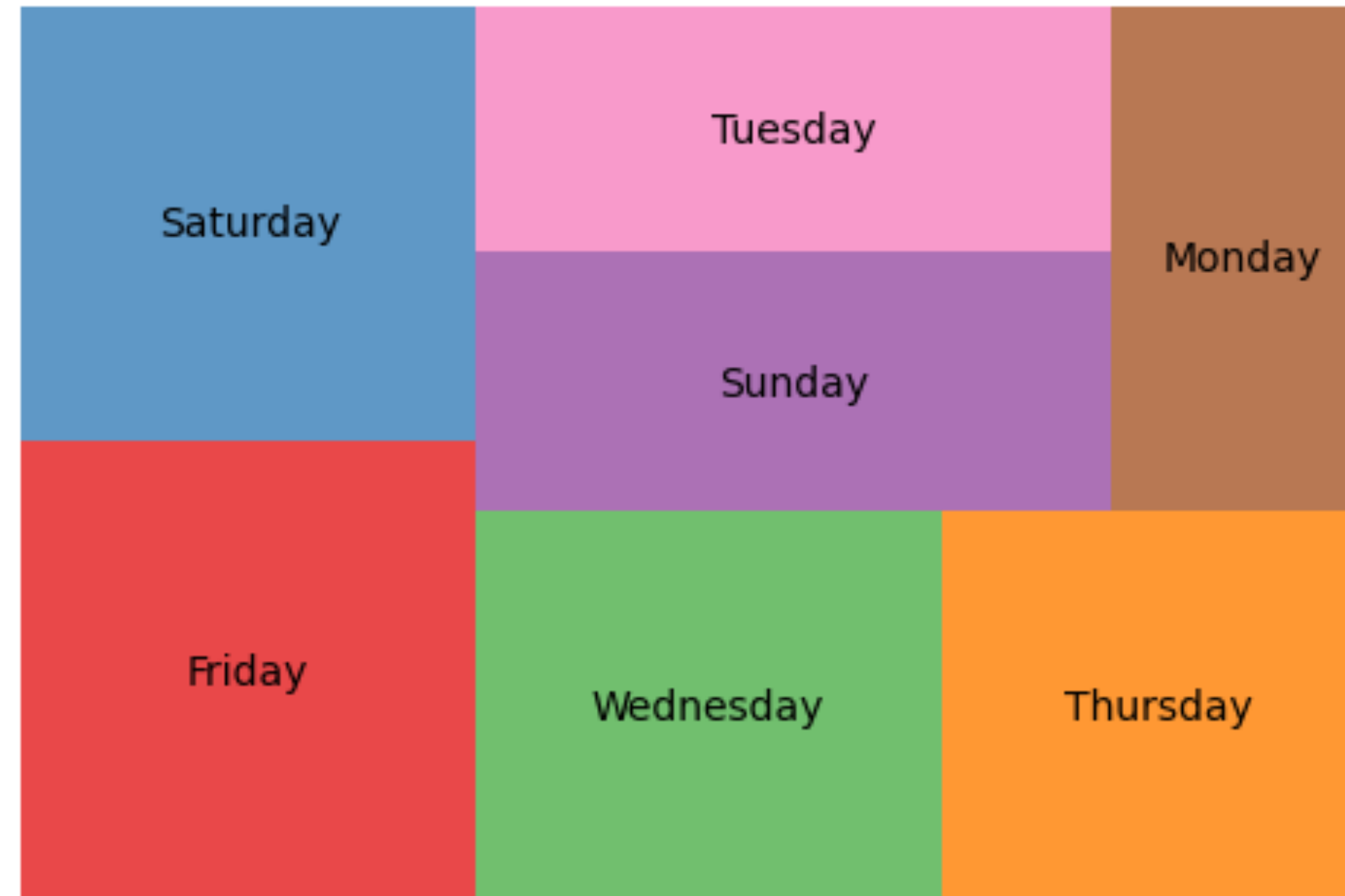
Répartition des trajets par jour de prise en charge



Composition

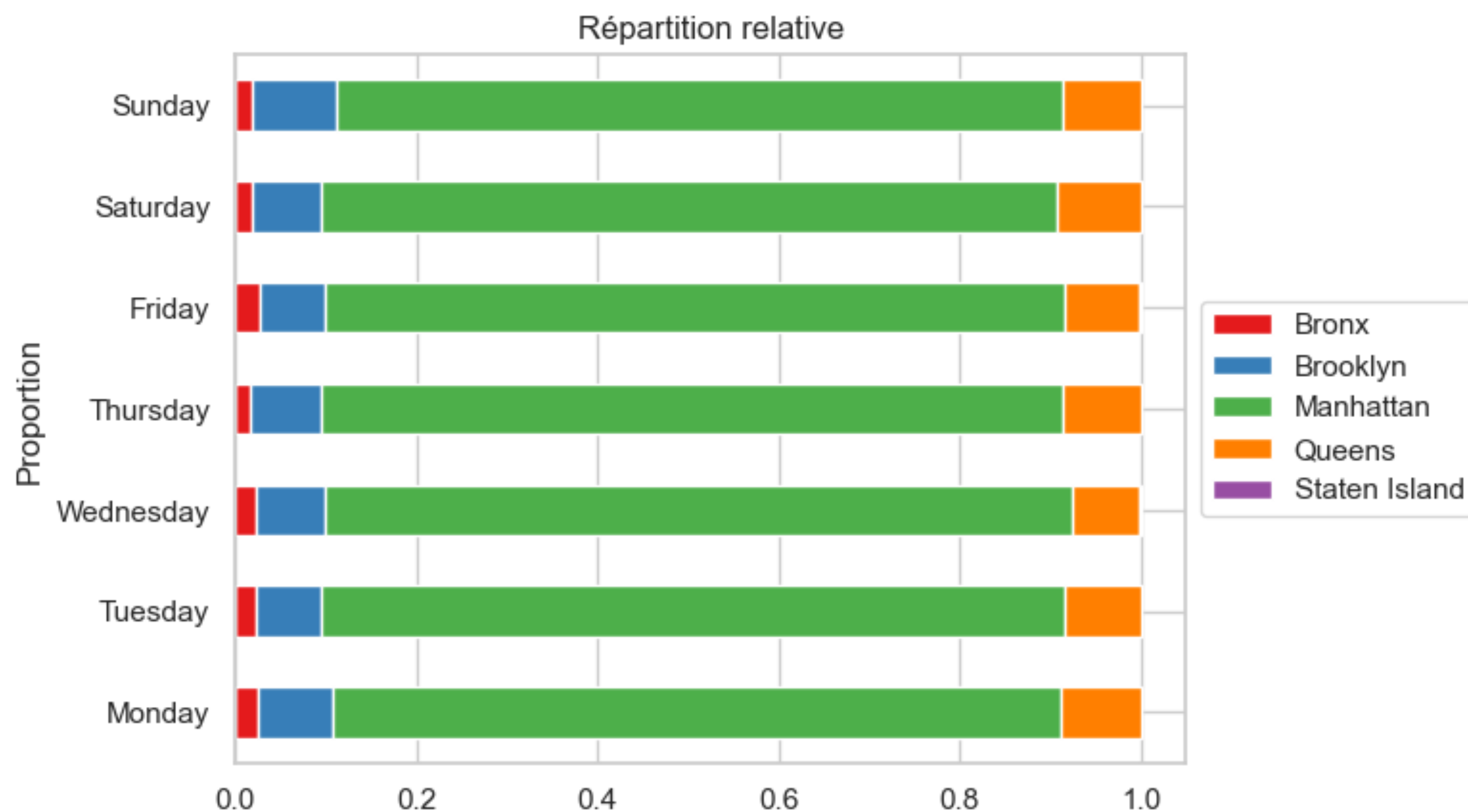
Treemap (carte arborescente) : permet de comparer des valeurs agrégées (moyenne, total, proportion) entre différentes catégories. Il est simple à lire et très efficace pour des comparaisons directes. L'aire de chaque rectangle est proportionnel à la part qu'il représente dans l'ensemble.

Répartition des trajets par jour de prise en charge



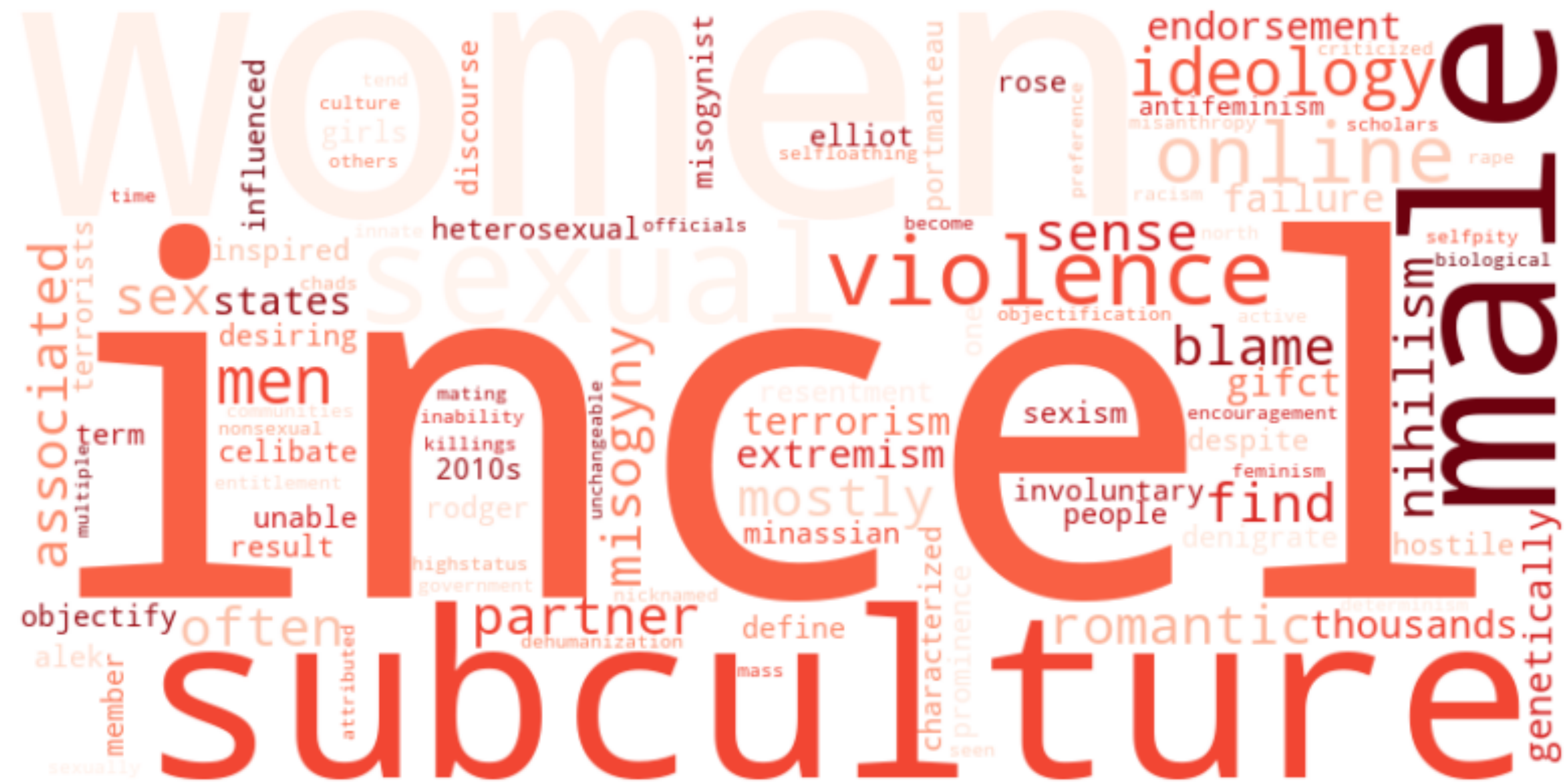
Composition

Diagramme à barres empilés : il permet de comparer la composition d'un total entre plusieurs catégories. Chaque barre représente 100 %. Les segments montrent la proportion de chaque catégorie dans le total.



Com

Wordcloud: Un word cloud est une **représentation visuelle du texte** où la taille de chaque mot reflète sa fréquence ou son importance. Il permet d'identifier rapidement les termes les plus présents et de donner une vue d'ensemble intuitive du contenu. Il est surtout utilisé dans des **contextes qualitatifs** plutôt que quantitatifs."



Relation

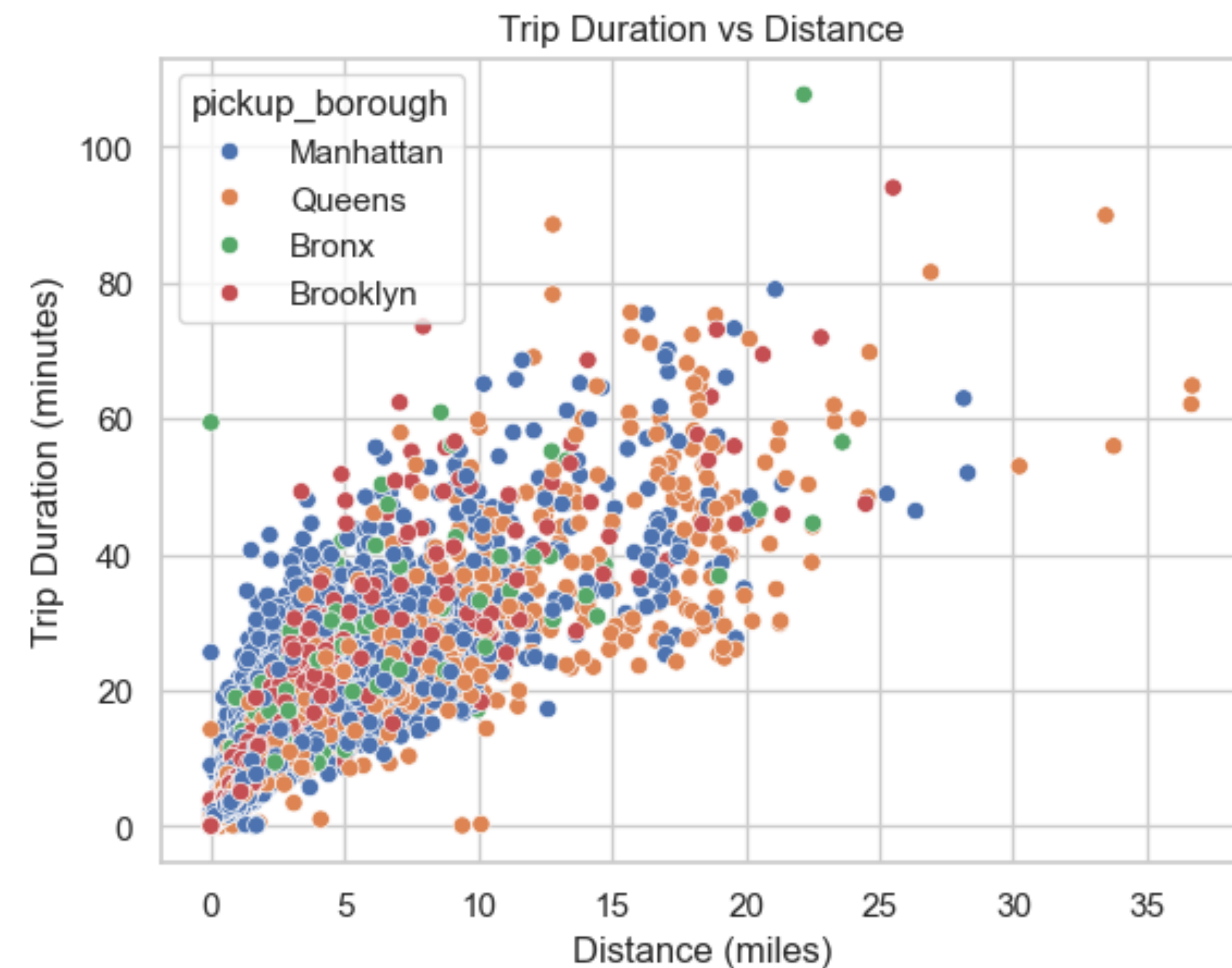
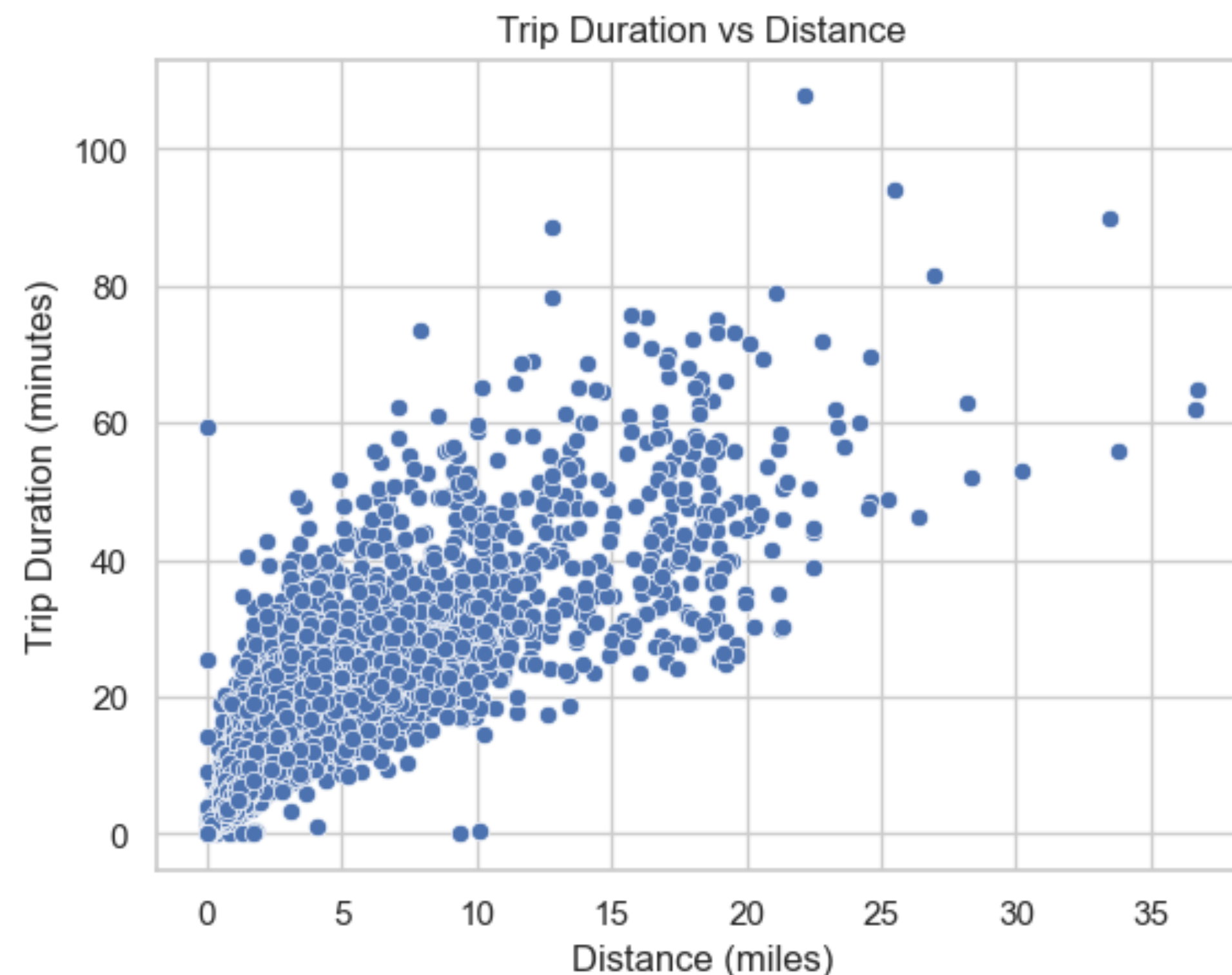
Les graphiques de relation servent à représenter les **liens entre deux ou plusieurs variables**, c'est-à-dire comment une variable peut influencer ou être associée à une autre. Ils sont très utilisés pour explorer et analyser différents types de données afin d'identifier des tendances, des corrélations, des regroupements ou des valeurs atypiques. Ils permettent de répondre à des questions comme : existe-t-il une relation entre les variables ?, quelles variables semblent liées ou indépendantes ?

Ils sont particulièrement utilisés pour :

- comparer des variables numériques (nuage de points, bubble plot) ;
- visualiser des corrélations ou matrices de données (heatmap) ;
- détecter des tendances, des regroupements ou des anomalies dans les données.

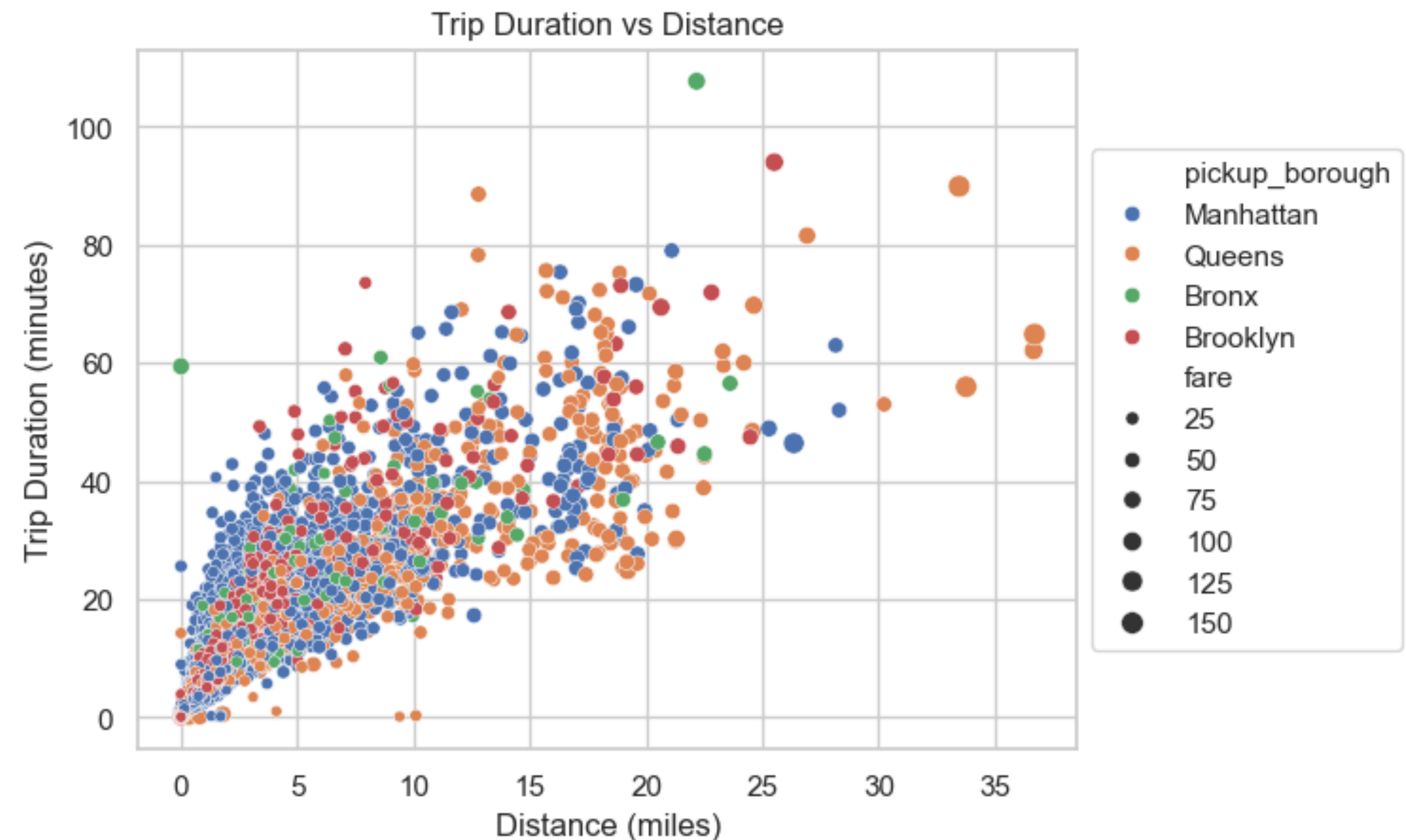
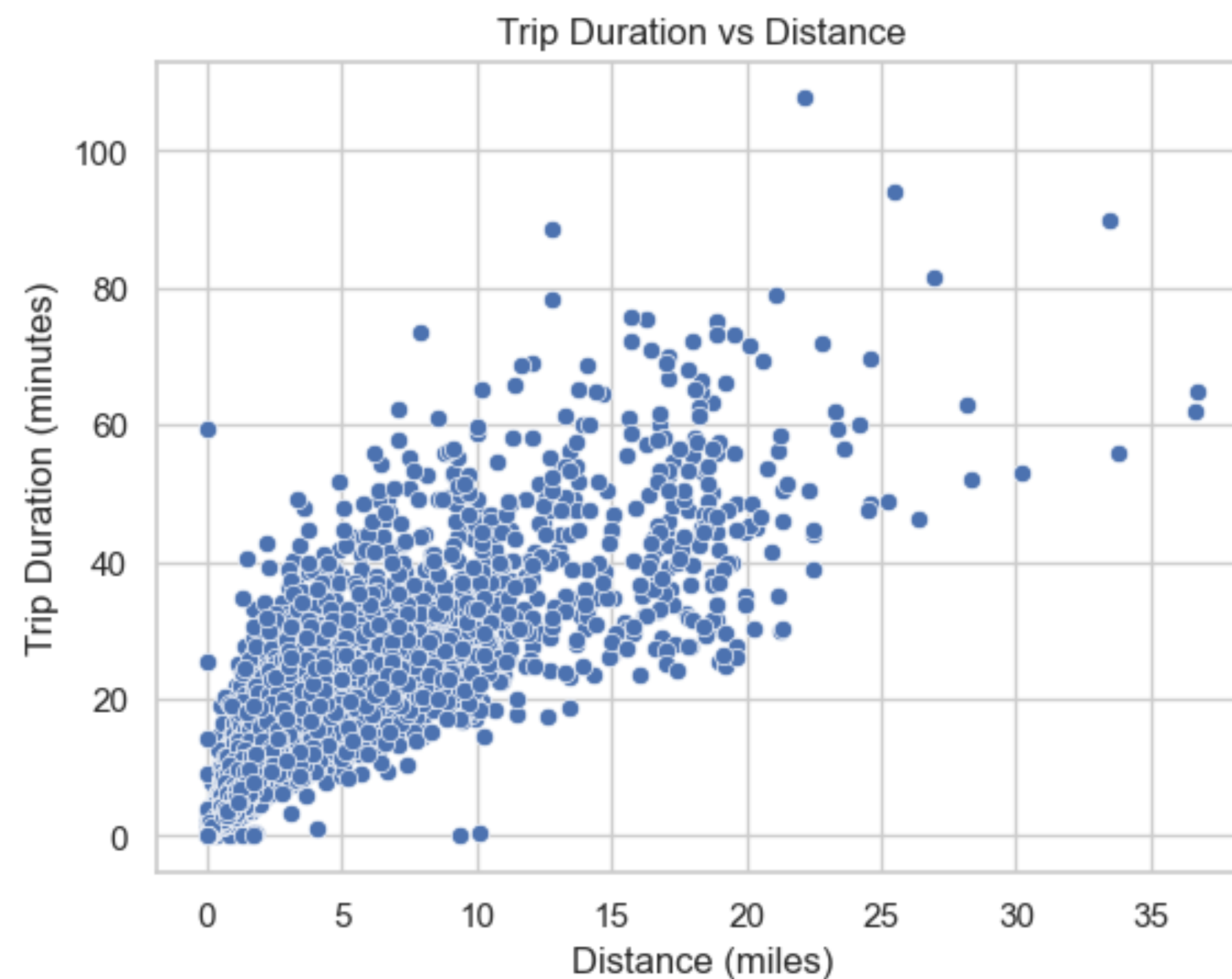
Relation

Nuage de points: Permet de visualiser la relation entre deux variables numériques. Il est simple à lire et très efficace pour détecter des tendances, des regroupements ou des valeurs atypiques. La position de chaque point reflète la valeur des deux variables considérées. Le **couleur** des points peut permettre d'identifier des sous-catégories. La **taille** des points peut montrer la relation avec une troisième variable.



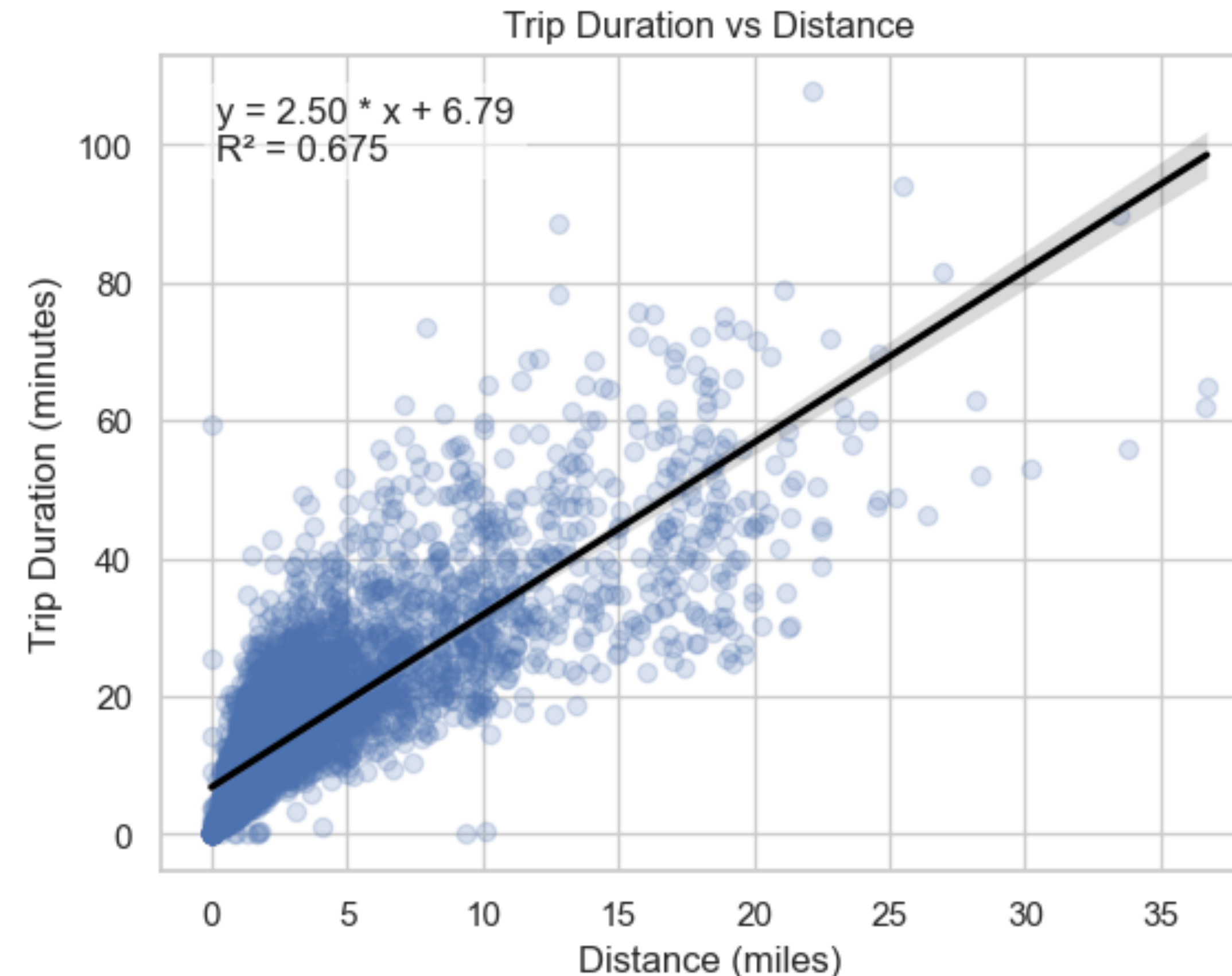
Relation

Nuage de points: Permet de visualiser la relation entre deux variables numériques. Il est simple à lire et très efficace pour détecter des tendances, des regroupements ou des valeurs atypiques. La position de chaque point reflète la valeur des deux variables considérées. Le **couleur** des points peut permettre d'identifier des sous-catégories. La **taille** des points peut montrer la relation avec une troisième variable.



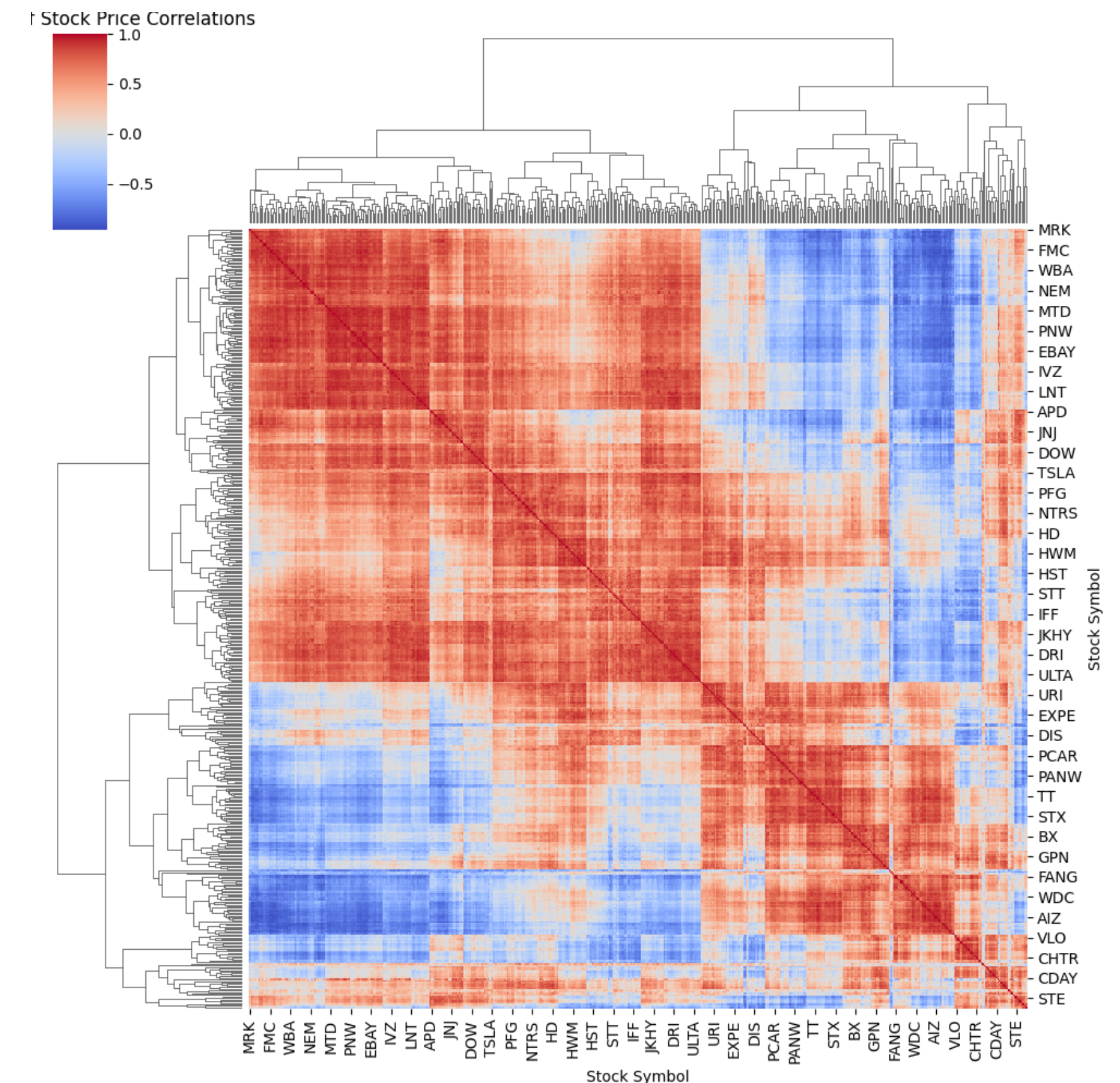
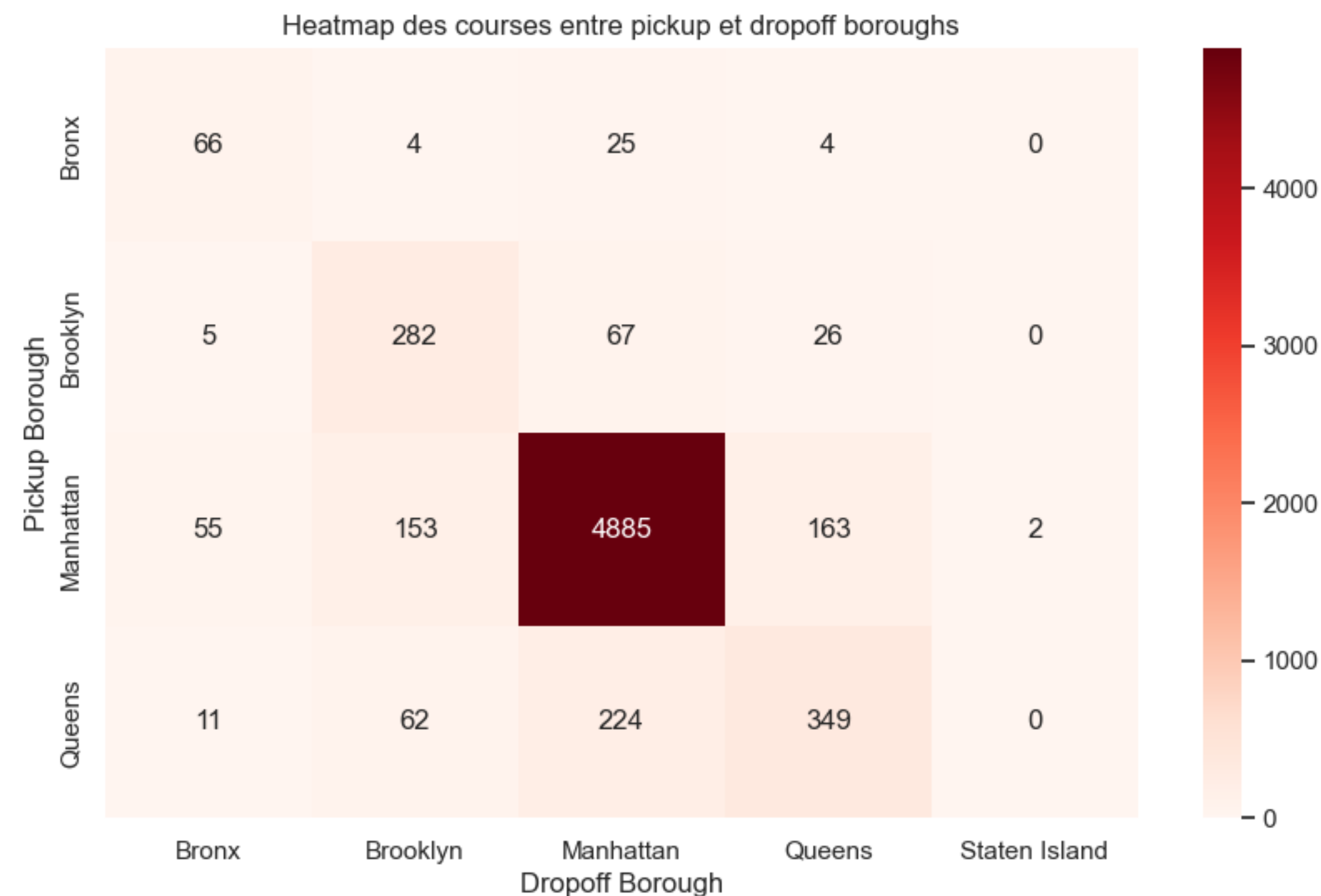
Relation

Graphique de régression: Permet de visualiser la relation entre deux variables numériques en ajustant une ligne de tendance (régression). Il est simple à lire et très efficace pour détecter des relations linéaires, des tendances ou des anomalies. La pente et l'interception de la ligne permettent d'estimer la force et la direction de la relation.



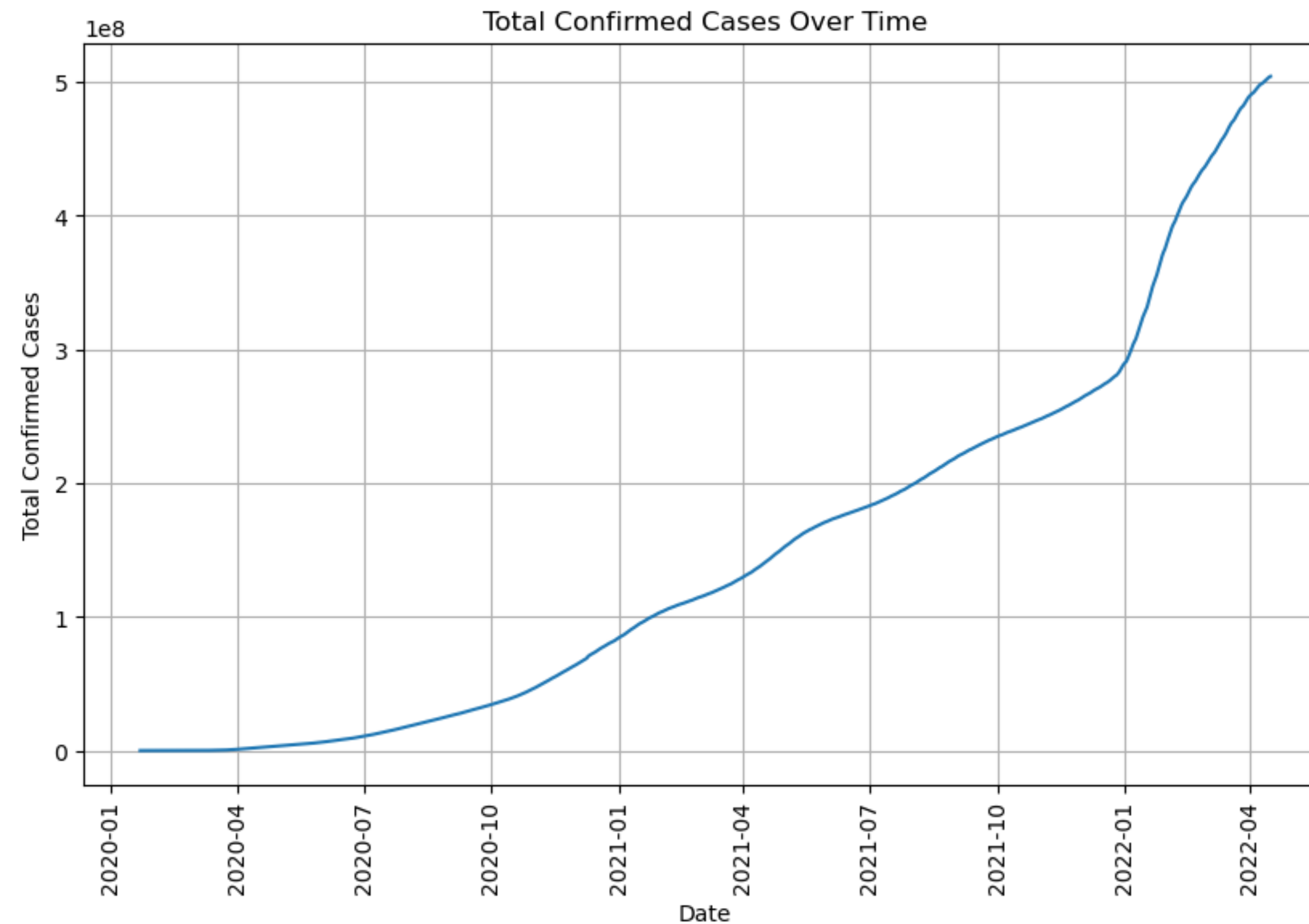
Relation

Heatmap: Permet de visualiser l'intensité ou la valeur d'une variable dans une grille de données, souvent sous forme de matrice. Il est simple à lire et très efficace pour détecter des patterns, corrélations, valeurs extrêmes, associations entre variables ou clusters. La couleur de chaque cellule reflète la valeur correspondante, permettant une interprétation rapide de grandes quantités de données.



Évolution temporelle

Les graphiques d'évolution temporelle servent à suivre l'évolution d'une ou plusieurs variables au fil du temps.



To be continued...

Types de visualisation

Catégorie de graphique	Exemples	Objectif / Ce qu'on montre	Type de variables
Distribution	Histogramme, Box plot, Violin plot, Density plot	Visualiser la distribution d'une variable : tendance centrale, dispersion, forme, outliers	Variable quantitative (continue ou discrète)
Comparaison	Diagramme à barres, Barres groupées, multiple boxplots	Comparer valeurs entre catégories ou groupes	Variable catégorielle + variable quantitative (valeur ou fréquence)
Composition	Barres empilées, Treemap, Pie chart, Aire empilée, WordClouds	Montrer comment un total se répartit entre ses parties / contributions relatives	Variable catégorielle + variable quantitative (part d'un total)
Relations entre variables	Nuage de points, Régression plot, Bubble plot, Heatmap	Identifier corrélations, tendances et patterns entre deux ou plusieurs variables	Deux (ou plus) variables quantitatives (ou mixte quanti + catégorielle pour groupes)
Évolution dans le temps	Graphique en ligne, Aire chart, Step chart	Observer les tendances, variations et saisonnalité au fil du temps	Variable temporelle + variable quantitative
Visualisations complexes / Hiérarchie et flux	Flow chart, Sankey diagram, Radial plot, Chord plot, Dendrogramme	Représenter processus, flux, relations hiérarchiques ou interactions complexes	Données de flux, réseaux ou hiérarchiques (souvent quantitatives + relations)
Visualisation géographique	Carte choroplèthe, Carte à points, Carte thermique	Représenter données spatiales et géographiques	Variable géographique/spatiale + variable quantitative ou catégorielle