

IUT Grand Ouest Normandie

Bachelor Universitaire de Technologie
Science des Données
Campus de Lisieux

Science des Données 1 - SAE 1.03
Préparation et synthèse d'un tableau de données

Thématique

Etude des locations de vélos citibike à New York



Auteurs

Diop Sidy
Houzier Baptiste

Année universitaire 2023-2024

Table des matières

Introduction	3
1 Préparation des données	4
1.1 Traitement des chaînes de caractères	4
1.2 Traitement des dates	4
1.3 Création de la variable duration	4
2 Analyse des données	5
2.1 Répartition des locations selon le type de vélo	5
2.2 Pourcentage de données aberrantes	5
2.3 Histogramme de la variable duration	6
2.4 Distribution de la variable duration	7
2.5 Distribution de la variable duration selon le type de vélo	7
2.6 Distribution du jour de location	8
2.7 Répartition des locations selon les heures	9
2.8 Etudes des trajets entre deux stations différentes	9
2.9 station de départ et station de fin (ordre décroissant)	9
Conclusion	11
3 Annexe 1	12
Table des figures	13

Introduction

Cette étude nous plonge sur l'étude d'un moyen de transport utilisé, les vélos. En septembre 2023, dans la ville de New York nous pouvons comptabiliser au plus de 3 millions de location de vélos citibike.

Ce jeu de données contiennent les informations suivantes :

- Ride ID
- Rideable type
- Started at
- Ended at
- Start station name
- Start station ID
- End station name
- End station ID
- Start latitude
- Start longitude
- End latitude
- End longitude
- Member or casual ride

En suivant la fiche projet, nous avons réalisé des analyses avec le langage R sur RStudio afin de de voir. Les données ont été sourcées de

[Citibike - New York.](#)

ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	
1	B0A0F1DEFA4B72FC	electric_bike	2023-09-03 10:20:41	2023-09-03 10:24:16	E 1 St & Bowery	5636.13	E 10 St & 2 Ave
2	2B26AB15647BF4EE	classic_bike	2023-09-27 15:44:23	2023-09-27 15:53:25	Pearl St & Hanover Square	4993.02	Allen St & Rivington St
3	9D2B5971CA4E513F	classic_bike	2023-09-19 13:40:48	2023-09-19 13:48:11	E 1 St & Bowery	5636.13	E 10 St & 2 Ave
4	17E6760596DC3ABE	classic_bike	2023-09-30 16:27:50	2023-09-30 16:56:35	Central Ave & Himrod St	4713.01	Mott St & Prince St
5	97EFF376A7E2DC70	classic_bike	2023-09-21 16:59:53	2023-09-21 17:07:36	St Marks Pl & 2 Ave	5669.10	Mott St & Prince St
6	D14BD5627029FFEA	classic_bike	2023-09-11 14:50:19	2023-09-11 14:54:47	Meserole Ave & Manhattan Ave	5666.04	Franklin St & Dupont St
7	D4AF7211BE4B6C54	classic_bike	2023-09-06 17:19:49	2023-09-06 17:23:40	Pearl St & Hanover Square	4993.02	Fulton St & William St
8	775E8EF8E42CF867	classic_bike	2023-09-26 17:48:10	2023-09-26 18:04:14	Adams St & Prospect St	4821.10	Washington Ave & Gree
9	2CC42BB00C45A968	classic_bike	2023-09-15 13:14:38	2023-09-15 14:18:07	West Broadway & Watts St	5569.07	Allen St & Rivington St
10	7FEB0C3990129D56	classic_bike	2023-09-06 03:29:02	2023-09-06 03:41:42	E 117 St & 1 Ave	7579.11	Tinton Ave & E 165 St

start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
5636.13	E 10 St & 2 Ave	5746.02	40.72486	-73.99213	40.72971	-73.98660	member
4993.02	Allen St & Rivington St	5414.06	40.70465	-74.00913	40.72020	-73.98998	member
5636.13	E 10 St & 2 Ave	5746.02	40.72475	-73.99212	40.72971	-73.98660	member
4713.01	Mott St & Prince St	5561.04	40.69671	-73.92293	40.72318	-73.99480	member
5669.10	Mott St & Prince St	5561.04	40.72842	-73.98714	40.72318	-73.99480	member
5666.04	Franklin St & Dupont St	5944.01	40.72694	-73.95300	40.73564	-73.95866	member
4993.02	Fulton St & William St	5137.11	40.70472	-74.00926	40.70960	-74.00655	member
4821.10	Washington Ave & Greene Ave	4419.03	40.70116	-73.98874	40.68650	-73.96563	member
5569.07	Allen St & Rivington St	5414.06	40.72323	-74.00314	40.72020	-73.98998	casual
7579.11	Tinton Ave & E 165 St	7991.01	40.79656	-73.93445	40.82480	-73.90242	member

1 Préparation des données

1.1 Traitement des chaînes de caractères

Avant de commencer l'analyse des données, nous avons dû d'abord préparer les données en changeant le type de certaines données. Nous avons commencé par changer le type de rideable type en factor, ce qui nous permettra par la suite de mieux la manipuler.

Combien dénombre-t-on de location de vélos pour le mois de juin ?

Nous démontrons 3 575 162 locations de vélos.

1.2 Traitement des dates

Par la suite nous avons changé le type de started_at en POSIXct car c'est une date. Nous avons dû également mettre le format.

```
# Traitement des chaînes de caractères

dataset$rideable_type = as.factor(dataset$rideable_type)
class(dataset$rideable_type)
# [1] "factor"

# Traitement des dates

dataset$started_at = as.POSIXct(dataset$started_at,
                                format = "%Y-%m-%d %H:%M:%S")
class(dataset$started_at)

dataset$ended_at = as.POSIXct(dataset$ended_at,
                               format = "%Y-%m-%d %H:%M:%S")
class(dataset$ended_at)

#création de la variable duration

dataset$duration <- as.numeric(difftime(dataset$ended_at,
                                         dataset$started_at,
                                         units = "min"))
```

1.3 Création de la variable duration

Nous avons créé cette variable car elle nous permet de voir le temps de location des vélos.

2 Analyse des données

Dans cette Partie, nous allons nous consacrer à l'analyse des données concernant la location des vélos mis à disposition de la population de New York aux Etats-Unis durant le mois de Septembre 2023. Nous allons : identifier le type de vélo le plus utilisé, le pourcentage de données aberrantes présentes dans notre jeu de données, l'évolution de la variable duration, identifier la densité de la variable duration selon le type de vélo utilisé, identifier le jour de la semaine dans lequel on a le plus de location de vélos et enfin identifier l'heure à laquelle les vélos sont le plus loués.

2.1 Répartition des locations selon le type de vélo

Ce graphique a pu être réalisé avec la fonction `barplot()`. Il montre la répartition des locations selon le type de vélo. En effet le classic bike est le type de vélo le plus utilisé par les New Yorkais en Septembre 2023. On peut observer (cf. Fig. 1) que pendant ce mois de septembre, le type de vélo le plus utilisé par les New Yorkais est le vélo classique (classique bike) avec 92,88% contre 7,12% pour le vélo électrique.

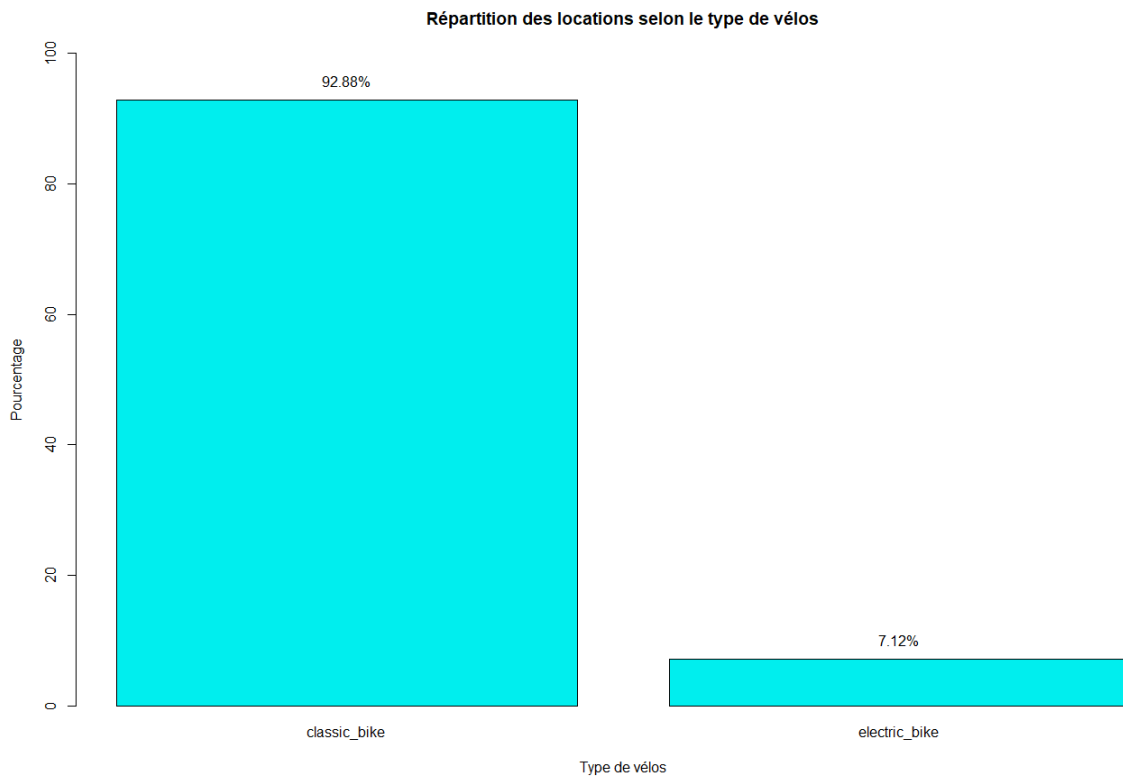


Fig. 1. Répartition des objets trouvés selon le département

2.2 Pourcentage de données aberrantes

Ce graphique a pu être réalisé avec la fonction `barplot()`. Il montre le pourcentage de valeurs aberrantes présent dans notre fichier de données. On peut observer (cf. Fig. 2) que l'on a à 99.96% de bonne valeur soit 0.04% de valeur aberrante. Ces valeurs aberrantes sont négligeables. En effet nous avons un jeu de données contenant exactement 3 575 162 valeurs et les 0.04% de valeurs aberrantes représentent 1430 personnes que l'on peut négliger.

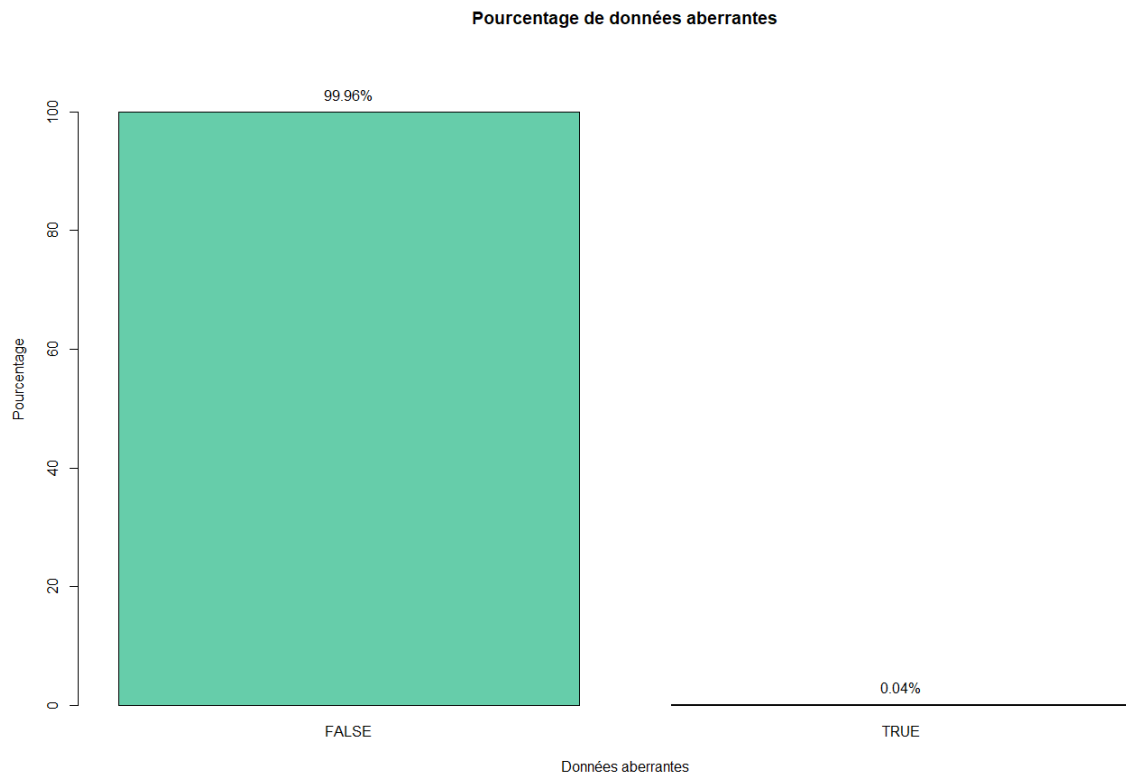


Fig. 2. Pourcentage de données aberrantes

2.3 Histogramme de la variable duration

Ce graphique a pu être réalisé avec la fonction `hist()`. Il représente (cf. Fig. 3) la durée de location des vélos citibike par rapport au nombre de vélos loués. On peut donc observer le temps pendant lequel les clients louent les vélos. Selon ce graphique, le temps de location d'un vélo le plus fréquent tourne autour des 10 minutes. Puis 5 minutes. On peut donc en déduire que les usagers n'utilisent ces vélos que pour les courts trajets.

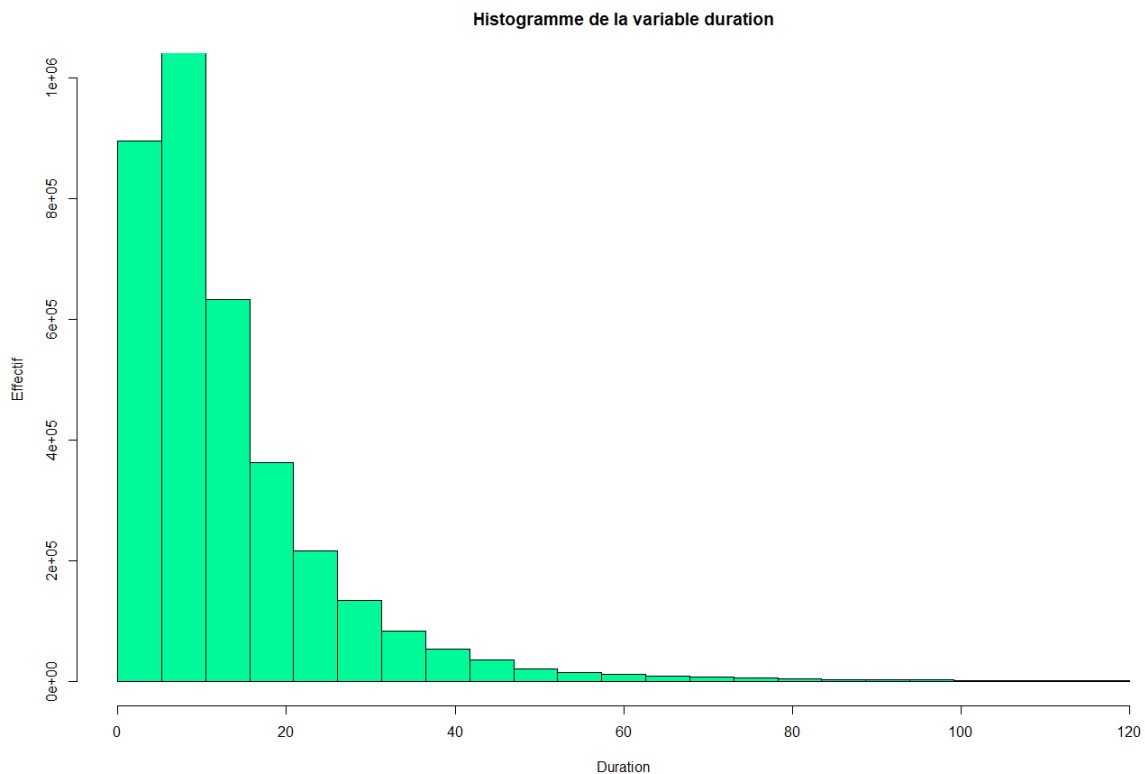


Fig. 3. Histogramme de la variable duration

2.4 Distribution de la variable duration

Ce graphique a été réalisé via la fonction `hist()`. Il permet (cf. Fig. 4) de comprendre la distribution de la variable duration. Ce graphique est en rapport avec le graphique précédent. Cela nous montre la répartition de la variable duration, c'est-à-dire la distribution du temps de location de vélo par les utilisateurs. Nous pouvons observer que les utilisateurs utilisent les vélos en grande majorité entre 0 et 20 minutes.

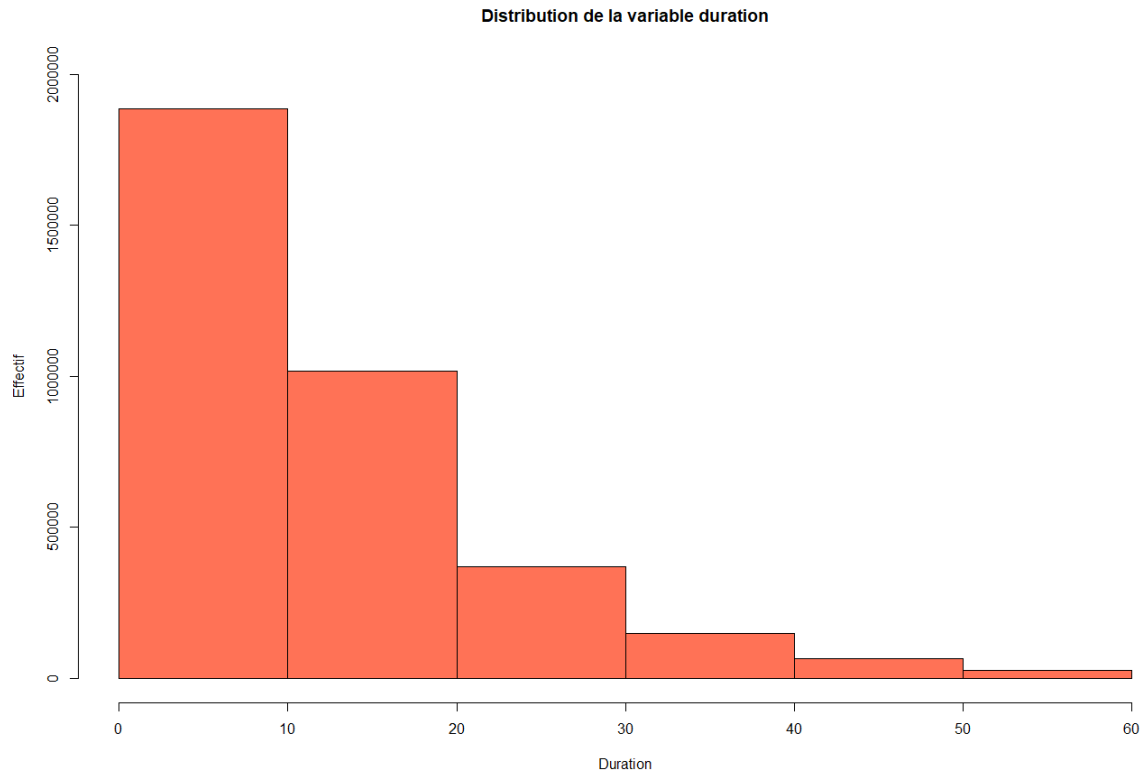


Fig. 4. Distribution de la variable duration

2.5 Distribution de la variable duration selon le type de vélo

Afin de faire ce graphique, nous avons utilisé les fonction `plot()` et `lines()`. Ce graphique (cf. Fig. 5) nous explique la distribution de la variable duration selon le type de vélo. Comme le graphique vu ci-dessus il est en rapport avec les graphiques précédent. Nous pouvons voir que même avec un vélo classique ou un vélo électrique, l'utilisateur a le même temps de location qui est le plus élevé à 10 minutes. La durée de location diminue énormément jusqu'à 20 minutes. De 30 minutes à 60 minutes cela stagne.

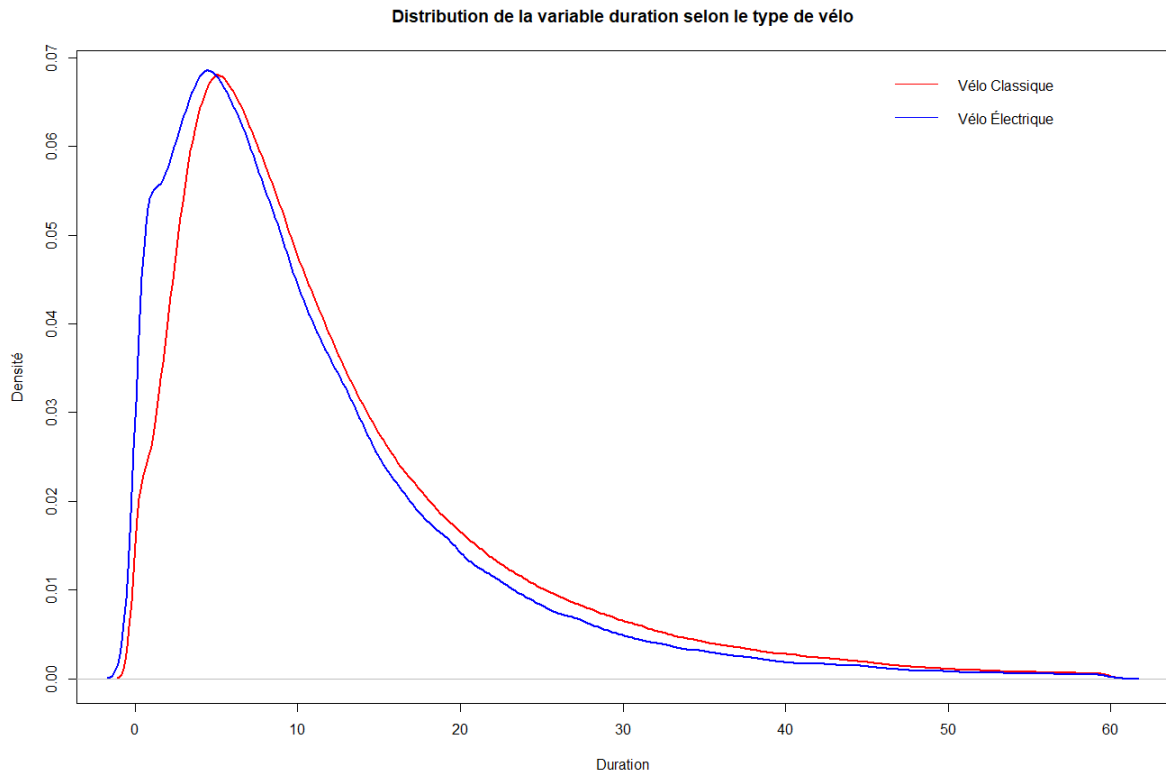


Fig. 5. Distribution de la variable duration selon le type de vélo

2.6 Distribution du jour de location

Pour ce graphique, nous avons d'abord créé une variable "weekdays" avec les fonctions `weekday()` et `ordered()` pour mettre les jours de la semaine et dans le bon ordre.

Par la suite nous avons manié la fonction `barplot()`. Ce graphique (cf. Fig. 6) explique la distribution du jour de location, c'est-à-dire la répartition des jours de locations des vélos. Comme nous pouvons le voir sur ce graphique le jour avec le plus de locations de vélos est le dimanche avec 609 672 jours. Le jour le moins utilisé est le mercredi avec un peu plus de 330 000 jours.

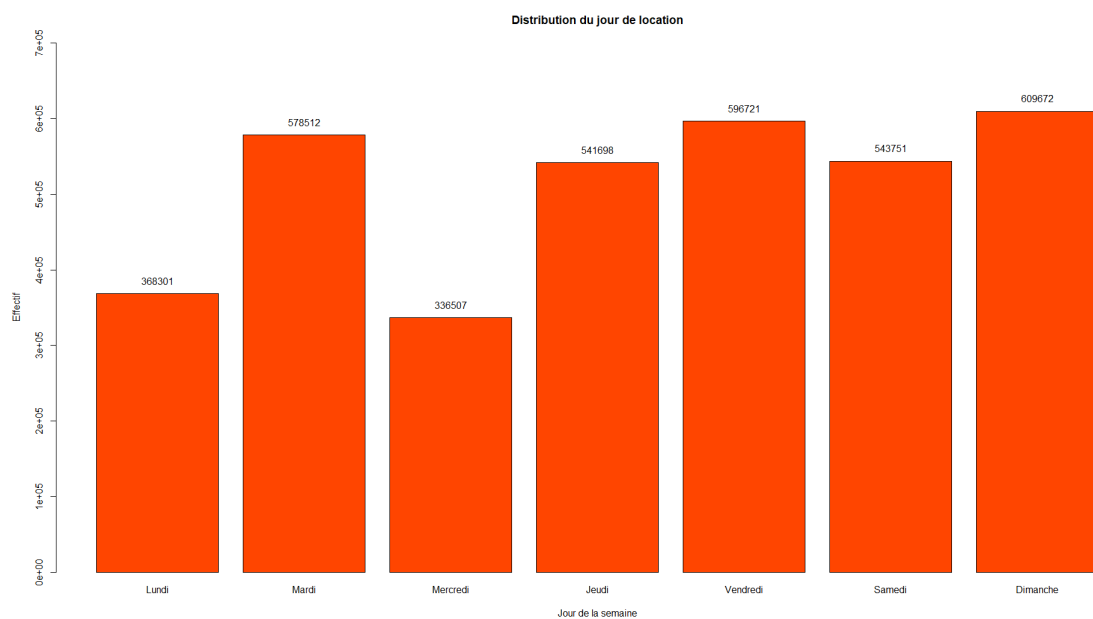


Fig. 6. Distribution du jour de location

2.7 Répartition des locations selon les heures

Pour ce graphique nous avons du manier la librairie lubridate qui permet de manipuler les heures ainsi que les dates. Ensuite nous avons crée un facteur avec des étiquettes pour chaque jour de la semaine. Nous avons utilisé la fonction `barplot()` pour ce graphique. Il (cf. Fig. 7) montre la répartition des locations des vélos selon les heures. Nous pouvons constater que la location de vélos est élevé vers 17h avec un pourcentage de 8.84 pourcents. Le seconde est juste après, c'est-à-dire 18h avec 8.67 pourcents. les heures avec le moins de locations de vélo sont de minuit à 6h du matin avec 0.37 pourcents à 1.9 pourcents. Cela est compréhensible car à ces heures les gens sont généralement en train de dormir.

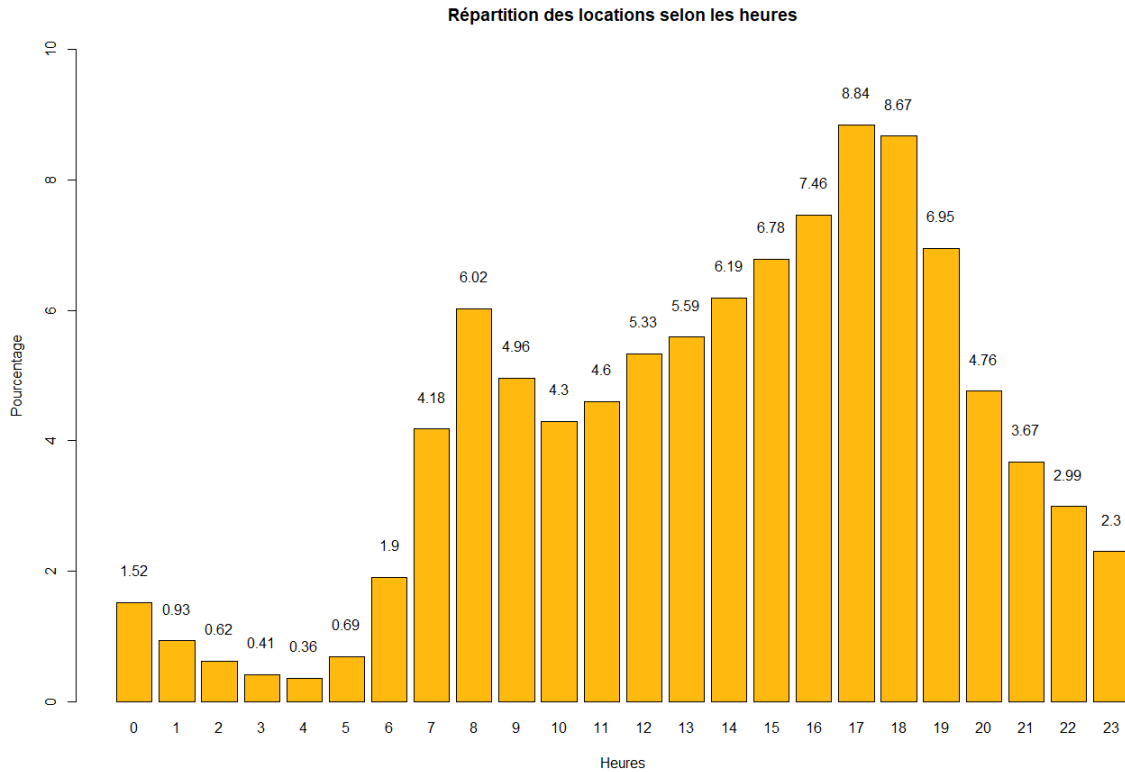


Fig. 7. Répartition des locations selon les heures

2.8 Etudes des trajets entre deux stations différentes

Maintenant, on désire étudier les trajets. Dans ce but, et à l'aide de la fonction `paste()`, nous avons réalisée une concaténation de la station de départ et de la station d'arrivée dans une variable nommée `travel`. Dans un deuxième temps, nous avons utilisée la fonction `table()` et la fonction `sort()` pour éditer un top 10 des trajets les plus fréquents.

Comme nous pouvons le voir (cf. Fig. 8), le trajet le plus utilisé est Central Park S et 6 AveCentral Park S et 6 Ave avec 1799 trajets.

Dock St & Front St	Old Fulton St,	Old Fulton St	Dock St & Front St,
1096	927		
North Moore St & Greenwich St	Essey St & Church St,	E 10 St & Avenue A	E 11 St & Avenue B,
664	664		646
Broadway & W 51 St	Broadway & W 53 St,	St Marks Pl & 2 Ave	St Marks Pl & 1 Ave,
619	619		602
6 Ave & W 34 St	Broadway & W 36 St,	W 41 St & 8 Ave	Broadway & W 41 St,
567	567		566
Central Park S & 6 Ave	5 Ave & E 87 St,	West St & Chambers St	Pier 40 - Hudson River Park,
561	561		558

Fig. 8. Etudes des trajets entre deux stations différentes

2.9 station de départ et station de fin (ordre décroissant)

A présent, On souhaite se concentrer sur les trajets entre deux stations différentes. Pour cela, nous avons du créer une variable `test3` binaire (TRUE ou FALSE) via la fonction `ifelse()` pour détecter deux stations différentes. Via les fonctions `table()` et `prop.table()`. Ensuite nous avons déterminé le pourcentage de trajets entre deux stations différentes et réalisée un filtre pour ne conserver que les données relatives à un trajet entre deux stations différentes. on a ensuite

enregistré ces données dans un objet nommé temp6. Enfin, à l'aide des fonction table() et sort(), nous avons également fait un top 10 des trajets entre deux stations différentes.

```

Central Park S & 6 AveCentral Park S & 6 Ave,
1799
1324
Dock St & Front StOld Fulton St,
1096
Old Fulton StDock St & Front St,
927
West St & Chambers StWest St & Chambers St,
837
7 Ave & Central Park South7 Ave & Central Park South,
836
Broadway & W 58 StBroadway & W 58 St,
825
11 Ave & W 41 St11 Ave & W 41 St,
733
Grand Army Plaza & Central Park SGrand Army Plaza & Central Park S,
729
12 Ave & W 40 St12 Ave & W 40 St,
691
> |

```

Fig. 9. station de départ et station de fin (ordre décroissant)

Conclusion

Après une analyse approfondie des données sur les locations de vélos Citibike à New York en septembre 2023, plusieurs conclusions significatives peuvent être tirées.

Premièrement, cette étude a révélé que le vélo classique était le type de vélo le plus largement utilisé par les habitants de New York, représentant environ 92,88% des locations, tandis que les vélos électriques ne comptaient que pour environ 7,12% des locations.

Deuxièmement, concernant la qualité des données, le jeu de données présentait une fiabilité élevée, avec seulement 0,04 pourcents de valeurs aberrantes, représentant un nombre négligeable par rapport à la taille totale du jeu de données, soit 3 575 162 valeurs.

Troisièmement, en ce qui concerne la durée des locations, la majorité des utilisateurs louaient les vélos pour des trajets courts, avec une durée moyenne de location autour de 10 minutes, indiquant une utilisation principalement pour de petits trajets.

Quatrièmement, l'analyse a montré une similarité dans la durée de location, indépendamment du type de vélo (classique ou électrique), avec une concentration significative de locations autour des 10 premières minutes, suivie par une diminution marquée pour les durées supérieures.

Cinquièmement, en termes de répartition des locations selon les jours de la semaine, une tendance a été observée, indiquant un pic d'utilisation à certains jours par rapport à d'autres, mais avec une utilisation généralement stable tout au long de la semaine, à l'exception du mercredi qui a enregistré un peu plus de 330 000 locations, inférieur aux autres jours.

Sixièmement, en ce qui concerne les heures de location, un pic d'utilisation a été observé autour de 17h et 18h, représentant environ 8,84% et 8,67% respectivement, tandis que les heures les moins utilisées étaient entre minuit et 6h du matin, suggérant une baisse d'utilisation pendant les heures de sommeil.

Enfin, l'étude des trajets les plus fréquents a permis de déterminer les 10 trajets les plus courants entre différentes stations, fournissant ainsi des informations supplémentaires sur les itinéraires les plus populaires empruntés par les utilisateurs.

En somme, cette analyse détaillée des données de location de vélos Citibike à New York en septembre 2023 offre une vision approfondie des habitudes d'utilisation, des préférences et des tendances des utilisateurs de vélos partagés dans cette ville, permettant ainsi des insights précieux pour les améliorations futures du système de partage de vélos et pour mieux répondre aux besoins des citoyens.

3 Annexe 1

System Data

Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular? What days of the week are most rides taken on? We've heard all of these questions and more from you, and we're happy to provide the data to help you discover the answers to these questions and more. We invite developers, engineers, statisticians, artists, academics and other interested members of the public to use the data we provide for analysis, development, visualization and whatever else moves you.

This data is provided according to the [Citi Bike Data Use Policy](#).

Citi Bike Trip Histories

We publish [downloadable files of Citi Bike trip data](#). The data includes:

- Ride ID
- Rideable type
- Started at
- Ended at
- Start station name
- Start station ID
- End station name
- End station ID
- Start latitude
- Start longitude
- End latitude

Ce rapport se base sur les données relatives aux vélos loués par le réseaux citibike dans la ville de New York en Septembre 2023. Ces données sont directement accessibles via le lien suivant

[Données citibike](#)

Ces données concernent les années civiles de 2017 à 2023. Vous trouverez ci-dessous une copie d'écran du site

Index of bucket "tripdata"
























Name	Date Modified	Size	Type
 201306-citibike-tripdata.zip	Apr 30th 2018, 03:18:55 pm	16.79 MB	ZIP file
 201307-201402-citibike-tripdata.zip	Jan 18th 2017, 11:23:25 pm	178.26 MB	ZIP file
 201307-citibike-tripdata.zip	Jan 18th 2017, 11:23:27 pm	27.07 MB	ZIP file
 201308-citibike-tripdata.zip	Jan 18th 2017, 11:23:27 pm	32.09 MB	ZIP file
 201309-citibike-tripdata.zip	Jan 18th 2017, 11:23:27 pm	33.16 MB	ZIP file
 201310-citibike-tripdata.zip	Jan 18th 2017, 11:23:28 pm	33.07 MB	ZIP file
 201311-citibike-tripdata.zip	Jan 18th 2017, 11:23:28 pm	21.62 MB	ZIP file
 201312-citibike-tripdata.zip	Jan 18th 2017, 11:23:28 pm	14.31 MB	ZIP file
 201401-citibike-tripdata.zip	Jan 18th 2017, 11:23:29 pm	9.70 MB	ZIP file
 201402-citibike-tripdata.zip	Jan 18th 2017, 11:23:29 pm	7.25 MB	ZIP file
 201403-citibike-tripdata.zip	Jan 18th 2017, 11:23:29 pm	14.13 MB	ZIP file
 201404-citibike-tripdata.zip	Jan 18th 2017, 11:23:29 pm	21.41 MB	ZIP file
 201405-citibike-tripdata.zip	Jan 18th 2017, 11:23:29 pm	27.59 MB	ZIP file
 201406-citibike-tripdata.zip	Jan 18th 2017, 11:23:29 pm	29.90 MB	ZIP file
 201407-citibike-tripdata.zip	Jan 18th 2017, 11:23:30 pm	30.89 MB	ZIP file
 201408-citibike-tripdata.zip	Jan 18th 2017, 11:23:30 pm	30.63 MB	ZIP file
 201409-citibike-tripdata.zip	Jan 18th 2017, 11:23:30 pm	30.25 MB	ZIP file
 201410-citibike-tripdata.zip	Jan 18th 2017, 11:23:30 pm	26.15 MB	ZIP file
 201411-citibike-tripdata.zip	Jan 18th 2017, 11:23:31 pm	16.83 MB	ZIP file
 201412-citibike-tripdata.zip	Jan 18th 2017, 11:23:31 pm	12.72 MB	ZIP file
 201501-citibike-tripdata.zip	Jan 18th 2017, 11:23:31 pm	7.01 MB	ZIP file
 201502-citibike-tripdata.zip	Jan 18th 2017, 11:23:31 pm	4.82 MB	ZIP file
 201503-citibike-tripdata.zip	Jan 18th 2017, 11:23:31 pm	8.43 MB	ZIP file

Table des figures

1	Répartition des objets trouvés selon le département	5
2	Pourcentage de données aberrantes	6
3	Histogramme de la variable duration	6
4	Distribution de la variable duration	7
5	Distribution de la variable duration selon le type de vélo	8
6	Distribution du jour de location	8
7	Répartition des locations selon les heures	9
8	Etudes des trajets entre deux stations différentes	9
9	station de départ et station de fin (ordre décroissant)	10