INGÉNIERIE STATISTIQUE & FINANCIÈRE
UNIVERSITÉ PANTHÉON-ASSAS

Prédiction des durées de trajet des taxis new-yorkais

Projet R

Réalisés par :

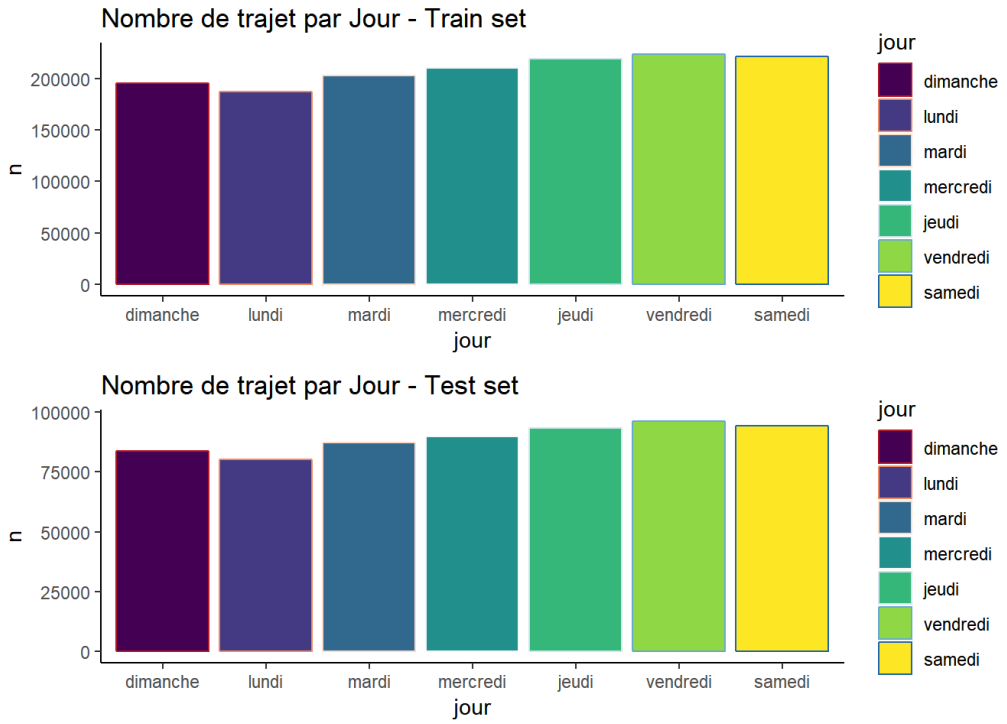
Abdul Fazila, Brehier Hugo, Ndione Anta Salimata

Avril 2019

Introduction

Nous souhaitons prédire les durées de trajet des taxis new-yorkais de janvier à juin 2016 à partir des heures de prise en charge, de dépôt, des jours et en ajoutant également les conditions météorologiques.

- Analysez la cohérence entre les deux jeux de données train et test nombre de trajets par jour/mois, position géographique des prises en charge



Nous avons, tout d'abord analysé les deux bases train et test, elles présentent les mêmes distributions des jours et positions de prise en charge. Nous avons ensuite effectué toutes nos analyses sur la base train.

- Donnez les principaux indicateurs statistiques de la durée de prise en charge.

```
summary(train$trip_duration)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1     397     662     959    1075 3526282
```

La durée des voyages est formatée en secondes, avec une moyenne de 15 minutes.

La moitié des trajets se situe entre 6 minutes et 18 minutes.

Avec un maximum dans les centaines d'heures, on peut penser qu'il y aura des valeurs aberrantes dans le dataset.

- Analysez les données aberrantes

```
#NAs ?
sum(is.na(train))
```

```
## [1] 0
```

```
sum(is.na(test))
```

```
## [1] 0
```

On remarque tout d'abord qu'il n'y a aucune valeur manquante dans le train ou le test.

```
#points abberants en durée ? 22hrs !
train %>% arrange(desc(trip_duration)) %>%
 mutate(trip_hr = (trip_duration /3600)) %>% select(trip_hr)
```

```
## # A tibble: 1,458,644 x 1
##    trip_hr
##      <dbl>
##  1   980.
##  2   619.
##  3   569.
##  4   539.
##  5    24.0
##  6    24.0
##  7    24.0
##  8    24.0
##  9    24.0
## 10    24.0
## # ... with 1,458,634 more rows
```

Il y a cependant des trajets de 24 heures, voire quelques-uns de centaines d'heures !

Ceux-ci sont aberrants.

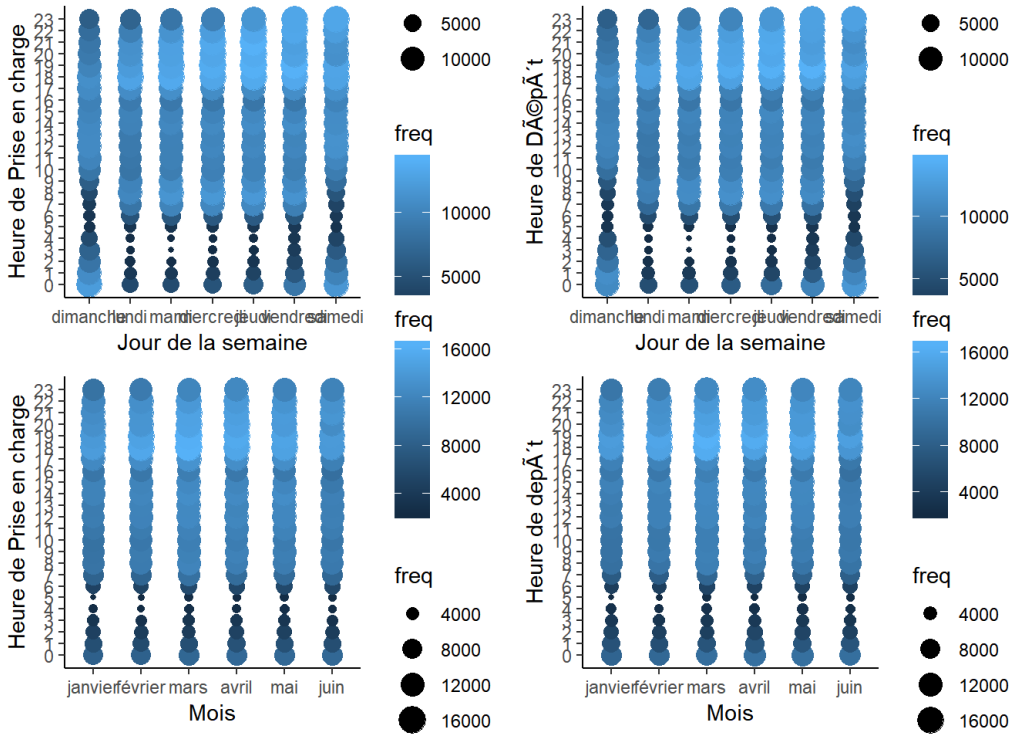
```
train2 = train %>% filter(passenger_count == 0)
glimpse(train2)
```

```
Observations: 60
Variables: 11
$ id                <fct> id3917283, id3645383, id2840829, id3762593, id2154895, id0796773, id2091096, ...
$ vendor_id         <int> 2, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 2, 1, 1,...
$ pickup_datetime   <dttm> 2016-06-06 16:39:09, 2016-01-01 05:01:32, 2016-02-21 01:33:52, 2016-01-04 12...
$ dropoff_datetime  <dttm> 2016-06-07 16:30:50, 2016-01-01 05:01:36, 2016-02-21 01:36:27, 2016-01-04 13...
$ passenger_count   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ pickup_longitude  <dbl> -73.77637, -73.99313, -73.94624, -73.81522, -73.86163, -73.95494, -73.99365, ...
$ pickup_latitude   <dbl> 40.64525, 40.75747, 40.77290, 40.70008, 40.70503, 40.68787, 40.75705, 40.7383...
$ dropoff_longitude <dbl> -73.77636, -73.99329, -73.94677, -73.95070, -73.86163, -73.95474, -73.91887, ...
$ dropoff_latitude  <dbl> 40.64526, 40.75754, 40.77484, 40.75522, 40.70503, 40.68786, 40.75779, 40.7383...
$ store_and_fwd_flag <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, Y, N, N, N, N, N, N, N, N, N, N,...
$ trip_duration     <int> 85901, 4, 155, 2251, 8, 9, 2072, 15, 41, 15, 7, 1556, 22, 105, 13, 5, 4, 49, ...
```

Il y a également une soixantaine de trajets sans passagers.

On peut les enlever.

- Analysez les heures de prise en charge et de fin de la course

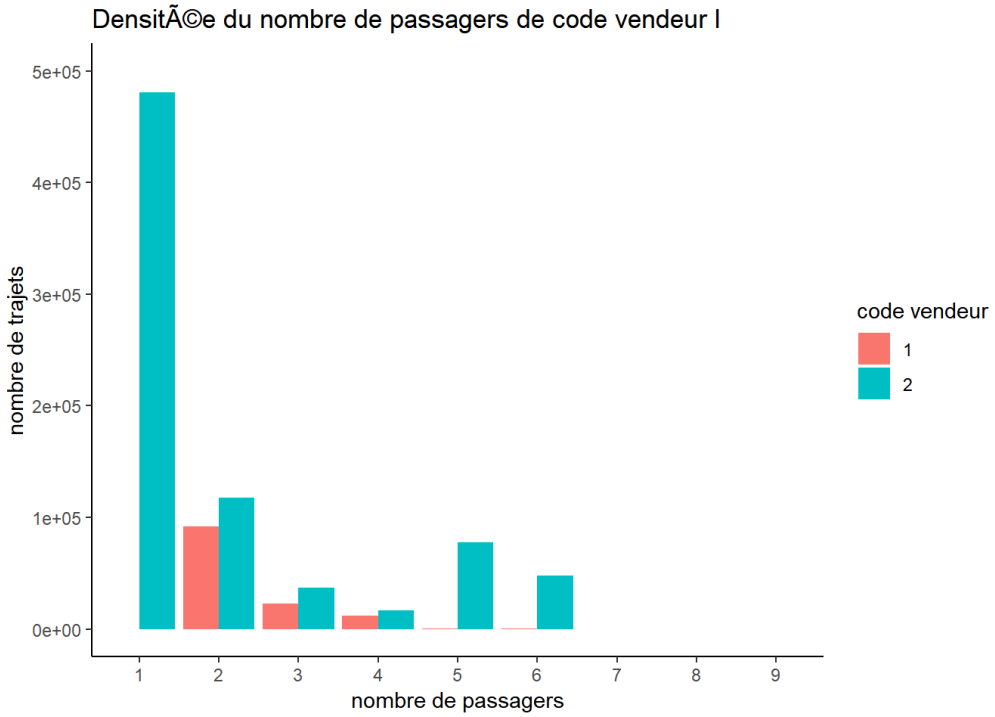


Nous remarquons, que très peu de trajet ont lieu entre 2h et 6h en semaine. Les trajets commencent aux alentours de 7h du lundi au vendredi, tandis que les week-ends, nous constatons qu'il y a des trajets entre minuit et 5h du matin et très peu de trajet aux alentours de 5h-6h du matin. En week-end les trajets commencent aux alentours de 9h.

Il y a énormément de trajet durant le reste de la journée et plus particulièrement en fin d'après-midi. Cependant, nous constatons que le mardi, nous avons très peu de dépôts entre 4h et 5h du matin.

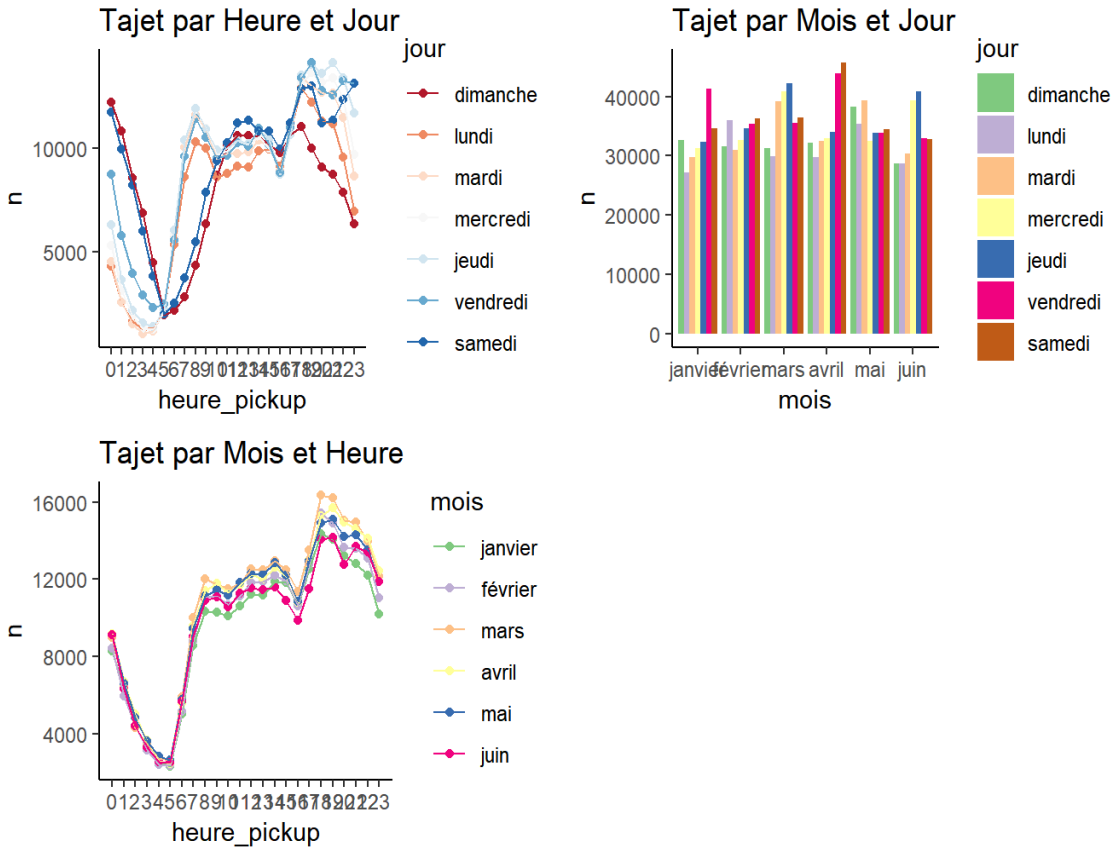
Nous n'avons pas de différences significatives entre les heures de prise en charge et dépôt mensuellement.

- Analysez le nombre de passagers par trajet. Tracez la distribution ou histogramme. Prendre en compte le code compagnie.



Le vendor_id 2 réalise le plus de trajet avec un pic important pour les trajets avec un passager. Globalement, c'est également lui qui réalise le plus de trajet peu importe le nombre de passagers. Il est important de souligner que le vendor_id 1 ne fait aucun trajet avec un seul passager.

- Analysez le nombre de prise en charge selon le jour de la semaine et selon l'heure de la journée
    - Ajoutez le mois comme variable discriminante : analysez le nombre de total de course par heure de la journée / journée dans la semaine en faisant dépendre du mois de l'année dans le graphique.

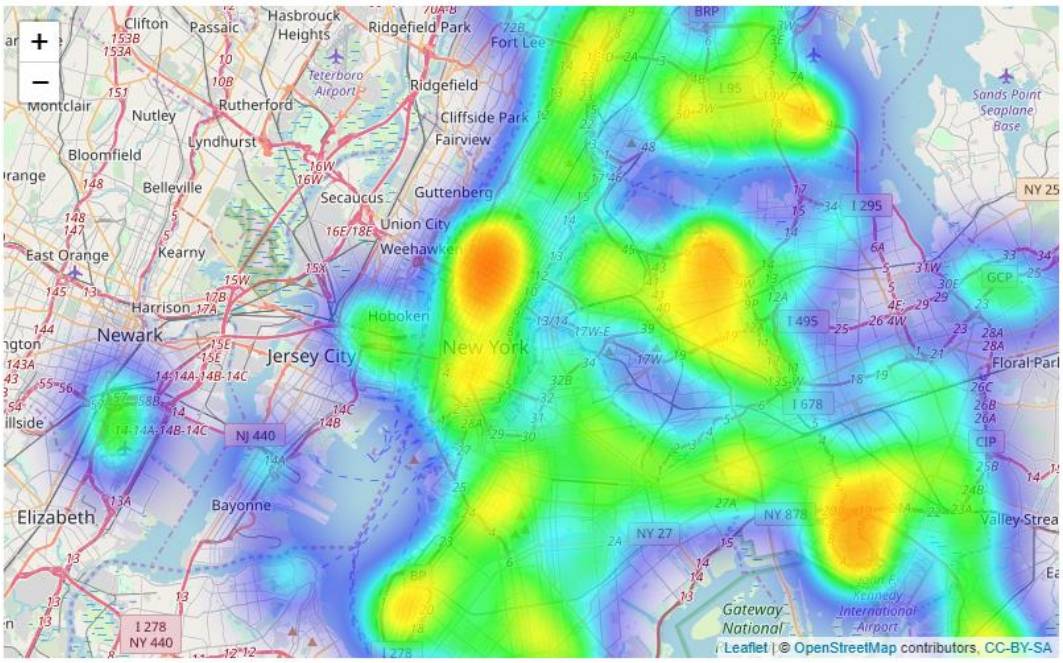


Le nombre de trajet par heure et par jour, nous montre bien une baisse des trajets à 5h du matin avec une reprise progressive. Nous remarquons le même effet pour tous les mois.

Nous avons le plus de trajet en fin de semaine vendredi et samedi et en mars et juin, nous avons un pic le mercredi/jeudi.

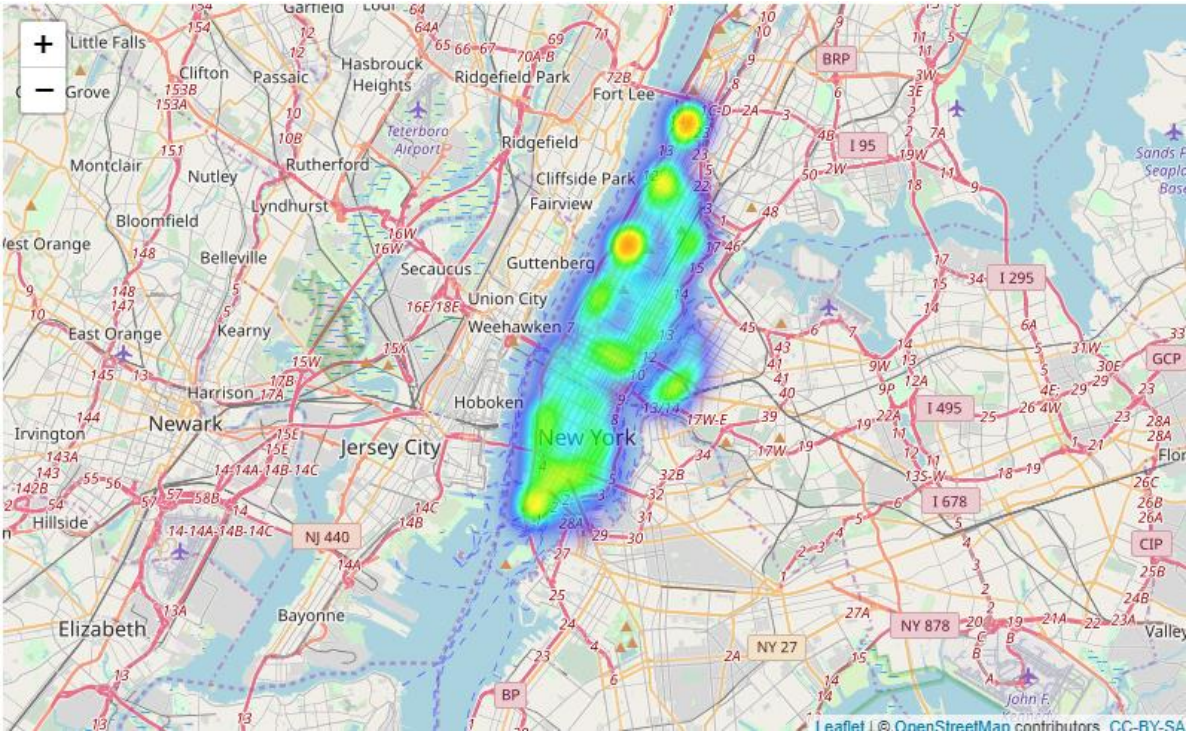
- Analysez les distributions des lieux de prise en charge et d'arrivée
    - Commentez les différences entre arrivées/départs
    - Y a-t-il des valeurs aberrantes ?

Carte des prises en charge

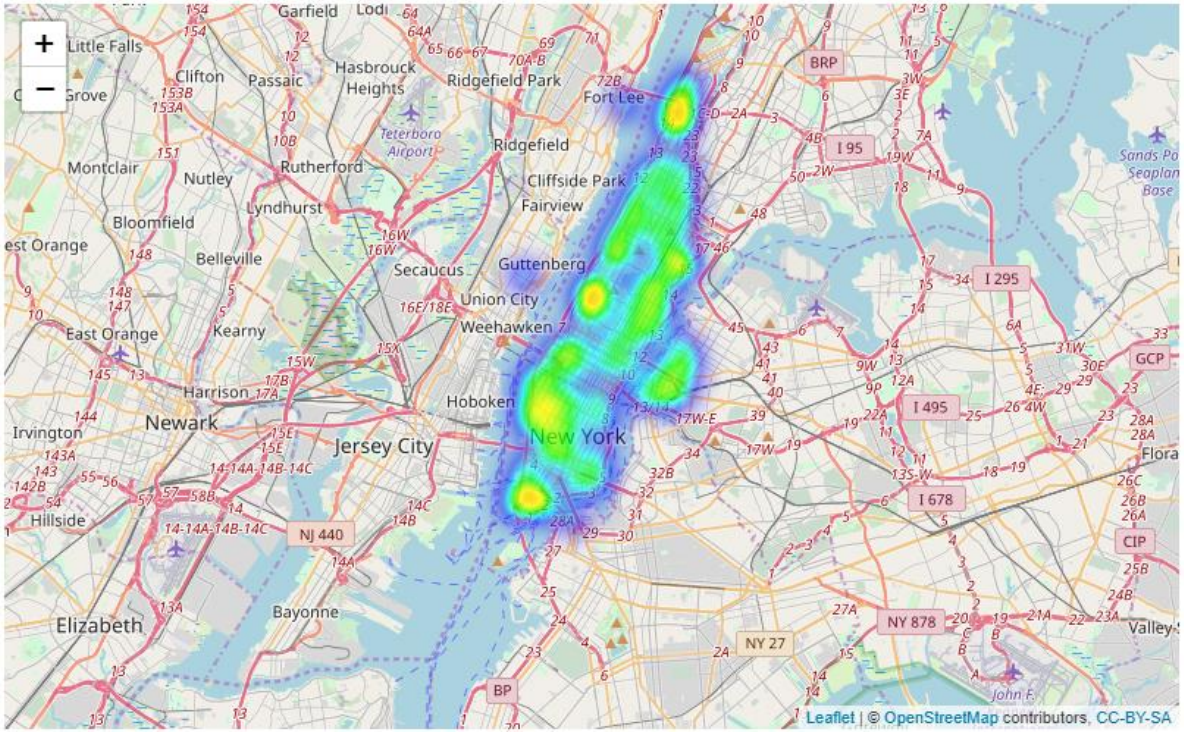


Carte des dépôts

On remarque la présence de 3 centres névralgiques : Manhattan et les deux aéroports de JFK et La Guardia.

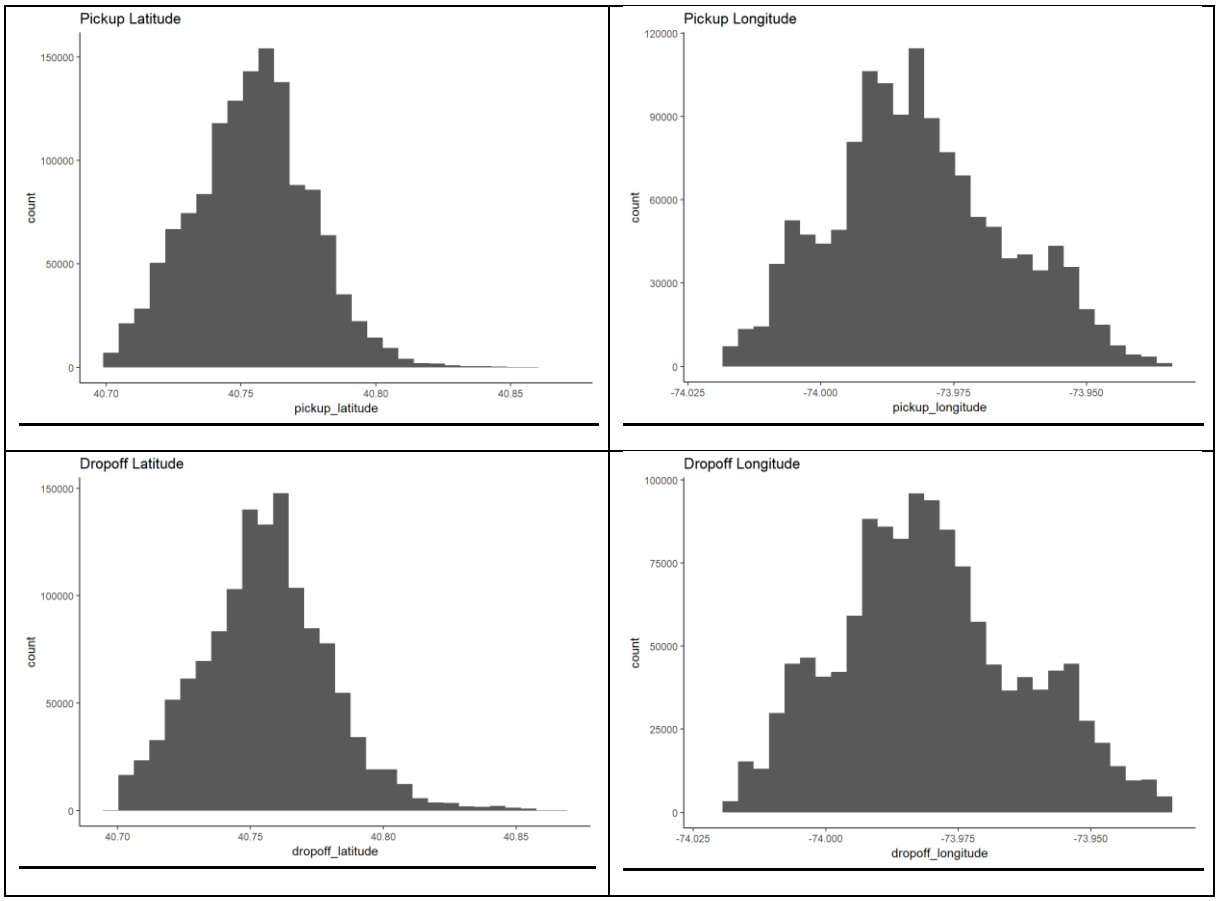


Carte des prises en charge à Manhattan



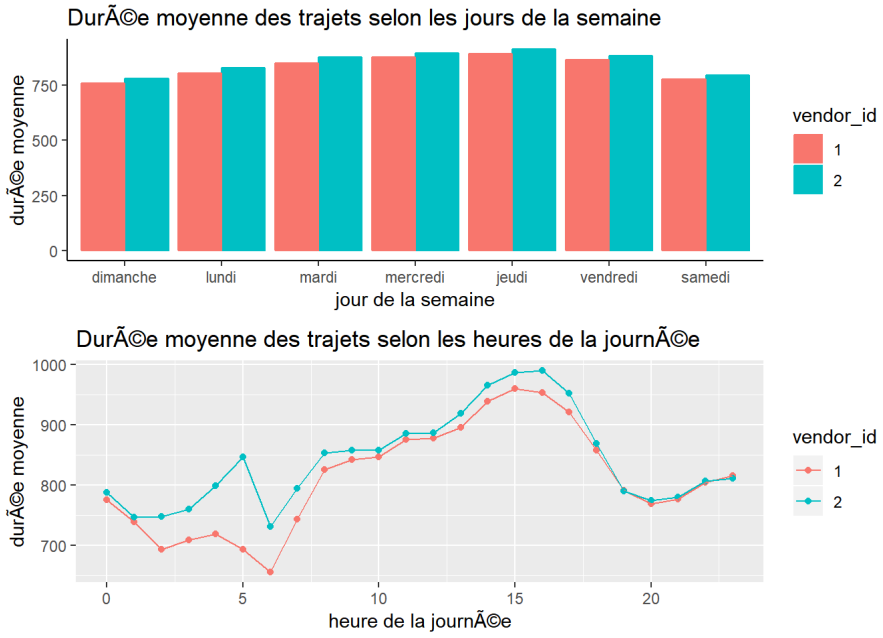
Carte des dépôts à Manhattan

En restreignant les données à Manhattan, on remarque un centre de prises en charges à la pointe Sud de l'Ile, lieu de travail intense (Time Square et Wall Street).



De manière plus méthodique, avec des histogrammes de latitude/longitude des prises en charges et dépôts, on remarque que les distributions sont semblables.

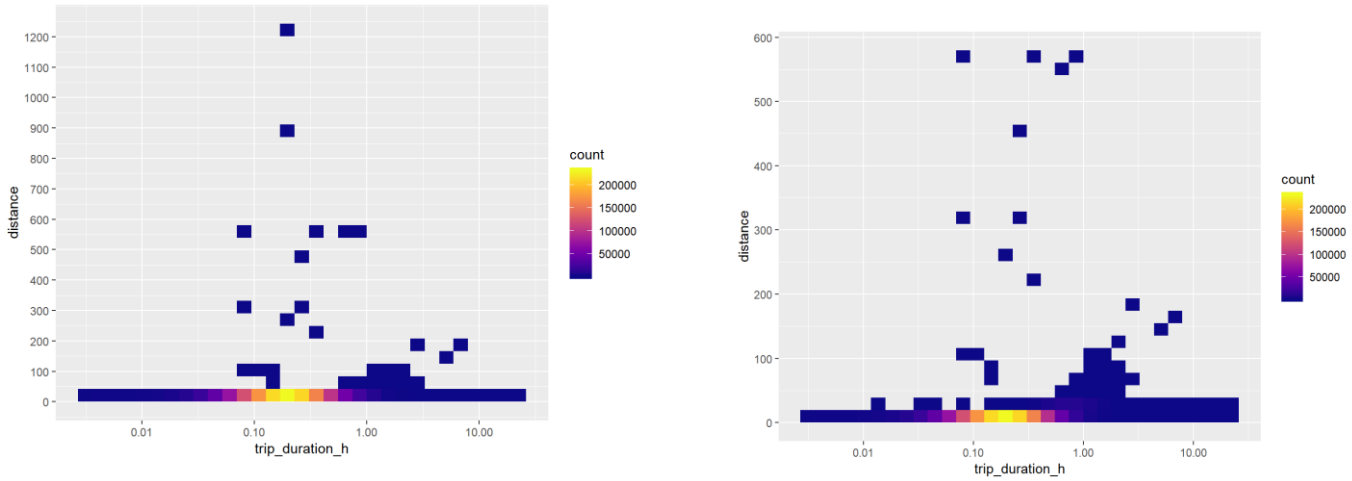
- Analysez la durée de prise en charge selon les autres variables (vendor_id par exemple)



La durée du code vendeur 2 est légèrement plus élevé que celle du code vendeur 1. la durée moyenne est plus importante le jeudi et est plus faible le week-end.

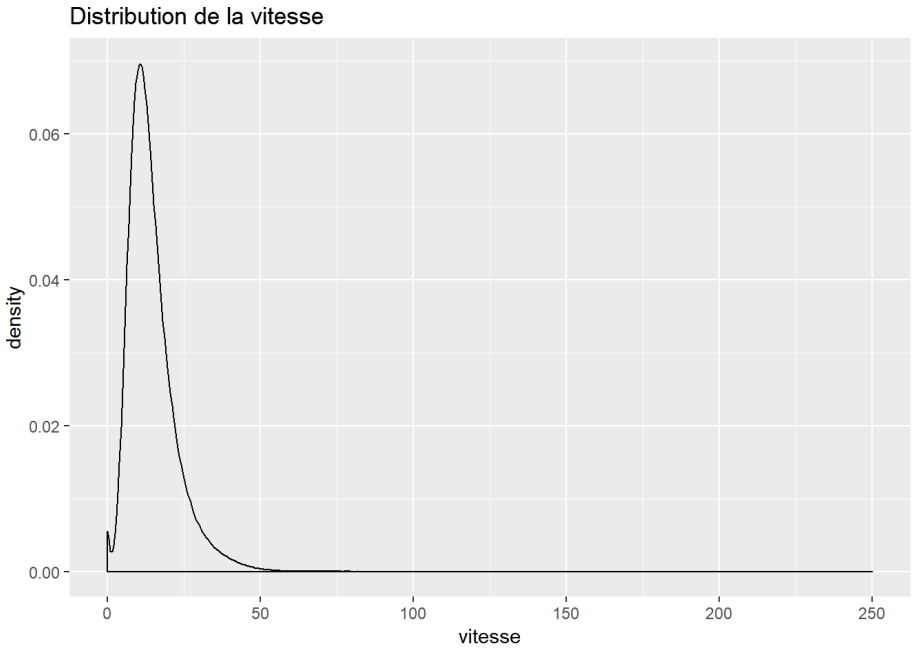
La durée moyenne des trajets est plus élevée vers 15h-16h pour les deux codes vendeurs. cependant, nous observons un pic à 5h du matin pour le code vendeur 2.

- Tracer une heatmap (et enlever les points qui vous semblent aberrants)



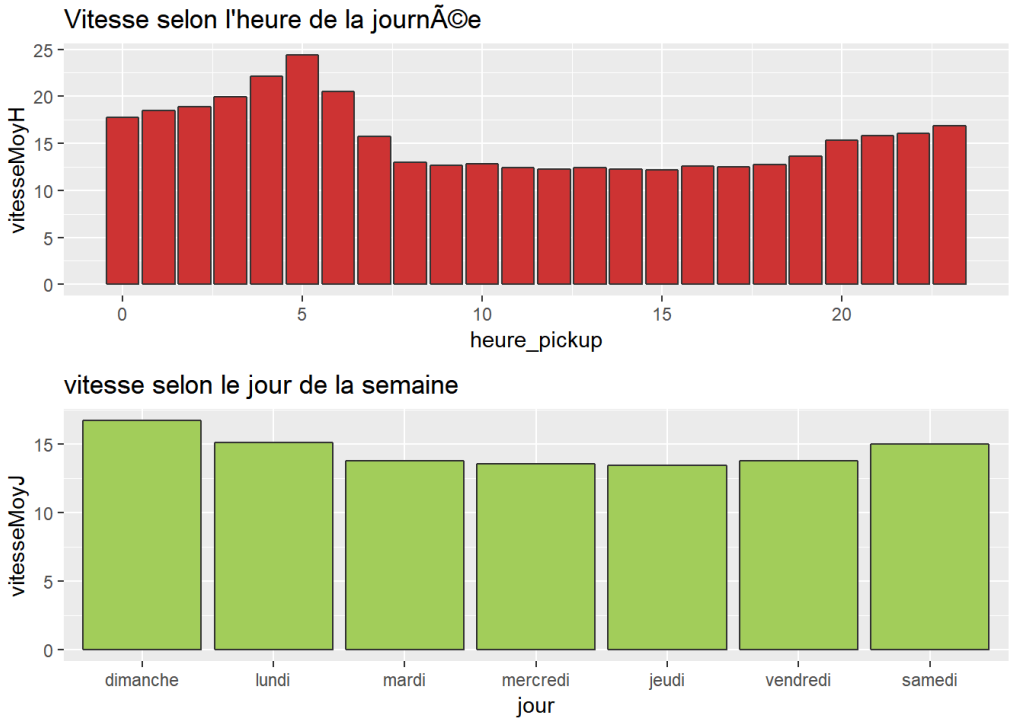
Les courtes distances sont plus concentrées entre 1 heure et 2h22 de durée de trajet. Nous avons également des points aberrants avec une distance supérieur à 700km pour une durée prise entre 1 et 2 heures.

- Afficher les statistiques descriptives de la vitesse ainsi que la distribution



La majorité des trajets ont une vitesse prise entre 0 et 50 km/h.

- Analysez la vitesse moyenne selon plusieurs axes : heures de la journée, journée de la semaine



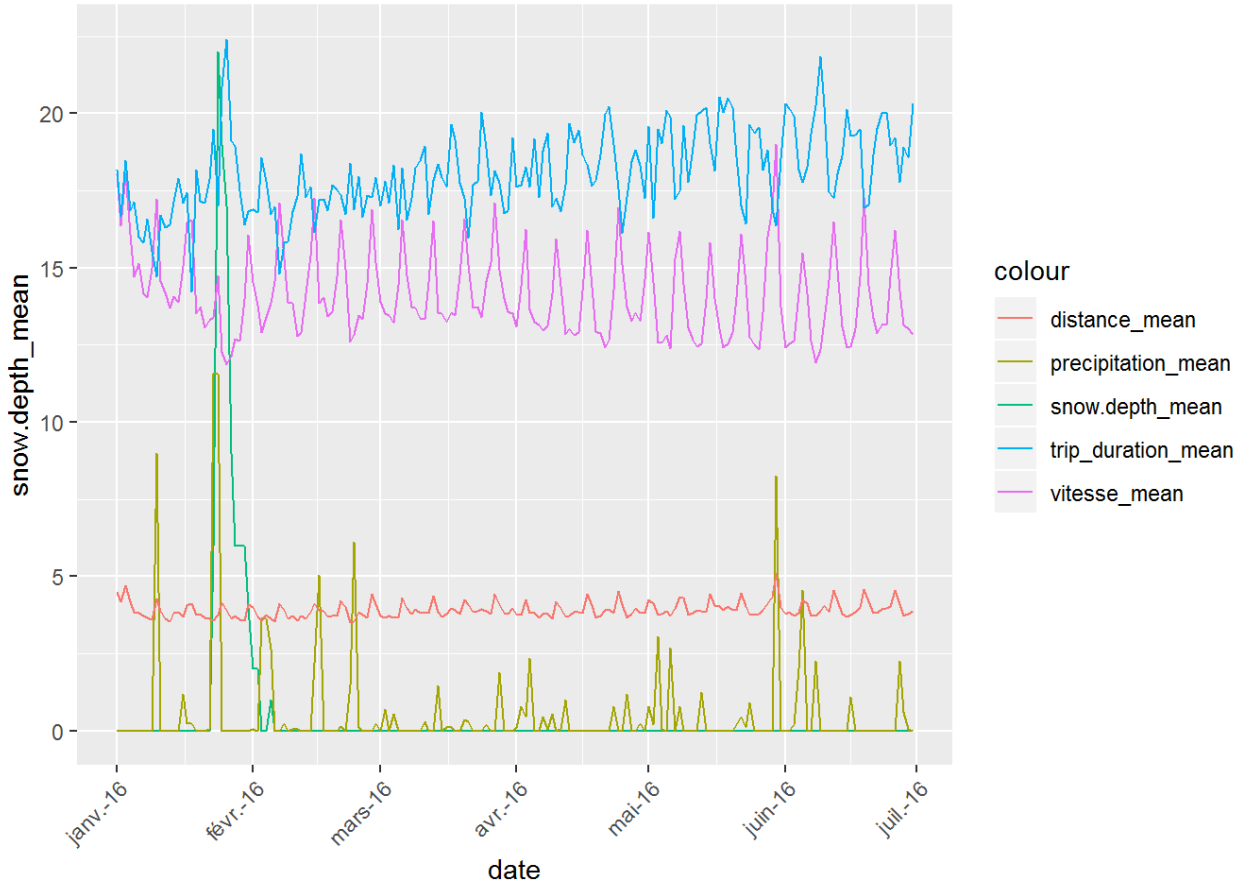
La vitesse moyenne croît à partir de minuit afin d'atteindre son maximum à 5h du matin. puis, elle décroit et reste stable tout au long de la journée et augmente très légèrement le soir.

La vitesse moyenne est beaucoup plus importante en week-end qu'en semaine.

- Tracer les prises en charges les moins vraisemblables sur une carte. Supprimez-les des données (cf. notebook ou shiny)

- Enrichissez les données (join table) avec les données météo (infos importantes : présence de pluie, présence de neige, hauteur de la pluie/neige, température)

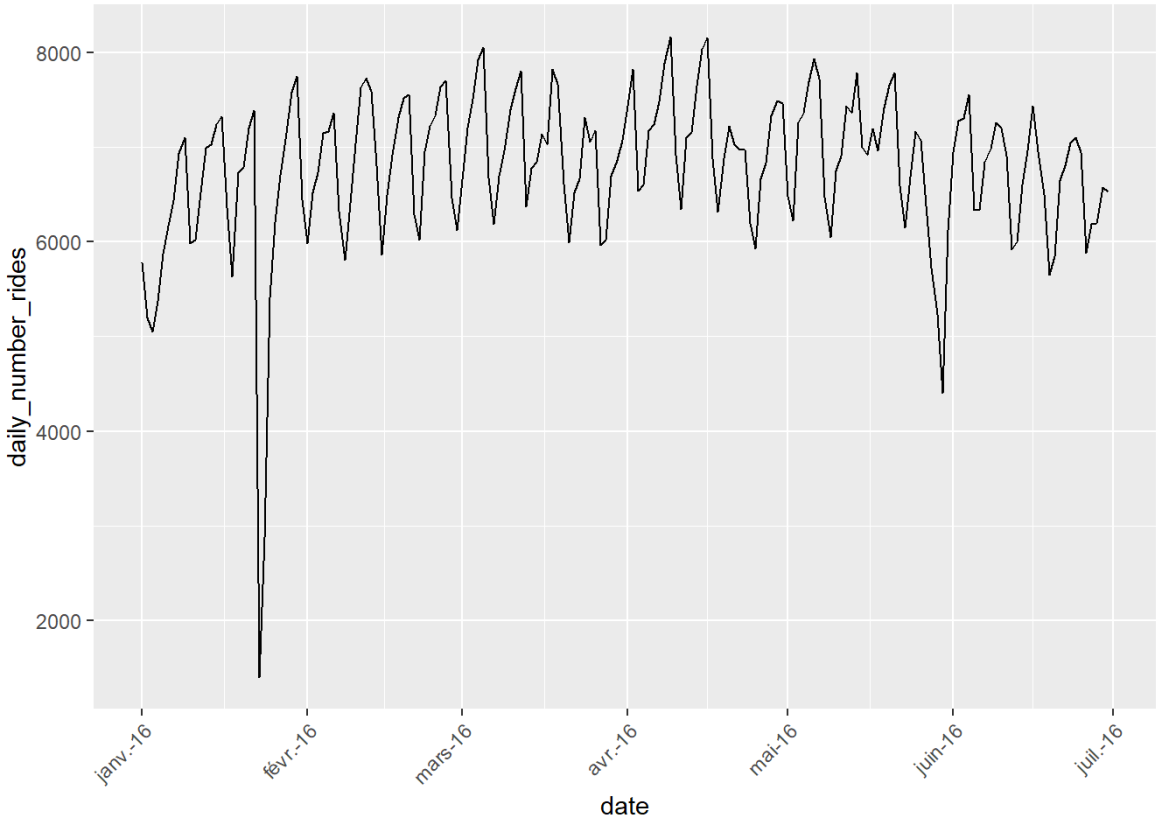
Voici les informations météorologiques sur la période couvrant le train et le test.



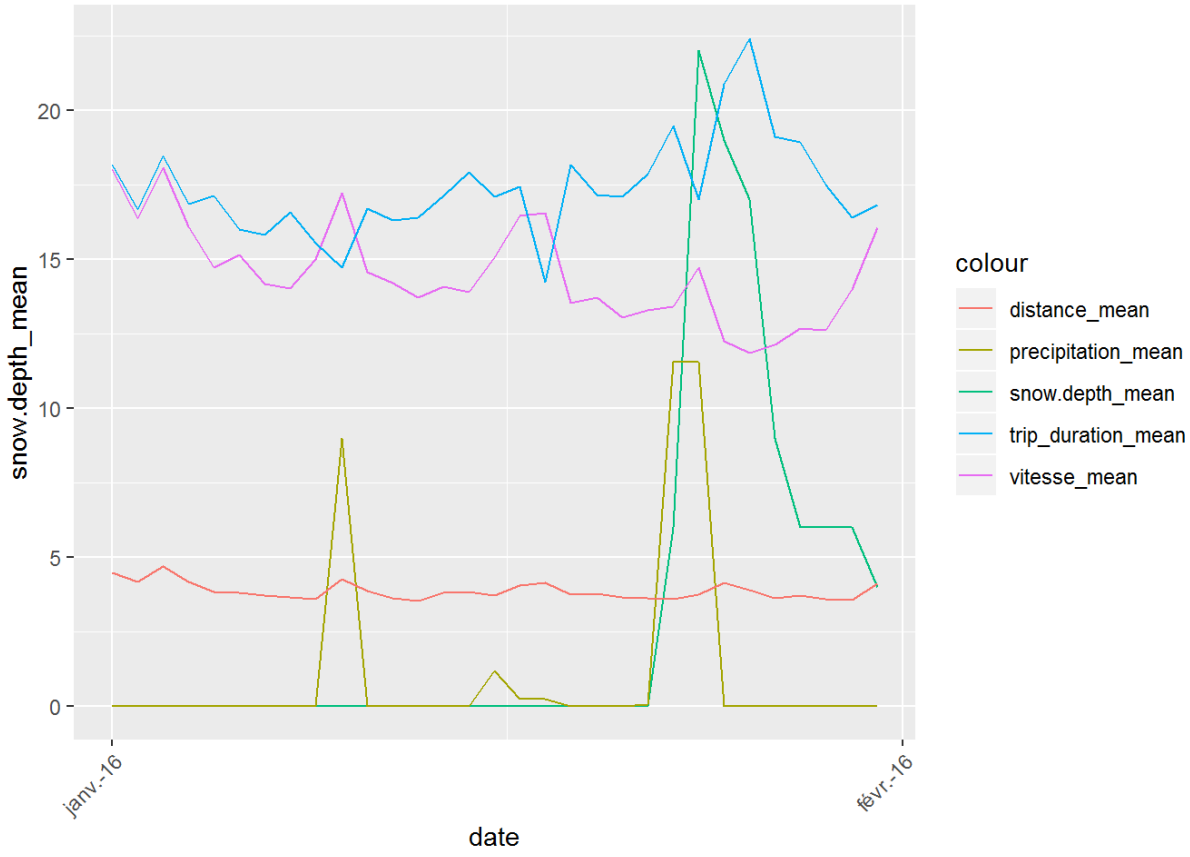
On remarque qu'il y a un gros blizzard fin-janvier. La profondeur de neige connait un pic. La vitesse des taxis baisse alors fortement et les durées augmentent.

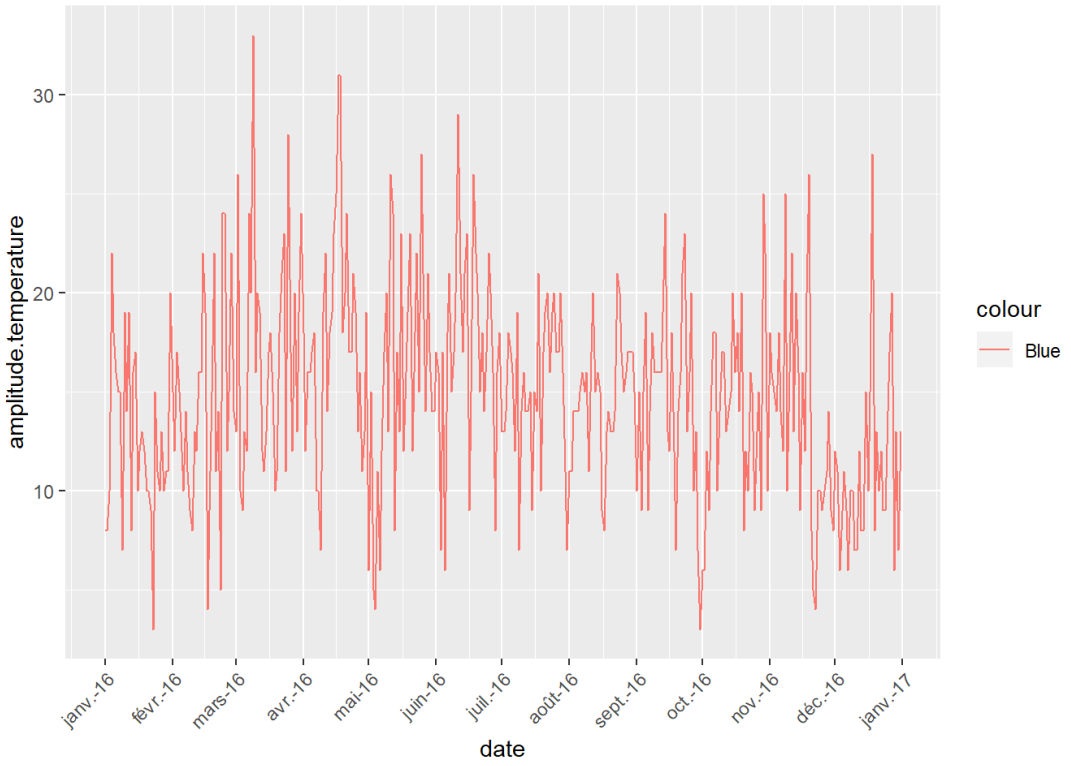
Les précipitations connaissent elles un pic fin mai.

Ces deux évènements se retrouvent dans le compte journalier de trajets, qui chute lors de ces évènements.



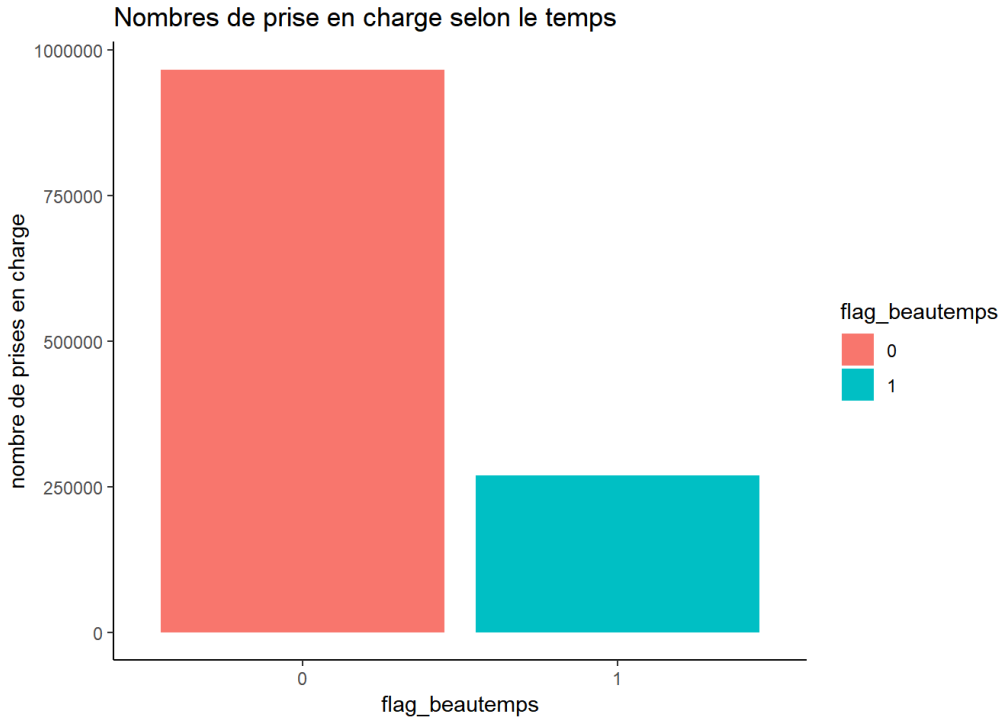
On observe sur le mois de janvier que la vitesse a quatre pics mensuels, les week-ends.





L'amplitude journalière des températures est la plus forte au printemps tandis qu'elle est la plus basse en hiver.

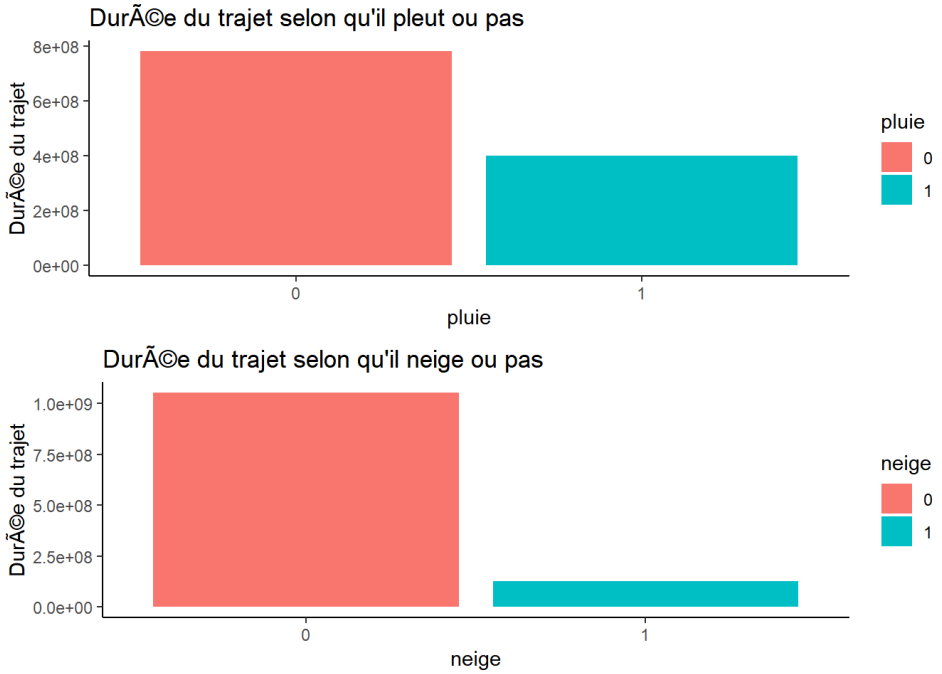
- Analysez le nombre de prise en charge par jour selon le fait qu'il y a eu mauvais temps ou pas ce jour là



Le flag_beautemps vaut 1 s'il n'a pas plu, pas neigé et il n'y a aucune neige au sol et si la température est supérieure à 16°C, sinon le flag prend la valeur 0.

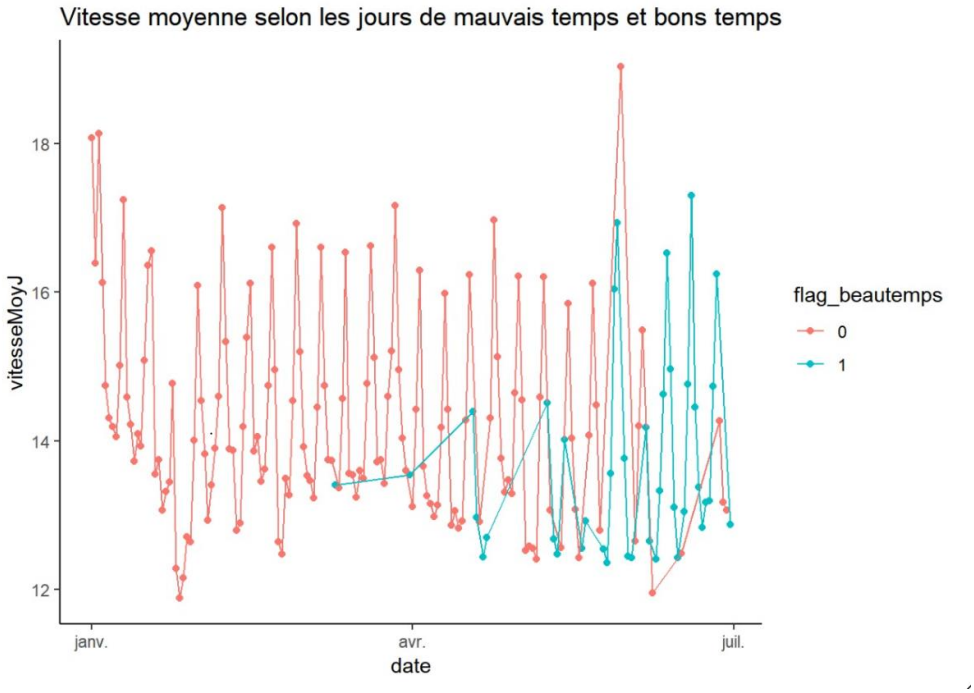
Le nombre de prise en charge est 3 fois plus élevé en période de mauvais temps qu'en période de beau temps.

- Analysez la durée selon le fait qu'il neige/pleuve ou pas



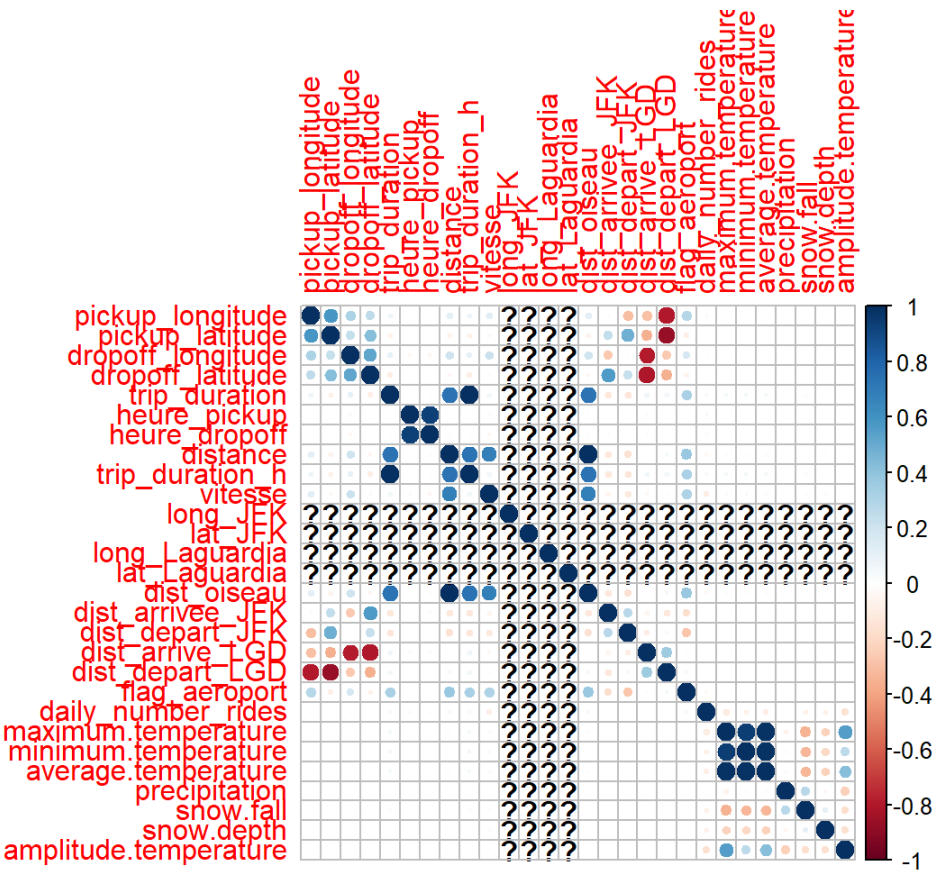
A notre grande surprise, nous remarquons que la durée de trajet n'est rallongée pas quand il pleut ou bien lorsqu'il neige. Ceci peut être expliqué que durant ces phénomènes météorologiques, il n'y a pas beaucoup de long trajet et donc plutôt des trajets courts.

- Analysez la vitesse moyenne selon les jours de mauvais / beaux temps

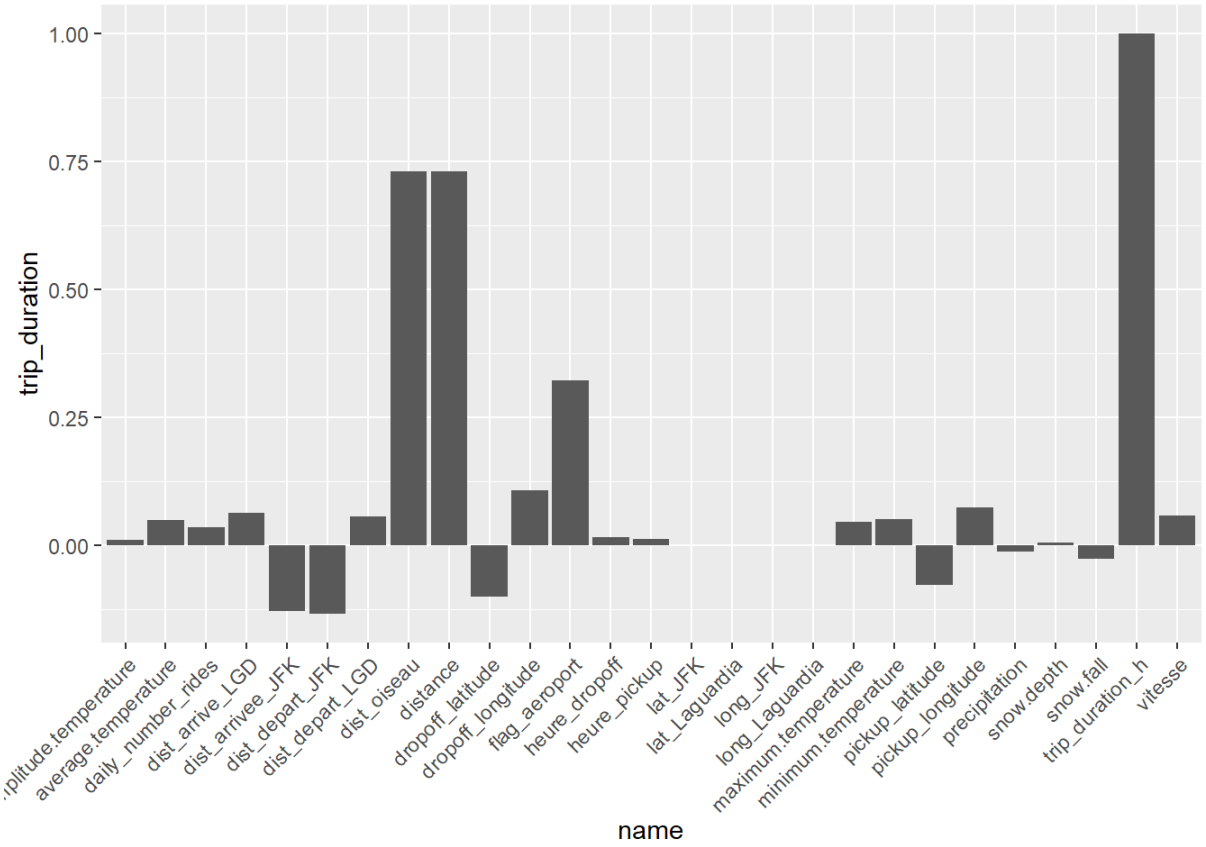


La vitesse moyenne fluctue beaucoup, elle est moins élevée en période de beau temps entre avril et juin et devient beaucoup plus élevée après juin.

- Faites une analyse de corrélations entre toutes les variables (corrélogramme)



Les données géodésiques sont corrélées. C'est également le cas des différentes mesures de température. Celles-ci sont aussi corrélées avec le mois et plus faiblement avec la neige.



Les corrélations les plus fortes avec la durée du trajet sont de l'ordre de 0.1 en valeur absolue.

Ce sont les latitudes de prise en charge et de dépôt. Puis viennent la longitude des dépôts, les températures puis le mois et le jour.

- Modélisation

Nous pouvons commencer notre modélisation par une simple régression linéaire, avant d'entamer une estimation par des modèles plus complexes comme une forêt aléatoire ou un gradient boosting.

Les données venant d'une compétition Kaggle, le test ne contient pas de trip_duration sur laquelle une erreur serait calculable. La meilleure manière de procéder est de constituer une estimation de l'erreur test par validation croisée sur le train.

On peut utiliser plusieurs techniques de selection de variable :

- Par sélection des variables les plus corrélées à la Target (trip_duration)
- Par sélection stepwise
- Par régularisation en norme L1 (Lasso)

Commençons par la 1ere méthode.

```
linearReg <- lm( trip_duration ~ dropoff_latitude + dropoff_longitude +
                                 mois.x + snow.fall + maximum.temperature , data=train2)

crossval = cv.lm(data = train2 , m = 3 , form.lm = linearReg)
```

Malheureusement, nous n'avons pas eu le temps de faire tourner les modèles.

En cause, la taille du train…

Conclusion

Cet examen-projet a été l'occasion pour nous de mieux connaitre R.

Aux débuts douloureux se sont substitués des facilités déconcertantes.

Plus particulièrement, la découverte du Tidyverse restera, pour nous, une connaissance très utile dans l'utilisation professionnelle de R.

Nous avons également pu découvrir un peu plus l'analyse exploratoire de données, tirées d'un dataset reconnu et aux types de données diversifiées (cartes, séries temporelles, catégories).

Nous regrettons de ne pas avoir eu le temps de faire plus de modélisation, mais ce n'est pas peine perdu !

Nous sommes maintenant armés pour maîtriser R en profondeur avec plus de travail.