Projet
Passeport-Montréal V+
2018

BD8

Qiaoling He
Dominique Tessier

Table des matières

Objectif

Développer et utiliser une approche d'apprentissage machine sur les données des déplacements avec Bixi.

Exploration des données

Description des données

Stations Bixi

1 fichier csv qui liste les stations bixi en activité pour l'année 2017

- ID, Nom
- Localisation (latitude, longitude)
- ***** Code postal (trouvé à partir de la localisation)*
- ***** Altitude (trouvé à partir de la localisation)*

ex : 015,LaSalle / 4e avenue,45.43074022417498,-73.5919108253438, H2S 2B6, 16

Source : https://www.bixi.com/en/open-data

*Note : Les paramètres précédés de « **** » ont été ajoutés par ETL*

Déplacements en Bixi

8 fichiers csv qui listent les déplacements en Bixi pour l'année 2017 (d'avril à novembre)

- Départ (date, heure et station)
- Arrivée (date, heure et station)
- Durée du déplacement en seconde
- Membre

ex : 2017-04-15 00:00,7060,2017-04-15 00:31,7060,1841,1

Source : https://www.bixi.com/en/open-data

Météo

8 fichiers CSV qui donnent des mesures sur le temps qu'il faisait en 2017 (d'avril à novembre)

- Date et heure
- "Température (°C)"
- "Direction et vitesse du vent
- Visibilité et la pression atmosphérique
- Temps

Ex : "2017-04-01 00:00","2017","04","01","00:00","0,5","","-0,5","","93","","7","","23","","2,4","","101,01","","","","","","","Neige"

Source : http://climat.meteo.gc.ca/historical_data/search_historic_data_f.html

Stations STM

1 fichier csv qui liste les stations de métro pour l'année 2017

- Stop_id, stop_code, stop_name,
- stop_lat, stop_lon,
- stop_url, location_type,
- parent_station,
- wheelchair_boarding
- ex : 11-01,10146,Station Berri-UQAM - Édicule Sainte-Catherine,45.514851,-73.559654,http://www.stm.info/metro/M11.htm,2,11S,2

- *Note : Les stations de métro ont été extraites de toutes les stations (entrée de métro et arrêt d'autobus)*

Source : http://www.stm.info/fr/a-propos/developpeurs

Prétraitement des données

Les fonctions pmv_loadAndPrepareBixiStations(), pmv_loadAndPrepareBixiDeplacements() et pmv_loadAndPrepareMeteo() permettent de charger les données des 3 datasets et de faire un certain nettoyage.

Note : Ces fonctions se retrouvent dans le script PMV_BD8_Init.R et servent de base aux autres scripts.

Stations Bixi :

- Le code postal a 3 caractères et a 2 caractères est ajouté au dataset

```
[1] "code"      "name"      "latitude"  "longitude" "altitude"  "postal"    "postal3"
[8] "postal2"
```

ILLUSTRATION 1: DESCRIPTEURS UTILISÉS POU LES STATIONS BIXI

Déplacements Bixi :

- L'heure est convertie en format numérique.
- Une clé est ajoutée pour faire le lien avec les données de la météo (date + heure)
- La période de la journée est ajoutée (période de 4 heures)

Note : Bixi fournit un dataset par mois. Toutes les données sont regroupées dans un seul.

```
 [1] "Date.Heure"  "Annee"       "Mois"        "Jour"        "Heure"
 [6] "Temperature" "DirVent"     "VitVent"     "visibilite"  "Pression"
[11] "Temps"       "bixiLink"
```

ILLUSTRATION 2: DESCRIPTEURS UTILISÉS POUR LES DÉPLACEMENTS BIXI

Météo :

- Certaines valeurs sont manquantes. On les remplacera par la dernière valeur valide. Traitement fait pour le temps, la pression atmosphérique, la direction du vent, la visibilité, la température et la vitesse du vent
- Une clé est ajoutée pour faire le lien avec les données des déplacements Bixi
- Les colonnes inutiles ou sans aucune donnée sont enlevées .
- Un nom significatif est donné au colonne
- L'heure est converti en format numérique.

Note : Météo Canada fournit un dataset par mois. Toutes les données sont regroupées dans un seul. On prendra les mêmes mois que pour les déplacements Bixi

```
 [1] "Date.Heure"  "Annee"       "Mois"        "Jour"        "Heure"
 [6] "Temperature" "DirVent"     "VitVent"     "visibilite"  "Pression"
[11] "Temps"       "bixiLink"
```

ILLUSTRATION 3: DESCRIPTEURS POUR LA MÉTÉO

Stations de métro :

- Les données sont récupérées sans aucun changement

```
[1] "stop_id"       "stop_code"     "stop_name"     "stop_lat"      "stop_long"
[6] "location_type" "parent_station" "wheelchair"
```

ILLUSTRATION 4: DESCRIPTEURS POUR LES STATIONS DE MÉTRO

Statistiques descriptives

Les analyses suivantes sont extraites du script PMV_BD8_Descriptive.R

Nombre de déplacements en Bixi : 4740357

Nombre de stations Bixi : 546

Nombre d'entrées de stations de métro : 628 (Il y a plusieurs entrées par station)

Répartition des déplacements en fonction des membres

is_member	freq
0	895551
1	3844806

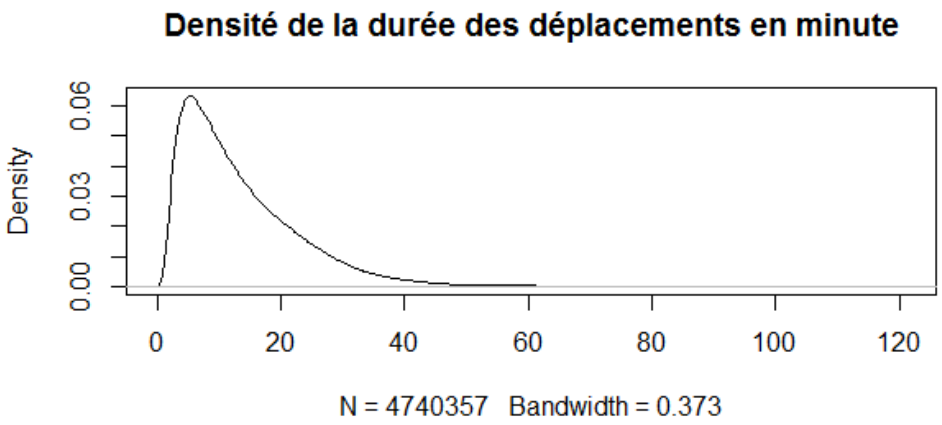

ILLUSTRATION 5: DENSITÉ DE LA DURÉE DES DÉPLACEMENTS

	temps	freq
28	Généralement nuageux	1614187
27	Généralement dégagé	1262537
26	Nuageux	786547
25	Dégagé	640074
24	Averses de pluie	159298
23	Pluie	156294
22	Brouillard	42038
21	Pluie,Brouillard	22371
20	Orages,Averses de pluie	19468
19	Bruine	10202
18	Orages	4884
17	Brume sèche	4655
16	Bruine,Brouillard	4263
15	Averses de pluie modérées,Brouillard	2652

ILLUSTRATION 6: FRÉQUENCE DES DÉPLACEMENTS EN FONCTION DU TEMPS

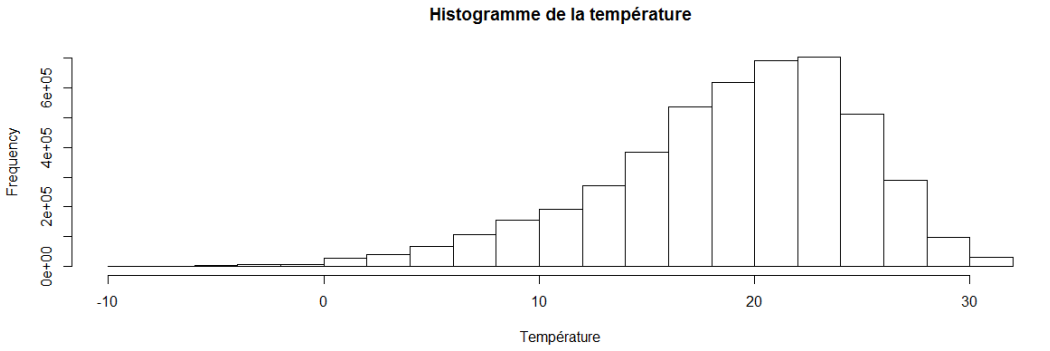

ILLUSTRATION 7: FRÉQUENCE DES DÉPLACEMENTS EN FONCTION DE LA TEMPÉRATURE

```
    distance        duration_sec
Min.   :    0   Min.   :   61.0
1st Qu.:  852   1st Qu.:  372.0
Median : 1489   Median :  651.0
Mean   : 1849   Mean   :  818.5
3rd Qu.: 2504   3rd Qu.: 1094.0
Max.   :13857   Max.   : 7199.0
```

ILLUSTRATION 8: STATISTIQUES SUR LA DISTANCE ET LE TEMPS DE DÉPLACEMENT

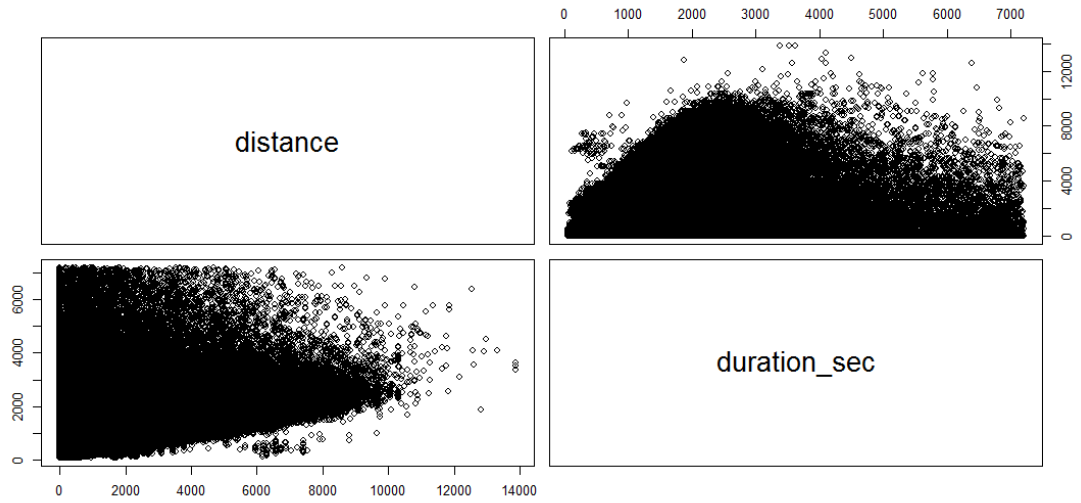

ILLUSTRATION 9: RELATION ENTRE LA DISTANCES ENTRE 2 STATIONS ET LE TEMPS PARCOURU (SUR UN ÉCHANTILLON DE 40 % DE TOUS LES DÉPLACEMENTS)

Note : Les analyses suivantes sont extraites du script PMV_BD8_Deplacement.R

Le graphique suivant permet de visualiser l'emplacement des stations Bixi par rapport aux stations de métro

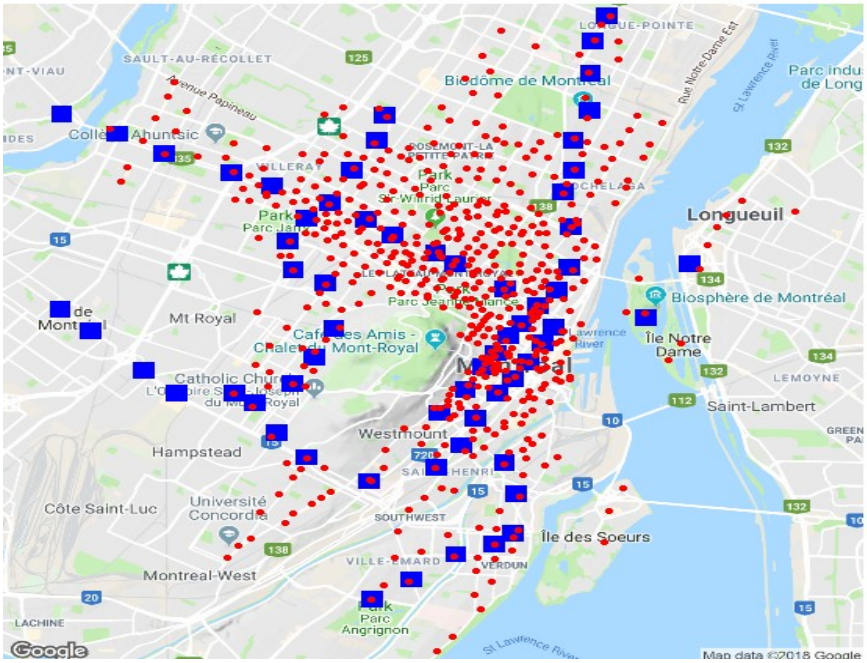

ILLUSTRATION 10: EMPLACEMENTS DES STATIONS BIXI ET DU MÉTRO

La cartes suivante montre la fréquence des départs en fonction de la période de la journée (bloc de 4 heures)

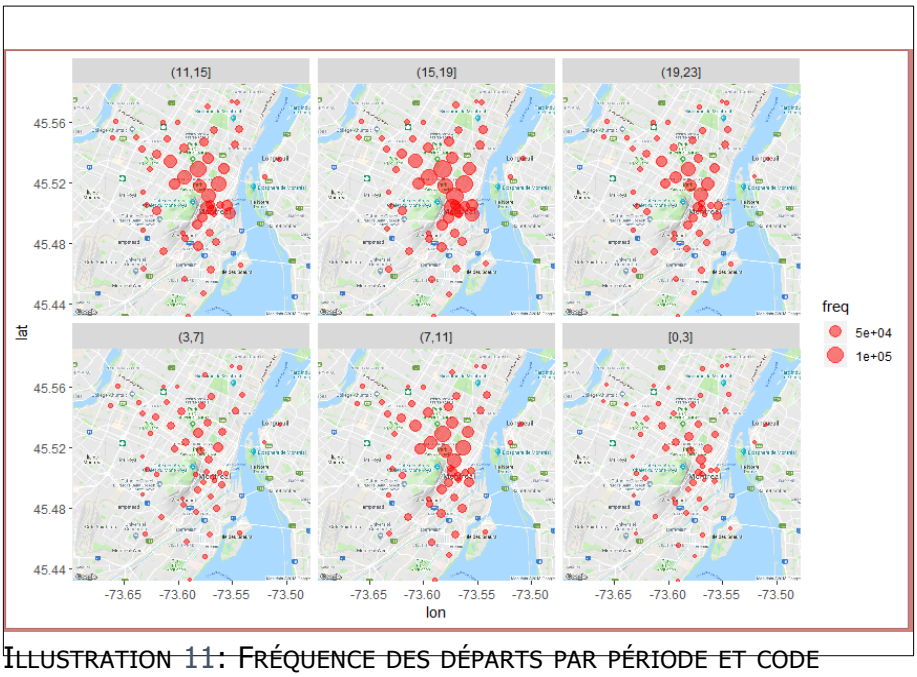

ILLUSTRATION 11: FRÉQUENCE DES DÉPARTS PAR PÉRIODE ET CODE

La carte suivante montre la fréquence des départs en fonction de l'heure de la journée et par quartier (identifié par le code postal)

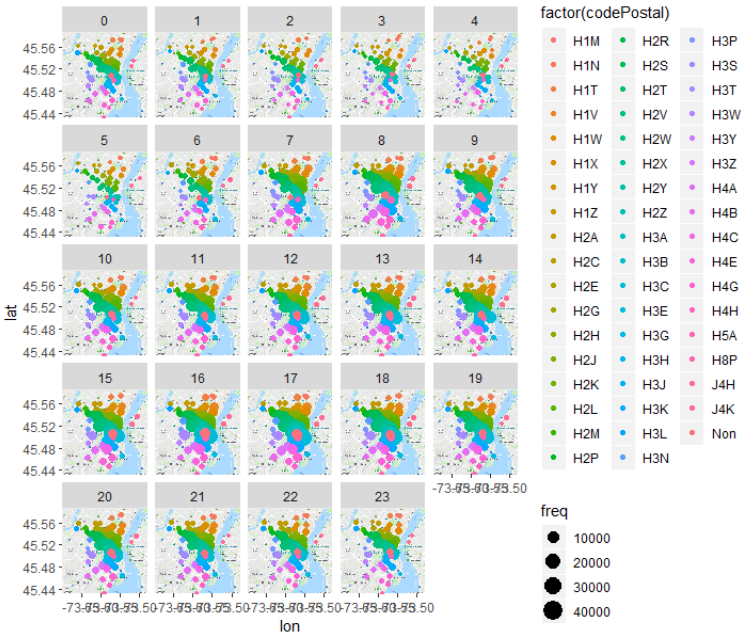

ILLUSTRATION 12: FRÉQUENCE DES DÉPARTS PAR HEURE ET CODE POSTAL

Temps de parcours (supervisé)

Construire un modèle pour évaluer le temps de parcours entre 2 stations.

Modélisation : Supervisée. Utilisation de Naïves Bayes

classe : temps de parcours (par période de 15 minutes). 5 périodes ("moins de 15","[15-30)","[30-45)","[45-60)","Plus de 60")

Note : référer au script PMV_BD8_Modelisation.R

Identification des variables

Les datasets (après le pré-traitement de base) donnent les paramètres suivants ;

Pour la météo :

```
 [1] "Date.Heure"  "Annee"       "Mois"        "Jour"        "Heure"       "Temperature"
 [7] "DirVent"     "VitVent"     "visibilite"  "Pression"    "Temps"       "bixiLink"
```

Pour les déplacements :

```
 [1] "start_date"         "start_station_code" "end_date"           "end_station_code"
 [5] "duration_sec"       "is_member"          "hourStart"          "meteoLink"
 [9] "period"
```

Pour les stations Bixi:

```
 [1] "code"       "name"       "latitude"   "longitude" "altitude"   "postal"     "postal3"
 [8] "postal2"
```

Après fusion des datasets, l'ajout et la suppression de certaines variables, on obtient les paramètres suivants :

```
      Mois            Jour           Heure         is_member       Temperature
 Min.   : 4.000   Min.   : 1.00   Min.   : 0.00   Min.   :0.0000   Min.   :-9.60
 1st Qu.: 6.000   1st Qu.: 9.00   1st Qu.:10.00   1st Qu.:1.0000   1st Qu.:15.70
 Median : 7.000   Median :16.00   Median :15.00   Median :1.0000   Median :19.90
 Mean   : 7.455   Mean   :16.17   Mean   :14.17   Mean   :0.8111   Mean   :18.99
 3rd Qu.: 9.000   3rd Qu.:24.00   3rd Qu.:18.00   3rd Qu.:1.0000   3rd Qu.:23.30
 Max.   :11.000   Max.   :31.00   Max.   :23.00   Max.   :1.0000   Max.   :31.90
    DirVent          VitVent        visibilite       Pression        distance
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.40   Min.   : 97.17   Min.   :    0
 1st Qu.:15.00   1st Qu.:10.00   1st Qu.:24.10   1st Qu.:100.61   1st Qu.:  852
 Median :22.00   Median :15.00   Median :24.10   Median :101.01   Median : 1489
 Mean   :20.03   Mean   :16.07   Mean   :33.39   Mean   :101.04   Mean   : 1849
 3rd Qu.:25.00   3rd Qu.:21.00   3rd Qu.:48.30   3rd Qu.:101.41   3rd Qu.: 2503
 Max.   :36.00   Max.   :57.00   Max.   :80.50   Max.   :103.30   Max.   :13857
  diffAltitude      period_duration     temps2
 Min.   :-111.000   Min.   :1.000   Min.   : 1.00
 1st Qu.: -12.000   1st Qu.:1.000   1st Qu.:13.00
 Median :  -1.000   Median :1.000   Median :14.00
 Mean   :  -3.263   Mean   :1.428   Mean   :13.88
 3rd Qu.:   5.000   3rd Qu.:2.000   3rd Qu.:14.00
 Max.   : 107.000   Max.   :5.000   Max.   :28.00
```

ILLUSTRATION 13: PARAMÈTRES POUR LA MODÉLISATION KMEANS

Note : La distance (à vol d'oiseau) et le dénivelé entre 2 stations ont été ajoutés ; Le temps de parcours (en seconde) a été transformé en 5 périodes (par tranche de 15 minutes) ; Le temps (ensoleillé, nuageux, …) a été « factorisé »

On remarque que la distance est parfois égale à zéro. Après investigation, on remarque qu'il y en a 96983.
Il faut aussi filtrer les enregistrements dont la distance est inférieure à 50m.
Note : Ce doit être les personnes qui ont eu un problème avec un vélo ou avaient le temps de faire l'aller-retour.

Choix des variables

Afin d'identifier les paramètres qui serviront dans le modèle, on affiche la corrélation entre les descripteurs.

	Mois	Jour	Heure	is_member	Temperature	DirVent	VitVent	visibilite	Pression	distance	diffAltitude	period_duration	temps2
Mois	1.00	-0.17	-0.03	0.04	-0.12	0.05	-0.15	-0.07	0.34	-0.01	-0.01	-0.05	-0.04
Jour	-0.17	1.00	-0.01	-0.04	0.06	-0.08	0.02	-0.01	-0.07	0.00	0.00	0.01	-0.06
Heure	-0.03	-0.01	1.00	-0.04	0.17	0.05	0.02	-0.02	-0.08	-0.05	0.06	0.02	-0.02
is_member	0.04	-0.04	-0.04	1.00	-0.11	-0.03	0.01	-0.05	0.02	-0.04	-0.02	-0.27	0.03
Temperature	-0.12	0.06	0.17	-0.11	1.00	0.10	0.05	0.07	-0.33	0.02	0.02	0.09	-0.14
DirVent	0.05	-0.08	0.05	-0.03	0.10	1.00	0.04	0.11	-0.16	0.00	0.01	0.02	-0.14
VitVent	-0.15	0.02	0.02	0.01	0.05	0.04	1.00	0.11	-0.38	-0.02	0.00	0.00	-0.03
visibilite	-0.07	-0.01	-0.02	-0.05	0.07	0.11	0.11	1.00	0.15	0.01	0.00	0.06	-0.07
Pression	0.34	-0.07	-0.08	0.02	-0.33	-0.16	-0.38	0.15	1.00	0.00	-0.01	-0.02	-0.03
distance	-0.01	0.00	-0.05	-0.04	0.02	0.00	-0.02	0.01	0.00	1.00	-0.13	0.59	-0.01
diffAltitude	-0.01	0.00	0.06	-0.02	0.02	0.01	0.00	0.00	-0.01	-0.13	1.00	-0.02	0.00
period_duration	-0.05	0.01	0.02	-0.27	0.09	0.02	0.00	0.06	-0.02	0.59	-0.02	1.00	-0.02
temps2	-0.04	-0.06	-0.02	0.03	-0.14	-0.14	-0.03	-0.07	-0.03	-0.01	0.00	-0.02	1.00

ILLUSTRATION 14: CORRÉLATION ENTRE LES DESCRIPTEURS

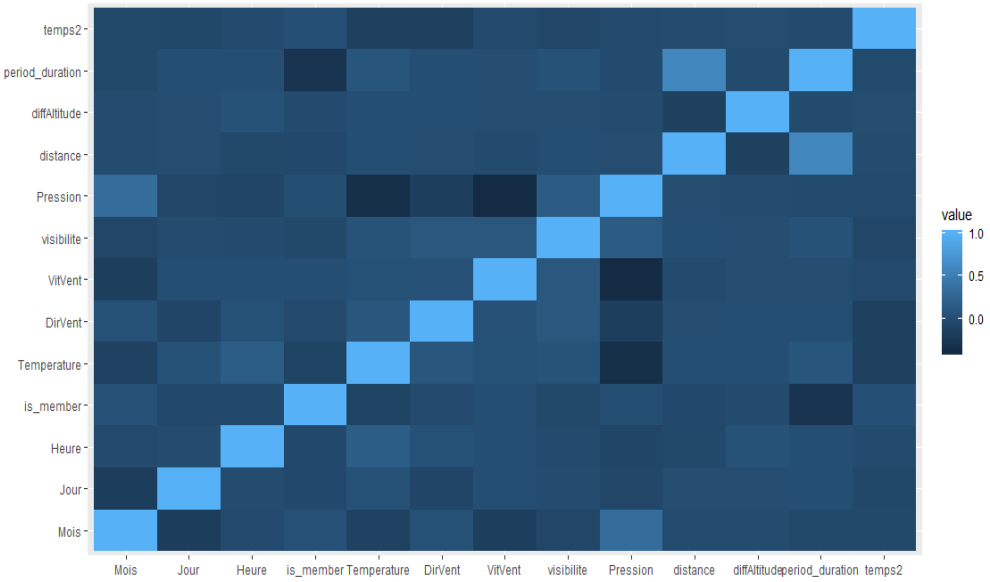

ILLUSTRATION 15: CORRÉLATION ENTRE LES DESCRIPTEURS

Note : Les paramètres ont été normalisés avec la fonction scale

Le graphique montre une bonne corrélation entre la distance et notre classe (Period_duration). Il y en a aussi une dans une moindre mesure avec la température et si l'usager est un membre ou non.

Partage des données en training et test

Les données de l'année 2017 seront traitées.

80% serviront au « training » et 20 % pour le test, soit :

3792286 enregistrements pour le « training » et 948071 pour le test.

Note : On pourrait aussi tester le modèle avec les données partielles de l'année 2018

Apprentissage Naïves Bayes

La modélisation se fera avec les données suivantes :

dt.distance	dt.is_member	dt.Temperature	dt.period_duration
Min. : 52	Min. :0.0000	Min. :-9.60	1:3075774
1st Qu.: 891	1st Qu.:1.0000	1st Qu.:15.70	2:1277518
Median : 1518	Median :1.0000	Median :19.90	3: 235332
Mean : 1888	Mean :0.8162	Mean :18.97	4: 32744
3rd Qu.: 2537	3rd Qu.:1.0000	3rd Qu.:23.30	5: 21804
Max. :13857	Max. :1.0000	Max. :31.90	

Les 3 premiers seront les prescripteurs et le dernier sera la classe.

modele	list [4] (S3: naiveBayes)	List of length 4
apriori	integer [5] (S3: table)	2460570 1021644 188692 26217 17415
tables	list [3]	List of length 3
levels	character [5]	'1' '2' '3' '4' '5'
call	language	naiveBayes.default(x = scale(train[, 1:3]), y = train[, 4])

Comparaison du modèle avec la portion test

	prediction				
actuel	1	2	3	4	Row Total
1	560986	53172	55	0	614213
	0.604	0.057	0.000	0.000	
2	67650	184250	4695	0	256595
	0.073	0.198	0.005	0.000	
3	6991	29724	10280	0	46995
	0.008	0.032	0.011	0.000	
4	1512	3549	1498	1	6560
	0.002	0.004	0.002	0.000	
5	1529	2200	536	6	4271
	0.002	0.002	0.001	0.000	
Column Total	638668	272895	17064	7	928634

Cela donne un résultat de 81.3 %

Le modèle donne peut de résultat pour les périodes 4 et 5 (plus de 45 minutes).Cela représente 65620 déplacements.

Si on regarde la relation entre la durée et la distance, on remarque qu'il n'y a plus trop de relation.

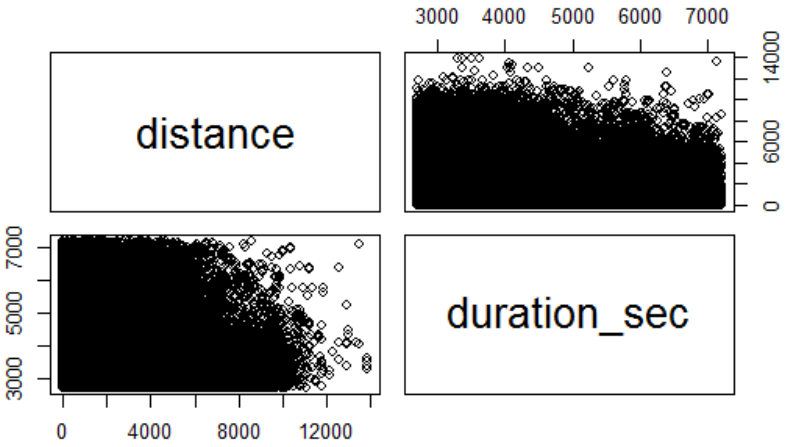

ILLUSTRATION 16: RELATION ENTRE LA DURÉE ET LA DISTANCE (POUR UNE DURÉE DE PLUS DE 45 MINUTES)

Modélisation C5.0 de Quinlan

En utilisant les mêmes prédicteurs mais en utilisant la modélisation C5.0 de Quinlan on obtient les résultats suivants :

	prediction					
actuel	1	2	3	4	5	Row Total
1	574621	39064	353	69	106	614213
	0.619	0.042	0.000	0.000	0.000	
2	72654	177566	6065	176	134	256595
	0.078	0.191	0.007	0.000	0.000	
3	7590	26403	12653	242	107	46995
	0.008	0.028	0.014	0.000	0.000	
4	1612	2972	1592	327	57	6560
	0.002	0.003	0.002	0.000	0.000	
5	1620	1799	535	92	225	4271
	0.002	0.002	0.001	0.000	0.000	
Column Total	658097	247804	21198	906	629	928634

ILLUSTRATION 17: MODÉLISATION 5.0: TEST DU MODÈLE

Ceci nous donne un taux d'exactitude de 82.4 %

Arbre de décisions

Bâtir un arbre de décision en se servant de la distance comme paramètre et le temps de déplacement (en minute) comme classe.

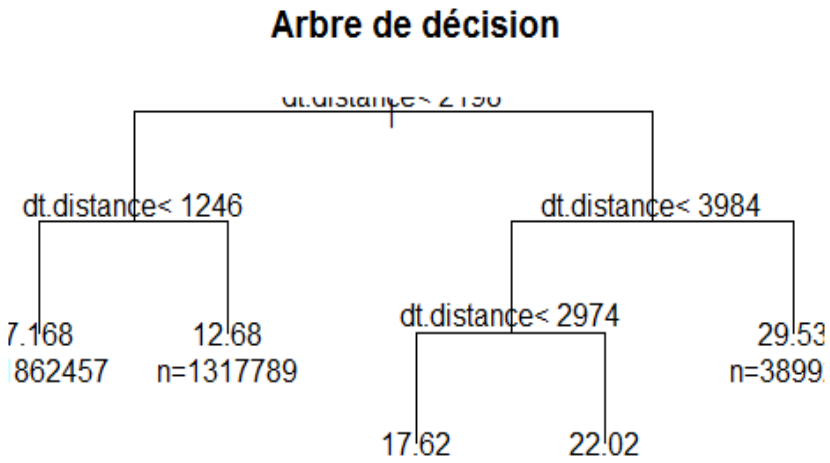

ILLUSTRATION 18: ARBRE DE DÉCISION DU TEMPS DE DÉPLACEMENT (EN FONCTION DE LA DISTANCE

Avec les données de 2017, on peut faire 5 clusters

Regroupement des stations (non supervisé)

Déterminer le nombre de « clusters » pour le regroupement des stations (basé sur la longitude et la latitude de la station).

Construire un modèle pour évaluer le temps de parcours entre 2 stations.

Modélisation : Non supervisée. Utilisation de k-means

Le calcul du nombre de clusters pour la localisation des stations est : 3

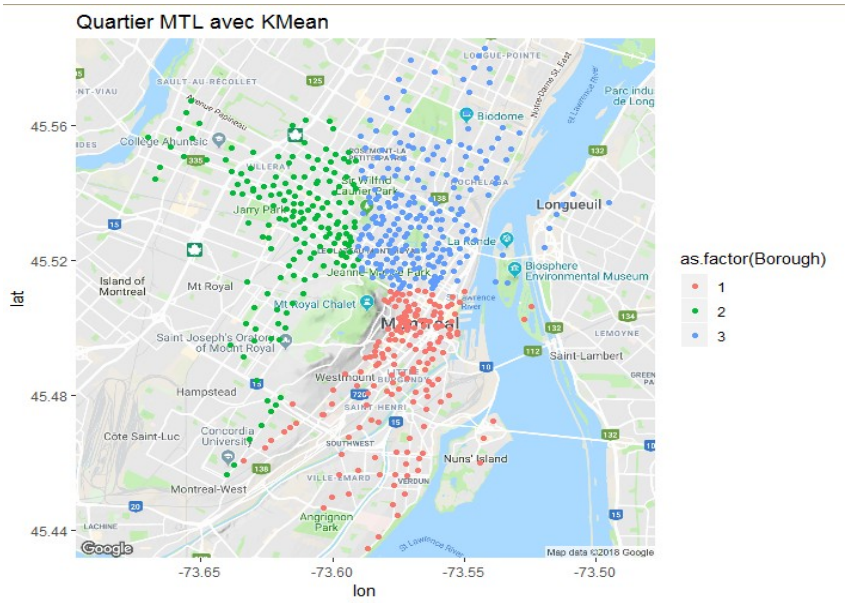

ILLUSTRATION 19: RÉPARTITION DES STATIONS PAR CLUSTER

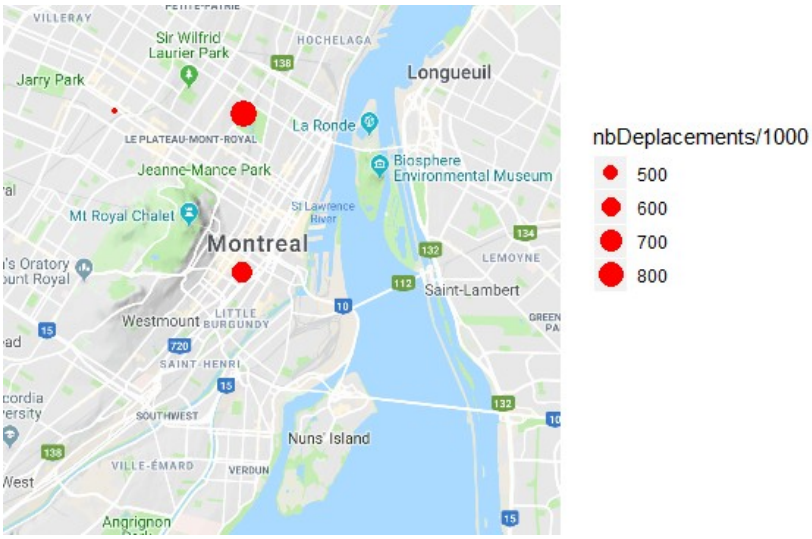

ILLUSTRATION 20: FRÉQUENCE DES DÉPLACEMENTS PAR CLUSTER

Annexe

Scripts R

- PMV_BD8_Init.R : Script qui contient les fonctions permettant de charger et nettoyer les datasets.
- PMV_BD8_Descriptive.R : script qui fait l'analyse descriptive des datasets

- PMV_BD8_Modelisation.R : script pour la modélisation.Naïves Bayes
- PMV_BD8_kmeans.R : script pour la modalisation kmeans
- PMV_BD8-deplacement.R : Script pour analyser les déplacements en fonction de l'heure de la journée.