Université de Montpellier, France

Projet Final HMMA 307 - M2 MIND

Analyse du nombre de passages de cyclistes à Seattle

Santinelli Emma

Département de Mathématiques

30 Octobre 2020

Résumé

L'objectif de ce projet est de faire de l'analyse descriptive du nombre de vélos passants sur le Pont Fremont de Seattle. A partir de ces données, nous voulons apprendre des habitudes de travail de ces utilisateurs de vélo. Ainsi, sans faire d'hypothèses de modèles mathématiques, nous allons extraire le maximum d'information disponibles dans ces données de cyclistes. Pour ce faire, nous utiliserons le langage de programmation Python.

1) Présentation des données

Nous sommes en présence de données issues d'un appareil de comptage automatique du nombre de vélo installé sur le pont de Fermont à Seattle. Cet appareil, installé fin 2012, a pour objectif de compter à chaque heure de la journée le nombre de vélos qui passe sur l'un des deux côtés du pont (Est ou Ouest). Ainsi, notre jeu de données est constitué de 136334 heures de données d'observation et de trois variables : le nombre de vélos passés sur le côte Est du pont, celui passés sur le côté Ouest et enfin le nombre total de vélo passés sur le pont.
Voici un apperçu de notre jeu de données :

Date	Total	East	West
2012-10-03 00:00:00	13.0	4.0	9.0
2012-10-03 01:00:00	10.0	4.0	6.0
2012-10-03 02:00:00	2.0	1.0	1.0
2012-10-03 03:00:00	5.0	2.0	3.0
2012-10-03 04:00:00	7.0	6.0	1.0

Nous pouvons voir dans ce tableau que le 3 octobre 2012 à 1h du matin, 4 vélos sont passés par le côté est du pont et 9 sont passés par le côté ouest Ainsi, au total 13 vélos sont passés sur le pont.

Nous voulons maintenant avoir un aperçu global de ces données, pour cela nous affichons une description statistique de celles-ci. Voici le résultat obtenu:

	Total	East	West
count	136334.000000	136334.000000	136334.000000
mean	112.957663	51.506125	61.451538
std	143.602975	66.212433	89.403054
min	0.000000	0.000000	0.000000
25%	14.000000	6.000000	7.000000
50%	61.000000	28.000000	30.000000
75%	148.000000	69.000000	75.000000
max	1097.000000	698.000000	850.000000

Nous sommes donc en présence de 136334 données d'heures d'observation. En moyenne, 113 vélos passent sur ce pont par heure depuis 2012, dont 51 par le côté Est et 61 par le côté Ouest. Le maximum du nombre de vélos passés en une heure sur ce pont est de 1097.

Pour finir, nous traçons le graphique représentant le nombre de vélos total par jour passant sur le pont en fonction des mois des années 2012 à 2020.

On constate qu'il y a un pic au mois de Juillet 2014. De plus, on voit qu'il y a une variation du nombre de cyclistes en fonction de la saison : lors de l'hiver il y a moins de 15000 vélos par jour qui passent sur ce pont (avec un minimum de 4500 atteint en Janvier 2014), alors qu'en été ils sont plus de 20000 par jour. Cette tendance saisonnière peut s'expliquer par les conditions climatiques et météorologiques.

2) Transformation des données

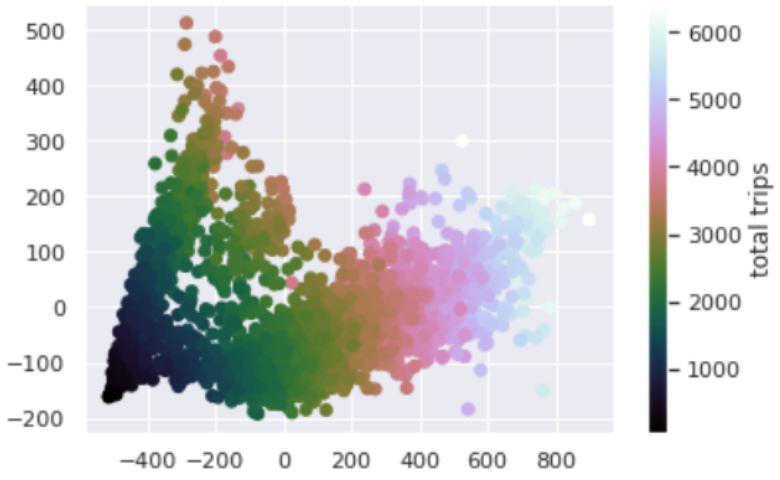
Dans le but d'analyser au mieux ces données, nous allons les transformer. En effet, nous allons réunir chaque donnée collectée pour le même jour dans une ligne en fonction de son heure. Nous allons ainsi avoir une matrice à deux dimensions: la première pour les vélos arrivant par le côté Est, et la deuxième pour ceux arrivant par le côté Ouest. Chaque ligne de cette matrice correspond à un jour de l'année et chaque colonne correspond a une heures de la journée (pour un total de 48 heures). Voici un aperçu du tableau de données obtenu après transformation:

	East										...	West								
Date	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	23
2012-10-03	4	4	1	2	6	21	105	257	291	172	...	51	92	182	391	258	69	51	38	25	12
2012-10-04	7	3	3	0	7	15	91	230	284	147	...	56	74	161	353	241	107	56	39	21	30
2012-10-05	4	4	4	2	7	18	68	218	251	131	...	62	84	190	290	209	73	41	31	26	16
2012-10-06	8	10	7	1	4	3	12	17	58	59	...	114	96	76	73	55	38	18	15	20	19
2012-10-07	6	12	2	4	1	6	9	14	43	67	...	115	109	93	73	45	23	36	35	9	11

Nous sommes maintenant en présence d'un jeu de données consitué de 2920 jours et 48 variables (les 48 heures d'une journée).

3) Visualisation des données

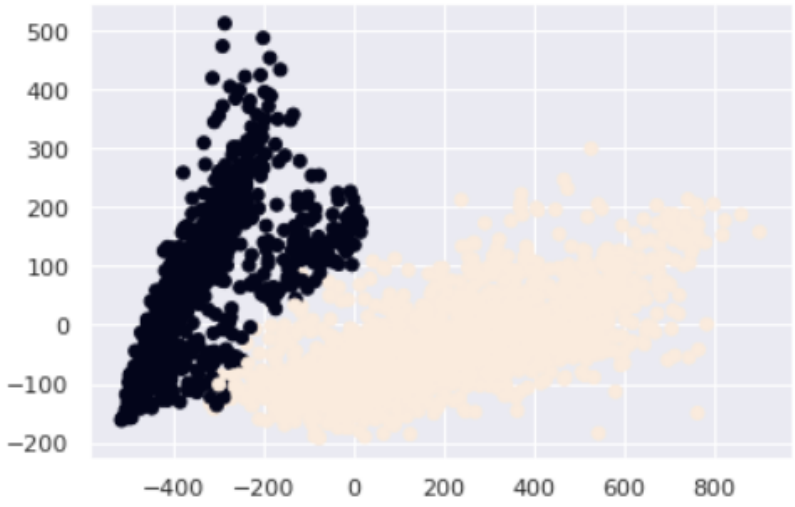
Nous pouvons penser que ces données représentent 2920 objets distincts qui vivent dans un espace de dimension 48, ou la valeur de chaque dimension est le nombre de vélos compté à une heure particulière sur une partie du pont (Est ou Ouest). Cependant, cette visualisation en 48 dimensions rendrait nos données difficilement visualisables. C'est pourquoi nous faisons le choix d'utiliser une technique de réduction de dimension qui utilise la méthode de l'analyse en composante principale (ACP). Cette méthode est une projection linéaire des données qui préserve le maximum de variance. Pour nos données, nous choisissons de garder 90% de la variance totale. Après application de cette méthode sur nos données, nous sommes en présence d'un objet à 3 dimensions, ce qui signifie que ces trois composantes projetées decrivent au moins 90% de la variance totale des données. Maintenant que nous avons des données en trois dimensions, nous pouvons plus facilement les représenter dans un graphique. Voici le graphique obtenu :



Sur ce graphique, chaque point est coloré en fonction du nombre total de vélos par jour compté. Ainsi, plus le point est clair et plus le nombre de vélo compté ce jour est grand. On voit que les données se séparent en deux groupes distincts et que le nombre de vélos compté augmente en fonction de la longueur de de chaque composante. Nous avons deux types de jour : les jours avec un grand nombre de vélos passés sur le pont et les jours avec un petit nombre. Enfin, les deux groupes sont de moins en moins dissociable lorsque le nombre de vélo compté sur le pont par jour diminue.

4) Classification non supervisée des données

Nous aimerions maintenant séparer les deux groupes que nous avons trouvés plus haut. Pour cela, nous allons utiliser l'algorithme de mélange gaussien. Nous allons appliquer ce modèle à nos données de cyclistes et représenter graphiquement le résultat obtenu.



Notre classification semble avoir fonctionné et séparé les deux groupes que nous avions repérés sur le graphique précédent. Nous appliquons cette classification à notre tableau de données et nous obtenons un résumé de celle-ci :

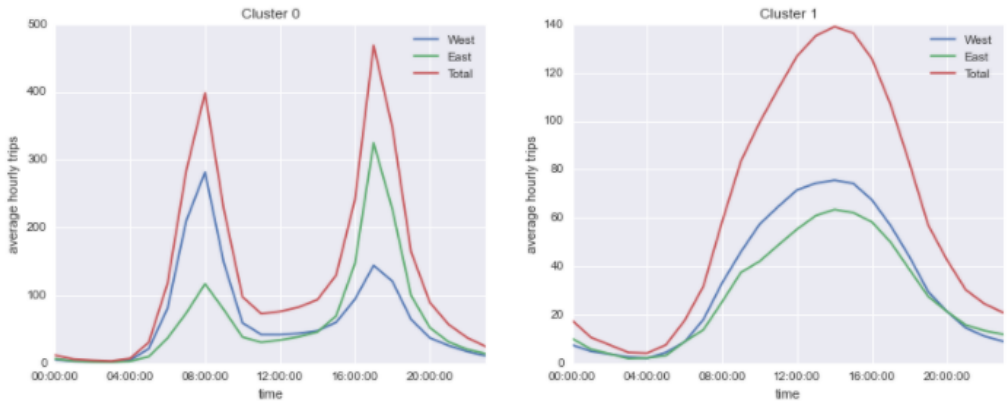
Date	West	East	Total	Cluster
2012-10-03 00:00:00	9.0	4.0	13.0	1
2012-10-03 01:00:00	6.0	4.0	10.0	1
2012-10-03 02:00:00	1.0	1.0	2.0	1
2012-10-03 03:00:00	3.0	2.0	5.0	1
2012-10-03 04:00:00	1.0	6.0	7.0	1

Ce tableau nous indique que les 5 premières heures de la journée du 3 octobre 2012 appartiennent au groupe 1. Nous finissons par afficher les tendances moyennes d'appartenance aux groupes 1 ou 2 des différentes heures de la journée.

Cluster		West	East	Total
0	**00:00:00**	8.312780	5.907509	14.220288
	01:00:00	4.545500	3.648434	8.193933
	02:00:00	2.841372	2.489806	5.331179
	03:00:00	1.576827	1.539035	3.115863
	04:00:00	2.184983	1.696171	3.881154

Nous constatons que les premières heures de la journée ont plus tendance à appartenir au groupe 0.

Pour finir, nous affichons le graphique représentant le nombre de passages de vélo par heure en fonction de l'heure pour les deux groupes. Voici les graphiques obtenus :

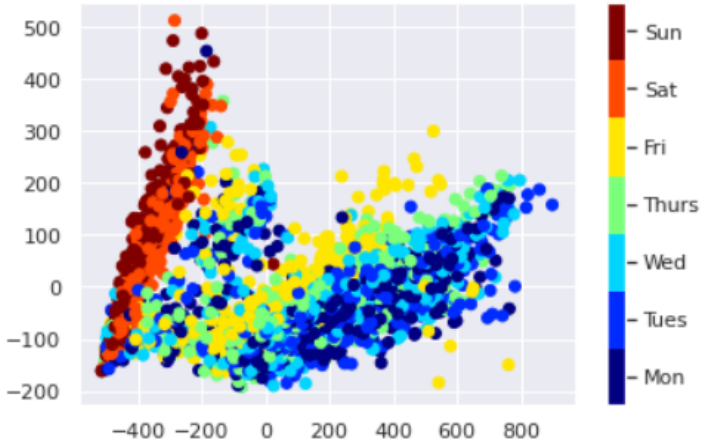


Nous constatons qu'il y a deux pics dans le groupe 0. En effet, le nombre de vélos par heure atteint son maximum aux environs de 18h, avec une prépondérance pour les vélos passant par le côté Est du pont. Ce phénomène pourrait s'expliquer par le fait que les personnes rentrent du travail. En outre, nous pouvons voir qu'il y a un deuxième pic le matin autour de 8h, avec un nombre plus important de vélos arrivant par le côté Ouest du pont. Ainsi, ces deux pics pourraient expliquer le fait que les cyclistes vont au travail par le côté Ouest et rentrent du travail par le côté Est du pont. Pour le deuxième groupe, nous pouvons constater que le nombre total de vélos par heure ne cesse d'augmenter entre 6h et 15h, avec un pic à 14h, puis descend en fin

de journée. Nous pouvons donc penser que le groupe 0 prend en compte les jours de la semaine, et donc de travail, alors que le groupe 1 prend en compte les jours du week-end. Nous allons confirmer nos hypothèses construites à l'aide de ce graphique en analysant les habitudes de travail des habitants de Seattle.

5) Les habitudes de travail à Seattle

Nous allons entrer plus dans les détails et essayer de voir comment nous pouvons extraire des informations sur les habitudes de travail des habitants de Seattle à partir des données que nous disposons. Nous commençons par tracer le graphique des données en distinguant les points en fonction du jour de la semaine auxquels ils appartiennent. Voici le graphique obtenu :



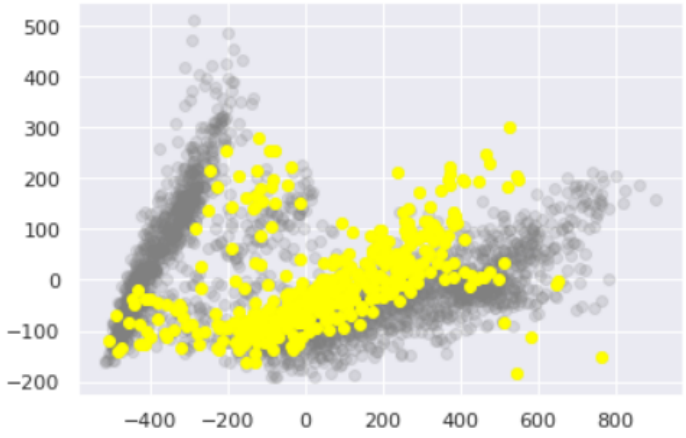
Nous constatons que les jours du week-end (samedi et dimanche) sont regroupés sur le côté gauche du graphique, tout comme l'était les données du groupe 1 sur le graphique précédent. De plus, les autres jours de la semaine sont regroupés au centre du graphe. Ainsi, notre intuition était bonne et les groupes différencient les données en fonction du jour de la semaine. Nous vérifions cette idée en affichant les données assignées aux différents groupes :

	cluster	is_weekend	weekday
2012-10-03	1	False	Wed
2012-10-04	1	False	Thu
2012-10-05	1	False	Fri
2012-10-06	0	True	Sat
2012-10-07	0	True	Sun

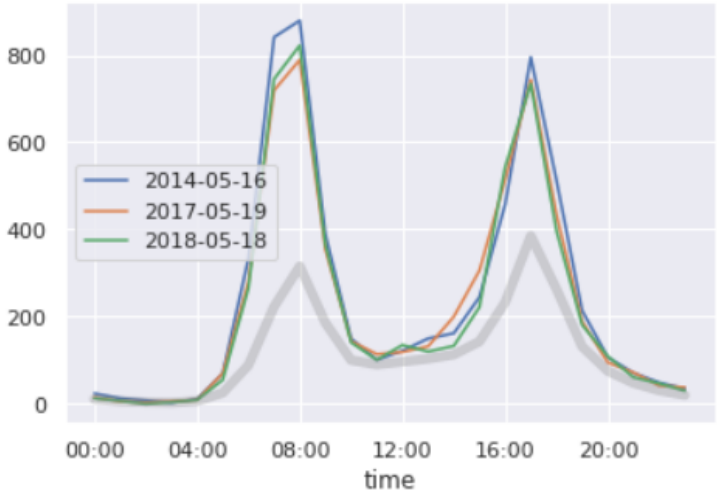
Nous constatons que notre idée était bonne: les jours de la semaine appartiennent au groupe 1 et les jours du week-end au groupe 0. Cependant, nous constatons que les vendredis sont assez éparpillés sur le graphique. Nous allons donc regarder ce jour de la semaine avec une attention particulière.

6) Que ce passe-t-il les Vendredis?

Nous commençons par tracer le graphique des données obtenu après l'ACP en surlignant les points qui ont été enregistré un vendredi.



Nous constatons que les données relevées les vendredis ne rentrent pas totalement dans une des deux classes. Pourtant, le vendredi étant un jour de la semaine, ces données devraient donc appartenir à la classe 1. Nous portons une attention particulière aux trois vendredis situés après 600 sur l'axe des abscisses. Après analyse de ces trois points, nous savons qu'il s'agit des vendredis : 16 Mai 2016, 19 Mai 2017 et 18 Mai 2018. Ces trois vendredis sont au milieu du mois de Mai, est-ce une coincidence ? Nous allons visualiser ces trois vendredis sur un graphique.



Nous constatons que le nombre de cyclistes ayant passé le pont ces trois vendredis est trés grand. En effet, lors des pics du départ au travail et celui du retour du travail, environ 800 personnes sont passées par heure. Ce nombre est nettement au-dessus des données journalières habituelles. Après quelques recherches, nous avons appris que ces trois vendredis étaient les journées annuelles du mouvement "Aller au travail au Travail" dans la ville de Seattle.

Conclusion

Après avoir analysé et décrit statistiquement les données sur le nombre de vélos qui passe sur le pont Fermond ainsi qu'utiliser des méthodes de classification non supervisée, nous avons dégagé certaines informations sur celles-ci. En effet, nous avons appris sur les habitudes de travail des habitants de Seattle qui passent en vélo sur le pont Fermond pour aller au travail. En résumé, nous avons appris que :

- Les cyclistes de Seattle ont tendance a poser un jour de congé pour les fêtes nationales telles que : Le nouvel an, Thanksgiving, le jour de Noël, le jour de la fête d'indépendance et le Memorial Day.
- Les cyclistes de Seattle ont tendance a aller travailler les jours de fête nationale moins communes telles que : le jour de Columbus, le jour de Martin Luther King Jr., le jour des presidents, le jour des vétérans.
- Les cyclistes de Seattle font leur maximum poour ne pas être retenu au travail pendant le week-end.