Bike Challenge

Sophie Manuel

Nous cherchons dans ce challenge à prédire le nombre de vélos qui passeront à Albert 1er à Montpellier, le 2 Avril entre minuit et 9h. Pour cela on utilisera les données collectée depuis 1 an sur lesquelles nous ferons une régression linéaire.

1 Le jeu de donées

Pour exploiter le jeu de données il fallait le modifier. Nous devons débuter avec le téléchargement des données avec le package `download`, grâce auquel nous pouvons avoir une mise à jour des données immédiatement.

Puis, nous avons pu retirer les lignes et les colonnes vides ou inintéressantes pour ce challenge. Nous avions donc 3 colonnes : "Date", "Heure", "Total de la journée".

Une fois que tout cela est fait, nous pouvons procéder à la sélection des données en prenant que celles étant entre minuit et 9h comme demandé.

Ensuite, grâce au package `pandas` nous avons changer le format des dates pour les mettre au format international afin de pouvoir les traiter plus facilement. De plus, il a permit de grouper les données par jour avec la fonction `groupby`du package. Pour grouper ces variables, nous avons choisi de ne prendre que la dernière données du jour dans la colonne "Total de la journée", car elle représente le nombre de vélos passés entre minuit et l'heure en question. Nous avons pu utiliser `.last()` pour grouper les données car elles étaient rentrées dans l'ordre chronologique, si ce n'était pas le cas nous aurions dû utiliser le maximum.

2 Régression et prédiction

Enfin, nous avons pu faire une régression linéaire avec la fonction `lingress` du module `scipy.stats`. Nous avons trouvé $\hat{\beta}_0 = 67.12$ pour l'ordonnée à l'origine et $\hat{\beta}_1 = 0.49$ pour la pente à 10^{-2} près. Donc :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times x = 67.12 + 0.49 \times x$$

Il permet de prévoir qu'il y aurait 157 vélos qui passeraient le 2 avril entre minuit et 9h, arrondi au vélo près. Cependant, lorsque nous regardons le R^2 on prévoit que ce modèle n'est pas très bon pour de la prédictions sur nos données. Il est de 6%.

Voici le graphique que nous avons obtenu:

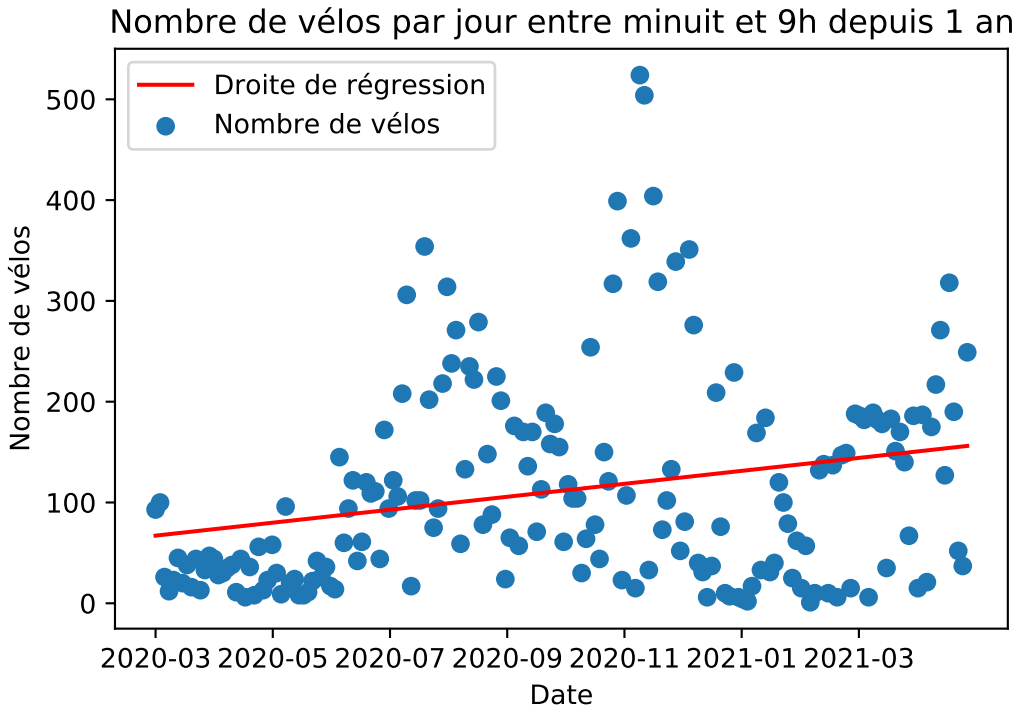

Figure 1: Graphique regression lineaire

2.1 Critiques du modèle

`On constate que la variance des données est très grande au cours du temps. Effectivement entre juillet`

De plus, la courbe est clairement faussée par les données avant le mois de juin 2020 car c'était durant le premier confinement qui était assez strict. Celles-ci on un effet de levier sur la droite.

Pour conclure, ce modèle (bien qu'imparfait) prédit qu' il y aura 157 vélos qui passeront à Albert 1er à Montpellier, le 2 Avril entre minuit et 9h.

Lien du dépot Github : https://github.com/SophieManuel/Bike_Challenge