Projet — Partie 1

Échéancier: au plus tard le dimanche 13 octobre à 23h55.

BIXI Montréal est un organisme à but non lucratif qui gère un système de vélopartage dans la région de Montréal, voir `https://bixi.com/fr/` pour plus de détails. L'objectif de cette première partie de projet est d'analyser les données en libre accès sur l'utilisation de BIXI et les déplacements de la saison 2021.

Données:

Les données brutes contiennent tous les déplacements à Bixi de l'année 2021. Chaque observation comporte des informations sur un déplacement individuel: la date et l'heure de départ, la station de départ, la date et l'heure de fin, la station de fin, la durée totale du déplacement et une variable indiquant si l'utilisateur est un membre BIXI. Seules les trajets d'une durée de moins de deux heures dans les mois de mai à octobre sont considérés à des fins d'analyse statistique. En 2021, les membres de Bixi pouvaient emprunter un vélo sans frais additionnels pour une durée de 45 minutes, et bénéficiaient de rabais pour les trajets excédants cette durée ou pour la location de vélos avec assistance électrique. En plus de ces informations, des données météorologiques ont été fusionnées afin d'inclure la température moyenne quotidienne (en °C) et la quantité quotidienne de précipitations (en mm). Le jeu de données avec lequel vous allez travailler inclut les variables suivantes:

`dep`	heure et date de départ du trajet
`dur`	durée du trajet (en secondes)
`mem`	variable indicatrice binaire pour les membres, (`mem=1` pour une personne avec un abonnement, `mem=0` sinon)
`jour`	jour de la semaine, du `dimanche` au `samedi`
`temp`	température quotidienne moyenne (en °C)
`prec`	précipitation cumulativ journalière (en mm)
`pointe`	variable catégorielle, `1` pour les trajets entre 7:00 et 10:00 (pointe du matin), `2` pour ceux entre 16:00 à 18:00 (pointe du soir), `3` sinon.

Important: Chaque équipe se verra assigner un échantillon aléatoire stratifié de la base de données. Assurez-vous de travailler avec les données assignées à votre équipe.

Mandat:

L'objectif de cette première partie du projet est d'explorer les facteurs qui influencent la durée des trajets à vélo en répondant aux questions ci-dessous. Tout au long du projet, assurez-vous que vos analyses vous permettent de répondre aux questions de recherche d'une manière appropriée et adéquate. Commentez les résultats et discutez des principales conclusions de ces analyses d'un point de vue d'affaires, en fournissant des informations intéressantes et pertinentes. Le code doit être fourni dans un script séparément de votre rapport. Chaque fois qu'un modèle statistique est employé, veillez à

- rapporter les coefficients ou les différences estimées (avec des unités), avec une estimation de l'incertitude,
- fournir des interprétations des paramètres sur une échelle adéquate,
- tirer des conclusions qui reflètent le contexte,
- discuter de la validité de vos analyses,
- discuter de toute lacune ou limitation de vos modèles.

Bien que vous puissiez essayer plusieurs modèles pour votre analyse, votre rapport devrait contenir les conclusions et analyses pour le(s) plus approprié(s) afin de répondre aux questions de recherche. Assurez-vous de justifier votre choix de modèle et de fournir des diagnostics pour la validité de votre modélisation. En particulier, vérifiez si une transformation adéquate de la variable réponse permettrait de mieux expliquer la loi conditionnelle de $Y \mid \boldsymbol{X}$.

Avant de débuter, effectuez une analyse exploratoire des données. Un **maximum de 2 pages** est alloué pour cette partie: ne conservez que les portions **pertinentes** et les résumés dans votre rapport.

Questions de recherche:

1. En moyenne, les membres de BIXI effectuent-ils des trajets plus courts que les non-membres? Les résultats sont-ils les mêmes si l'on tient compte de l'utilisation en fin de semaine ou en semaine?
2. Est-ce que la durée des trajets est influencée par la météo? Au vu du résultat que vous obtenezj, est-ce que vos modèles initiaux devraient être revisités?
3. Les durées de trajets sont-elles différentes selon que l'on se trouve aux heures de pointe ou non en semaine? Existe-t-il des différences entre l'utilisation pour les heures de pointes en semaine le matin ou le soir? *Indication: considérez la spécification de contrastes.*

Évaluation:

Chaque portion du projet sera évaluée selon les critères suivants:

(a) Clarté et qualité du rapport [3 pts]:
- la structure et la présentation du rapport,
- le respect des règles d'orthographe, de grammaire et de syntaxe,
- la clarté et concision de l'écriture.

(b) Pertinence de la discussion [5 pts]:
- le caractère adéquat des interprétations et des trouvailles,
- la pertinence des conclusions.

(c) Rigueur de l'analyse [12 pts]:
- la pertinence des modèles employés,
- la validité et la justesse des interprétations,
- l'exhaustivité de l'analyse pour répondre aux questions.

Instructions:

- Projet en équipe (minimum trois, maximum quatre personnes).
- Un(e) seul(e) membre de l'équipe doit soumettre les rendus via ZoneCours
- Vous êtes encouragés à créer votre rapport avec Quarto ou R Markdown.
- Les livrables incluent
  - un rapport en PDF **d'au plus 10 pages**.
  - le code **R** utilisé pour générer les résultats (utilisez la fonction `knitr::purl` pour extraire le code si nécessaire)
  - le fichier Quarto ou R Markdown utilisé pour générer le rapport, le cas échéant.
- Utilisez la convention `MATH60604_P1_matricule.extension`, où `matricule` est le matricule HEC de la personne qui soumet le rapport, et `extension` est une de `pdf`, `qmd`, `R`, ou `Rmd`.
- Chaque équipe doit également fournir le nom des membres une brève description de la contribution de chaque membre de l'équipe au travail. Cette description peut être directement dans le rapport.
- En effectuant les analyses, vous pouvez créer de nouvelles variables (par exemple, des transformations de variables) sur l'ensemble des données attribuées à votre équipe, mais vous ne pouvez pas fusionner des données auxiliaires d'autre provenance.
- vos analyses doivent être **reproductibles**, c'est-à-dire que nous devons être en mesure d'exécuter votre code pour obtenir les mêmes résultats que ceux fournis dans votre rapport.
- Suivez les instructions relatives à l'utilisation de l'IA générative détaillées dans le plan du cours.

Remarques importantes:

- Politique concernant les soumissions tardives:
  - 24 heures ou moins de retard: -15%
  - entre 24 – 48 heures de retard: -30%
  - plus de 48 heures: note de zéro
- Toute portion de votre rapport copiée mot pour mot à partir du matériel de cours ou d'autres sources sans citation adéquate sera considérée comme du plagiat et recevra une note de zéro. Citez correctement les sources.