Project — Part 2

Due date: November 9 at 11:55 PM

In this second part of the project, we will continue to explore data from BIXI Montréal (https://bixi.com/en/).

Data:

The raw data consist of Bixi usage records from the 2024 season, wherein each observation consists of an individual trip and includes details on the start time and place, as well as the end time and place. Note that, as in the first part of the project, only trips under 2 hours in the months extending from May to October, inclusively, are considered. In this second part of the project we will explore **aggregated data** from the 2024 season; specifically, the **daily** records are aggregated at the **station level**. In addition, weather information (temperature and precipitation) have been merged into the data. Each observation in the modified data consists of the total number of daily BIXI trips leaving from each station, and includes the following variables:

`station`	departure station name
`arrondissement`	borough where the station is located
`mm`	departure month, ranging from 5 (May) to 10 (October)
`jj`	departure day, ranging from 1 (Monday) to 7 (Sunday)
`n_tot`	total number of trips
`n_AM`	number of trips leaving during morning rush hour (7:00 AM to 10:00 AM)
`n_PM`	number of trips leaving during evening rush hour (4:00 PM to 7:00 PM)
`n_rush`	number of rush hour trips (defined as `n_AM + n_PM`)
`temp`	average daily temperature (in °C)
`prec_ind`	binary variable indicating moderate to heavy rainfall, (defined as a day with a total precipitation amount exceeding 10 mm)

Important: Each team will be assigned a distinct stratified sample of the data. Be sure to work with the specific dataset assigned to your team.

Mandate:

The goal of this second part of the project is to explore the factors which affect daily BIXI usage by addressing the questions given below. This second part of the project will focus on modeling the **proportion** of daily trips departing during rush hour. Throughout, be sure that your analyses allow you to answer the business questions in an appropriate and adequate manner. Comment on findings and discuss the main takeaways from these analyses from a business perspective, providing interesting and relevant insights. Whenever a statistical model is used, be sure to

- report estimated coefficients, along with uncertainty measures,
- provide relevant parameter interpretations pertaining to the business question,
- provide relevant conclusions that reflect the context,
- discuss the validity of the analysis carried out,
- discuss any shortcomings or limitations of the analysis carried out.

While you may explore several models in analyzing the data, your report should only include the most appropriate model for answering the questions below. You may use a single model to answer all of the questions. Be sure to justify your choice of model, showing only relevant output.

Before beginning, carry out an exploratory data analysis. A **maximum of 2 pages** is allotted for the exploratory analysis in your report. Be sure to include only **relevant** output and findings.

Business questions:

1. Does the odds of a rush hour departure vary significantly across the months?
2. Which day of the week has the highest odds of a rush hour departure? Which day of the week has the lowest odds of a rush hour departure? Are the results affected by moderate / heavy rainfall?
3. Explain any estimation issues that come up in the results.

Evaluation:

Each part of the project will be graded according to the following criteria:

(a) Quality and clarity of the report :
- the structure and presentation of the report,
- the syntax and grammar of the writing,
- the clarity and conciseness of the writing.

(b) Relevance of the discussions :
- the appropriateness of the interpretations and insights discussed in the report,
- the relevance of the conclusions given in the report.

(c) Correctness of the analysis :
- the appropriateness and adequacy of the model(s) considered,
- the validity of the interpretations and conclusions,
- the completeness of the analyses in addressing the questions.

Submission instructions:

- This project is teamwork (minimum three, maximum four students).
- A single student should submit online through ZoneCours
- Deliverables include
  - your PDF report, **at most 7 pages** (you are encouraged to create your report using Quarto or R Markdown)
  - the **R** code (use `knitr::purl` to extract the code if necessary)
- Use the naming convention `MATH60604A_P2_id.extension`, where `id` is the HEC identifier of the student submitting the report and `extension` is one of `pdf`, `R`.
- The assignment report must include a **coverpage** including the names of team members and a brief description of each team members' contribution to the work (note that the coverpage does not count toward the 7 page limit).
- In carrying out the analyses, you may create new variables (e.g., variable transformations) based on your team's assigned dataset, but you cannot merge in complementary data.
- Your analyses should be **reproducible**: I should be able to run your code to obtain the same output provided in your report.
- Please be sure to follow the instructions regarding the use of generative AI detailed in the course outline.

Important remarks:

- Policy on late submissions:
  - 24 hours or less late: -15%
  - 24 – 48 hours late: -30%
  - over 48 hours late: not accepted (grade of 0)
- Any part of your report that is copied verbatim from course material, or other sources, will be considered plagiarism and given a grade of zero. Provide proper attribution of sources and citations for any reference.