Project — Part 1

Due date: at the latest on Sunday, October 13, 2024 at 11:55 PM

BIXI Montréal is a non-profit organization that manages a bike sharing system in the metropolitan area, see `https://bixi.com/en/` for details. The first part of the project will consist in an analysis of open access data on BIXI bike rentals.

Data:

The raw data consist of the records of every Bixi rental for the 2021 season. In particular, each observation consists in an individual trip and includes the following information: the start date and time, the start station, the end date and time, the end station, the total trip duration, and a variable indicating whether the user is a BIXI member. Only trips under 2 hours in the months extending from May to October, inclusively, are considered for the statistical analysis here. Note that in the 2021 season, members could use a regular Bixi for up to 45 minutes for free, and obtained rebates for electric bike rentals or longer trips. In addition to Bixi usage, weather information was merged with the BIXI data to provide the daily average temperature (in °C) and the daily cumulated amount of rainfall (in mm). The data you will be working with includes the following variables on each individual BIXI trip:

`dep`	trip departure date and time
`dur`	trip duration (in seconds)
`mem`	binary variable indicating whether user is BIXI member (`mem=1` if the user has a BIXI membership and `mem=0` otherwise)
`wday`	trip weekday (derived from `dep`, taking on values `Sunday` through `Saturday`)
`temp`	average daily temperature (in °C)
`prec`	total amount of daily rainfall (in mm)
`rushhour`	categorical variable, `1` for morning rush hour (7AM to 10AM), `2` for afternoon rush hour (4PM to 6PM), `3` otherwise.

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 first part of the project is to explore the factors which affect trip duration by addressing the questions given below. 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. Code should be provided as a standalone file. Whenever a statistical model is used, be sure to

- report estimated coefficients or differences (with units), along with uncertainty measures.
- provide appropriate parameter interpretations,
- 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(s) for answering the questions below. Be sure to justify your choice of model(s), showing only relevant output. Throughout, perform model diagnostics to assess the adequacy of the model(s) considered. In particular, verify whether a suitable transformation of the response variable may be more adequate to model the conditional distribution of $Y \mid \boldsymbol{X}$.

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. On average, do BIXI members have shorter trips than non-members? Are the results the same if one adjusts for weekend vs. non-weekend usage?
2. Are trip durations impacted by weather factors? In light of the results you obtain, should your initial model(s) be revisited?
3. Do rush hour trip durations differ from those during non peak hours (i.e., not rush hour) during weekdays? Are there differences between the AM and PM rush hour weekday usage, respectively? (Hint: consider using contrasts).

Evaluation:

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

(a) Quality of the report **[3 pts]**:
- 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 **[5 pts]**:
- the appropriateness of the interpretations and insights,
- the relevance of the conclusions drawn.

(c) Correctness of the analysis **[12 pts]**:
- the appropriateness and adequacy of the models 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
- You are encouraged to create your report using Quarto or R Markdown.
- Deliverables include
  - your PDF report, **at most 10 pages**.
  - the **R** code (use `knitr::purl` to extract the code if necessary)
  - the Quarto or R Markdown file used to generate the report, if applicable.
- Use the naming convention `MATH60604A_P1_id.extension`, where `id` is the HEC identifier of the student submitting the report and `extension` is one of `pdf`, `qmd`, `R`, `Rmd`.
- The assignment report must include the names of team members and a brief description of each team members' contribution to the work.
- 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**: we 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.