Project — Part 2

Due date: at the latest Sunday, November 24, 2024 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 2023 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 2023 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)
`dd`	departure date, ranging from 1 to 31
`wday`	weekday, taking on values `Sunday` through `Saturday`
`AM`	number of trips leaving in the morning (7:00 AM to 10:00 AM)
`tot`	total number of trips
`temp`	average daily temperature (in °C)
`prec`	total amount of daily rainfall (in 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 both in terms of the proportion of daily trips leaving in the morning hours, as well as the counts (number of trips) themselves. Throughout, be sure that your analyses allow you to answer the specific 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, along with uncertainty measures,
- provide appropriate parameter interpretations, as pertaining to the question,
- provide relevant conclusions that reflect the context, as pertaining to the question,
- 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. Any time a statistical test is carried out, be sure to clearly state the underlying hypotheses in terms of the model parameters, provide the value of the test statistic and resulting p-value, and provide a relevant conclusion which reflects the context.

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. First explore the factors that influence the **proportion** of daily trips that depart in the morning (AM). Consider a **single model** which allows to adequately address each of the following questions:

   (a) Is there significant variation in the odds of an AM departure across the different boroughs? Carry out an appropriate statistical test, using $\alpha = 1\%$. Provide a relevant conclusion, in the context of the problem.

   (b) According to the fitted model, which month has the highest odds of an AM departure? Which has the lowest? Justify your answer based on the estimated regression coefficients, and interpret the results. Are the results are surprising?

   (c) BIXI suspects that the odds of a morning departure decreases on weekends in comparison to weekdays, and that this decrease is further accentuated when there is rain. Does the fitted model support this? Explain your answer, based on the estimated regression coefficients.

   (d) Does the impact of temperature on the odds of a morning departure significantly differ on weekends in comparison to weekdays? Carry out an appropriate statistical test using $\alpha = 1\%$.

2. Suppose it is of interest to model BIXI usage in terms of the **total number of daily trips at the station level**. Discuss the difficulties that arise in attempting to fit such a model from the data available.

   Some helpful considerations:

- Every BIXI station has a different capacity in terms of the number of bicycles that are available for use, and this capacity varies randomly in time. (Has it ever happened to you that there were no bikes available at your nearest station?)
- The raw data records each individual trip. In tabulating the total number of daily trips at the station level, there are no records for a station that was not used on a given day. (Hint: explore descriptive statistics of the variable `tot` in the dataset).

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)
- 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`, `R`.
- The assignment report must include the names of each team member 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 the report that extends beyond the page limit will not be considered (not graded).
- 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.