Project

BIXI Montréal is a non-profit organization that manages a bike sharing system in the Montreal area, see `https://bixi.com/en/` for details. The goal of this project is to analyse open access data on BIXI usage and trips from the 2021 season.

<u>**The Data:**</u> The raw data, available at `https://bixi.com/en/open-data-2/`, consists of details on each bixi rental for the 2021 season. In particular, each observations provides the following details on the individual trip: the start date and time, the start station, the end date and time, the end station, the total trip duration, and an indicator for whether the user is a BIXI member. Here, **a subset** of the original data was used, **including only short trips of under 60 minutes**. The data was then **aggregated to yield daily usage information at each station**. In addition, weather information (found here `https://climate.weather.gc.ca/historical_data/search_historic_data_e.html`) was merged with the BIXI data to provide the daily average temperature (in °C) and the daily amount of rainfall (in mm). The final dataset (which you will be working with) provides information on the daily usage of BIXI trips leaving from each station, and includes the following variables:

station	the start station
mm	the month (ranging from 4: April, to 11: November)
dd	the day of the month (ranging from 1 to 31)
wday	the weekday (from Sunday – Saturday)
mem	member indicator (1= BIXI member, 0=non-member)
holiday	holiday indicator (1=holiday, 0=non-holiday)
dur	total duration (in minutes) of all trips leaving from the given station on the specified day
avg	average duration (in minutes) of trips leaving from the given station on the specified day
rev	total revenue generated by trips leaving from the given station on the specified day (**note:** the revenue is only available for non-members*)
n_AM	total number of trips leaving from the specified station in the morning, specifically, between 4:00 AM and 12:00 PM
n_PM	total number of trips leaving from the specified station in the afternoon, specifically, between 12:00 PM and 8:00 PM
n_tot	total number of trips leaving from the specified station throughout the day
temp	average daily temperature (in °C)
rain	total amount of rainfall on specified day (in mm)

<u>Important</u>: Each team will be assigned a specific subset of the data. Be sure to work with the specific dataset assigned to your team.

[*] *As noted above, the revenue generated by the BIXI trips is only available for non-members. This follows from the limited information available in the raw data. In particular, the raw data does not identify individual members but rather just individual trips with a variable indicating whether the user is a BIXI member. The pricing of BIXI usage for members consists of a flat monthly, or seasonal, fee and then per-minute charges for each minute exceeding 45 minutes. Since the data does not have identifiers for the unique members, we cannot accurately incorporate the membership fee into the revenue generated by BIXI members. For non-members, on the other hand, there is a flat rental fee for each trip, and then a usage fee calculated per minute. The revenue here was manually calculated using the pricing scheme found here:* `https://bixi.com/en/pricing/`.

Instructions:

Throughout this project, the goal is to explore the factors which affect trip lengths (duration), the revenue generated, and usage for everyday BIXI usage (i.e., for trips under 60 minutes). The project is broken down into four parts, each of which are detailed below. Note that the instructions leave a lot of room for creativity and you have complete flexibility in how you choose to approach the task. You can also modify the data, as relevant for your analysis - that is, you can create new variables, transform variables, merge other relevant data, etc.!

Part 1: Exploratory analysis (due September 22 before 11:55 PM)

- Carry out an exploratory analysis of the data.
- Explore variables individually as well as relationships between variables, keeping in mind the response variables of interest in this project and potential covariates.
- Comment on findings and discuss the main takeaways from this analysis, from a business perspective. Be sure to provide appropriate and relevant data summaries to support your discussion.

Part 2: Linear regression models (due October 20 before 11:55 PM)

- Explore linear regression models for the response variables of interest, specifically, for trip lengths (duration) and revenue.
- Be sure that your analyses allow you to answer well formulated business / research questions that you wish to explore through these models. The goal is to use linear regression models to provide interesting and relevant insights from the data.
- Comment on findings and discuss the main takeaways from these analyses from a business perspective. Be sure to provide relevant model outputs that support your discussion.
- Discuss any shortcomings or limitations of the analyses carried out.

<u>Part 3: Generalized linear models</u> (due November 10 before 11:55 PM)

- Explore various generalized linear models for the response variables of interest, specifically, for the number of rentals (total, AM, and PM). In addition, create a new variable indicating whether the average daily trip duration exceeds 15 minutes, and explore models for this new variable.
- Be sure that your analyses allow you to answer well formulated business / research questions that you wish to address. The goal is to use generalized linear models to provide interesting and relevant insights from the data.
- Comment on findings and discuss the main takeaways from this analysis from a business perspective. Be sure to provide relevant model outputs that support your discussion.
- Discuss any shortcomings or limitations of the analyses carried out.

<u>Part 4: Linear mixed models</u> (due December 8 before 11:55 PM)

- Revisit the linear regression models explored in Part 2 of your project, this time allowing to capture the potential intra-station correlation.
- Comment on findings and discuss the main takeaways from this analysis from a business perspective. Be sure to provide all relevant model outputs that support your discussion.
- Discuss any shortcomings or limitations of the analyses carried out.

Evaluation:

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

(a) Clarity and quality of the report:

- the structure and presentation of the report,
- the clarity and conciseness of the writing.

(b) Creativity and originality:

- the creativity of the analyses explored,
- the insights discussed in the report,
- the relevance of the conclusions given in the report.

(c) Quality and completeness of the analysis:

- the appropriateness of the methods considered,
- the completeness and rigour of the analyses,
- the validity of the interpretations and statistical conclusions given in the report.

Part 1 will be graded on **10 points** with weighting:

(a) 2 pts (b) 3 pts (c) 5 pts

Parts 2–4 will each be graded on **20 points** with weighting:

(a) 3 pts (b) 5 pts (c) 12 pts

Submission instructions:

- Each part of your project should be no more than 15 pages long (not including the coverpage). Any additional pages over the limit will not be graded.
- Projects must be submitted through ZoneCours in a ZIP file which includes:
  - the written report in PDF format
  - the corresponding R code (either as an R code file, or as an RMD file)

    * *important note: your analyses should be reproducible, that is, I should be able to run your code to obtain the output provided in your report.*
- Each team must also provide a very brief description of each team members' contribution to the work. This can be given in the written report.

A few important remarks:

- The project should be done in groups of 5 students.
- You can change teams for each part of the project if you wish.
- Please notify me of your group members by email **no later than two weeks prior the due date**. Anyone who has not provided me with group information by this date will be assigned a group.
- Policy on late submissions:
  - within 24 hours late: -20%
  - 24 – 48 hours late: -50%
  - over 48 hours late: not accepted (grade of 0)
- As always, plagiarism, in any form, will not be tolerated. **Any part of your report that is copied verbatim from course material, or other sources, will be considered plagiarism and given a grade of 0!**