# 42184 Data Science for Mobility
# 42577 Introduction to Business Analytics course
# Challenge statement

Welcome to this year's challenge!

The topic this year is *sustainable mobility*. At a time when the world is facing unprecedented challenges of different kinds, including climate change, pandemics, social inequality, and degrading biodiversity, we need to be conscious of the impact that mobility has on our society. How can we make emission-free transport systems more efficient and attractive? In this project, we invite you to appropriate this question and use your best Data Sciences skills to explore it. We do not expect you to discover revolutionary knowledge and save the world with a single Data Sciences project, instead, we want you to address the mandatory questions (below) but also seek yourself for new questions, new data, new insights.

You have access to data from Bluebikes (Boston), a bike-sharing system in the United States. The dataset includes more than 400 stations and 4000 bikes, and it contains over 2 million bike rides observed during 2022 (January to August). This dataset has the general objective of helping Bluebikes operating at its best and of making bike sharing more attractive. You can also download additional data (such as station data)[1] here. The data is in itself an interesting Data Sciences exploration.

## Project

The project has three components:

- *Prediction challenge* (30%): All groups need to address the same problem (30%).
- *Exploratory component* (40%): Each group is invited to choose their own research question and explore the data accordingly.
- *Report (30%) -* Each group should deliver one or more jupyter-notebooks, that should be self-explanatory in each step (or block). This will function as a report, so it should have introduction and conclusions, besides comments and reflections. However, there are some rules about the structure of the report, which should follow the 4-part outline shown below:

  Section 1: Introduction + Data analysis and visualization

  Section 2: Prediction Challenge

  Section 3: Exploratory Component

  Section 4: Conclusions

At the end of this document you will find a list of practical informations, which will include details on what is expected in each task, and how these aspects contribute to the final grade.

[1] https://www.bluebikes.com/system-data

## Introduction to the data

Figure 1 shows the variables you will have in this dataset. The data is provided as a CSV file. Notice that the variables require a lot of treatment in order to be usable (e.g. Dates, categorical, strings, different scales, IDs).

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Unnamed: 0** | 0 | 1 | 2 | 3 | 4 | 5 |
| **tripduration** | 597 | 411 | 476 | 466 | 752 | 339 |
| **starttime** | 2022-01-01 00:00:25.1660 | 2022-01-01 00:00:40.4300 | 2022-01-01 00:00:54.8180 | 2022-01-01 00:01:01.6080 | 2022-01-01 00:01:06.0520 | 2022-01-01 00:01:08.5000 |
| **stoptime** | 2022-01-01 00:10:22.1920 | 2022-01-01 00:07:32.1980 | 2022-01-01 00:08:51.6680 | 2022-01-01 00:08:48.2350 | 2022-01-01 00:13:38.2300 | 2022-01-01 00:06:47.5310 |
| **start station id** | 178 | 189 | 94 | 94 | 19 | 107 |
| **start station name** | MIT Pacific St at Purrington St | Kendall T | Main St at Austin St | Main St at Austin St | Park Dr at Buswell St | Ames St at Main St |
| **start station latitude** | 42.359573 | 42.362428 | 42.375603 | 42.375603 | 42.347241 | 42.3625 |
| **start station longitude** | -71.101295 | -71.084955 | -71.064608 | -71.064608 | -71.105301 | -71.08822 |
| **end station id** | 74 | 178 | 356 | 356 | 41 | 68 |
| **end station name** | Harvard Square at Mass Ave/ Dunster | MIT Pacific St at Purrington St | Charlestown Navy Yard | Charlestown Navy Yard | Packard's Corner - Commonwealth Ave at Brighto... | Central Square at Mass Ave / Essex St |
| **end station latitude** | 42.373268 | 42.359573 | 42.374125 | 42.374125 | 42.352261 | 42.36507 |
| **end station longitude** | -71.118579 | -71.101295 | -71.054812 | -71.054812 | -71.123831 | -71.1031 |
| **bikeid** | 4923 | 3112 | 6901 | 5214 | 2214 | 6877 |
| **usertype** | Subscriber | Subscriber | Customer | Customer | Subscriber | Subscriber |
| **postal code** | 02139 | 02139 | 02124 | 02124 | 02215 | 02139 |

*Figure 1. Dataframe view*

For the *prediction challenge*, you are expected to predict the **demand for the bike-sharing system (number of pickups).** You can do the predictions at a city level (i.e. ignoring the geographical representation – e.g., start and end station).

- You are expected to predict the total demand for the bike sharing system 2 hours in the future (e.g., given demand data until 9 am, predict the number of pickups for the interval 10-11 am). You should **not shuffle the data**. You should instead use the data from January to July (included) to train your model, and the data for August as a test set. As a benchmark, we expect you to be able to predict the test set with an $R^2$ of at least 0.60. You can use any sklearn regression model you want, including those not taught in the class.
- An important aspect to consider in the prediction challenge is how the data are aggregated. For example, what is the impact of aggregating trips every 60, 30, or 15

minutes? Students should compare how different levels of aggregation impact the outputs of the model.

In the *exploratory component*, each group needs to address at least one new research question. Here, we expect you to formulate your own question, and follow the data sciences cycle. The project will be positively valued with one or more of the following extensions:

- Extension of the dataset with other relevant data (weather data, national holidays, special events, etc.)
- Generation and analysis of insightful visualizations;
- Usage of the breadth of techniques from the class beyond regression and data preparation (e.g. dimensionality reduction, clustering, classification, time series)

Some example research questions:

- How to make predictions for each station? What about a cluster of stations?
- What is the correlation between data? (e.g. are the predictions of tomorrow influenced by the data of 5 months ago? When should we stop?)
- Are there periodic and seasonal trends (winter, summer, …) and how can we model them?
- What is the impact of land use (e.g. proximity to bus/metro station, shops, residential area vs business district.)?

**Note:** The ordering of tasks we mention is **not** mandatory. In other words, if you prefer to start with the exploratory component, and then go to the prediction challenge, this is very acceptable. You can mention that in the report (or invert Sections 2 and 3). Similarly, data analysis might appear after the introduction (if relevant). However, please be aware that a simple descriptive analysis of the data is not sufficient to complete the task. Make sure to go one step forward and try at least one of the techniques discussed in the course.

## Evaluation

The evaluation of the report will be based on the following criteria:

- Clarity - self-explanatory nature of the notebooks
- Thoroughness - Each research question deserves to be explored to the right amount of depth
- Insightfulness - It's important to go beyond the surface of the conclusions
- Technical aspects: Data have been properly analyzed (data cleaning data preparation, data pre-processing).
    - Which model has been used (only one model, multiple models, only linear models, or non-linear models)?
    - Is the model appropriate?
    - Which performance metrics were used (how performances were evaluated)? Were they appropriate?
    - How was the approach benchmarked (how conclusions were drawn)?

- Honesty - While it's fine to use others' code (as a starting point), these shouldn't generally be the actual deliverable **and** the appropriate ethical practice is to **always** reference the source of that code you used.

## Rules

- Each group should consist of 3 to 4 students.
- The submission of the project shall be a zip file with all the notebooks. This zip file should contain the surnames of the group members (for example, for Pablo, Anders, Suarez, and Mila, it should be Pablo_Anders_Mila.zip).
- At the end of the report, there must be a section where **individual contributions are clearly clarified**. In case of doubts on individual contributions or authenticity of the report, the teachers will call the group for an oral defence. This section should **not be part of the page counts.**
- Meeting the deadlines for the milestones is important, including for non-evaluated milestones. A penalty of 10% is given for each extra day of delay

**PLEASE INDICATE NAME, SURNAME, and STUDENT NUMBER IN THE REPORT**

## Report length

The report must be in the form of a jupyter notebook. The structure should be the one described on page 1. There is no overall page limit. However, the project (description of the research questions and results) should not exceed 4 pages. This limit does not apply to figures and codes.

To be more precise, the report can include unlimited figures, and there is no limit to the length of the code. The 4 pages limit only applies to markdown cells. As a reference, you can use this code[2] to make the word count of your markdown cells. One document page is about 500 words (3000 characters including space). The project should be approximately 2000-2500 words. Again, this applies only to markdown. **While this is not a strong constraint**, **excessively long reports will be penalized**.

## Important dates

- October 18 – Announcement of this challenge statement
- October 31 – Communication of group members (through DTU learn)
- December 2 – Final submission – all materials, including report notebook. Submit through DTU learn

[2] https://stackoverflow.com/questions/71194571/word-count-of-markdown-cells-in-jupyter-notebook