# CSC5003 - Project
# Data Exploration and Analysis

The main goal of this project is to use the technologies learned in class to conduct data analysis on a topic of your choice.

For this purpose, you need to choose a dataset (see suggestions below) on which you need to identify an interesting business objective and implement a series of analyses of that data to achieve that objective.

A business objective could be:

- *A decision-making problem.* For example, deciding whether I can use a bike in Paris safely (see http://www.geobythecloud.fr/2014/06/la-carte-du-dimanche-de-lopen-data-qui.html)
- *A prediction problem.* For example, organizing police rounds according to the spatial distribution of past criminology (see data.gouv.fr/fr/organizations/observatoire-national-de-la-delinquance-et-des-reponses-penales-ondrp/)
- *A user study.* We want to better understand user behavior. For example, measuring travel average time and zone of Vélib' usage in Paris (see http://matthieuctvt.github.io/)

More specifically, you will have to:

1. Choose a dataset: Try to find a topic you are interested in and check if there are datasets related to that topic. You can use the list of data sources presented below to find datasets.
2. Choose a business objective. You must define the main problem you are trying to solve. Express it as a general question that you will decompose in small questions.
3. Discover the data: We want to know what the data contains, what it is about, its potential usage, and general metrics such as max, min, or average.
4. Clean up the data: The dataset you chose may be noisy, incomplete, or may contain irrelevant information. Clean it and choose a relevant subset for the chosen objective.

5. Possibly enrich the data with other complementary datasets: You might need additional datasets to complete your analysis. For example, we often need demographic information that we can find on the government website.

6. Build the dataset to be processed: Write a complete pipeline that takes as input one or several datasets and outputs your final dataset on which you will work.

7. Choose the algorithms to apply: Identify the relevant algorithms and methods that will allow you to reach your business objective.

8. Implement and apply the algorithms

9. Interpret the results: The results alone mean nothing. You need to provide an interpretation and explain how they solve your business questions or sub-questions.

10. Build dashboard visualizations explaining the results: A picture is worth a thousand words. Choose relevant visual representations to explain your results (plots, maps, charts, histograms, simple animations, ...). You can use the tools we suggest below.

11. If necessary, re-loop if the results do not meet your expectations

**Work organization:** Groups of 2 or 3 people

**Project Delivery:** You must upload the following materials into Moodle before **January 29th**

- A presentation including a demo (5 mn per person),
- A short report that explains your choices and your results (Pdf)
- Your program's source code

**Data sets sources (non-exhaustive list):** Open Data!

- yelp.com/dataset
- ncdc.noaa.gov
- https://www.kaggle.com/datasets
- opendata.aws
- datasetsearch.research.google.com
- data.unicef.org
- opendata.paris.fr

- data.gouv.fr
- data.gov
- data.worldbank.org
- data.fivethirtyeight.com
- who.int/data/gho
- data.europa.eu

### Tools for Visualization

- https://dash.plotly.com and plotly.com/graphing-libraries/
- https://umap.openstreetmap.fr/fr/
- https://cibotech.github.io/evilplot/