


## Background

Bike sharing represents one of the most promising solutions emerged in recent years to tackle road congestions and air pollution in big cities. The service can also benefit the economy (impact on businesses, reduced expenditure on healthcare) [\[1\]](#) and the personal customers' health [\[2\]](#). Among the European metropolises, London is the one that believes most in the bike sharing service, investing a lot in its development in recent years.

This project stems from a business needs of "Santander cycles", company that has held the monopoly of the bike-sharing service in the city since 2010, which today has more than 11,000 bicycles available in over 130 areas. [\[3\]](#) 

Currently, the service management keeps track of a large number of data regarding bike rentals usage, stored in the company's centrally located databases. The company is aware that data management is not being exploited in an efficient manner, hence the company has expressed the need for a tool for handling and monitoring bike-usage data. The main purpose of the company is to improve the whole system, involving more and more citizens and tourists.

## Scope

The end result of the project is a customised on-premise tool for the company, used by the company managers to solve the previously described problems and improve the quality of the service offered.

The project phases that will be undertaken to achieve the required results are described below:

- Enrich bike-usage data with external data: data concerning the bike-sharing usage is retrieved directly from the company databases and will be pre-processed and cleaned. At this point data consists of hourly bicycles usage among the different areas of the city. Those information is then integrated with external data retrieved from open source services, resulting in an enriched data representation that combines hourly bicycle usage among different areas with additional information like period of the day, period of the year, weather condition and temperature.
- Data visualization to support decision making: the enriched model obtained in the previous point will be analysed and used to produce visualizations (in term of graphs and plots) that will highlight and discover patterns in the data, useful to support decision making (examples of visualizations: bike-shares distribution over the hours, period of the year, weather condition, temperature, wind speed, etc.)
- Organising the arrangement of bicycles at the various stations: the tool will offer the functionality to predict the number of bikes that will be rented in each city area in the short term (up to five days), which will allow the company to better organise the availability of the bicycles at the various bike-stations and satisfy a greater number of customers.

For this purpose, a ML regression model will be built, which will be trained on the basis of past enriched data and able to predict future bike-sharing usage in each area of the city, based on information about the specific day, hour and future weather forecasts.

In summary, the end result will be a tool that will offer the company the following functionalities: display data visualization graphs of past data enriched with additional external information, show the forecast of bicycles rented in each area over the next five days.

The scope of the project does not include the suggestions of actions to be taken in a specific moment of the day/year (e.g. move bicycles between stations), which must be the responsibility of the tool's user.

## Goals

The project aims to create a tool to improve the company's bike-sharing service, providing the functionalities described in the previous section. Once the tool is completed and used by the company, the following goals are identified:

- Increase the annual number of customers of the service
- Increase the annual company revenue
- Decrease the number of times stations do not have bicycles available
- Correctly predict the number of bicycles rented in a given area at a given hour

A Proof Of Concept (PoC) will be developed to demonstrate the feasibility and capability of the project.

## Metrics

The following metrics are evaluated from the moment the tool is adopted by the company

Qualitative metric	Quantitative metric	Evaluate performance
Increase annual number of customer of the service	Increase number of bikes rented in a year by 15%	Monitor the annual number of bikes rented
Increase the annual company revenue	Achieve a 10% increase in revenue with respect to previous year	Monitor the overall revenue of the company
Decrease the number of times stations do not have bicycles available	Reduce the unavailability of bicycles by 20%.	Monitor the number of times stations fail to meet customer demand
Correctly predict the number of bicycles rented in a given area at a given hour	Obtained a ML regression model that predicts future bike-shares with RMSE < 20	Evaluate the ML regression model and calculate the Root Mean Squared Error of the predictions

## Personnel involved

People involved in the project:

- Us: Project Manager, Data Scientist, 2 Software Developers, Software tester, Consultant
- The client: Sales director, Marketing Manager, Data administrator

## Key stakeholders

Client	Santander Cycles
Sponsor	Santander Cycles
Project Manager	Davide Cremonini
Project Team Members	Data Scientist: Iwan Rheon Software Developers: Lena Headey, Ed Skrein Software Tester: Ian Glen Consultant: Diana Rigg

## Project Milestones

Milestone	Date
Start of the project	15 <sup>th</sup> of February 2021
Data visualization of enriched data produced	23 <sup>rd</sup> of March 2021
Predictive model developed and evaluated	7 <sup>th</sup> of April 2021
Customised on-premise tool developed	3 <sup>rd</sup> of May 2021
Project end	11 <sup>th</sup> of May 2021

## Project Budget

This section presents an initial estimate of the budget needed to implement the project: around 75.000€.

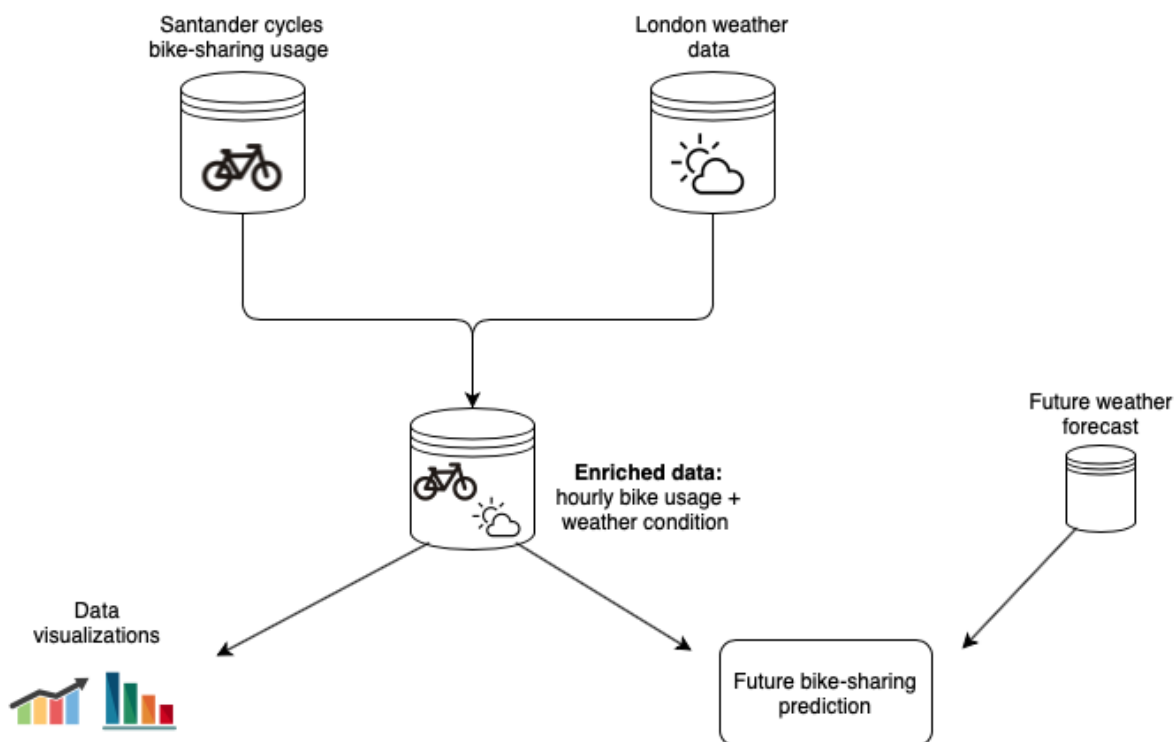
## Data Architecture

For the realization of the project, three different types of data will be used:

- Company's data: data concerning the bike-sharing usage is retrieved directly from the database of the company. The company stores bike-usage information into several csv files of different sizes and length, covering different time ranges. The tool will be responsible for retrieving the data and combining it together, pre-processing and cleaning it. In addition, the system will be updated in real time as the database changes.
- Weather data: data concerning the past weather conditions regarding the city of London will be retrieved from the open source service provided from OpenWeatherMap.org, which grants a free commercial use.
- Future weather data: future weather information will be retrieved, again, from OpenWeatherMap.org, granting a free commercial use.

All three types of data, as well as all software solutions and tools used, come from open source services that grants a free-commercial. However, in order not to have a time limit on requests to the OpenWeather service, a monthly subscription is required.

The general architecture of the project, in terms of the data used and their relationship, is presented below. This architecture will be made available to the company as an independent and functioning on-premise tool.



The PoC implemented and shown to the company will be based on data made available in advance by the company, covering the years from 2012 to 2014.

### Constraints, Assumptions, Risks and Dependencies

Constraints	The implementation of the project is dependent on the company's permission to use their real-time data contained in their databases, via VPN access. For this reason, the project cannot start until such access is guaranteed.
Assumptions	It is assumed that the company will ensure that the data contained in the database are real and always up-to-date . It is assumed that the company remains in continuous contact with the development team. It is assumed that programming tools (Anaconda environment, etc) and external data (OpenWeather) remain of free-commercial use and stable.
Risk and Dependencies	Risk that the implemented ML model does not achieve satisfactory results. Dependence on weather data provided by an external service. Their quality and availability affects the performance of the ML model.

### Approval signatures

---

Project Client

---

Project Sponsor

---

Project Manager