# Introduction

Touring a big city such as the New York City in the United States of America is definitely a must do for any tourist visiting the state from abroad. However, for tourists on a limited time vacation, this can be a little bit stressful given how huge the city is, and the relatively little time to have to make the most out of this experience. Luckily, the field of Data Science is here to make our lives, and most importantly the tourist’s, a lot easier.

In this project, we will mainly assist a hypothetical tourist trying to effectively select the best places to visit given the little time they have. This will save the extensive research and decision-making time especially for indecisive tourists! We will adopt a structured top-down analysis, starting with the selection of the best NYC Borough, to the identification of the top venues of the potential neighborhoods to be selected.

The report will be structured as follows:

First, the topic and the problem are introduced in the introduction section of this report. Second, the data that will be used in this project and the source of the data will be described in the Data section. Third, we will elaborate the methodology adopted in this project in the methodology section. In this section, we will discuss and describe the exploratory data analysis that we did, and what machine learnings we used and why. Fourth, we will present the obtained results, followed by a discussion section. And finally, we will conclude our findings and elaborate on some future works.

# Data

## Data Background

The project extensively relies on the dataset of the New York City made available by the course instructor in the labs (Week 3 of the capstone project course). The Data is first obtained as a JSON file that was later loaded and read into a Pandas Dataframe. Moreover, critical data about the venues in each neighborhood in the NYC will be retrieved from the FourSquare API.

The acquired data will be manipulated via exploratory data analysis tools such as pandas library tools, and will be visualized using matplotlib (for plots) and folium (for maps).

The data is to be cleaned and filtered to acquire only the necessary information and then used and manipulated for a clear data analysis.

This of course will necessitate importing the necessary libraries and dependencies which will be elaborated in the methodology section of this report.

## **Data** Content

The data used in this project and acquired from the available NYC Dataset on the IBM server mainly consist of the different neighborhoods of the city, categorized under 5 main Boroughs, along with their corresponding latitude and longitude. There are 306 neighborhoods in the City of New York, distributed over 5 Boroughs, as we will see in the next sections of this report.

With respect to the data obtained from the Foursquare API, it mainly comprises the different venues in each neighborhood of each borough in the city, along with the venue’s coordinates, name, and category.

The obtained data is extremely rich in content which will allow for an extensive data analysis.