Data Science Capstone Project – Scoping Report

# Comparison of London and Madrid Venues for Tourists

## 1. INTRODUCTION

London and Madrid are two major capital cities in Europe and account for a large proportion of tourism and commerce. It can be challenging to choose between destinations, market research suggests that people would find a comparison tool outlining certain areas and venues would help to select hotels, bookings and plan activities.

Explore London the capital and largest city of both England and the United Kingdom, as well as the largest city within the Europe. Standing on the River Thames in the south-east of England, at the head of its 80 km estuary leading to the North Sea, London has been a major settlement for two millennia.

Explore Madrid the capital and largest city of Spain. The population of the city is roughly 3.3 million with a metro area population of almost 6.5 million. Madrid is best known for its great cultural and artistic heritage, a good example of which is the El Prado Museum. Madrid also boasts some of the liveliest nightlife in the world.- World Tourism Portal

## 2. BUSINESS CHALLENGE & OBJECTIVE

The objective of this study is to provide a statistically robust analysis of Madrid and London zones and to provide an analysis of the venues available within the two cities. By characterising the most numerous sites by zone this will give the tourist / traveller an insight into zones within the city and enable them to decide if it suits their travel desires. Data science methods including clustering will be used to robustly present the differences and opportunities between these two cities

The primary target are tourists and travellers that are seeking information on the two cities venues to help plan their stay and to understand the main differences in venues between London and Madrid. Further uses for this analysis may include:

- Businesses seeking out niches i.e Speciality Restaurants etc
- Students seeking to understand the attractions of host cities
- Individuals looking to work in another city and understand the general venues in certain city areas

## 3. DATA

This section will discuss the sourcing of data, the applications used to process it and a tentative outline for analysis pending on the quality of data once analysed.

The data must provide relevant, unbiased and statistically meaningful data for analysis. After a detailed public domain review, the following was identified:

- List of Zones / Districts / Areas within the Cities of London and Madrid
- Geographic Coordinate Data (Latitude and Longitude in this case) of the Zones and their respective venues
- Venue data containing hotel, restaurant, bar and transport data to provide an insight into neighbourhood demographics

Data Sources

- Wikipedia – Districts of London

- Wikipedia – Districts of Madrid
- FourSquare API

Software

- Anaconda 3.6
- Jupiter Notebook
- Pandas
- ESRI
- Folium

Data Manipulation

It is highly likely that there will need to be some sorting and filtering of Wikipedia and FourSquare data. This will be undertaken using the Pandas data frame functions and other statistical modules. The project will likely follow the following workflow: identify data > clean / process / wrangle data > reference data to geographic coordinates > clean / process combined data and undertake data / statistical analysis. The current plan for statistical processing will be the use of K-means clustering, this will be tested and if an appropriate method used in the analysis. .