**PROJECT PROPOSAL**

* Your goal

L’objectiu principal és poder obtenir la informació principal de qualsevol ciutat (superfície, població, coordenades, ...) per tal d’explotar-la i representar-la de forma fàcil i ràpida.

* The data source and web technology it is based upon

La font que utilitzarem serà Wikipedia. Buscarem l’enllaç que fa referencia a diverses ciutats d’Espanya i extraurem tota la informació que hi ha a la taula de la dreta de la pàgina (tots els enllaços referents a ciutats contenen aquesta taula informativa).

* The approach/technology that you will use for downloading

Per obtenir les dades de Wikipedia, utilitzarem Rstudio i un package anomenat XLS.

* How you plan to clean the data (scraping)

Primer de tot, caldrà esborrar els primers registres que fan referencia a les imatges de la ciutat i de la seva bandera i que, per tant, apareixen buits o com NA’s. També caldrà esborrar totes aquelles columnes que apareixen buides un cop eliminats els registres esmentats anteriorment. Per últim, caldrà validar que s’han llegit bé les dades. Per exemple, en el cas de les coordenades de la ciutat, és llegeix molta més informació de la necessària. Caldrà extreure únicament els dos números referents a les coordenades, ja que és l’única informació rellevant per a nosaltres.

* How you would analyze the extracted information

Intentarem obtenir les coordenades de les ciutats per tal de representar-les en un mapa. A més, a partir de les distàncies entre ciutats (per exemple, per a Barcelona també obtenim informació sobre la distància a Madrid, Bilbao, València, ...) intentarem representar-les en un esquema amb nodes i arcs.