IBM – Coursera

IBM Professional Certificate

Capstone project - Final report

**Estimate price of house based on location and square meter**

Mayur Naik – 2020

Table of content:

[**I.** **Introduction:** 2](#_Toc529925814)

[**II.** **Data description:**](#_Toc529925815)

[**III.** **Methodology:**](#_Toc529925816)

[**IV.** **Results:**](#_Toc529925820)

[**V.** **Discussion:**](#_Toc529925821)

[**VI.** **Conclusion:**](#_Toc529925822)

[**References:**](#_Toc529925823)

[**Table of Figures:**](#_Toc529925824)

1. **Introduction:**

This report will give an overview on Data Science Capstone project which is a part of IBM professional Certificate hosted on Coursera platform. The main goal of the project to estimate the cost of the house based on the location and square meter values.

The target audience for this project can be a person who is willing to buy the house or a bank who will be giving mortgage to an individual who is planning to buy the house in Dublin City, Ireland. The main idea behind the project is to cluster some of the properties that are found on the daft site based on the location, Once we get the clusters of the properties we will find out the price per square meter for each of the clusters. This price per square meters will than be used to estimate the price of the house on unseen data provided the location and sqm value.

This project will not only be used to get the rough estimations of prices but will also help us to understand similar properties in Dublin city, Ireland based on location and accessibility. This will help a future buyer or investor understand which locations are having similar properties.

1. **Data Description:**

The City of Dublin, Ireland is chosen as the observation target due to following reason:

* There are lots of property prices available on Daft which can be easily scrapped.
* Property prices and property sales in Dublin are rapidly increasing day by day.
* The avaiblity of the geo data which can be used to visualize the dataset onto a map.

The dataset will be composed of following two main data sources:

* Daft Ireland: <https://www.daft.ie/> which will provide the details such as property price, square meter value, latitude, longitude, Eircode and property address
* FourSquare API which provides the surrounding venues of a given coordinates

The process of collecting the data is as follows:

* Scrape the Daft Ireland website based on the location. As daft contains around 17k properties in Dublin we will limit the scraper to collect the data for only 2500 properties which is the maximum limit of our Foursquare API.
* For each properties, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
* Count the occurrence of each venue type in a given location. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
* Identify and remove the outlies if any.