CPS3235 Data Science: From Data to Knowledge

Study Unit Assignment

Marc Ferriggi (286397M)

Lecturer: Dr. Jean-Paul Ebejer

L-Università
ta' Malta

Department of Computer Science
University of Malta
January 2019

Contents

1 Data Storage - Computer Science Papers Dataset

2 Data Extraction and Visualization - EU Stats

3 Data Science Project - Dataset Analysis

3.1 The Data Supplied

The data supplied was taken directly from *Times of Malta* classified listing page on some of the days when the listings would go live from the 23rd of April 2015 to the 24th of October 2018. For mostly every week in that time frame, the data was either collected on a Monday or on a Wednesday or on both those days of the week. Each file contains the property listings for a whole week, thus there are duplicate entries in the dataset. Note that due to the differences in the time frame between each data point, if this data is going to be used for time series analysis then a sample needs to be taken with even time intervals to ensure accuracy of the tests.

The data collected is simply the *html* code used to display the web page. This is available to anyone that has access to the website, thus anyone can have access to this data given the page was accessed on the specified date. Since anyone can book a classified advert on *Times of Malta*, there is no guarantee that the information is correct or accurate, thus we cannot be certain that the prices of the properties listed on the classified page reflect the true market value of property in Malta set by experts in the field. However, this could still give us a good indication of the trend.

With regards to data quality, the data provided is not complete as not all listings have information on the area of the house, whether or not the house has a garage or information on the type of property. However, from glancing at the raw data, it would seem that all entries have data on the location and price of the property. Since the data is extracted from the same system it must be consistent. Finally, when doing a study on the current property situation in Malta, data from 2015, 2016 or even 2017 will not be relevant or timely as the property marked has changed drastically in these past years in Malta.

3.2 Features of Interest

The features of interest from the raw data were extracted with the task in mind of building a predictive model that would predict the expected price of the property given several features. Thus a sample was chosen from the provided dataset, in particular data from the beginning of August, September and October was chosen since these would reflect most accurately the current prices in the property market.

- Property ID: This will be extracted from the *name* variable in the *html* code in order to have a unique ID referencing each entry.

- Location: This categorical variable will store the location of the property. Since the price of property depends on the location, this would make a good predictor for the model.

- Property Type: Another categorical variable which lists the type of property for sale. The categories are the following: house, penthouse, maisonette, apartment, farmhouse, villa, house of character, block or unknown.

- Plot Area: This continuous quantitative variable stores the area of the land in square meters (thus its measured in a ratio scale). The plot area of the land should have a direct effect on the final price of the property, thus it should make a good predictor.

- Has Pool: This dichotomous variable takes a value of 1 if the property listed has a pool and 0 otherwise.

- Has Garage: This categorical variable has 3 categories; yes (1), no (0) and optional (2).

3.3 Feature Extraction

The data was extracted from the provided *html* files by using the *Python* package *Beautiful Soup*. Once the property listings were found, the features were extracted using regular expressions or python's string manipulation libraries. The data in the location category was then arranged to ensure there's only 1 category per location. Finally, the duplicated entries were dropped, leaving a total of 1422 unique entries.