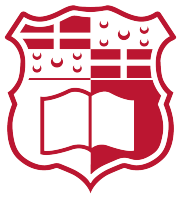

CPS3235 Data Science: From Data to Knowledge

Study-Unit Assignment

MARC FERRIGGI

Supervised by Dr Jean Paul Ebejer

Department of Computer Science

Faculty of ICT

University of Malta

January, 2019

A assignment submitted in partial fulfilment of the requirements for the degree of B.Sc. (hons.) Conputer Science AND Statistics and Operations Research.

Statement of Originality

I, the undersigned, declare that this is my own work unless where otherwise acknowledged and referenced.

Candidate Marc Ferriggi

Signed _____

Date January 9, 2019

Contents

1 Introduction **1**

2 Data Storage - Computer Science Papers Dataset **2**

3 Data Extraction and Visualization - EU Stats **3**

4 Data Science Project - Dataset Analysis **4**

4.1 The Data Supplied 4

4.2 Features of Interest 5

4.3 Feature Extraction 5

4.4 Some Interesting Results 6

4.5 Identifying a Correlation 7

4.6 Statistical Analysis 8

5 Conclusion **9**

List of Figures

4.1 Pools 6
4.2 Top Localities 6
4.3 Price Densities 7
4.4 Correlation Plot 7

List of Tables

1

Introduction

2

Data Storage - Computer Science Papers Dataset

3

Data Extraction and Visualization - EU Stats

4

Data Science Project - Dataset Analysis

4.1 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.

4.2 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).

4.3 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 and 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.

4.4 Some Interesting Results

Following some analysis on the extracted data, some interesting statements were initially made. Note that at this stage no statistical tests were done as this would be done later on in the analysis.

The first interesting conclusion was made on data to do with properties having pools. Firstly, the average value of properties with pools seems to be significantly larger than those without pools (as expected). In fact the average value of a property with a pool (regardless of location or size) was found to be €818,616.07 while the average value of a property without a pool was found to be €342,582. Furthermore, as can be seen in Figure 4.1, at the time of the study, the locality selling the most properties with pools was Zejtun.

Another interesting result found during the data exploration stage was that currently on the market, Sliema is the locality that has the larges number of properties for sale. The top 10 localities selling properties can be seen in Figure 4.2.

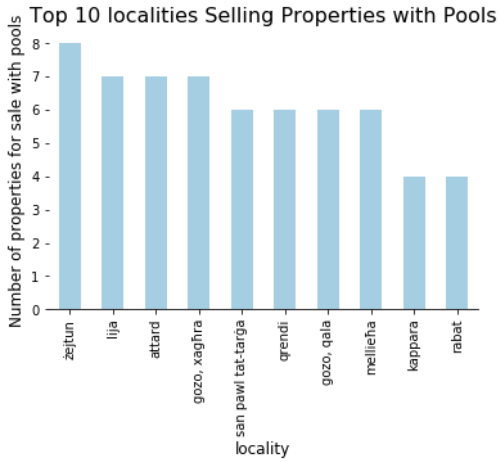

Figure 4.1: Localities Selling the most amount of Pools

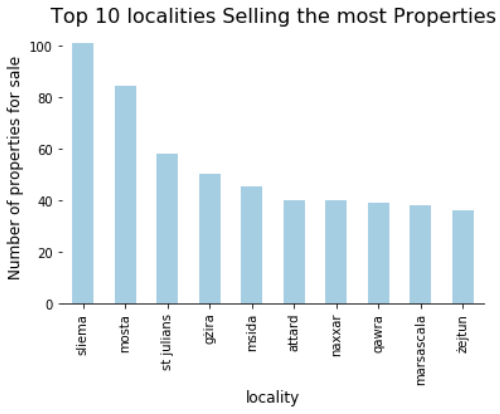

Figure 4.2: Localities Selling the most Properties

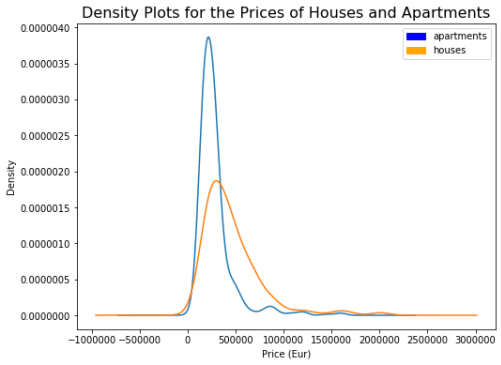

Figure 4.3: Density plots for the Prices of Houses and Apartments

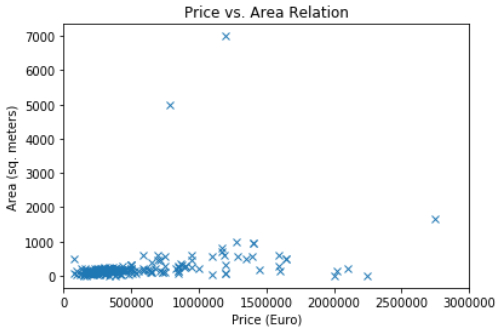

Figure 4.4: Scatter Plot of Price vs. Area

One of the questions that would be interesting to look at was *"which locality has the most expensive land?"*. In order to answer this, the price per square meter of land for each property on the market was worked out by dividing the price by the area. The average was then worked out for each locality using the *groupby* function and it was concluded that the most expensive land on the market right now is in Pender Gardens with an average of €21,818.18 per square meter! Note that at this stage, properties with less than 35 square meters of land were removed as they were considered to be outliers.

The last interesting conclusion was made when analysing the property types. In general, as expected, the most expensive property type for sale was found to be the block of apartments which have an average price of €1,414,214. However it was interesting to note from looking at the density plots for the prices of apartments and houses (Figure 4.3) that prices for houses have greater variability than those of apartments.

4.5 Identifying a Correlation

The two variables that would seem to have an obvious correlation would be *price* and *area*. As the area of the land increases we would expect the price to increase too. In fact, having a quick look at the scatter plot in Figure 4.4 (after limiting the x-axis to remove some outliers) seems to confirm this. The outliers are points that are affected by the locality, it's obviously important not to remove these when building the model as the locality will be a key feature in the model.

A Pearson correlation test was then run and the data was found to have a correlation coefficient of 0.4861. The p-value of the test was found to be 1.84252e-15 which is less

than 0.01, thus we are 99% confident that the value is significant.

4.6 Statistical Analysis

Initially, tests for normality were done on the continuous variables, these tests are important as they would decide weather to use parametric tests or non-parametric tests in the future, as well as decide what kind of model to build to predict the prices of properties. In order to ensure the accuracy of the normality tests, two different tests were run; the Shapiro-Wilk normality test and the D'Agostino-Pearson omnibus test. Both these tests test the null hypothesis that the data is sampled from a normal distribution. Both these tests on both the continuous variables gave us the result that our data does not in fact come from an underlying normal distribution.

After running the normality tests, a test to see if there's a significant difference in the area of the property if it has a garage was run. It was decided to use the Mann Whitney-U Test with continuity correction as this is a non-parametric alternative to the independent sample t-test which tests for a significant difference in the medians of two independent groups. The test results showed with 99% confidence that as expected, the average area of properties with a garage was significantly larger to those without a garage. Note that more details on the test can be seen in the markdown cells in the JuPyter notebook file '03 - Statistical Analysis'.

5

Conclusion