**University of Malta**

**PLAS Tech Units**

**LAS3004 Data X - An Introduction to Data Science: Storage, Visualization and Analysis**

**Lecturer: Dr Jean Paul Ebejer**

**Name: George Bonanno**

**ID: 320086M**

**Date of Submission: 23/01/2016**

Contents

[Question 1: Visualization 2](#__RefHeading__1948_36555078)

[Dependecy Note: 2](#__RefHeading__1950_36555078)

[A. Visualizations 3](#__RefHeading__1952_36555078)

[A.1 Kingbase chess games visualizations 3](#__RefHeading__1954_36555078)

[B. Bad Visualization 8](#__RefHeading__1956_36555078)

[Question 2: Data Science 10](#__RefHeading__1958_36555078)

[A) Data Supplied 10](#__RefHeading__1960_36555078)

[B) Features of Interest 12](#__RefHeading__1962_36555078)

[C) Visualizations 14](#__RefHeading__1964_36555078)

[C.1. Property type price Count Distribution 14](#__RefHeading__1966_36555078)

[C.2. Property type area count Distribution 15](#__RefHeading__1968_36555078)

[C.2. Property type – price Box Plot 16](#__RefHeading__1970_36555078)

[C.3. Boxplot of Property Type vs. Area 16](#__RefHeading__1972_36555078)

[C.4. Most popular locations 17](#__RefHeading__1974_36555078)

[C.5 Geo-visualisation of the mean price 18](#__RefHeading__1976_36555078)

[C6. Advertisment count on Sundays and other days 18](#__RefHeading__1978_36555078)

[C.6. Mean Price on Advertising Days 19](#__RefHeading__1980_36555078)

[D) Statements 20](#__RefHeading__1982_36555078)

[E) Statistical Analysis 21](#__RefHeading__1984_36555078)

[F) Predictive Model 23](#__RefHeading__1986_36555078)

# Question 1: Visualization

## Dependency Note:

The R scripts mentioned in this section make use of the following libraries:

* RSQLite
* stringr
* RcolorBrewer
* ggplot2

These packages can be installed using the following statement:

install.packages(c("RSQLite", "stringr”, ”RcolorBrewer",”ggplot2”))

## 1. Visualizations

### 1.1 Kingbase chess games visualizations

The Kingbase chess data set available on (http://www.kingbase-chess.net/) was explored through the use of a number of visualisations. The visualisations helped in getting an understanding of:

* the distribution of result outcomes (i.e. number of white wins, black wins and draws)
* the variation of the number of moves required to complete the game with time.
* The most common starting move of the winner.
* The number of games played each year

The size of the data set was larger than 1GB. This implied that the files could not be loaded and processed at one go in memory. A buffer of 10000 lines was hence iteratively used to read the file contents. The string inside this buffer (concatenated with any unparsed string left from the previous buffer) was parsed using the pgn format specification found in [1]. It was assumed that the files obeyed the pgn file format. A number of features were extracted from each parsed game, namely:

* event name
* date of game
* site
* result
* first move of the winner

Each of these features were stored in an sqllite database (question1/db/chess.db). This file was not included since its size exceeded 120MB. This database was then queried to extract the necessary information that allowed the following visualisations to be created. Given that the \*.pgn files are in question1 directory, the sqllite database can be populated again with this data by running the following statement in a bash shell. This will however take a considerable amount of time. It is advised to skip the execution of this statement and continue with the next steps to retrieve the visualisations.

cd question1

./loadChessData.sh \*.pgn

The required data for this visualisation **has been already extracted** and stored in question1/.Rdata folder. The number of games loaded from the pgn files amount to 1861460 rows. The visualisations described below can be generated by running the chessVisualisation.R script in an interactive R shell:

source(chessVisualisation.R)

Running the following functions will generate the visualisation that describe aspects of the loaded chess games:

plotYearlyGames()

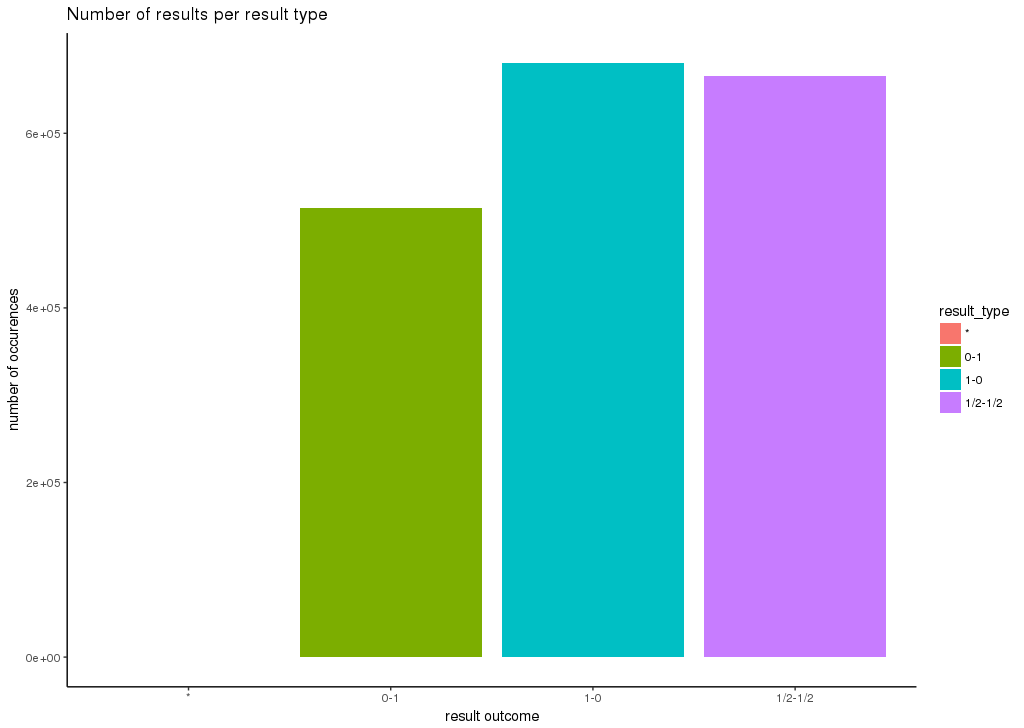
winningMovesHeatMap()

moveInGameBoxPlot()

resultCountBarPlot()

The visualisations generated by these function will be discussed in more detail.

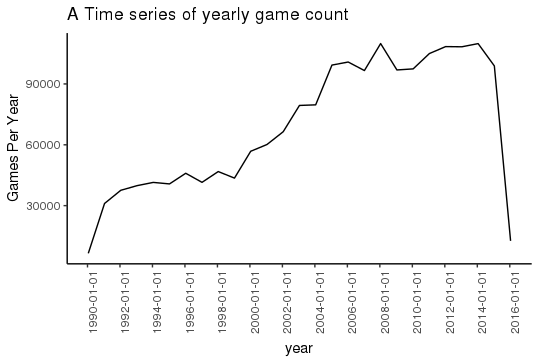
#### 1.1.2 Distribution of Game Outcomes

Chart 1: Distribution of game outcomes

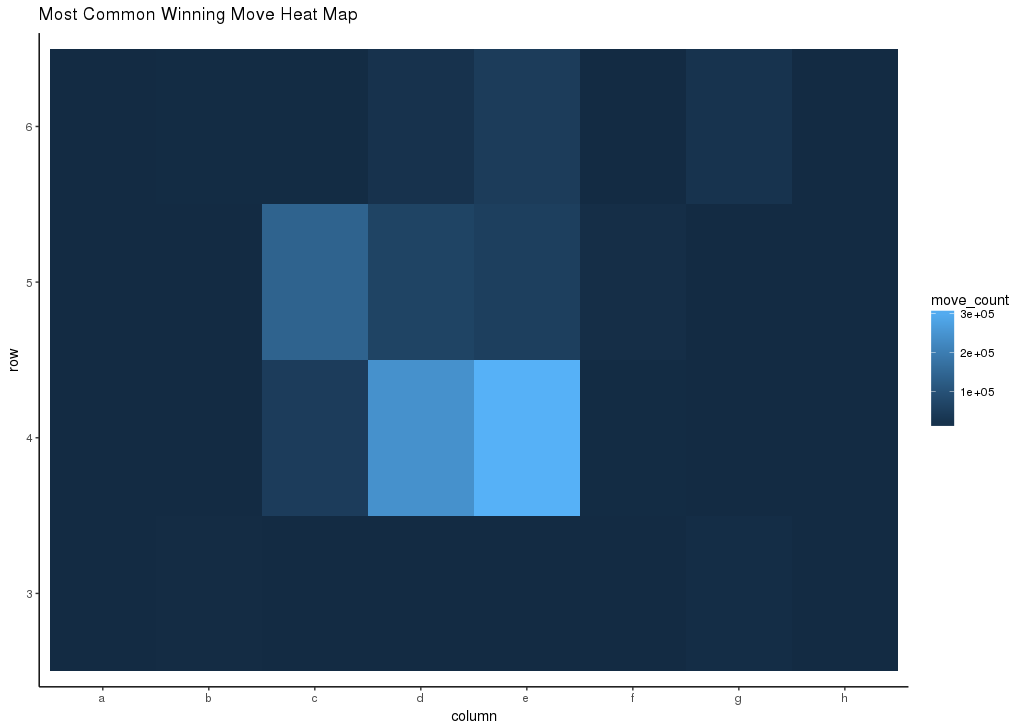
The outcome of the 1861460 games is depicted in chart 1. One can note that the number of of white wins and draws is considerably higher than black wins. This may suggest that a white player might have a better chance of winning the game. The number of games whose outcome is described with (\*) is negligible when compared with the counts of the other result outcomes.

#### 1.1.2 Number of yearly Games

The processed kingston data set contains chess games ranging from 1990 to 2016 (ie. 2016/02). As seen in chart 2, one can note that the number of chess games increases to over 90,000 games per year between 2006 to 2014. The number of yearly games decreases drastically in 2015 and 2016.

Chart 2: Number of yearly games

#### 1.1.3 Heatmap showing the starting moves of winners

Chart 3: Most common starting move of winners

The above heatmap allows the viewer to observe the most common starting moves that led the player to win the game. The heat map shows that the most common starting position is e4, followed by c5, d4, c4 and d5. The most common move is consistent with the previously discussed fact that white wins are higher than black wins since only the white player can start with c4.

#### 1.1.4 Number of game moves box plot

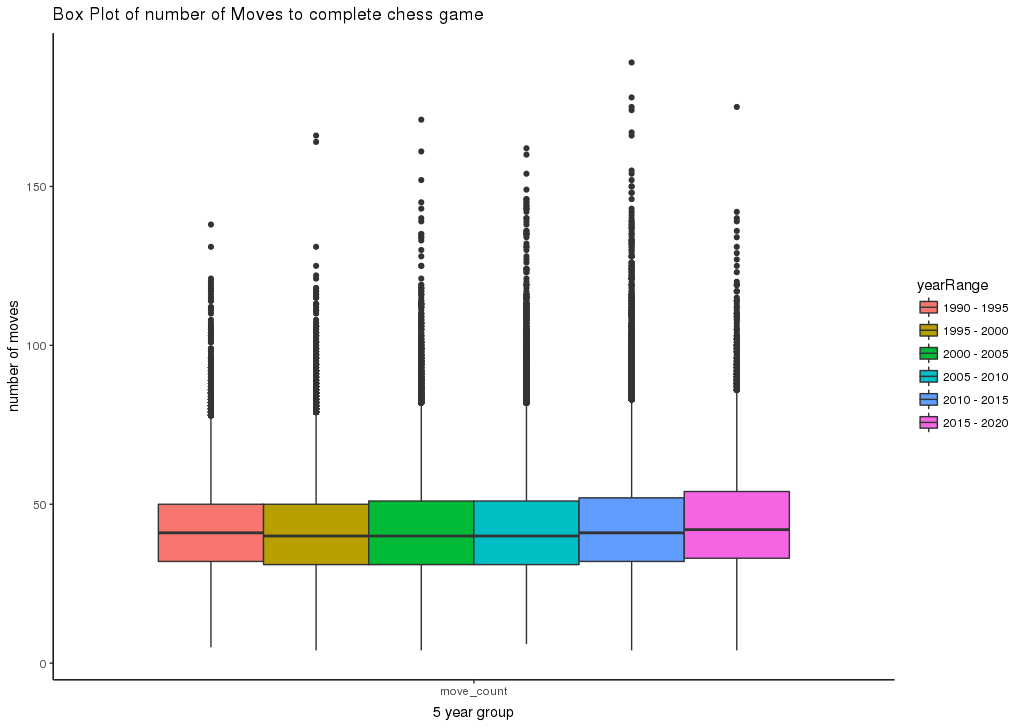
Chart 4: Number of Game Moves box plot

Chart 4 shows a box plot of the number of moves found in chess games grouped every 5 years. One can note that the mean and interquartile range are almost equal for the 5 year groups. Another interesting fact is that the outliers consists of games that have more than 75 moves.

## B. Bad Visualization

The following visualisation (Illustration 5) was taken from a *Times of Malta* Article 'Malta ranked second safest place in the world for natural disasters' that appeared on 4th June 2016 [1]. This article used a pie chart to view the natural risk chance of Malta when compared with the 3 three highest natural disaster risk considered countries.

This visualisation does not help in identifying whether Philipines and Vanuatu's natural disaster risk is higher since their size is almost the same. The chart does not attach the numerical value (such as percentage risk) to the slices. Furthermore, the slices are not ordered by size, making it diffucult for the reader to compare.

This visualisation was re-implemented in R as a bar plot in Chart 6. One can note that is is easier to compare the risk percentages for the Philipines and Vanuatu. The risk percentage can be determined easily for each graph. It is much easier to note Malta's low risk when compared to the other countries since a comparison by height is easier than a comparison by angle. This visualisation was generated by the script extractPieChartInfo.R.

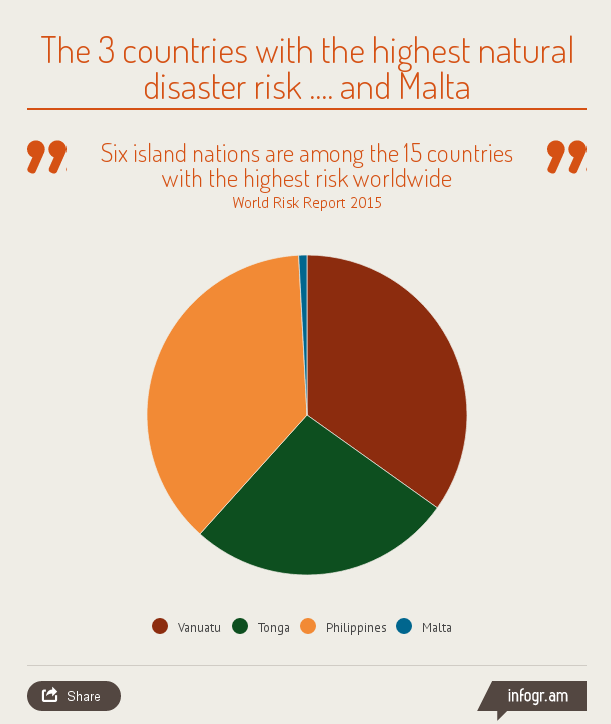
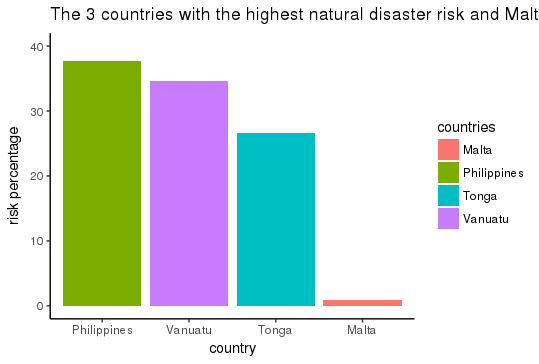
Chart 5: A 'terrible' visualisation

Chart 6: Re-implementation of the 'terrible' visualisation

# Question 2: Data Science

## 2.1 Data Supplied

The supplied text files contains the html markup of the Times of Malta property advertisments published between 2015-04-16 to 2016-11-22. A word frequency analysis, in fact, shows that the most frequently occuring words are 'property' and 'sale'. These advertisments feature from 539 different editions. The date range for which no advertistment was found on those days can be found in Table 1. One can note that the largest date range that for which no advertisments feature is 6 days. This information was extracted using advertismentsDates.R.

The number of advertisments on Sundays is much higher than the number of advertisments on other days. The number of such adverstiments on Sunday ranges from 200 to 600. A typical week day has less than 40 advertisments. This will be discussed in more detail in the coming sections.

|  |  |  |
| --- | --- | --- |
| Date from | Date To | Date Difference |
| 2015-04-29 | 2015-05-05 | 6 |
| 2016-06-29 | 2016-07-04 | 5 |
| 2016-11-03 | 2016-11-08 | 5 |
| 2015-10-01 | 2015-10-05 | 4 |
| 2016-02-25 | 2016-02-29 | 4 |
| 2016-08-04 | 2016-08-08 | 4 |
| 2016-10-30 | 2016-11-02 | 3 |
| 2015-07-21 | 2015-07-24 | 3 |
| 2015-10-30 | 2015-11-02 | 3 |
| 2015-12-24 | 2015-12-27 | 3 |
| 2016-03-04 | 2016-03-07 | 3 |
| 2016-07-11 | 2016-07-14 | 3 |
| 2015-05-26 | 2015-05-28 | 2 |
| 2015-08-07 | 2015-08-09 | 2 |
| 2015-10-07 | 2015-10-09 | 2 |
| 2016-01-04 | 2016-01-06 | 2 |
| 2016-03-24 | 2016-03-26 | 2 |
| 2016-05-16 | 2016-05-18 | 2 |
| 2016-06-02 | 2016-06-04 | 2 |
| 2016-06-11 | 2016-06-13 | 2 |
| 2016-06-13 | 2016-06-15 | 2 |
| 2016-06-20 | 2016-06-22 | 2 |
| 2016-06-27 | 2016-06-29 | 2 |
| 2016-08-21 | 2016-08-23 | 2 |
| 2016-08-30 | 2016-09-01 | 2 |
| 2016-09-15 | 2016-09-17 | 2 |
| 2015-10-25 | 2015-10-26 | 1 |

Table 1: The range of days for which no advertisment was found.

A closer look at the content shows that most of the advertisments have the same structure:

* start with the locality of the property. Apart from a handful of properties found in France or Sicily (in Ragusa), most properties are located on the Maltese islands with the exception of few places. However a high number of Maltese localities were spelt incorrectly. For example, "GĦOCHARGĦOCHUR" rather than "GĦARGUR" or "BIROCHŻEBBUĠA" rather than "BIRŻEBBUĠA". In some cases, the locality was less specific than in other adverts. For example, cottonera was used in certain adverts. On the other hand, other adverts specifically mentioned a city in Cottonera (e.g. Vittoriosa).
* mention the property type (e.g. Apartment, villa, maisonette)
* have one (or more) contact numbers
* a price given in Euro. Sometimes, the price is expressed per square meter. Large values were expressed as kilo (e.g 20K) or as million (e.g 2.5m). Occasionally, the price was incorrect due to a typing error (250,000 instead of 250m000 ).
* Area. This is expressed normally in square meters. However, in the case of fields or plots the area can be also found expressed in hectares (ha) or tumoli (this for property located in Gozo). Occasionally, the area is expressed by length and breadth.
* A sea view or other advantage that this property has due to its position.
* Other property features (such as pool and garages).

There are some advertisments placed by developers or real estate brokers (such as '*PROPERTIES for sale on www'* or '*PRICE REDUCTIONS this week on properties*'. In general, such advertisments do not contain the properties mentioned previously such as locality, area or even price.

## 2.2. Features of Interest

As previously mentioned, most of the adverts describe the property in terms of its **location**, **area**, price as well as the **type of property** (such as apartment, maisonette or villa). These are in fact the features of interest that capture the reader's attention. The other features (whether the property has a garage, pool, located by the sea) strike first the attention of a smaller subset of readers (such as wealthier people or people looking for a second residence). For this reason, the following attributes were selected from each advert.

* **Advert Date**: The date when the property was advertised.
* **Contact Number:** The contact number was extracted since it allows the particular property to be identified when compared to other properties. In this way, an advertisment of an 20K apartment of Sliema with contact number 123 would be considered different from an advertisment of a 20K Sliema apartment whose contact number is 456. The absence of a contact number would make these adverts equivalent and considered as duplicates.
* **Area:** The area in square meters (if it exists). Regular expression were used to express the area that is possibly expressed in square meters, hectares or tumoli. The value are then converted accordingly.
* **Price:** The property's price is extracted in euros. Any prices expressed as millions or thousands (ie. 'm' or 'k')
* **Property Type**: The type of property advertised such as flat, houses and apartments. The possible types were determined after extracting the most common nouns from the entire corpus and seeing which of these describe a property type. Adverts that do not have any of these nouns are removed.
* **Locality:** The locality is extracted from the first phrase of the advertisment. The possible correct locations were identified and placed in a text file (places.txt). The extracted location was first passed through a location mapper that attempts to correct (or generalise) the location. For example, the locations 'Birgu' and 'Vittoriosa' would be mapped to Cottonera. All localities in Gozo were marked as 'Gozo'. On the other hand, 'Tower Road' is mapped to 'Sliema'. If this mapping failed, the levenshtein distance of every location in places.txt is computed with the extact location. The word that has the least distance is then selected. Foreign locations (such as the previously mentioned Sicily) are removed since only adverts for local properties will be analysed.

The extracted features are then placed in a csv file. This csv file can contain duplicates since the same property could be advertised in more than one edition. Filtering only unique csv files through a bash shell allowed duplicate adverts to be removed without any effort:

cat extracted\_features.csv | sort | uniq

The feature extraction, data cleaning and duplicate removal just mentioned above is acheived by executing the following script. The R script invoked from this scripts reads the html text files from under a data folder found in the same directory.

./extractPropertiesToCsv.sh

This will generate three csv files:

* extracted\_features.csv: the csv file with the extracted features of the adverts. This is the source file from where the next 2 files will be generated.
* extracted\_with\_date\_unique.csv: the unique adverts with the advert date.
* unique\_features.csv: the unique adverts without the advert date.

The last 2 files allow us to extract information related both to what has been advertised over time as well as the unique adverts. The visualisation, statistical analysis and model prediction will be built using data in these csv files.

The extracted adverts in the generated csv files contain missing information. There are instances were at least one attribute is missing. All adverts with a missing location are removed since it is difficult to infer the location from the other attributes at this stage of this data science project. Furthermore, the adverts with more than one missing attributes are removed due to high uncertainty introduced when determing more than one unknown attributes based on the other features.

Adverts without a property type were assigned the 'apartment' attribute. The following imputation decisions were taken based on the assumption that the locality is the attribute that mostly affects the price type. It was further assumed that whilst the area varies with the vary property type, it does not vary with the locality (due to the small size of the Maltese islands). Considering these assumption, the following imputations are done:

* Missing prices are then imputed by taking the median price of the adverts of properties in the same locality.
* Missing areas were assigned the median area of the property type. Any property types with a missing median area (i.e. there is no advert with the property type that has a known area) is removed.

Following the imputation, box plots of the prices and areas for a given property type was visualised. The outliers were found were checked to see if there was an error in the data extraction. Some values needed to be corrected (such as in the case of the previously described 250m000 price. This was initially resolved by the extraction process to 250€). The outliers found after correcting the feature extraction process were found to be in these ranges:

* Prices less than 3000€ and greater than 3 million euros.
* Areas greater than 500 square meters.

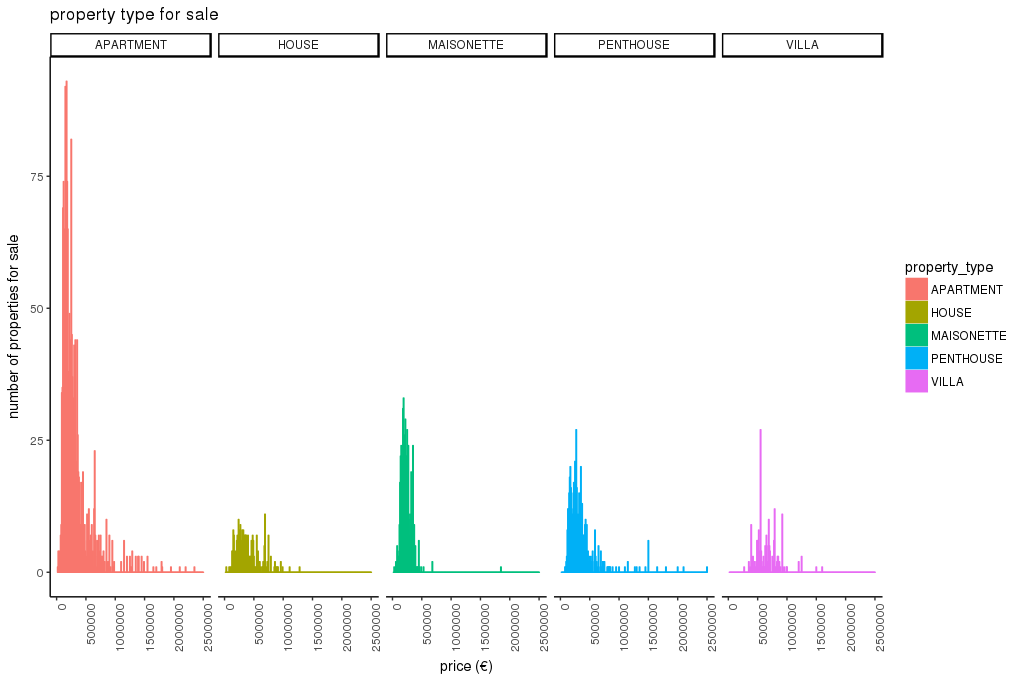
Adverts whose prices were in these ranges were removed based on these decisions.

The outlier removal, error correction and imputations were implemented in csvProcessing.R.

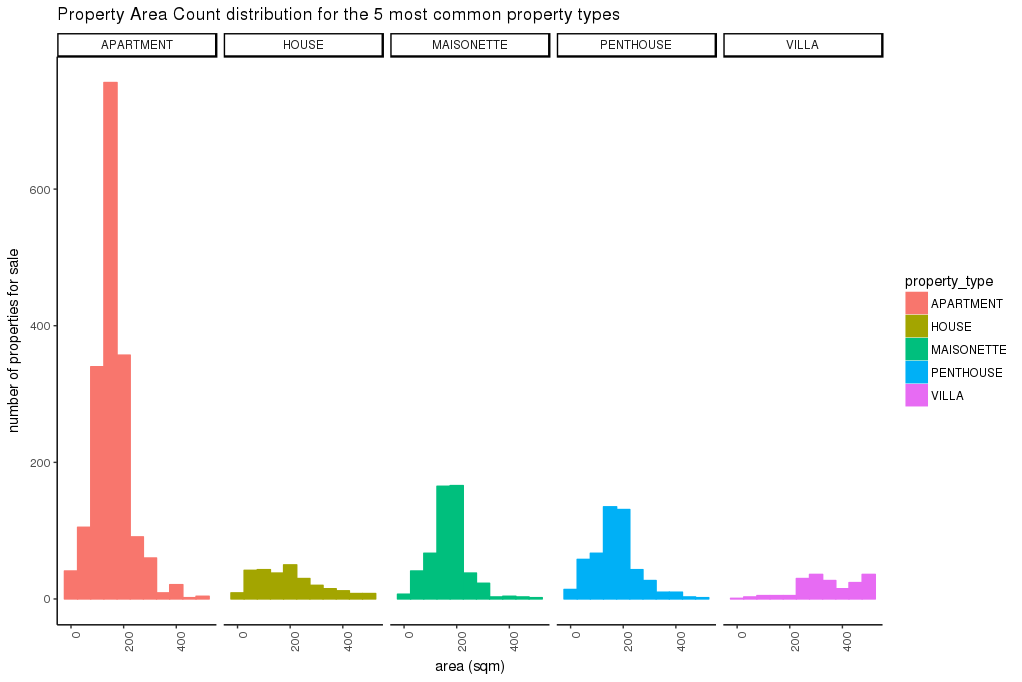
## 2.3. Visualizations

### 2.3.1. Property type price Count Distribution

The five most common property types are apartments, houses, maisonettes, penthouses and villas. The price - count distribution for these prices types can be found in Chart 7. All distributions are skewed to the left. The villa distribution is the less skewed than the others, suggesting that in general, villas are more expensive than the other property types.

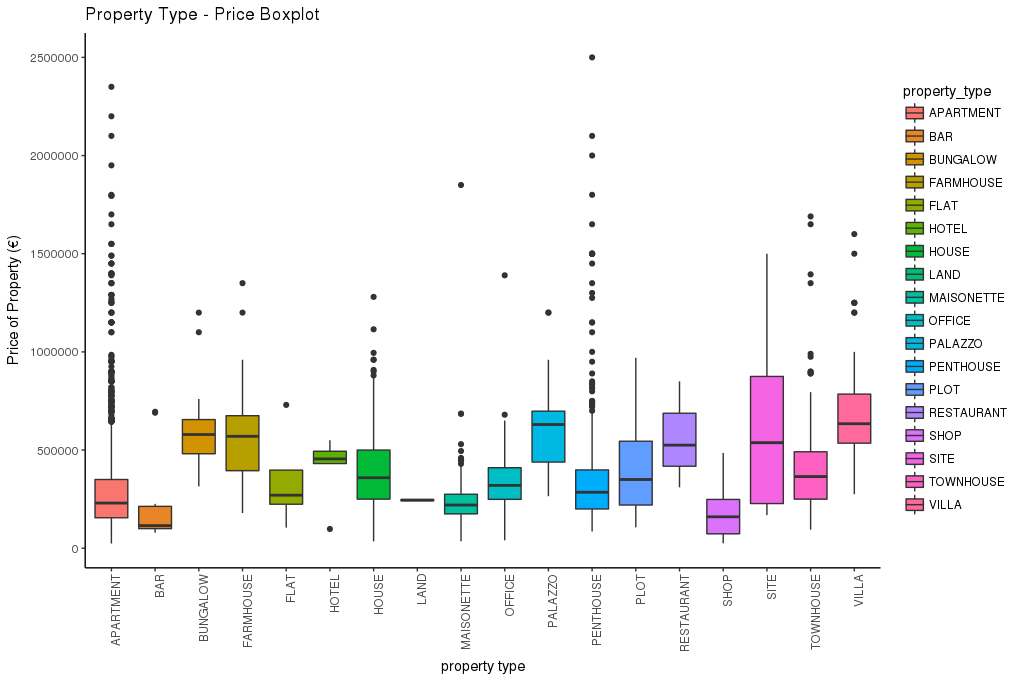
Chart 7: Price-count distribution for the 5 most common property types

### 2.3.2. Property type area count Distribution

Chart 8: Property Area Count distribution for the 5 most common property types

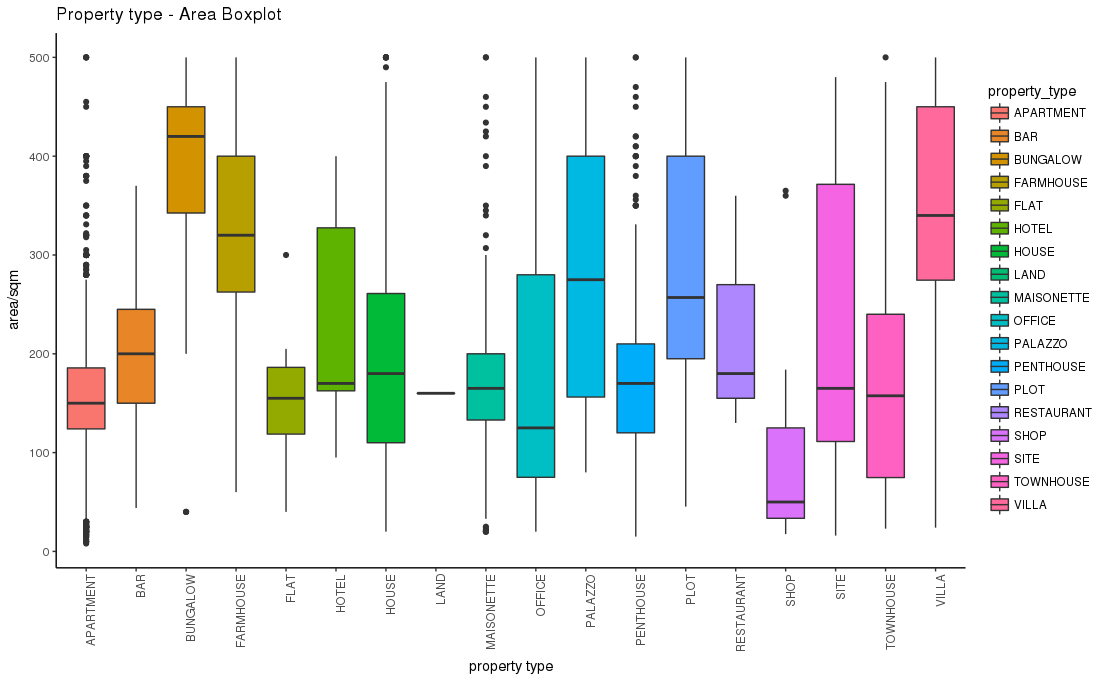
### 2.3.3. Property type – price Box Plot

Chart 9 depicts a property type – price box plot. This box plot helps to compare the mean, lower and upper quartile prices of the different types of property as well as the number of adverts that do not fall within these quartiles.

Chart 9: Property type-Price Box Plot

### 2.3.4. Boxplot of Property Type vs. Area

Following the previous box, a box plot for the area of different property types is shown in chart 10.

Chart 10: A property - area box plot

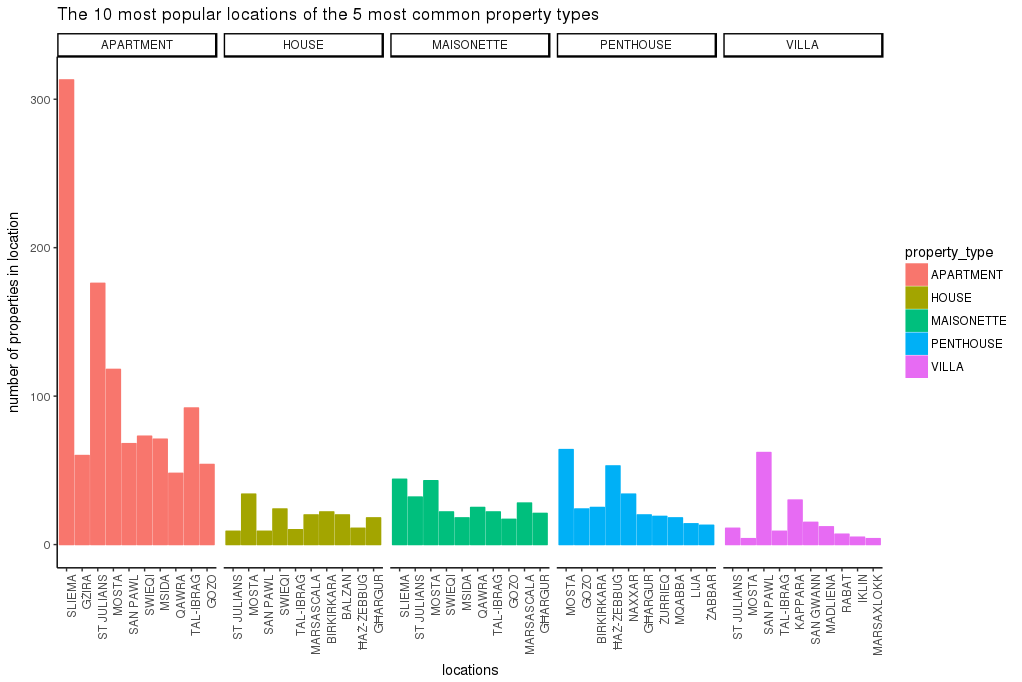
### 

### 

### 

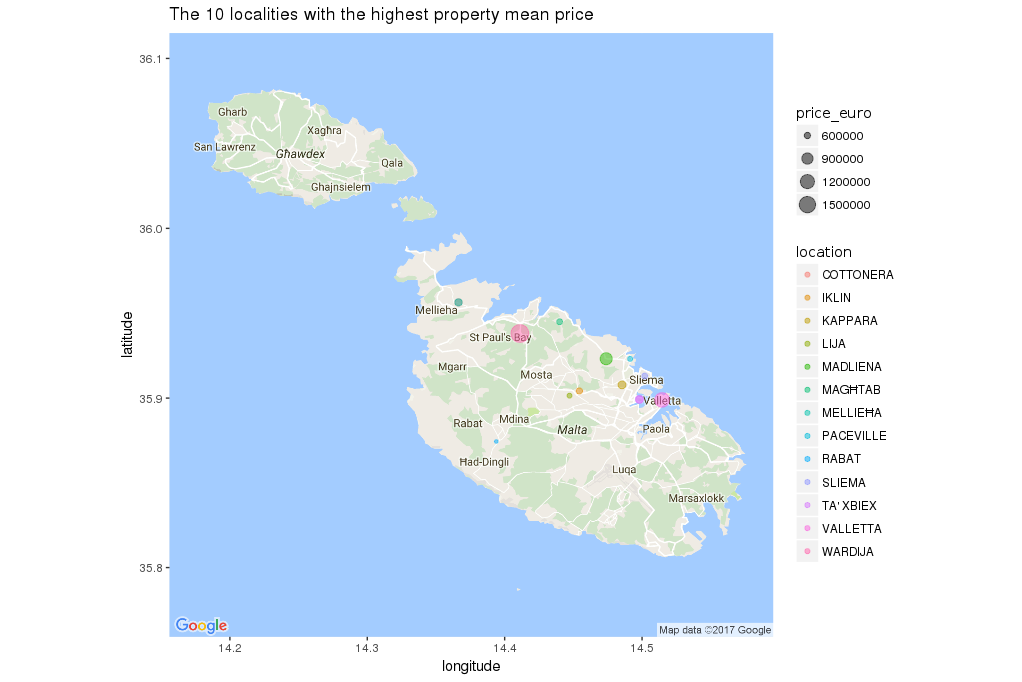
### **2.3.5. Most popular locations**

Chart 11 shows the 10 most popular locations for the 5 most common property types.

Chart 11: The 10 most common popular location of the 5 most common property types

### 2.3.6. Geo-visualisation of the mean price

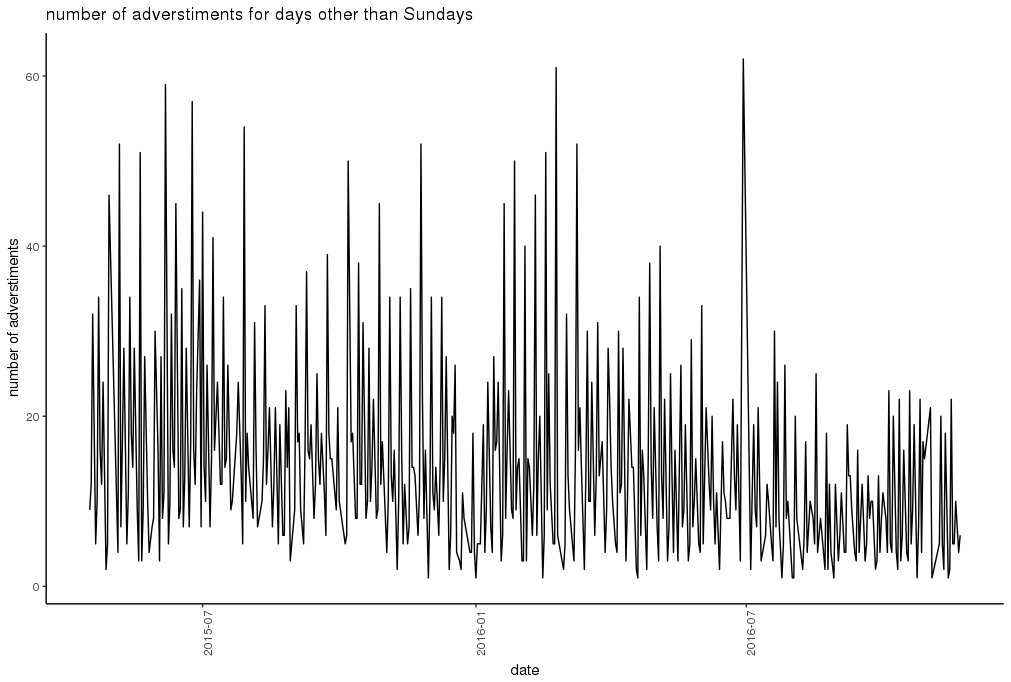
Chart 12 shows a geo-visualisation of the mean price of the ten most common locations. The size of the point marking locality is proportional to the mean price. One can note how the mean price varies with location.

Chart 12: The 10 localities with the highest property mean price

### 

### 2.3.7. Advertisment count on Sundays and other days

Chart 13 shows the number of adveristments for different property types on Sundays. Chart 14 shows the total number of advertisments for days other than Sundays. The latter does not show the counts of the individual property types since this visualisation resulted very cluttered and diffucult to read. As mentioned in section A, one can note how the number of advertisments for different property types is very consistent. On the other hand, the number of advertisments on other days varies between 20 and 40, occasionally peaking up to 60.

Chart 14: Number of Advertisments on non-Sundays

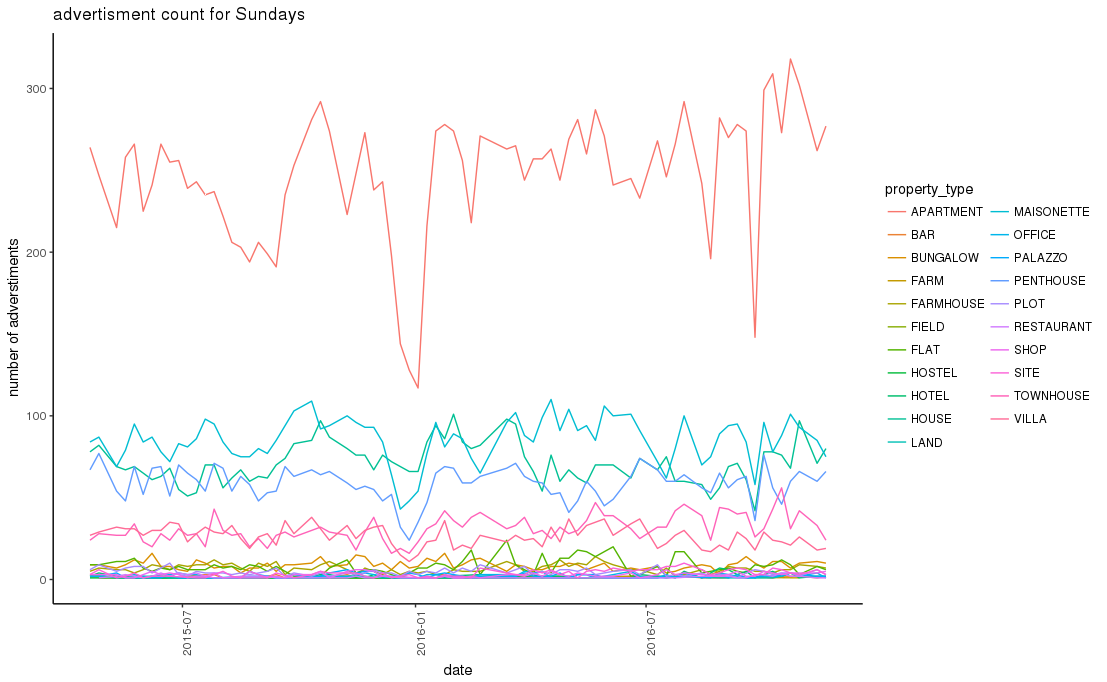


Chart 13: Number of advertisments on Sundays

### 

### 2.3.8. Mean Price on Advertising Days

The mean price on every advertisment day is shown in chart 15. One can note the occasional peak, attributed to the outliers.

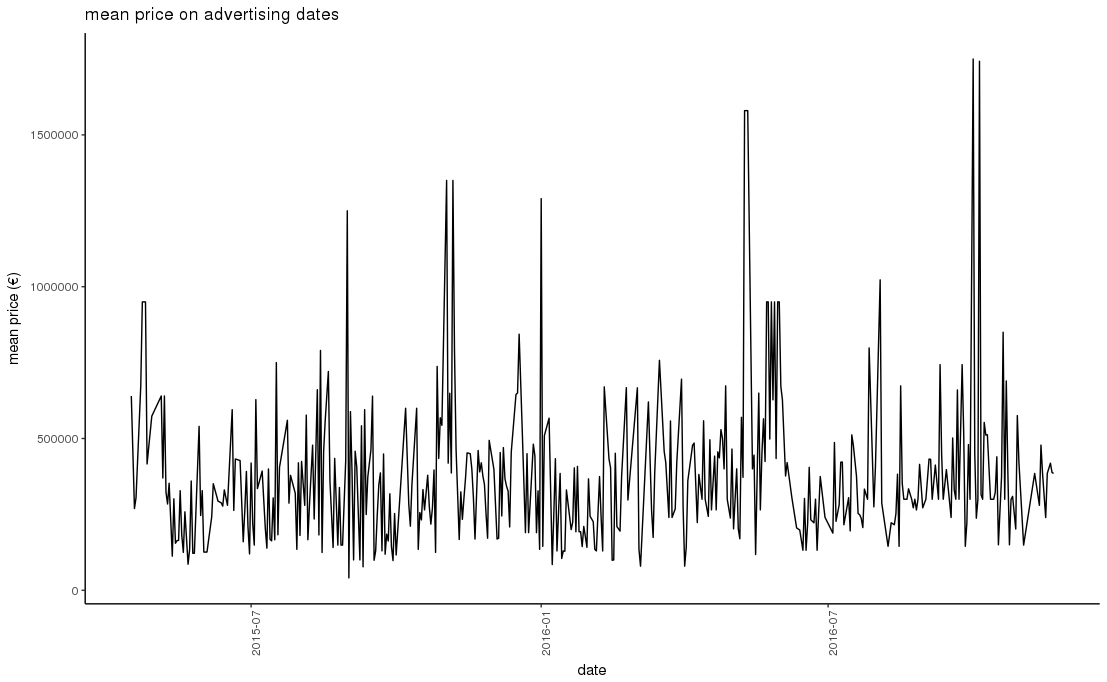
****

Chart 15: Advertising Day mean price

## 2.5. Statements

The visualisations generated in section C allows us to infer statements on the advertised property that was available for feature extraction:

**Apartments are the most advertised type of property.** Chart 7 gives a distribution of the different property types. One can note the striking difference of the apartment count with respect to the other counts. **Most of these apartments have an area of 100-150 metres squared**, as shown in Chart 8.Furthermore, the box plot in Chart 9, shows that for most property types used to live in (i.e. excluding shops, bars etc)**, the mean price decreases with frequency**.

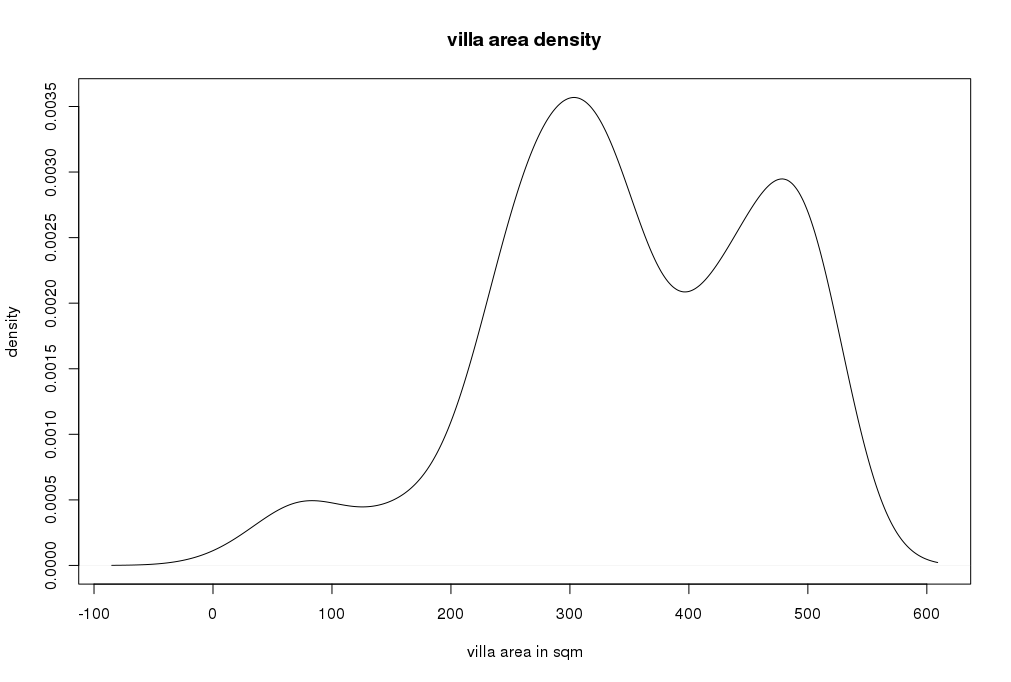
Chart 11 shows a **high concentration of apartments for sale at Sliema**. On the other hand, **houses and maisonettes for sale seem to be distributed quite uniformly** between their 10 most common locations. **Mosta and Ħaż-Żebbuġ have a strikingly higher amount of penthouses** when compared to other localities.

The geovisualisation in chart 12 shows that the **Valletta and St' Paul's bay have the highest property mean price.** These localities are followed by Wardija and Mellieha. The high mean price of St. Paul's Bay could be attributed to the inclusion of localities such as 'Tal-Fjuri' and 'Kennedy Grove'.

The rather uniform amount of adverts show in charts 13 and 14 both on Sundays and other days shows that there has not been any change in availability on property market. This may explain the minimal mean price fluctuation observed in chart 15.

## 2.6 Statistical Analysis

A statistical analysis is performed to check whether there is a significant difference between the area of maisonettes and villas. The actual calculations for this check can be found in statisticalAnalysis.R.

The t-test will be used to achieve this since the standard deviation is unknown. The t-test requires the samples to be normally distributed. As shown below, the fact that their distribution curve differs from the normal distribution curve will impair the validity of the result obtained from the t-test.

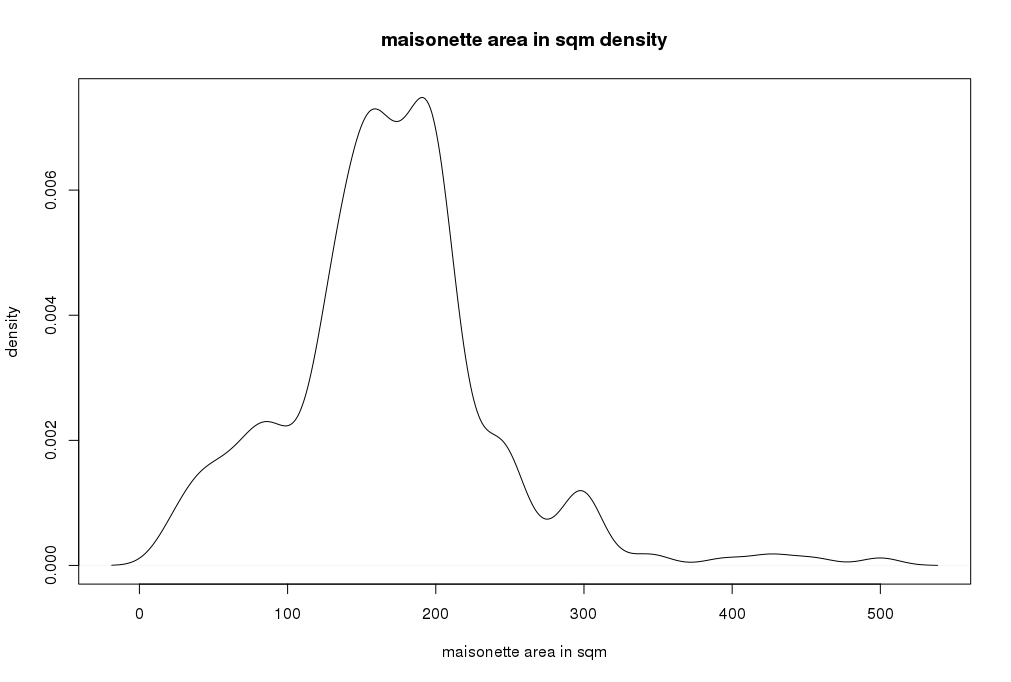


Chart 16: The density distributions of the areas of maisonettes and villas

With this in mind, we can proceed with defining the hypothesis to test:

Null hypothesis: mean area between villas and maisonettes is equal.

Alternative hypothesis: mean area between villas and maisonettes is not equal.

The F-test of these sample is used to compare the variances of these two samples (using the var.test function). In this case, the F-test value is 2.2e-16. This implies that the Welch test should be used since the samples have unequal variances. The t-test can then be computed as follows:

t.test(villas$area\_sqm,maisonette$area\_sqm,var.equal=FALSE, paired = FALSE)

The p-value obtained in less than 2.2e-16 which is less than 0.05. This shows that the mean of maisonette area is **significantly different** form the mean of area of villas.

## 2.7. Predictive Model

A naive-bayes classifier is used to predict the price range of a given property for sale described by the locality, property type and area. The implementation can be found in predictiveModel.R.

In this model, it is assumed that these features (location, property type and area) independently contribute towards the probability that the property falls in a certain price range [3]. The price ranges are taken in intervals of 5000€ since it is common to mention prices in these ranges (e.g. 20-25K, 50-55K etc). The price range for the properties for sale propDetails was hence calculated for every advertisment after finding the lowest and highest prices (maxPrice and minPrice) that are multiple of 5000:

propDetails$priceRange <- factor(cut(propDetails$price\_euro,

seq(from=minPrice,to=maxPrice, by=5000),

right = TRUE))

For testing purposes, 10% of the data was removed from the training set (trainingData). This 10% was removed randomly. This sample is called priceRangesToPredict. The model was built using trainingData:

model <- naiveBayes(priceRange ~ .,data=trainingData)

The model is used to predict the price range of the advertisments in the test data to obtain predictedPriceRanges:

predictedPriceRanges <- predict(model,priceRangesToPredict)

A confusion matrix is used to compare the predicted price ranges with the actual price ranges:

cMat <- confusionMatrix(testData$priceRange,predictedPriceRanges)

As seen in the below confusion matrix's description, the accuracy of the training set is very low (7%).

Accuracy Kappa AccuracyLower AccuracyUpper

0.07065217 0.05709598 0.04666839 0.10181011

A reason that might this extremely poor accuracy is that the assumption stated initially does not hold.

**References**:

[1] <https://en.wikipedia.org/wiki/Portable_Game_Notation>

[2] <https://www.quora.com/How-and-why-are-pie-charts-considered-evil-by-data-visualization-experts>

[3] https://eight2late.wordpress.com/2015/11/06/a-gentle-introduction-to-naive-bayes-classification-using-r/