**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: 15/01/2016**

Contents

[Question 1: Visualization 2](#_Toc440486814)

[A. Outstanding Visualizations 2](#_Toc440486815)

[A.1. Time Series 2](#_Toc440486816)

[A.2. Bar Graph 2](#_Toc440486817)

[A.3. Stacked Bar Graph of Population 3](#_Toc440486818)

[B. Terrible Visualizations 4](#_Toc440486819)

[B.1. Pie Chart 4](#_Toc440486820)

[B.2. Side by Side Bar graph 4](#_Toc440486821)

[B.3. Bar graph 5](#_Toc440486822)

[C. Re-implementation of Dynamic Representation 6](#_Toc440486823)

[Question 2: Data Science 7](#_Toc440486824)

[A) Data Supplied 7](#_Toc440486825)

[B) Features of Interest 8](#_Toc440486826)

[C) Visualizations 9](#_Toc440486827)

[C.1. Histogram of Prices 9](#_Toc440486828)

[C.2. Boxplot of Price vs Type 9](#_Toc440486829)

[C.3. Stacked Bar chart of Properties by Type and District 10](#_Toc440486830)

[C.4. Side by Side Bar chart of Mean Prices by Type and District 10](#_Toc440486831)

[C.5. Scatter Plot of SQM vs Price 10](#_Toc440486832)

[C.6. Line Graph of Mean Price throughout Time 11](#_Toc440486833)

[C.7. Geo-Visualisation of Location vs Nos of Properties 11](#_Toc440486834)

[D) Statements 13](#_Toc440486835)

[D.1. The more frequent properties are, the less expensive they are: 13](#_Toc440486836)

[D.2. The most expensive apartments are in the northern harbour 14](#_Toc440486837)

[D.3. Most apartments are found in the North 14](#_Toc440486838)

[D.4. Largest Property type is Palazzo 15](#_Toc440486839)

[D.5. Dates most Properties Feature is on a Sunday 16](#_Toc440486840)

[E) Statistical Analyses 17](#_Toc440486841)

[E.1. T Tests 17](#_Toc440486842)

[E.2. Z Tests 18](#_Toc440486843)

[F) Predictive Model 19](#_Toc440486844)

# Question 1: Visualization

## Dependecy 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”))

## A. Visualizations

### A.1 Kingbase chess games visualizations

The Kingbase chess data set available on (http://www.kingbase-chess.net/) was explored through the user 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 it was over 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 amout 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()

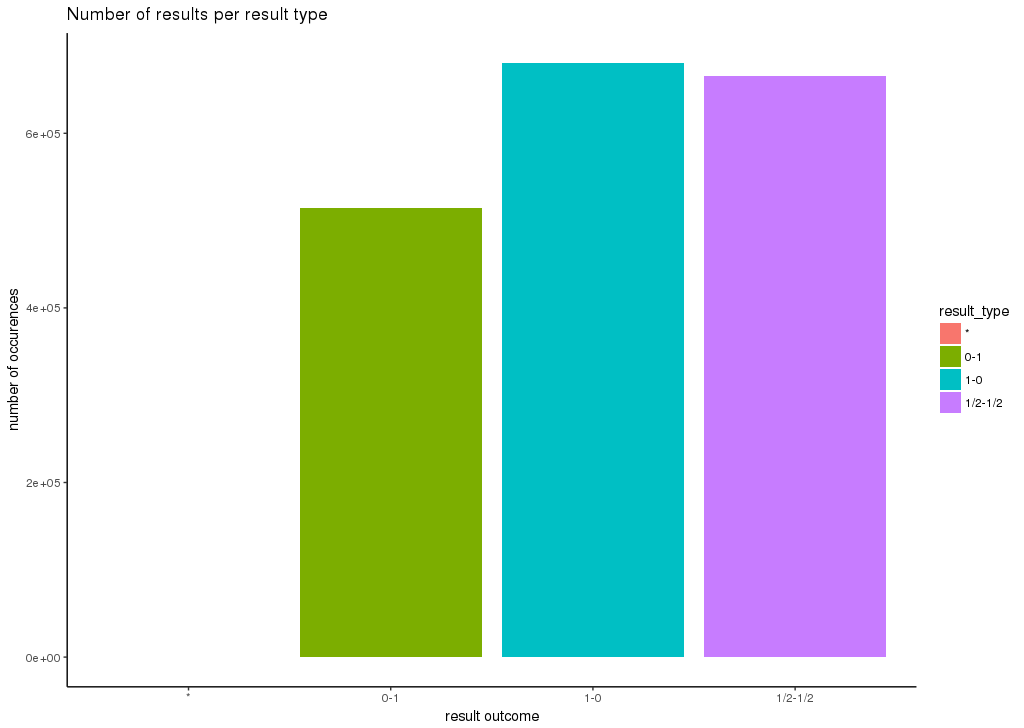
winningMovesHeatMap()

moveInGameBoxPlot()

resultCountBarPlot()

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

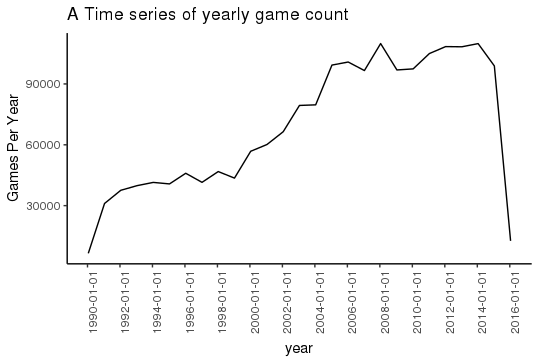
#### Distribution of Game Outcomes

Illustration 1: Distribution of game outcomes

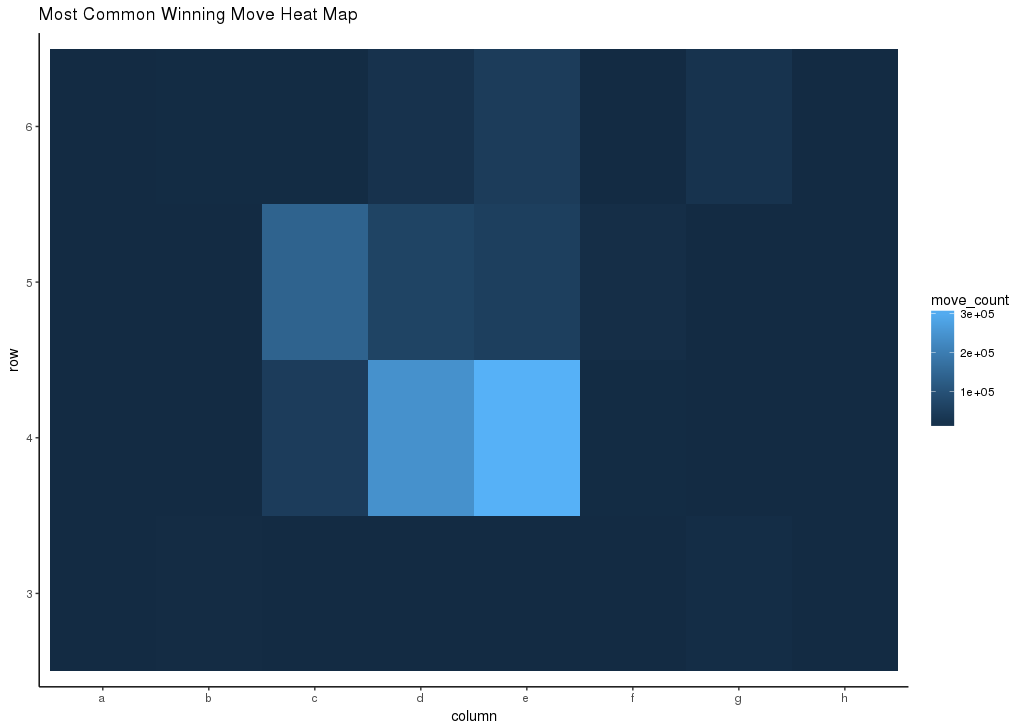
The outcome of the 1861460 games is depicted in chart 1. One can note that the number of of whie winws and draws is considerably higher than black wins. This may suggest that a white player might have a better chance of winning the game. The number of games whose outcome is described with (\*) is negligible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

#### Number of yearly Games

The processed kingston data set contains chess games ranging from 1990 to 2016 (acutually 2016/02). 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.

Illustration 1: Number of yearly games

#### Heatmap showing the starting moves of winners

Illustration 1: 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.

#### Number of Game Moves box plot

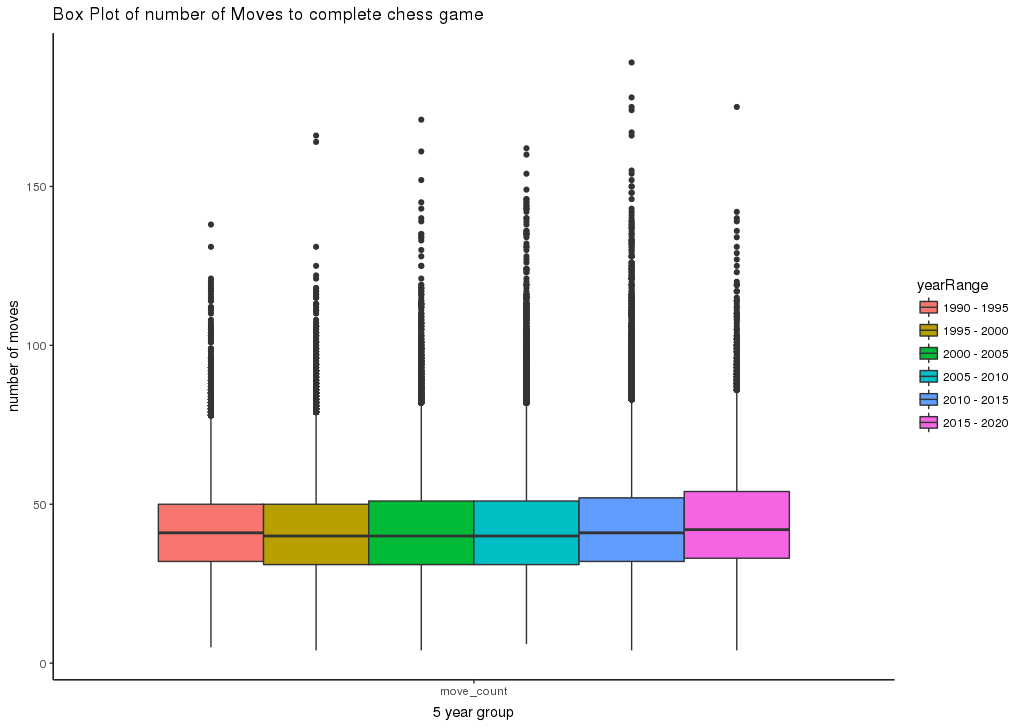
Illustration 1: Number of Game Moves box plot

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. This article used a pie chart to minute compare the risk 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 slize 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 Illusation 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 almost much easier to note Malta's low risk when compared to the other countries since a comparison by height is easier than by angle. This visualisation was generated by the script

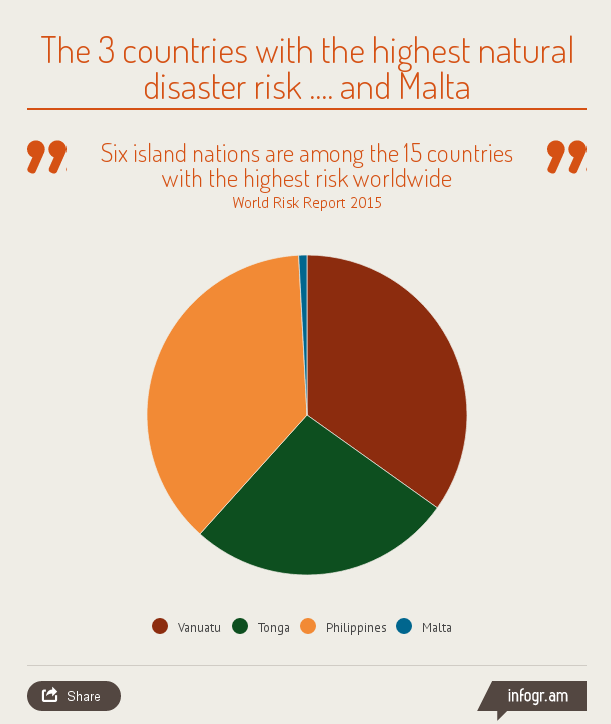
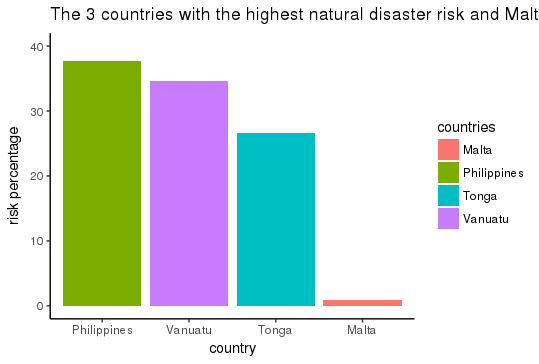
Illustration 1: A 'terrible' visualisation

Illustration 1: Re-implementation of the 'terrible' visualisation

# Question 2: Data Science

## A) Data Supplied

The supplied text files contain the html markup of the Times of Malta property advertisments published between 2015-04-16 to 2016-11-22. A word frequency analysis 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 is ranges from 200 to 600. A typical week day has less than 40 advertisments. This will be discussed in more detail further on.

|  |  |  |
| --- | --- | --- |
| 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 previosuly such as locality, area or even price and are quite vague.

## B) 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). 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 property type. 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.
* **Location:** The location 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 attempt to correct (or generalise) the location. For example, the locations 'Birgu' and 'Vittoriosa' would be mapped to Cottonera. 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 extract location. The word that has the least distance is selected.

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 unqiue csv files through a bash shell allowed duplicate adverts to be removed:

cat extracted\_features.csv | sort | uniq.

The feature extraction, data cleaning and duplicate removal 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. This is the source file from where the next 2 files will be generated.
* extracted\_with\_date\_unique.csv: the unique features with the advert date.
* unique\_features.csv: the unique features 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 over

The extracted adverts in the generated csv files contain missing information. There are instances were at least one the attributes 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 assignment 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 value 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:

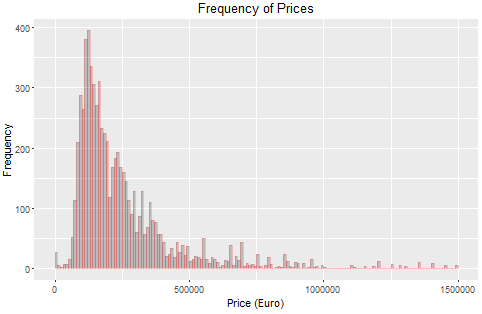
* Prices less than 3000€ and greater than 3 million euros.
* Areas greater than 500 square meters.

Adverts whose features were in these ranges were removed.

## C) Visualizations

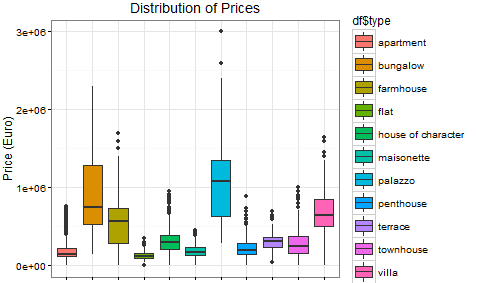
### C.1. Histogram of Prices

From the below histogram, we can note that prices are skewed to the left from the normal distribution. This is because all the properties are presented in one graph. The least expensive properties (apartments) are most frequent, while the most expensive once are at the right hand side tail.



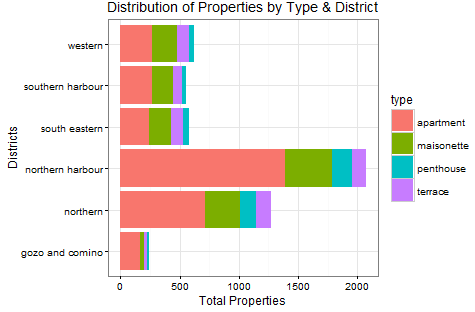
### C.2. Boxplot of Price vs Type

A box plot on the above histogram’s representation shows the distribution of prices for each property type side by side. The boxes ranges are quartiles, the middle line is the mean, whiskers are the variability outside the quartiles and individual points are the outliers.



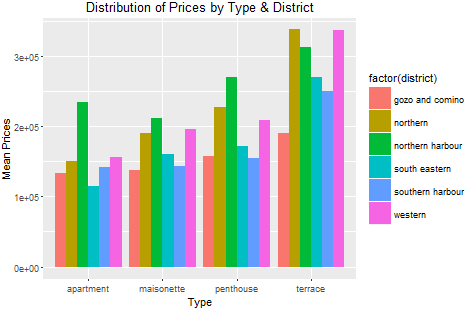
### C.3. Stacked Bar chart of Properties by Type and District

A stacked bar chart groups four property types by sub categories divided by districts and presents them on each other to show the number of properties.



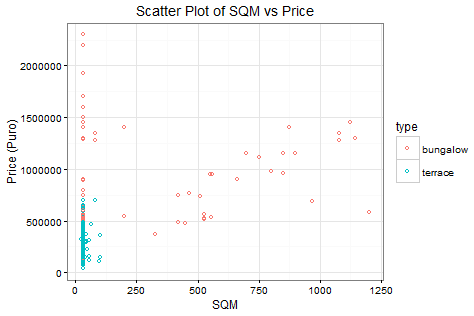
### C.4. Side by Side Bar chart of Mean Prices by Type and District

A side by side bar chart compares prices located in separate districts. It is noted that from the chosen properties, the number of terrace houses is always greater than the rest.

****

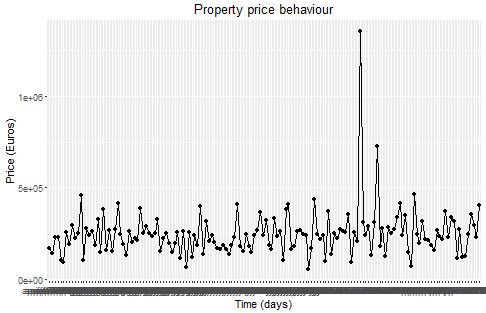
### C.5. Scatter Plot of SQM vs Price

The scatter plot inspects the relationship between the continuous variables price and sqm for two property types. Terrace houses are in the lower left side of the graph while bungalows are distinctively larger. The common sizes of properties in the straight line pattern are due to imputation problems arising from a large number of unavailable sqm.



### C.6. Line Graph of Mean Price throughout Time

The line graph shows that more or less prices have remained in the same level of variability ranges. There are two outliers due to some highly expensive properties featuring with other few cheap once.



### C.7. Geo-Visualisation of Location vs Nos of Properties

The geographical chart is based on data of the south eastern district, aggregated by locality to present number of properties spread within this region. The ggmap library is used to retrieve the longitude and latitude of each place so that these may be identified in the map to display the point of properties, using the dot sizes to show prices.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Location** | **Mean Price** | **Total Properties** | **Longitude** | **Latitude** |
| birzebbuga | 173339.8 | 61 | 14.524746 | 35.8136 |
| ghaxaq | 359401 | 44 | 14.516009 | 35.84404 |
| gudja | 182729.2 | 48 | 14.502904 | 35.84698 |
| kirkop | 154023.8 | 21 | 14.484347 | 35.84085 |
| marsascala | 207625.3 | 272 | 14.556788 | 35.86036 |
| marsaxlokk | 262992.3 | 26 | 14.53931 | 35.84117 |
| mqabba | 228380 | 25 | 14.469419 | 35.84441 |
| qrendi | 245588.2 | 17 | 14.454862 | 35.83285 |
| safi | 133875 | 8 | -9.227203 | 32.30082 |
| zejtun | 180423.6 | 123 | 14.536397 | 35.85487 |
| zurrieq | 203734.8 | 79 | 14.481065 | 35.82163 |

The map of Malta is displayed such that the above data may be identified and plotted below.



Unfortunately, I could not get this to work properly due to some technical issues which I did not manage to figure out how to solve.

## D) Statements

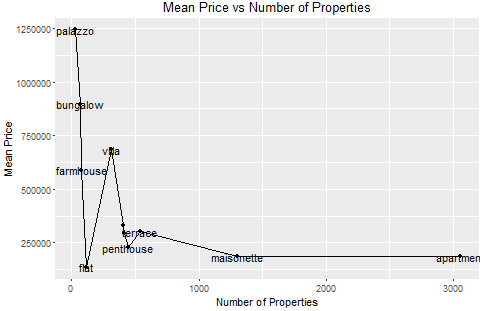
General observations are derived by calculating counts and means of quantitative variables grouped by some nominal category and sorted according to some order (cheapest, larges, etc.). This distributes measures according to separate groups for comparison. The general format used is similar to the sql approach: select <calculation> <quantity>, <category> where <category> in [<category values>] group by <category> ordered by <calculation> <quantity> <order>, eg select type, count(\*) from df group by type.

### D.1. the more frequent the properties are, the less expensive they are:

The table shows that the most expensive property (palazzo) is the least available while the least expensive (apartment) is in fact the most common.

|  |  |  |
| --- | --- | --- |
| type | nos types | mean price |
| apartment | 3054 | 184410.7 |
| maisonette | 1300 | 185033.5 |
| terrace | 541 | 301556.4 |
| penthouse | 443 | 227683.5 |
| townhouse | 415 | 291234.4 |
| house of character | 402 | 330313.5 |
| villa | 312 | 683740.6 |
| flat | 118 | 136053.7 |
| farmhouse | 78 | 589633 |
| bungalow | 71 | 897662 |
| palazzo | 26 | 1244346.2 |

A line graph is used to present the continuity of price variation between properties with respect to their amounts. In general, apart from flats, there is a decreasing pattern where less expensive property types are more available.

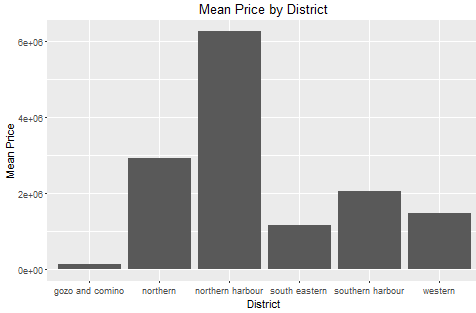


### D.2. The most expensive apartments are in the northern harbour

The head of the data frame grouped by location shows that the top expensive properties all lie in locations of the north harbour.

|  |  |  |  |
| --- | --- | --- | --- |
| district | location | nos\_apartments | mean\_price |
| northern harbour | tigné point | 13 | 612923.1 |
| northern harbour | portomaso | 13 | 474153.8 |
| northern harbour | fort cambridge | 14 | 409142.9 |
| northern harbour | the strand | 1 | 400000 |
| northern harbour | sliema | 351 | 344787.4 |
| southern harbour | senglea | 8 | 342000 |

The graph below shows the relative prices of districts:

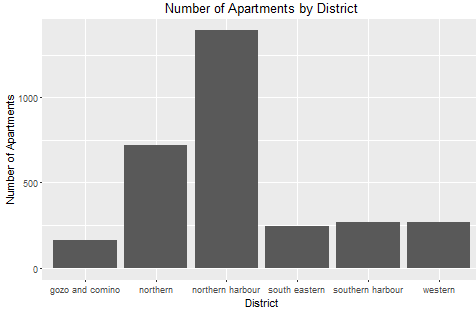


### D.3. Most apartments are found in the North

This time the frequency of presence of apartments is determined to be in the general north of the country:

|  |  |  |  |
| --- | --- | --- | --- |
| district | location | nos\_apartments | mean\_price |
| northern harbour | sliema | 351 | 344787.4 |
| northern harbour | st julians | 221 | 261082.1 |
| gozo and comino | gozo | 165 | 132965.5 |
| northern | qawra | 164 | 120896.9 |
| northern harbour | msida | 146 | 129355.5 |
| northern harbour | gzira | 126 | 170876.2 |

Unlike the above statement, the table does not give a very clear distinction within which of the North regions, the predominant number of apartments lies. However, the following bar chart shows that collectively the north harbour contains the most apartments.

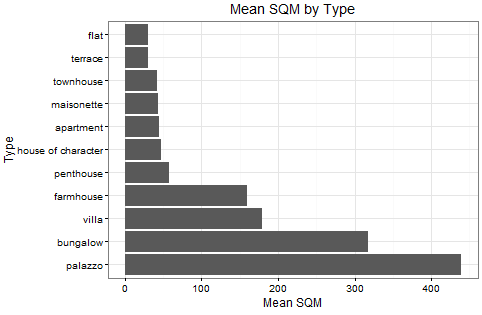


### D.4. Largest Property type is Palazzo

A group by property is done on the data frame to determine the aggregates below:

|  |  |  |
| --- | --- | --- |
| type | Mean sqm | Nos of properties |
| palazzo | 438.73077 | 26 |
| bungalow | 317.59155 | 71 |
| villa | 179.24038 | 312 |
| farmhouse | 159.87179 | 78 |
| penthouse | 57.45147 | 443 |
| house of character | 47.56716 | 402 |
| apartment | 45.0668 | 3054 |
| maisonette | 43.01846 | 1300 |
| townhouse | 42.23373 | 415 |
| terrace | 30.81885 | 541 |
| flat | 30 | 118 |

The results ordered by sqm shows the palazzo to be the largest property.

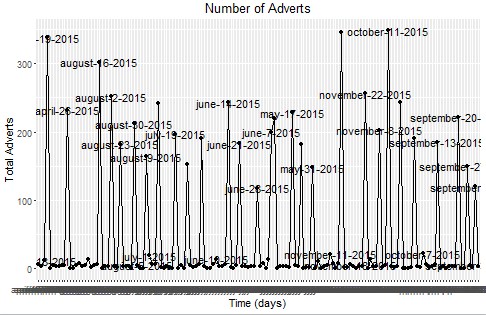


### D.5. Dates on Which Most Properties Feature is on Sundays

The below table is a grouping by first featured date for an aggregate count of the properties, ordered descending by the latter.

|  |  |
| --- | --- |
| first\_featured | total\_properties |
| october-11-2015 | 348 |
| november-15-2015 | 345 |
| april-19-2015 | 338 |
| august-16-2015 | 302 |
| november-22-2015 | 256 |
| august-2-2015 | 252 |

The dates where most adverts are are all Sundays, due to sellers adverting when more people read newspapers on this public day. In addition to the above frequency representation, below is a line graph of adverts throughout the several months of data sampling, showing peeks on the end of each week.



## E) Statistical Analyses

Statements from the sample are derived through Statistical Analyses by determining correlation between two aspects though hypothesis testing. The analysis plans used are the T test to study if there are any statistical differences between properties in different state and Z test to determine if certain characteristics influences price.

### E.1. T Tests

Two disjoint samples are picked on the basis of two different types. This method is applicable since we do not have the population available and we do not know the true standard deviation. The test is one of an independent nature since samples are unrelated and values in one group reveal no information about the other. This test assumes that the samples are taken from a normally distributed population, have equal variances and follow a normal distribution. If the samples do not have equal variances, the Welch t-test is used to adjust the number of degrees of freedom.

T Test 1: There is a significant difference is pricing between apartments and maisonettes:

* Null Hypnotises: mean pricing between apartments and maisonettes are equal.
* Alternative Hypnotises: mean pricing between apartments and maisonettes not equal.
* Evaluate sample variances by F-test to verify homogeneity of variances

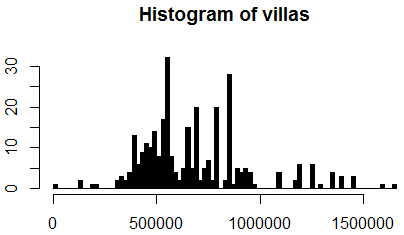
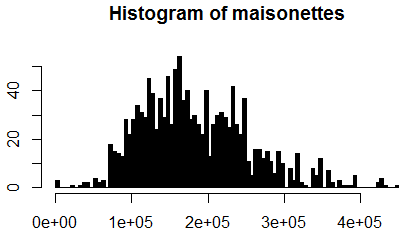
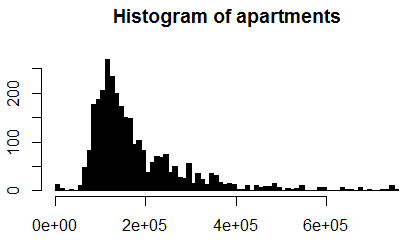
var.test(apartments, maisonettes) returns a p-value of 2.2e-16, which is < 0.05. This means that the two variances are non-homogeneous, so we use Welch (set var.equal=F).

* t.test(apartments, maisonettes, var.equal = FALSE, paired = FALSE) returns a p-value 0.8278, which is > 0.05. This means that there is **not enough evidence** to show that that averages are significantly different.

T Test 2: Pricing of villas is significantly larger then apartments:

* Null Hypnotises: mean pricing between apartments and villas are equal.
* Alternative Hypnotises: mean pricing of apartments is greater than maisonettes.
* F-test var.test(villas, apartments) returns a p-value of 2.2e-16, which is < 0.05. Thus variances are different, so we use Welch.
* t.test(villas, apartments, alternative="greater", var.equal=F) returns a p-value 2.2e-16, which is < 0.05. This means that mean price of villas are **statistically larger** then apartments.

Drawback: unfortunately, the histograms show that the samples above are not perfectly normal. This mean the normality assumption is not entirely followed and might pose a problem in the credibility of the tests.



**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