**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()

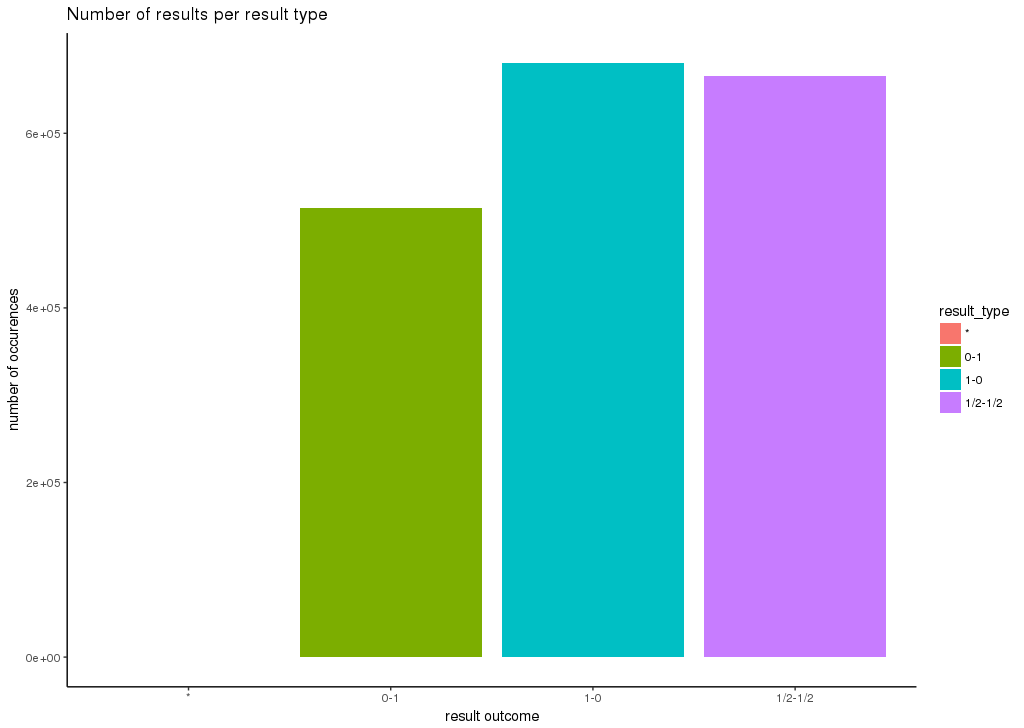
winningMovesHeatMap()

moveInGameBoxPlot()

resultCountBarPlot()

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

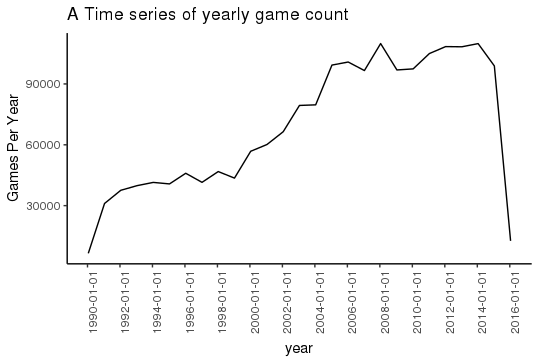
#### Distribution of Game Outcomes

Illustration 1: Distribution of game outcomes

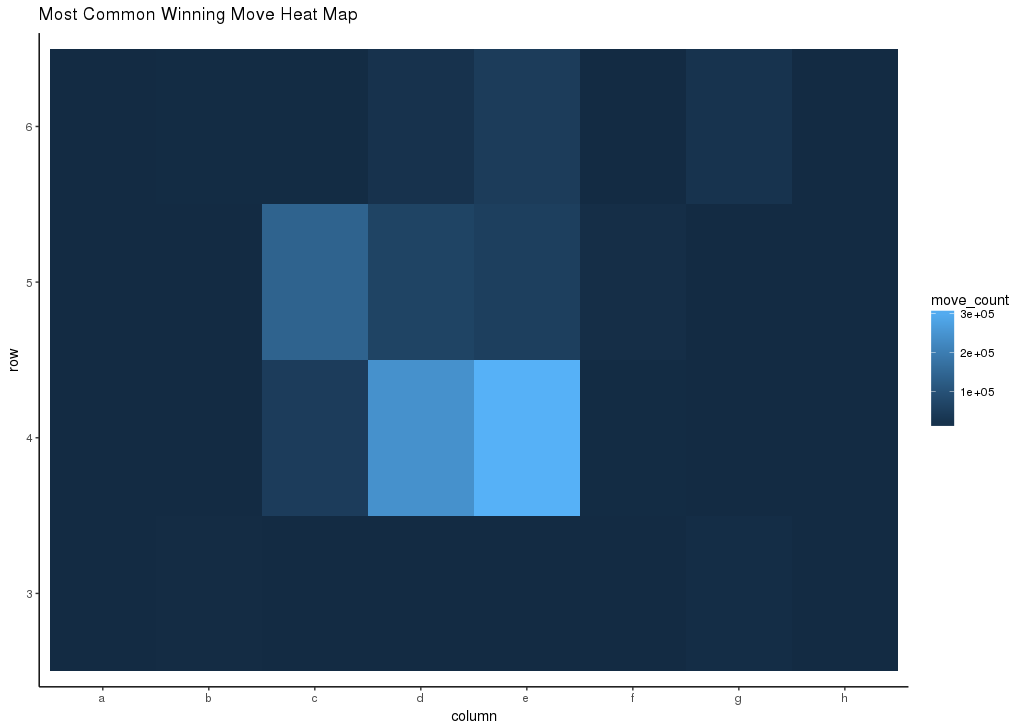
The outcome of the 1861460 games is depicted in chart 1. One can note that the number of of whie winws and draws is considerably higher than black wins. This may suggest that a white player might have a better chance of winning the game. The number of games whose outcome is described with (\*) is negligible 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 2: Number of yearly games

#### Heatmap showing the starting moves of winners

Illustration 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.

#### Number of Game Moves box plot

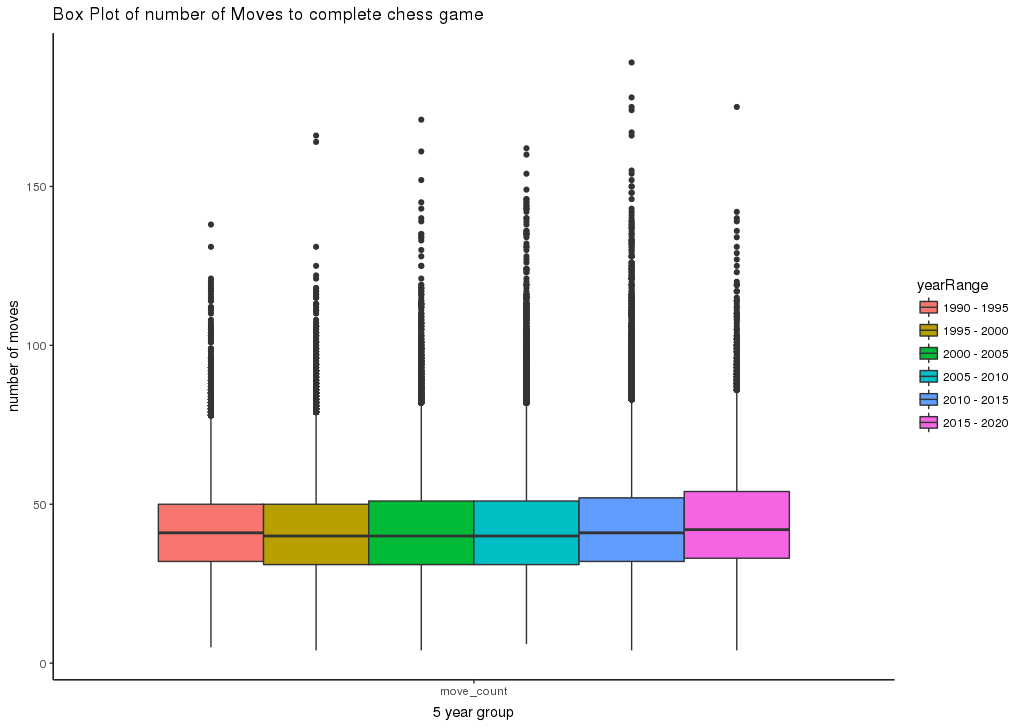
Illustration 4: Number of Game Moves box plot

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. 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

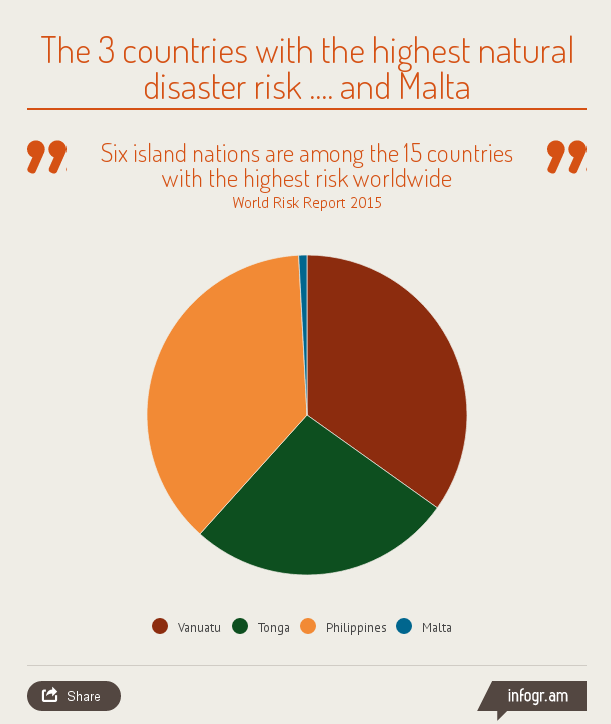
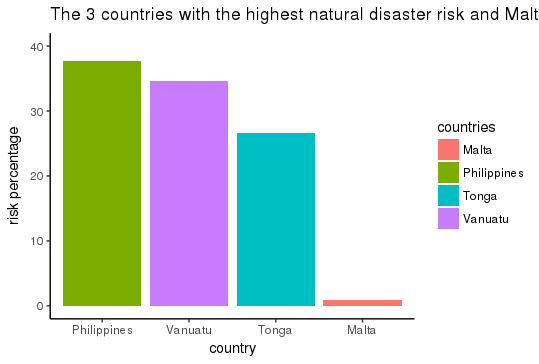
Illustration 5: A 'terrible' visualisation

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

# Question 2: Data Science

## A) Data Supplied

**Source**: the population dealt with in this assignment is the adverts of properties for sale, obtained from scrapped webpages of the Times of Malta.

**Domain**: the area of interest of the project is in properties and there various aspects they possess, e.g. rooms, price, etc. The aim of this study is to conclude statistic predicates, determine any relationship between characteristics and forecast attributes of interest through data science.

The information obtained is of type Quantitative and Qualitative. **Quantitative** Data is attributes which are measured e.g. prices. These may be objectively studied by means of averages and statistical correlations. **Qualitative** Data are descriptive entities, such as type of building. These may be classified by category types in frequency distributions. These influence decisions of potential buyers according to their personal preferences and are thus analysed in relation with other features of interest.

**Collection**: the selection of elements from the properties adverts population is one of a survey type. This is a probabilistic approach in a systematic fashion of determining representing participants upon where to perform studies.

**Interval**: adverts presented by date of publication are selected from some ad hoc day by a scheduled job, scrapping the website every week. The problem with such interval method is that it might miss periodicities. E.g. a villa seller might tend to advert on Sundays, when people are more available to look up places.

**Quantity**: the size of sample is one of a large nature, where data is consistently scrapped for seven months, capturing adverts featuring in several days. Thus, the sampling error mentioned above is addressed by large quantities of above 20,000, so that a more accurate representation of the true population is given.

**Quality**: a high quality dataset is ensured in the cleaning stages by addressing the following criteria:

1. **Timeliness**: data is readily available as: list of html pages, dataset of excel, cleaned structured csv. The cleaning process is automated such that it is re-runnable for any number of times to re-organise information. Data may be continually obtained from local news website and easily appended in the dataset. In addition, studies upon the provided data is relevant since it is recently obtained and property prices do not vary that much.
2. **Integrity**: the relationship between entities, such as owner (telephone nos), property, district, etc is established and can be related with each other.
3. **Accurate**: the sample corresponds to the real world, since it is obtained from publically available data. However, with regards to correctness, this may be an issue since errors due to misprints (eg 60,0000 euro) are found. These are addressed as much as possible in the cleaning by excluding outliers.
4. **Complete**: information is extracted as much as possible from the natural language descriptive data. Unfortunately, completeness is not always present for each entity, since there is missing information in adverts (eg nos of bedrooms for a finished house)

## B) Features of Interest

**Attributes**:

Natural language processing is used to semantically analyse text through smart search and concept deduction to determine attributes for study. Words are stemmed and sorted by frequency to find the most common keywords and provide an indication of potential data to look into.

**Sampling frame**: is organised by the following attributes:

* **Time**: first and last time advert featured
* **Quantity**: price, sqm, number of times advert featured, bedrooms, bathrooms,
* **Category ordinal**: floor
* **Category nominal of several groups**: location, district, telephone, type, view, back
* **Category nominal y/n options**: finished, new, garage, roof, ensuit, living, lift, balcony, yard, pool, garden, washroom.

All of these influence decisions of potential buyers according to their personal preferences and are thus analysed in relation with other features of interest.

**Cleaning**: unstructured raw data is processed into correct information for an accurate model to understand and base research upon. Text is parsed from html pages, loaded into data frames in R and transformed in the below steps.

**Removing duplicate data**: such adverts are determined by grouping data on all the above attributes extracted. If all features match exactly, it can be safely assumed that the matched rows refer to the same entity and repetitions are removed.

**Handling outliers**: abnormal differences in prices and sqm are determined to be 3 standard deviations from the mean for each property type. These are removed since 99.7% data should fall within this range for a large enough size which approaches the normal distribution. Such outliers are also removed for the purpose of this project since there are relatively few of them and these are not interesting for a collective study and hinder consistency in data. Also omission of real outlier data is done due to the fact that the data cleaning is automated and manual checking and excel adjustment should ideally be avoided.

**Handling missing data**: unknown values of rooms (bedrooms and bathrooms) are imputed by the mean of each property. Missing sqm is also estimated by the average of similar property, bedrooms, bathrooms, yard, pool and garden.

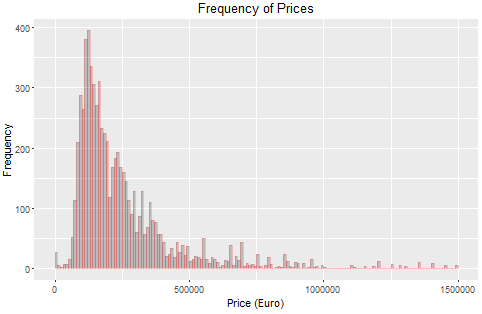
**Consistency amendments**:

* A new column for districts is included which maps the location into the regions: south eastern, southern harbour, northern harbour, northern, western, gozo and comino.
* Rows which are not mapped into the mentioned districts means that an invalid or foreign location is assigned and are thus not applicable and removed.
* Properties which do not fall into any type (villa, flat, etc) or are unclear are also removed since most of the data analysis is done on this feature and such cases are useless.

## C) Visualizations

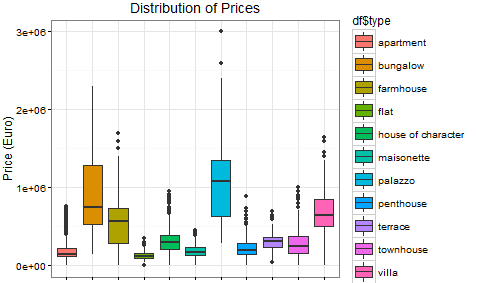
### C.1. Histogram of Prices

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.



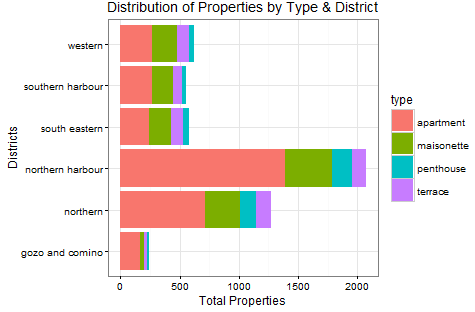
### C.2. Boxplot of Price vs Type

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.



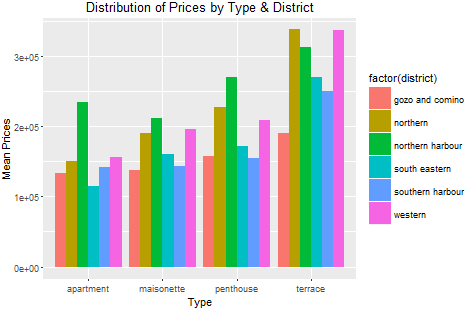
### C.3. Stacked Bar chart of Properties by Type and District

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.



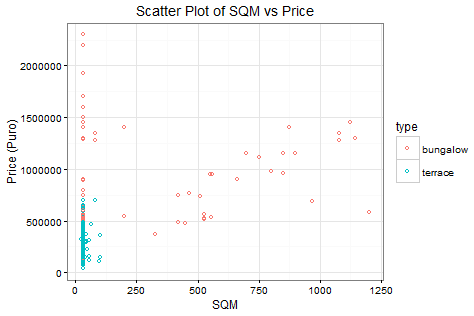
### C.4. Side by Side Bar chart of Mean Prices by Type and District

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.

****

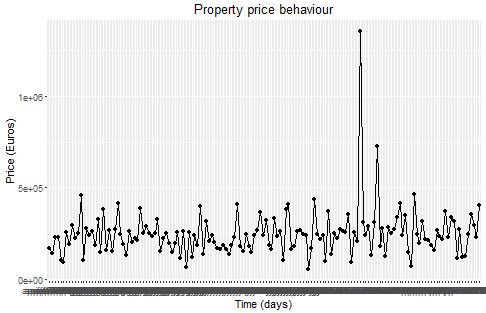
### C.5. Scatter Plot of SQM vs Price

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

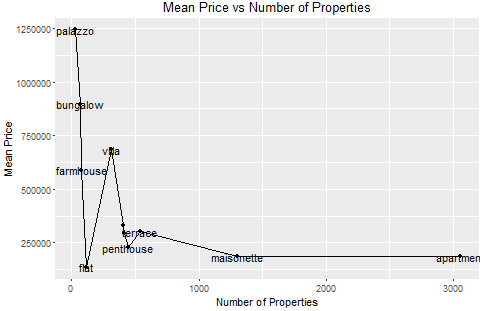
General observations are derived by calculating counts and means of quantitative variables grouped by some nominal category and sorted according to some order (cheapest, larges, etc.). This distributes measures according to separate groups for comparison. The general format used is similar to the sql 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.

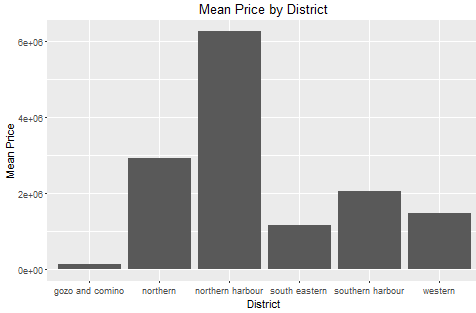


### D.2. The most expensive apartments are in the northern harbour

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:

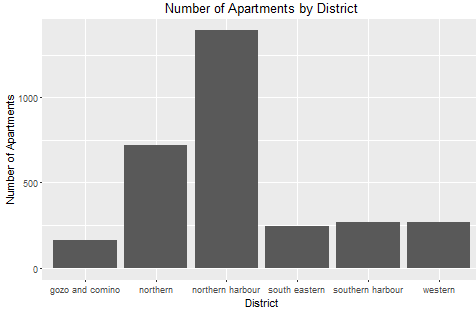


### D.3. Most apartments are found in the North

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.

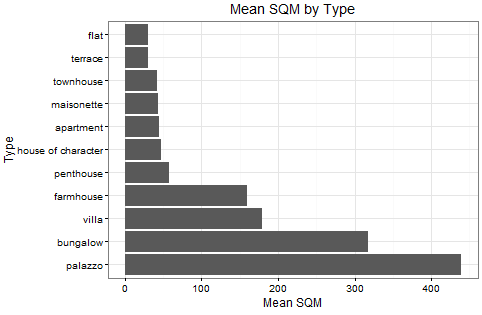


### D.4. Largest Property type is Palazzo

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.

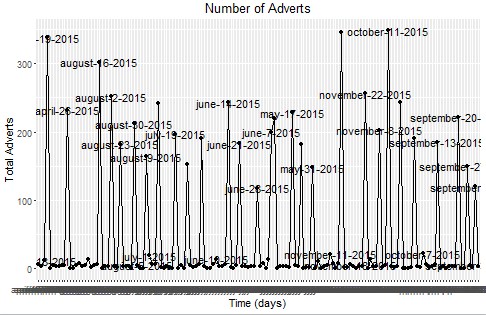


### D.5. Dates on Which Most Properties Feature is on Sundays

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

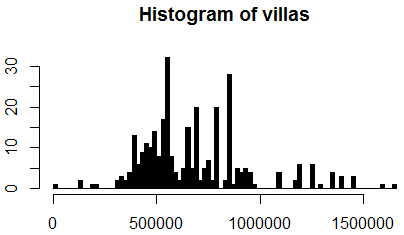
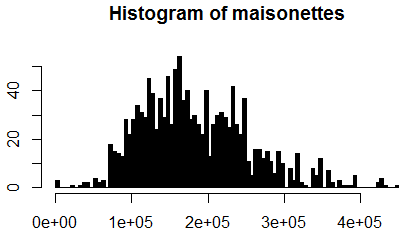
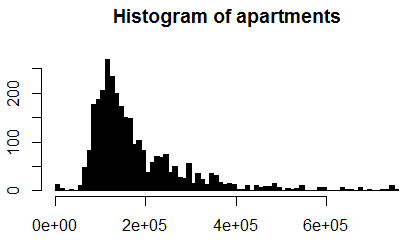
var.test(apartments, maisonettes) returns a p-value of 2.2e-16, which is < 0.05. This means that the two variances are non-homogeneous, so we use Welch (set var.equal=F).

* 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